SUBJECTIVE AND OBJECTIVE EVALUATION OF HDR VIDEO COMPRESSION

Martin Řeřábek, Philippe Hanhart, Pavel Korshunov, and Touradj Ebrahimi

Multimedia Signal Processing Group (MMSPG),
Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

ABSTRACT

Recent efforts by MPEG and research community to standardize high dynamic range (HDR) video compression require efficient objective metrics. This paper investigates how well the currently available metrics measure perceptual quality of HDR video. For this purpose, subjective tests were conducted using SIM2 HDR monitor and several video sequences provided by MPEG. The sequences were compressed with HEVC encoder in YCbCr420 10bit and YCbCr444 12bit formats. The results of subjective tests were correlated with objective metrics to identify suitable metrics to use for the objective evaluations of different HDR video coding solutions. A few variants of PSNR, as well as SSIM, MSE, VIFP, and especially HDR-VDP-2 metric, have high correlation with the subjective scores.

1. INTRODUCTION

Members of MPEG have recently initiated efforts to identify suitable quality assessment objective metrics to evaluate the performance of coding solutions for HDR video content [1]. This joint effort is the first of its kind, since previous work focused mostly on evaluation of tone-mapping algorithms for images and video [2, 3], backward compatible compression solutions [4, 5], and the effect of HDR technology on viewing experience [6, 7]. A few studies of objective metrics performance exist but for HDR images only [8, 9, 10].

This paper describes the initial set of subjective experiments conducted at EPFL to assess a set of selected objective metrics in terms of their performance to predict subjective quality of compressed HDR video. The subjective experiments focused on evaluating one encoder at different bitrates. Seven anchor video sequences were provided for the tests. Each sequence was encoded in YCbCr420 10bit and in YCbCr444 12bit formats to investigate the influence of the sub-sampling on the objective metrics performance. Four operational points were selected in such a way that for each sequence, the highest bitrate corresponds to the transparent visual quality (visually indistinguishable from the uncompressed original content), second highest bitrate to a near transparent quality, lowest bitrate to obvious highly degraded quality, and one bitrate in-between.

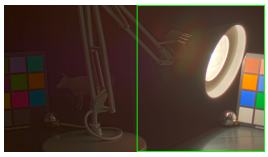
A SIM2 monitor was used in the subjective tests to display the content. Prior to the experiments, the monitor was calibrated to ensure linear transfer responses for each color channel. The provided video sequences were converted into OpenEXR frames, which in turn, were converted into specific bitmap format using a software provided by SIM2. Full paired comparison methodology was used in subjective visual quality evaluation experiments, with video pairs shown side-by-side on SIM2 HDR monitor. The resulted subjective raw scores were converted into corresponding Mean Opinion Score (MOS) values for each sequence using Thurstone Case V model. The results of objective metrics, provided to EPFL by participants of EE3 experiments, were plotted against MOS values with logistical fitting for each metric and each video sequence. Also, Pearson linear correlation coefficient (PCC), Spearman rank order correlation coefficient (SROCC), and root-mean-square error (RMSE) were computed for each objective metric.

2. SUBJECTIVE EVALUATION

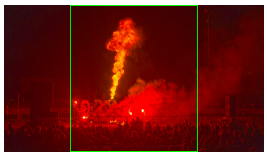
The experiments were conducted in the MMSPG test laboratory, which fulfills the recommendations for subjective evaluation of visual data issued by ITU-R [11]. The test room is equipped with a controlled lighting system with a 6500 K color temperature. The color of all the background walls and curtains present in the test area is mid grey. The laboratory setup is intended to ensure the reproducibility of the subjective tests results by avoiding unintended influence of external factors. In the experiments, the luminance of the background behind the monitor was about 20 cd/m2. The ambient illumination did not directly reflect off of the display. In every session, three subjects assessed the displayed test video content simultaneously. They were seated in one row perpendicular to the center of the monitor, at a distance of 3.2 times the picture height, as suggested in recommendation ITU-R BT.2022 [12].

This work has been conducted in the framework of the Swiss SERI project Compression and Evaluation of High Dynamic Range Image and Video, COST IC1003 European Network on Quality of Experience in Multimedia Systems and Services QUALINET, and FP7 EC EUROSTAR funded Project - Transcoders Of the Future TeleVision (TOFuTV).

