Crowdsourcing evaluation of high dynamic range image compression

Philippe Hanhart, Pavel Korshunov, and Touradj Ebrahimi

Multimedia Signal Processing Group, EPFL, Lausanne, Switzerland

ABSTRACT

Crowdsourcing is becoming a popular cost effective alternative to lab-based evaluations for subjective quality assessment. However, crowd-based evaluations are constrained by the limited availability of display devices used by typical online workers, which makes the evaluation of high dynamic range (HDR) content a challenging task. In this paper, we investigate the feasibility of using low dynamic range versions of original HDR content obtained with tone mapping operators (TMOs) in crowdsourcing evaluations. We conducted two crowdsourcing experiments by employing workers from Microworkers platform. In the first experiment, we evaluate five HDR images encoded at different bit rates with the upcoming JPEG XT coding standard. To find best suitable TMO, we create eleven tone-mapped versions of these five HDR images by using eleven different TMOs. The crowdsourcing results are compared to a reference ground truth obtained via a subjective assessment of the same HDR images on a Dolby 'Pulsar' HDR monitor in a laboratory environment. The second crowdsourcing evaluation uses semantic differentiators to better understand the characteristics of eleven different TMOs. The crowdsourcing evaluations show that some TMOs are more suitable for evaluation of HDR image compression.

Keywords: High dynamic range, image compression, JPEG XT, subjective quality assessment, crowdsourcing

1. INTRODUCTION

Recent advances in high dynamic range (HDR) capture and display technologies have further increased the need for a compression standard and a common storage format for HDR content. This situation motivated Joint Picture Expert Group (JPEG) in 2012 to initiate the development of a new JPEG backward-compatible compression standard and a file format, called JPEG XT.[1] With this standard, HDR images are encoded in two layers: the base layer, where a tone mapped version of the HDR image is encoded using the conventional JPEG standard; and a residual layer, where an extra HDR information is encoded. The advantage of such an approach is that any conventional JPEG decoder can extract the tone mapped image, keeping the backward compatibility and allowing for its display on a conventional LDR monitor. However, a dedicated JPEG XT decoder would be able to read the residual layer and reconstruct a lossy version of the original HDR image.

As with any visual technology created for human subjects, it is important to evaluate the performance of HDR compression algorithms in terms of quality of user experience. For this purpose, two alternative assessment methods are available: objective and subjective evaluations. Objective evaluation is often faster and easier to perform. However, it is known that most objective metrics do not correlate well with perceived quality. Moreover, the performance of objective metrics for HDR content, specifically for compressed images, has not been investigated in details yet.[2] Therefore, subjective evaluation, despite being time consuming and expensive, remains the most reliable assessment method of HDR.

To reduce the costs of subjective evaluations and also to consider more practical environments, researchers are investigating crowdsourcing platforms, which allow employing workers online from around the world. A comprehensive overview of crowdsourcing approaches for subjective evaluations of image and video content is given by Hossfeld *et al.* in Ref. 3. This work also emphasizes issues and limitations of crowdsourcing in the context of subjective evaluations. One of the constraints is the limited variety of display devices used by online workers. Due to this limitation, for example, a direct evaluation of HDR content is impossible, since only low dynamic range (LDR) displays are currently available on the consumer market. Therefore, it is necessary to use alternative representations of HDR content in crowdsourcing evaluations.

Further author information: (Send correspondence to Philippe Hanhart) E-mail: {firstname.lastname}@epfl.ch

One simple approach for representing HDR content on an LDR display is to tone map the HDR image to reduce its dynamic range. As the mapping from a large set of values to a smaller set of values is not unique, the tone mapped image might look quite different depending on the tone mapping operator (TMO) and its parameters. Therefore, the tone mapped image can have more emphasis on the darker or brighter areas, which might change the visibility of compression artifacts when compared to the HDR image.

In this paper, we investigate whether crowdsourcing approach combined with preprocessing by a TMO is suitable for evaluation of compressed HDR images. For that purpose, five HDR images were encoded with JPEG XT profile A at four different bit rates and evaluated using paired comparison methodology, which was selected for its high accuracy and reliability in constructing a scale of perceptual preferences. Eleven TMOs were used to convert the compressed HDR images to the corresponding tone-mapped LDR versions. In addition, JPEG LDR versions were also used in the evaluations, which were manually produced from original HDR images by JPEG ad hoc group. A reference ground truth was obtained via a subjective evaluation of the compressed HDR images and their manually produced JPEG LDR versions on a Dolby Research HDR RGB backlight dual modulation display (aka 'Pulsar') in a laboratory environment. To evaluate the suitability of crowd-based quality assessment of HDR coding, the results of the crowd-based evaluations were compared to the ground truth results of the lab-based evaluations. An additional subjective evaluation was conducted using semantic differentiators to better understand the characteristics of the different TMOs.

The remainder of the paper is organized as follows. The contents, compression parameters, and TMOs considered in the experiments are described in Sec. 2. Section 3 presents the details of the lab- and crowd-based subjective evaluations. Results are presented and analyzed in Sec. 4. Section 5 concludes the paper.

2. DATASET

2.1 Contents

Five HDR images[*] of different dynamic ranges, representing different scenes, were used in the experiments (see Fig. 1 and Table 1 for details). Originally, these images were selected by JPEG Committee for the verification tests of JPEG XT standard. JPEG Committee also provided LDR versions of these images that were manually tone-mapped using Adobe Photoshop from the original HDR images.