(a) Balloon



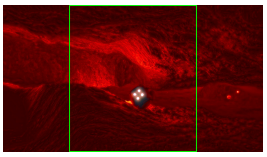
(b) Lux



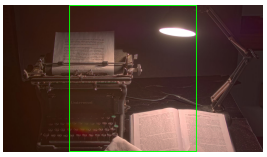
(c) FireEater2

(d) Market3

(e) PanoHD



(f) Tibul2



(g) Typewriter

Fig. 1: Frame examples of the dataset contents: training set (a-b) and testing set (c-g).

Sequence	fps	window		R4 QP	R4 bitrate [kbps] 420	R4 bitrate [kbps] 444	R3 QP	R3 bitrate [kbps] 420	R3 bitrate [kbps] 444	R2 QP	R2 bitrate [kbps] 420	R2 bitrate [kbps] 444	R1 QP	R1 bitrate [kbps] 420	R1 bitrate [kbps] 444
Balloon	25	836	1780	11	8887.90	9269.39	21	1035.14	1110.93	26	463.57	491.71	29	293.77	308.66
Lux	24	976	1920	5	1057.75	1119.45	22	145.88	155.58	26	97.50	102.89	33	48.03	50.04
FireEater2	25	400	1344	14	4612.85	5410.86	22	1494.72	1662.74	26	836.18	915.33	30	462.27	496.67
Market3	50	230	1174	19	10650.01	11216.34	25	4222.75	4381.97	28	2679.59	2759.06	30	1988.02	2034.45
PanoHD	60	488	1432	13	6333.07	6661.85	20	1978.88	2118.14	26	814.47	871.80	31	433.26	456.22
Tibul2	30	488	1432	15	10690.53	10117.50	23	2892.61	3118.48	29	936.12	1005.21	33	461.33	494.36
TypeWriter	24	488	1432	14	1057.15	1118.07	22	365.23	373.31	27	191.47	194.97	32	107.34	107.92

Table 1: Quality parameters (QPs) and bitrates of the training and testing sequences.

2.1. Dataset

The dataset used for the subjective evaluation tests consists of seven HD resolution HDR video sequences, namely, Balloon, Lux, FireEater2, Market3, PanoHD, Tibul2, and Typewriter. A typical frame example of each content in shown in Figure 1. Each content was encoded with HEVC using two different profiles: Anchor1 (based on the HEVC Main 10 profile, with 4:2:0 YCbCr chroma sampling and 10 bits per channel) and Anchor3 (based on the HEVC Main 12 profile, with 4:4:4 YCbCr chroma sampling and 12 bits per channel). All details about the anchor bitstreams generation are provided in [13]. Actual bitrates and quantization parameters (QPs) used in subjective tests for each content are listed in Table 1. Contents FireEater2, Market3, PanoHD, Tibul2, and Typewriter were used for the test, whereas the two remaining contents, Balloon and Lux, were used for training.

The subjective evaluation was performed on a SIM2 monitor, therefore, the provided bitstreams were first converted to OpenEXR format and then further converted to single 8 bit RGB 4:4:4 BMP files by using EasyHDRPlayer solution provided by SIM2 with default shader version 3.0 equations. Bitstreams were converted to OpenEXR using a proprietary solution from Technicolor in the following steps:

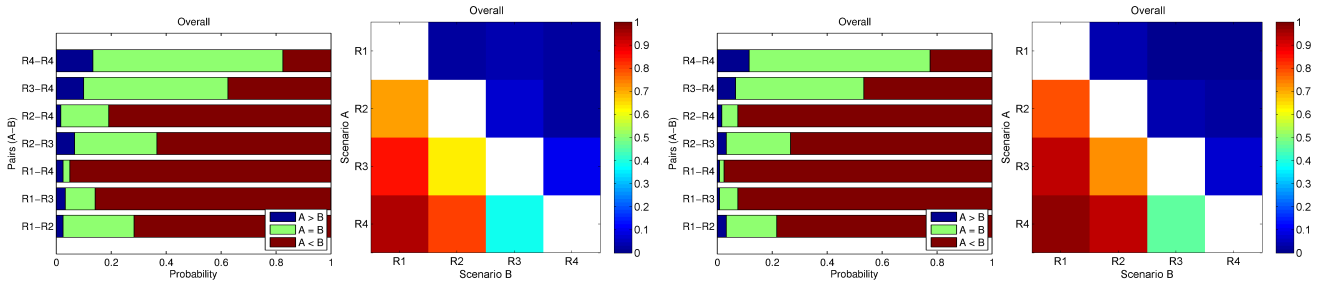
1. Decoding the bitstream by the HM,
2. Up-sampling from 420 to 444 (when needed),
3. Converting from YCbCr to R'G'B',
4. Applying the inverse TF to get RGB,
5. Writing the RGB in OpenExr files.