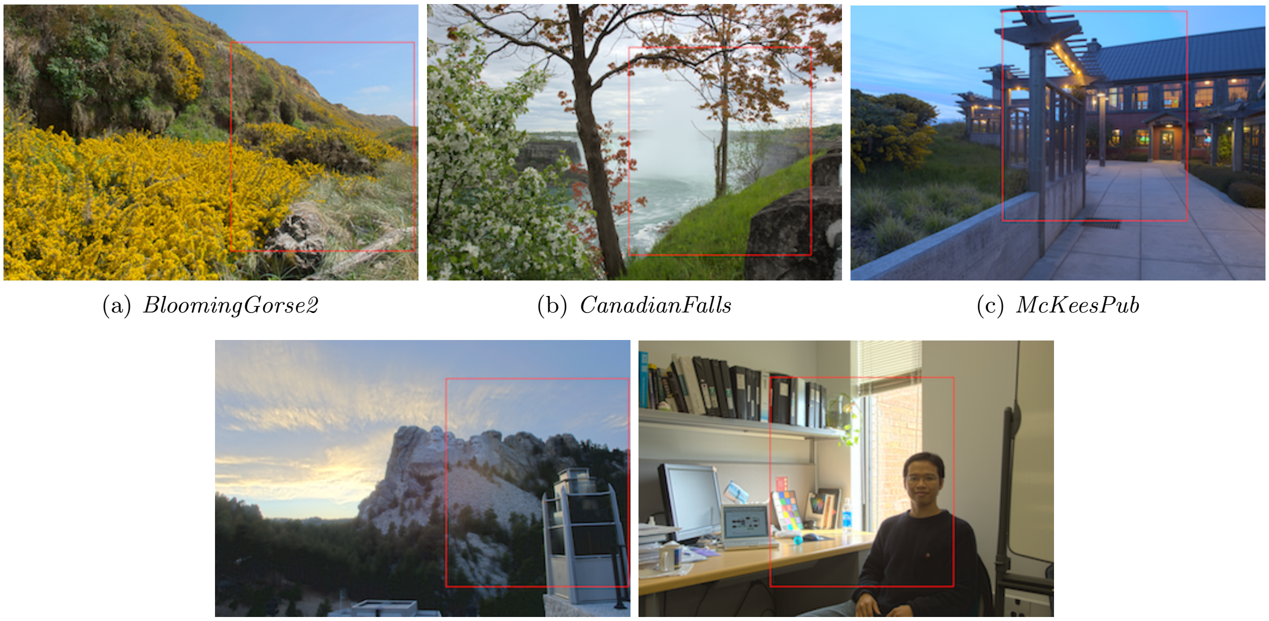
To prepare images for subjective experiments, both HDR and LDR versions were first downscaled by a factor of two with bicubic interpolation. The resulted images were then compressed using JPEG XT Profile A to four different bit rate values, ranging from a minimum of 0.3 bpp to a maximum of 2.2 bpp for different images. The bit rate values were selected for each content separately (see Table 1 for details) in such a way that there is a noticeable visual difference between images with different bit rates when they are displayed on an HDR monitor.

Compressed images were then cropped to 950 × 1080 pixels regions for side by side lab- and crowd-based subjective experiments (see Section 3 for details). The regions to crop were selected by expert viewers in such a way that cropped versions are representative of the quality and the dynamic range of the original images. Red rectangles in Figure 1 show the corresponding cropped regions. Downscaling together with cropping approach was selected as a compromise, so that a meaningful part of an image can be shown on the HDR monitor. Objective quality metrics were computed on the cropped versions of the images. For computation of the dynamic range, Banterle's HDR toolbox for MATLAB was used[†].

Eleven TMOs were used in the crowdsourcing subjective evaluation (see Table 2). These algorithms were selected based on their popularity in the research community and their visual characteristics. The HDR toolbox for MATLAB was used to apply the TMOs, as this toolbox implements many different operators. The TMOs were applied as follows. Each HDR original image was first compressed to four bit rates with JPEG XT producing four compressed HDR versions for each content. Each compressed HDR image was decoded and eleven TMOs were applied, producing eleven low dynamic range (LDR) images. These LDR images were compressed with high quality JPEG to be used in crowdsourcing evaluations.

[*] http://www.cis.rit.edu/fairchild/HDR.html

[†] http://www.github.com/banterle/HDR_Toolbox



(a) *BloomingGorse2* (b) *CanadianFalls* (c) *McKeesPub*

(d) *MtRushmore2* (e) *WillyDesk*

Figure 1: HDR images used in the experiments.

Table 1: HDR images information.

Image	Resolution	Dynamic range [dB]	Encoding parameters (q, Q)
	[pixels]	(cropped part)	q: base layer, Q: residual layer
BloomingGorse2	4288×2848	42	$(11, 12), (20, 13), (32, 15), (62, 15)$
CanadianFalls	4312×2868	41	$(16, 29), (30, 30), (65, 30), (80, 33)$
McKeesPub	4258×2829	60	$(5, 64), (15, 91), (48, 88), (83, 91)$
MtRushmore2	4312×2868	50	$(5, 20), (24, 82), (67, 80), (89, 78)$
WillyDesk	4288×2848	70	$(5, 63), (15, 79), (57, 90), (85, 91)$

Table 2: Tone mapping operators used in the subjective evaluation.

TMO	Description	Global / Local
Ashikimin[4]	A tone mapping algorithm for high contrast images	Local
Chiu[5]	Spatially nonuniform scaling functions for high contrast images	Local
Drago[6]	Adaptive logarithmic mapping for displaying high contrast scenes	Global
Durand[7]	Fast bilateral filtering for the display of HDR images	Local
Fattal[8]	Gradient domain high dynamic range compression	Local
Lischinski[9]	Photographic tone reproduction for digital images	Local
Linear	Simple linear scaling	Global
Logarithmic	Simple logarithmic scaling	Global
Reinhard[10]	Photographic tone reproduction for digital images	Local
TumblinRushmeier[11]	Two methods for display of high contrast images	Global
WardHistAdj[12]	A contrast-based scale factor for luminance display	Global

3. METHODOLOGY

The paired comparison methodology was selected to evaluate the quality of the different images, as this methodology provides a high accuracy and reliability in constructing a scale of perceptual preferences. Moreover, this methodology is very natural for test subjects to understand and use, which makes it perfectly suitable for crowdsourcing evaluations. The image pairs were presented in side-by-side fashion to minimize visual working memory limitations. Subjects were asked to judge which image in a pair ('left' or 'right') has the best overall quality. The option 'same' was also included to avoid random preference selections.

In the lab-based evaluation (see Sec. 3.1), one full HD (1920×1080 pixels) HDR monitor was used to display the images and a 20 pixels black border separated the pair of images. For the crowd-based evaluation (see Sec. 3.2), the images were downscaled by a factor two, using bicubic interpolation, so that workers were able to see the image pair in a web browser set for XGA (1024×768 pixels) monitor, which is one of the most common display resolutions.

3.1 Lab-based evaluation

Lab-based experiments were conducted at the MMSPG test laboratory, which fulfills the recommendations for subjective evaluation of visual data issued by ITU-R.[13] The test room is equipped with a controlled lighting system with a 6500 K color temperature, while the color of all the background walls and curtains present in the test area are mid grey. The laboratory setup is intended to ensure the reproducibility of the subjective tests results by avoiding unintended influence of external factors.

To display the test stimuli, a full HD (1920×1080 pixels) 42" Dolby Research HDR RGB backlight dual modulation display (aka 'Pulsar') was used. The monitor has the following specifications: full Rec. 709 color gamut, 4000 cd/m^2 peak luminance, low black level (0.005 cd/m^2), 12 bits/color input with accurate and reliable reproduction of color and luminance. In the experiments, the luminance of the background behind the monitor was about 20 cd/m^2. The ambient illumination did not directly reflect off of the display.

In every session, three subjects were assessing the displayed test video content simultaneously. They were seated in one row perpendicular to the center of the monitor, at a distance of about 3 times the picture height, as suggested in recommendation ITU-R BT.2022.[14]

Before the experiment, a consent form was handed to subjects for signature and oral instructions were provided to explain their tasks. All subjects were screened for correct visual acuity and color vision using Snellen and Ishihara charts, respectively. A training session was organized using additional contents to allow subjects to familiarize with the assessment procedure.

For each of the 5 contents, all the possible combinations of the 4 bit rates were considered. The HDR image decoded with JPEG XT, as well as the base layer decoded with JPEG, were evaluated, leading to a total of $5 \times 2 \times \binom{4}{2} = 60$ paired comparisons. To reduce contextual effects, the stimuli orders of display were randomized applying different permutations for each group of subjects and special care was taken for the same content not to be shown consecutively.

A total of 18 naïve subjects (11 females and 7 males) took part in the evaluation. They were between 20 and 34 years old with an average of 25.3 years of age.

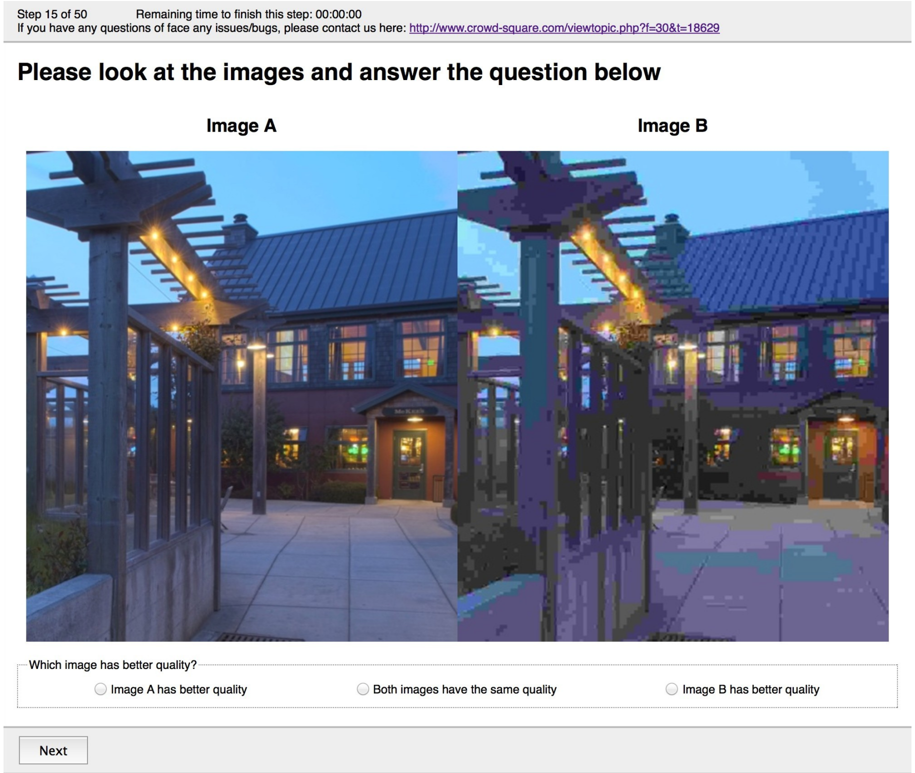
3.2 Crowd-based evaluation

The tone mapped images were compressed with JPEG at quality 95, which produces images that have visually lossless quality and file sizes still suitable for transmission to remote crowdsourcing workers. Figure 2 shows a screenshot of the crowdsourcing interface for JPEG XT compression experiment and Figure 3 shows a screenshot of the TMO characteristics evaluation. In both crowdsourcing experiments, a slightly modified version of the QualityCrowd 2 framework[15] was used.

3.2.1 Evaluation of JPEG XT compression

Before the experiments, short written instructions were provided to the workers to explain their tasks. Additionally, three training samples, with a different content, were displayed to familiarize workers with the assessment procedure. The training instructions and samples were presented using QualityCrowd 2.

For each of the 5 contents, all the possible combinations of the 4 bit rates were considered. In total, for each of 5 contents, 11 versions corresponding to 11 TMOs (see Table 2)) and the base layer decoded with JPEG were evaluated, leading to $5 \times 12 \times \binom{4}{2} = 360$ paired comparisons. As the number of pairs to be evaluated is very high, it is impossible for one worker to evaluate all pairs. Therefore, the pairs were randomly split into 9 batches of 40 pairs each. To reduce contextual effects, the stimuli orders of display were randomized and special care was taken for the same content not to be shown consecutively. Each worker was allowed to take only one batch.