Each video sequence was cropped to 944×1080 pixels, so that the video sequences were presented side by side with a 32-pixels separating black border. The cropped part of each content is shown in Figure 1 with a green rectangle. The coordinates of the cropped window are given in Table 1.

2.2. Display characterization

To display the test stimuli, a full HD 47" SIM2 HDR LCD display with individually controlled LED backlight modulation was used. Prior to the subjective tests, the HDR display was calibrated using the EasySolarPro software provided by SIM2. The R, G, and B primaries were measured at 1400 nit level since the measurement probe (X-Rite i1Display Pro) is limited to up to 2000 nit. All measurements and color calibration were done after the display was switched on for more than an hour.

Note that the native SIM2 display sampling format for



(a) Preference probabilities

(b) Preference probability of choosing "Scenario A" over "Scenario B"

(c) Preference probabilities

(d) Preference probability of choosing "Scenario A" over "Scenario B"

Fig. 2: Overall results: YCbCr420 10b (a-b) and YCbCr444 12b (c-d).

HDR is 4:2:2. This fact could have an impact on the quality of displayed content encoded in HEVC Main 12 profile, 4:4:4 YCbCr. However, the influence of color sampling difference between native SIM2 display format and displayed sequences was not studied and is not part of this paper.

2.3. Methodology

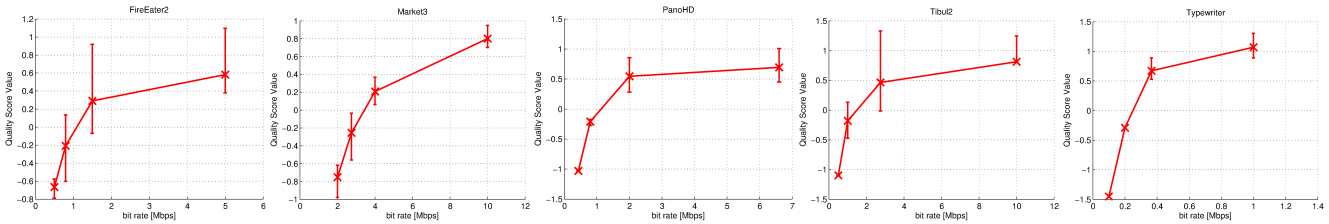
The full paired comparison evaluation methodology was selected for its high accuracy and reliability in constructing a scale of perceptual preferences. The video pairs were presented in side-by-side fashion to minimize visual working memory limitations. Subjects were asked to judge which video sequence in a pair ('left' or 'right') has better overall quality. The option 'same' was also included to avoid random preference selections. For each of the 5 contents, all the possible combinations of the 4 bitrates were considered, as well as an extra pair corresponding to R4 vs R4, i.e., 7 pairs for each content and color format, leading to a total of 70 paired comparisons for all contents.

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. 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 24 naïve subjects (8 females and 16 males) took part in the evaluation. They were between 19 and 28 years old with a mean of 23 years of age.