Step 15 of 50 Remaining time to finish this step: 00:00:00

If you have any questions of face any issues/bugs, please contact us here: http://www.crowd-square.com/viewtopic.php?f=30&t=18629

Please look at the images and answer the question below

Which image has better quality?

○ Image A has better quality ○ Both images have the same quality ○ Image B has better quality

Figure 2: Evaluation of HDR image compression: screenshot of the crowdsourcing interface.

Therefore, the evaluation required a relatively large amount of workers. Subjects were recruited through the Microworkers[‡] crowdsourcing platform. Only countries where English is a dominant language were chosen, with either more than 50% of population or more than 10 million of people speaking English, according to Wikipedia. Workers received a compensation of $0.50 for completing the 40 evaluation tasks.

To detect unreliable workers, two honeypots were inserted around the 1st and 3rd quarter of the evaluation task. The honeypots were related to the content of the image pair seen on the previous evaluation task.

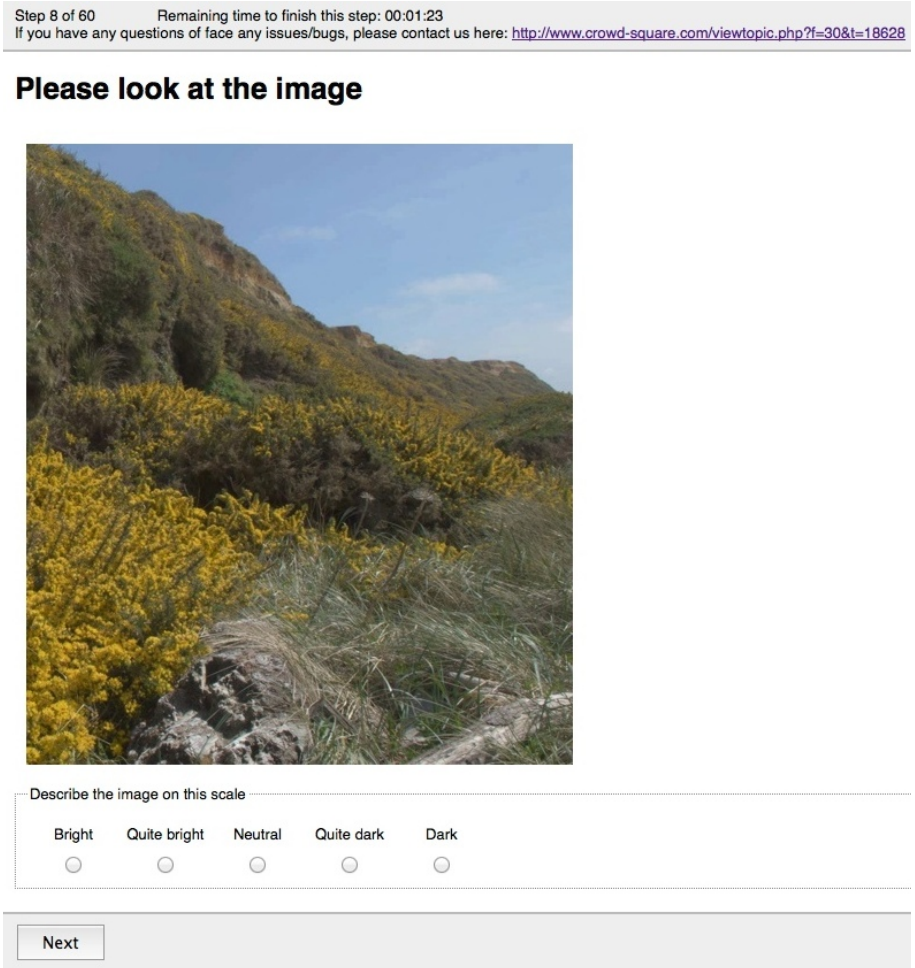
3.2.2 Evaluation of TMO characteristics

To evaluate the characteristics of the different TMOs, the semantic differential (SD) method was used considering the following bipolar adjective pairs:

1. Dark - Bright
2. Cold - Hot
3. Smooth - Sharp
4. Dirty - Clean
5. Dusty - Vivid
6. Static - Dynamic
7. Bad - Good
8. Unrealistic - Realistic
9. Unpleasant - Pleasant
10. Ugly - Beautiful

These adjective pairs were chosen based on the results from previous studies using SD on image characteristics[16–18] and the relevance of adjective pairs in the context of tone mapped images.

[‡] http://www.microworkers.com



Step 8 of 60 Remaining time to finish this step: 00:01:23
If you have any questions of face any issues/bugs, please contact us here: http://www.crowd-square.com/viewtopic.php?f=30&t=18628

Please look at the image

Describe the image on this scale

Bright ○ Quite bright ○ Neutral ○ Quite dark ○ Dark ○

Figure 3: Evaluation of TMO characteristics: screenshot of the crowdsourcing interface.

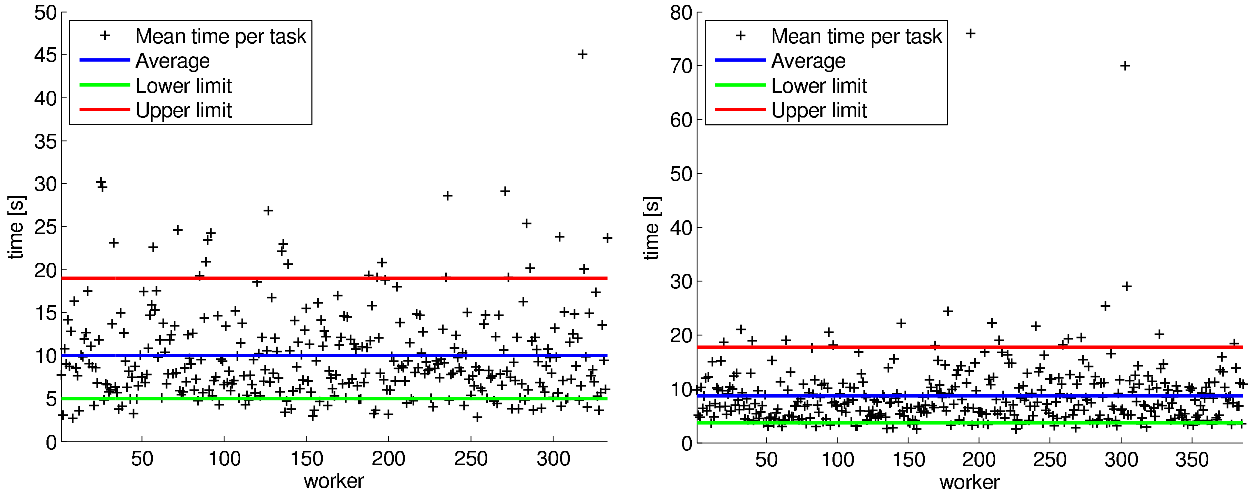
The crowdsourcing experiment to evaluate TMO characteristics was conducted in a similar way to JPEG XT compression experiment, except that instead of pairs of images, a single image per page was displayed to a worker. This resulted in $5 \times 12 = 60$ stimuli (five contents, 11 TMOs, and the base layer). The stimuli were randomly split into 12 batches of 5 stimuli each, one for each content, and the 10 bipolar adjective pairs were considered for each stimuli. Workers received a compensation of \$0.50 for completing the 50 evaluation tasks.

4. RESULTS

4.1 Unreliable workers detection

Unlike lab-based subjective experiment where all subjects can be observed by operators and its test environment also can be controlled, the major shortcoming of the crowdsourcing-based subjective evaluation is the inability to supervise participants behavior and to restrict their test conditions. When using crowdsourcing for evaluation, there is a risk of including untrusted data into the analysis due to wrong test conditions or unreliable behavior of some workers who try to submit low quality work to reduce their effort while maximizing their compensation.[3] For this reason, unreliable workers detection is an inevitable process in crowdsourcing evaluations. To identify a worker as 'trustworthy', the following four factors were considered in our experiments:

1. Mean observation time per question;
2. Honeypots, i.e., trap questions related to the content of the image seen in the previous question[3]
3. Post-screening according to the guidelines described in Section 2.3.1 of Annex 2 of Rec. ITU-R BT.500-13[13] (only for the evaluation of TMO characteristics).



(a) HDR image compression.

(b) TMO evaluation.

Figure 4: Mean time for one question for each worker.

Table 3: Number of reliable workers per batch depending on the considered outlier detection technique.

(a) HDR image compression.

Outlier detection	Batch								
	1	2	3	4	5	6	7	8	9
None	42	48	41	43	43	33	44	41	40
1st honeypot	29	32	32	27	27	24	23	29	27
2nd honeypot	28	31	25	34	29	25	30	28	28
Time	29	40	32	32	33	26	35	29	33
1st honeypot + time	26	29	28	22	24	21	22	25	24
2nd honeypot + time	23	28	21	27	29	22	28	22	25
2 honeypots	25	25	24	26	21	23	22	24	22
2 honeypots + time	22	22	20	21	21	20	21	20	20

(b) TMO evaluation.

Outlier detection	Batch											
	1	2	3	4	5	6	7	8	9	10	11	12
None	35	50	33	35	34	31	43	33	32	32	37	39
1st honeypot	28	23	27	25	24	26	21	23	27	24	28	22
2nd honeypot	26	34	28	23	25	21	34	27	22	23	24	30
Rec. ITU-R BT.500	30	44	29	29	29	28	34	29	29	29	29	37
Time	28	36	23	27	24	25	28	25	23	23	26	32
1st honeypot + Rec. ITU-R BT.500	28	23	27	25	24	25	21	23	27	24	28	22
2nd honeypot + Rec. ITU-R BT.500	26	34	28	23	25	21	34	27	22	23	24	30
1st honeypot + time	26	17	22	23	21	22	16	21	23	20	25	19
2nd honeypot + time	24	28	22	21	21	17	28	23	18	17	21	26
1st honeypot + Rec. ITU-R BT.500 + time	26	17	22	23	21	21	16	21	23	20	25	19
2nd honeypot + Rec. ITU-R BT.500 + time	24	28	22	21	21	17	28	23	18	17	21	26
2 honeypots	24	21	27	22	21	21	21	21	21	20	23	19
2 honeypots + Rec. ITU-R BT.500	24	21	27	22	21	21	21	21	21	20	23	19
2 honeypots + time	22	16	22	20	18	17	16	19	18	16	20	17
2 honeypots + Rec. ITU-R BT.500 + time	22	16	22	20	18	17	16	19	18	16	20	17

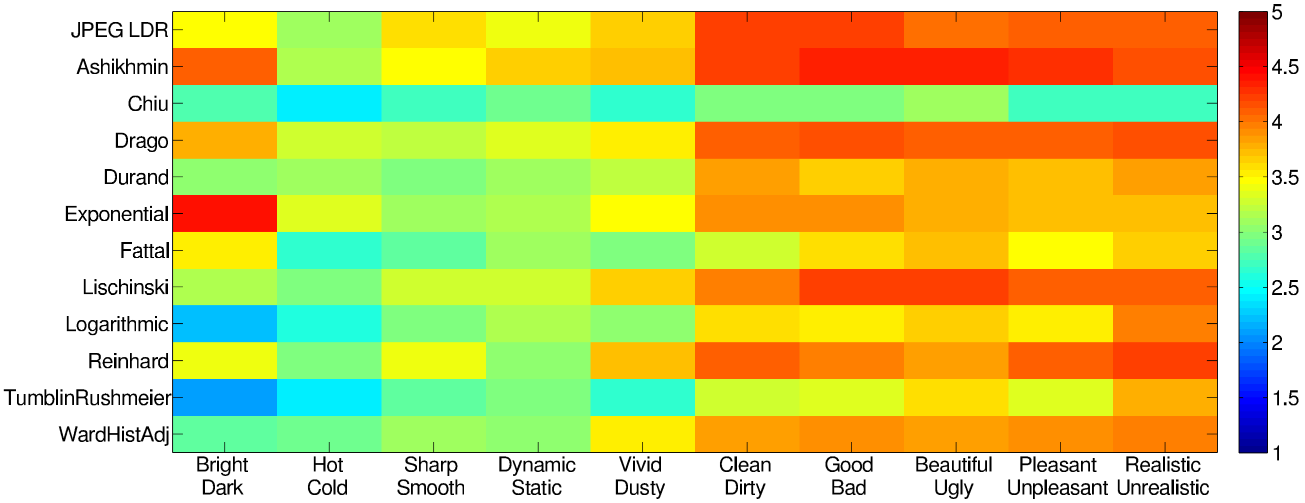


Figure 5: Semantic differentiator.

The observation time per question is measured as the time from when the question is displayed until the time the answer is given by the worker. The mean observation time can be calculated using this data. If the mean observation time per question is too short or too long when compared to the average of all workers, it can be deduced that the worker did not take the test seriously or was distracted during the task. Figure 4 depicts the mean response time per question for each worker for both evaluations. The mean observation time is between 9 and 10 seconds. As it can be observed, some workers demonstrate large values when compared to the corresponding mean, especially for the TMO evaluation campaign. To filter out unreliable workers, we set a lower and upper limit at −5 and +9 seconds from the mean respectively.

Table 3 reports the number of valid workers per batch for both evaluation campaigns. The results are reported for the individual outlier detection techniques, as well as for the combination of the different techniques. As it can be observed, the results are quite different from one batch to another. For example, 50 workers were necessary, considering all outlier detection techniques, to obtain 16 valid sets of scores for 2^{nd} batch of the TMO evaluation campaign, while 22 workers out of 33 produced reliable results in the 3^{rd} batch. In general, it can be observed that one to two third of the workers were considered unreliable according to the specified criteria.

For the JPEG XT image compression campaign, we targeted 18 reliable workers per batch, to reach the same number of subjects as in the lab-based evaluation (see Sec. 3.1). For the TMO characteristics evaluation campaign, we targeted 16 reliable workers per batch, which is the number of subjects commonly considered in subjective evaluations. In both cases, the most restrictive outlier detection was considered, since it includes mean time per question, 2 honeypots, and Rec. ITU-R BT.500. Originally, we ran 30 workers per batch. However, additional slots were opened to reach the required minimum number of valid workers. Therefore, in total, 375 and 434 workers took part in the JPEG XT image compression and TMO characteristc evaluation campaigns respectively. In the final results, only the first 18 and 16 workers were considered to be valid for the JPEG XT and TMO campaigns.

4.2 Evaluation of TMO characteristics

Figure 5 shows the SD profiles of the different TMOs considered in the experiment (see Table 2). Different patterns can be observed from this figure. The Dark-Bright pair seems to have the most diversity of results, whereas the other pairs have somehow similar results across the different TMOs, except for 'Chiu' TMO. This TMO was usually rated lower than the other TMOs. In particular, results show that 'Chiu' produces less pleasant and realistic images, that are not as good, clean, and beautiful than the other TMOs. Therefore, it can be concluded that this TMO does not produce good tone mapped images, which may affect the evaluation of HDR compression later. On the other hand, 'Ashikimin', 'Drago', 'Exponential', 'Lischinski', and 'Reinhard' seem to produce overall satisfactory tone mapped images, with 'Fattal' and 'WardHistAdj' being the next best subset of TMOs, since the corresponding characteristics tend to be towards the top of the scale.

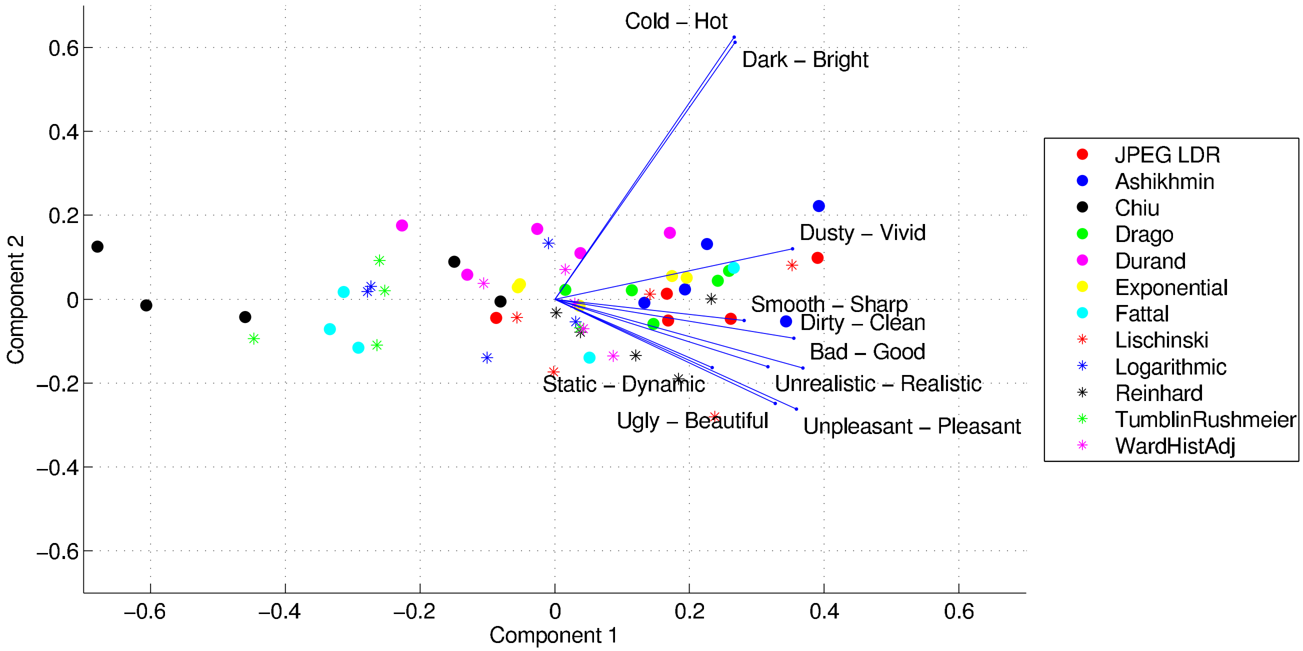


Figure 6: PCA of the SD profiles.

Table 4: Two factor analysis.

Adjective pairs	First factor	Second factor
Bad - Good	0.3689	-0.1642
Unpleasant - Pleasant	0.3591	-0.2624
Dirty - Clean	0.3550	-0.0934
Dusty - Vivid	0.3532	0.1200
Ugly - Beautiful	0.3275	-0.2487
Unrealistic - Realistic	0.3167	-0.1608
Smooth - Sharp	0.2813	-0.0510
Dark - Bright	0.2682	0.6117
Cold - Hot	0.2663	0.6245
Static - Dynamic	0.2339	-0.1628
Contribution ratio	62.03%	10.40%
Cumulative contribution ratio	62.03%	72.43%

The pairs Dirty-Clean, Bad-Good, Ugly-Beautiful, Unpleasant-Pleasant, and Unrealistic-Realistic seem to be quite correlated and most scores are above 3.5 (except for 'Chiu'). The pairs Smooth-Sharp, Static-Dynamic, and Dusty-Vivid are also quite correlated, but with lower scores (around 3). The pair Cold-Hot appears to correlate more with the pair Dark-Bright than the other pairs, but the values are lower and less spread. Intuitively, Cold-Hot can be related to Dark-Bright, but in the context of images, the second one is more obvious and easier to understand. This probably explains why the scores are more spread over the scale for Dark-Bright, whereas most values for Cold-Hot are around the neutral value (3).

Figure 6 depicts the results of the principal component analysis (PCA) applied on the SD profiles. The first component seems to be mostly related to the quality of the tone mapped image, although the contribution of the different related pairs is lower than 0.4 in all cases. The second dimension is mostly related to the brightness of the tone mapped image, as the contribution of the pairs Dark-Bright and Cold-Hot for this dimension is above 0.6. However, it appears that the brightness is also related to image quality. Table 4 reports the contribution of the different bipolar adjective pairs to the two principal components. The principal components explain 72.43% of the variance observed in the data. However, most of the contribution comes from the first component, which explains 62.03% of the variance.

Table 5: Number of significantly different pairs between crowd-based and lab-based evaluation of JPEG LDR.

BloomingGorse2	*CanadianFalls*	*McKeesPub*	*MtRushmore2*	*WillyDesk*	Average
1	1	0	0	1	0.6

Table 6: Number of significantly different pairs between JPEG LDR and HDR.

Evluation	Content					Average
	BloomingGorse2	*CanadianFalls*	*McKeesPub*	*MtRushmore2*	*WillyDesk*	
Laboratory	1	0	0	0	0	0.2
Crowdsourcing	0	2	0	0	0	0.4

4.3 Evaluation of HDR image compression

To determine whether the paired comparison results from the crowdsourcing evaluation are significantly different from the results of the ground truth lab-based evaluation, the Barnard's test[19] was used. This test is a statistical significance test of the null hypothesis of independence of rows and columns in a contingency table. It is claimed that the Barnard's test is more powerful than Fisher's exact test for 2 × 2 contingency tables. Therefore, this test was used to compare the results of a pair of images evaluated in the crowdsourcing evaluation to the ground truth results corresponding to the same pair. The test was repeated for all possible pairs, i.e., six pairs per content and per TMO.

Table 5 reports the number of significantly different pairs between crowd-based and lab-based evaluation of JPEG LDR. In this case, no tone mapping was applied to produce the images for the crowdsourcing evaluation and the same images were shown in both environments. Therefore, these results can be used to determine whether crowdsourcing produces reliable results when compared to the ground truth lab results. As it can be observed, the number of significantly different pairs is 0.6 on average for all contents or 3 out of 30 pairs, which means that 10% of the pairs obtained different results. The difference, though not very significant, is probably due to the influence of the different viewing conditions, which cannot be controlled in crowdsourcing evaluations, or the unreliability of workers, even though we used several mechanisms to detect outliers and potential cheaters.

Table 6 reports the number of significantly different pairs between JPEG LDR and JPEG XT for both lab- and crowd-based evaluations. Results show that, in the laboratory, only one pair was significantly different. This indicates that, for the compression parameters considered in this experiment, the relative quality difference between two LDR base layers was similar to that of the corresponding HDR images reconstructed with an additional enhancement layer. In the crowdsourcing evaluation, only two pairs were significantly different, but the differences occurred for a different content. This might indicate that content characteristics might influence the results in crowd-based evaluations. However, we have too few pairs per content to determine whether content characteristics have a statistically significant influence.

Table 7 reports the number of significantly different pairs between the tone mapped images evaluated in the crowdsourcing experiment and the HDR images evaluated in the laboratory. As it can be observed, most of the differences occur for contents *BloomingGorse2* and *CanadianFalls*, which have a limited dynamic range (see Table 1). Since the TMOs were essentially designed to handle images with higher dynamic range, they may not produce good tone mapped images when the input image has a limited dynamic range, which might influence the perception of artifacts. For example, if noise is present in a dark area, its intensity might be reduced by the compression of the dynamic range performed by the TMO. However, if the dynamic range is rather limited, this compression will be less and the noise might be more visible.

Results show that 'Chiu' resulted in 11 significantly different pairs out of 30 pairs, which means that this TMO is not suitable to assess the performance of HDR compression. As it was observed previously, this TMO produces worse tone mapped images, that were less pleasant and realistic, than the other TMOs. 'Ashikhmin', 'Fattal', 'Lischinski', and 'WardHistAdj' are the only TMOs that resulted in less than 6 significantly different pairs. These TMOs could be considered to assess the performance of HDR compression on LDR monitors. However, as the number of contents, bit rates, and subjects considered in this study are rather limited, it is impossible to draw general conclusion and recommend one particular TMO.

Table 7: Number of significantly different pairs between TMO and HDR.

TMO	Content					Average
	BloomingGorse2	*CanadianFalls*	*McKeesPub*	*MtRushmore2*	*WillyDesk*	
Ashikhmin	3	2	0	0	0	1.0
Chiu	4	3	2	1	1	2.2
Drago	3	2	1	1	0	1.4
Durand	1	3	0	1	1	1.2
Exponential	3	3	0	1	0	1.4
Fattal	0	3	0	0	0	0.6
Lischinski	0	2	0	0	1	0.6
Logarithmic	2	1	1	1	4	1.8
Reinhard	2	2	0	2	1	1.4
TumblinRushmeier	2	3	1	2	2	2.0
WardHistAdj	1	0	0	1	1	0.6
Average	1.7500	2.1667	0.4167	0.8333	0.9167	1.2167

5. CONCLUSION

In this paper, we investigated an approach to crowd-based quality assessment of high dynamic range content on low dynamic range displays by using tone mapping operators (eleven operators were studied) to reduce the dynamic range of the image. We conducted the crowdsourcing experiments using five HDR images encoded at four different bit rates with Profile A of the upcoming JPEG XT coding standard. The crowdsourcing results demonstrated that 'Ashikhmin', 'Fattal', 'Lischinski', and 'WardHistAdj' TMOs produce tone-mapped images that have high correlation with lab-based results and have high visual characteristics, which means these TMOs are the most suitable for the evaluation of HDR image compression. However, the number of contents and bit rates is too small to draw general conclusion and to recommend one particular TMO.

ACKNOWLEDGMENTS

This work has been conducted in the framework of the Swiss National Foundation for Scientific Research (FN 200021-143696-1), EC funded Network of Excellence VideoSense, Swiss SERI project Compression and Evaluation of High Dynamic Range Image and Video, COST IC1005 The digital capture, storage, transmission and display of real-world lighting HDRi, and COST IC1003 European Network on Quality of Experience in Multimedia Systems and Services QUALINET. The authors would like to thank Dolby Laboratories Inc. staff, and in particular Timo Kunkel, for providing Dolby Research HDR RGB backlight dual modulation display (aka 'Pulsar'). Help from Dion Tzamarias in generating the tone-mapped images used in subjective evaluations performed in this paper and scripts for analysis is also acknowledged.

REFERENCES

[1] Richter, T., "On the standardization of the JPEG XT image compression," in [*Picture Coding Symposium (PCS)*], 37–40 (December 2013).

[2] Narwaria, M., Perreira Da Silva, M., Le Callet, P., and Pepion, R., "On improving the pooling in HDR-VDP-2 towards better HDR perceptual quality assessment," in [*Human Vision and Electronic Imaging XIX*], **9014** (February 2014).

[3] Hossfeld, T., Keimel, C., Hirth, M., Gardlo, B., Habigt, J., Diepold, K., and Tran-Gia, P., "Best Practices for QoE Crowdtesting: QoE Assessment with Crowdsourcing," *IEEE Transactions on Multimedia* **16**(2), 541–558 (2014).

[4] Ashikhmin, M., "A tone mapping algorithm for high contrast images," in [*Proceedings of the 13th Eurographics Workshop on Rendering*], 145–156 (2002).

[5] Chiu, K., Herf, M., Shirley, P., Swamy, S., Wang, C., Zimmerman, K., et al., "Spatially nonuniform scaling functions for high contrast images," in [*Graphics Interface*], 245–245 (1993).

[6] Drago, F., Myszkowski, K., Annen, T., and Chiba, N., "Adaptive logarithmic mapping for displaying high contrast scenes," *Computer Graphics Forum* **22**(3), 419–426 (2003).

[7] Durand, F. and Dorsey, J., "Fast bilateral filtering for the display of high-dynamic-range images," *ACM Transactions on Graphics* **21**, 257–266 (July 2002).

[8] Fattal, R., Lischinski, D., and Werman, M., "Gradient domain high dynamic range compression," *ACM Transactions on Graphics* **21**, 249–256 (July 2002).

[9] Lischinski, D., Farbman, Z., Uyttendaele, M., and Szeliski, R., "Interactive local adjustment of tonal values," *ACM Transactions on Graphics* **25**, 646–653 (July 2006).

[10] Reinhard, E., Stark, M., Shirley, P., and Ferwerda, J., "Photographic tone reproduction for digital images," *ACM Transactions on Graphics* **21**, 267–276 (July 2002).

[11] Tumblin, J. and Rushmeier, H., "Tone reproduction for realistic images," *IEEE Computer Graphics and Applications* **13**, 42–48 (November 1993).

[12] Ward, G., "A contrast-based scalefactor for luminance display," *Graphics Gems IV* , 415–421 (1994).

[13] ITU-R BT.500-13, "Methodology for the subjective assessment of the quality of television pictures." International Telecommunication Union (January 2012).

[14] ITU-R BT.2022, "General viewing conditions for subjective assessment of quality of SDTV and HDTV television pictures on flat panel displays." International Telecommunication Union (August 2012).

[15] Keimel, C., Habigt, J., Horch, C., and Diepold, K., "QualityCrowd - A framework for crowd-based quality evaluation," in [*Picture Coding Symposium (PCS)*], (May 2012).

[16] Seetzen, H., Whitehead, L. A., and Ward, G., "A high dynamic range display using low and high resolution modulators," *SID Symposium Digest of Technical Papers* **34**(1), 1450–1453 (2003).

[17] Iwanami, T., Kikuchi, A., Kaneko, T., Hirai, K., Yano, N., Nakaguchi, T., Tsumura, N., Yoshida, Y., and Miyake, Y., "The relationship between ambient illumination and psychological factors in viewing of display images," in [*Color Imaging XIV: Displaying, Processing, Hardcopy, and Applications*], **7241** (January 2009).

[18] Kimoto, T. and Kato, C., "Novel evaluation of digital halftone image qualities by psychological analysis," *Advances in Image and Video Processing* **2**(2), 8–25 (2014).

[19] Barnard, G. A., "A new test for 2× 2 tables," *Nature* **156**, 177 (1945).