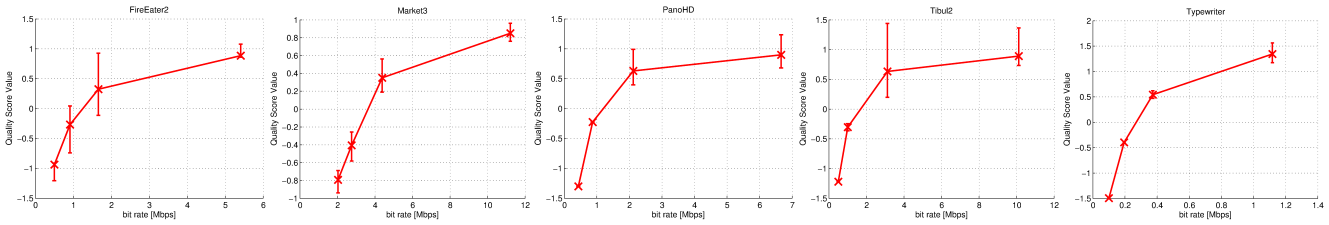
2.4. Results

Before estimating MOS values for paired comparison results, the winning w_{ij} and the tie t_{ij} frequencies were computed from the obtained subjective ratings for each pair of stimuli i and j. Note that $t_{ij} = t_{ji}$ and $w_{ij} + w_{ji} + t_{ij} = N$, where N is the number of subjects. This was done individually for each test video content and jointly over all contents. Figure 2 reports the preference probabilities and preference matrix computed over all video contents computed over all the contents. The preference probabilities graph is a histogram of the distribution of the scores for each individual pair. The preference matrix represents the probability that stimulus i is preferred over stimulus j, where i and j are the rows and columns of the matrix. Therefore, the preference matrix reports the winning frequency w_{ij} for each pair of stimuli i and j.

Then, the Thurstone Case V model was used to convert the ratings from the ternary scale to continuous-scale quality score values, which are equivalent to MOS, considering ties as being half way between the two preference options. The rating conversion from the ternary scale to continuous scale quality score values is described in [7]. The continuous scale quality score values can be interpreted as MOS, except that they are defined up to a scale factor and an offset. Therefore, there is no absolute relationship between two points. This means that the relative difference between two points is the most important and that two graphs corresponding to two different contents cannot be compared. All figures of MOS values demonstrate that the score value increases with the increase in peak luminance (see Figure 3). The quality score values tend to increase logarithmically. In most cases, the difference between individual bitrates seem to be significant, judging from confidence intervals that do not overlap, except for content FireEater2 and between rate points R3 and R4 for contents PanoHD and Tibul2. In most paired comparisons, higher bitrate was largely preferred and most ties occurred in pairs with consecutive rate points, i.e., R1-R2, R2-R3, and R3-R4. Considering the overall results for R4 vs R4, the option "same" was selected in about 70% of the pairs, which shows that there is a priori no statistically significant preference for one particular side of the display.



(a) YCbCr420 10b



(b) YCbCr444 12b

Fig. 3: Estimated MOS values from paired comparison subjective evaluations.

3. OBJECTIVE QUALITY METRICS

This session describes the results of correlations between perceived video quality and objective measurements. Objective metrics considered in this study are as follows:

- Metrics computed in linear domain
  - wPSNR-x: PSNR computed on x component
  - wPSNR_DEx: PSNR of mean of absolute value of deltaE2000 metric, derived with x as reference Y value
  - wPSNR_Lx: PSNR of mean square error of L component of the CIELab color space used for the deltaE2000 metric, derived with x as reference Y value
  - HDR-VDP-2
- Metrics computed in PQ-TF domain [14]
  - wtPSNR-x: PSNR computed on x component
- Metrics computed in perceptually uniform space [15]
  - MSE
  - SNR
  - SSIM
  - MS-SSIM
  - VIFP: VIF pixel based
- Metrics computed using multi-exposure [16]
  - wmfPSNRx0: mPSNR computed on x component

MSE, SNR, SSIM, MS-SSIM, and VIFP were computed using MeTriX MuX Visual Quality Assessment Package[1]. The MATLAB implementation of HDR-VDP-2[2] was used. Results for the other metrics were provided by MPEG members.

[1]MeTriX MuX v1.1: `http://foulard.ece.cornell.edu/gaubatz/metrix_mux/`

[2]HDR-VDP-2 v2.1.1: `http://hdrvdp.sourceforge.net/`

3.1. Performance indexes

Results of subjective tests can be used as ground truth to evaluate how well objective metrics estimate perceived quality. The result of execution of a particular objective metric is a video quality rating (VQR), which is expected to be an estimation of the MOS corresponding to the compressed HDR video. To be compliant with ITU recommendation for objective metrics performance evaluation [17], the following properties of the IQR estimation of MOS is considered: accuracy, monotonicity, and consistency. Consistency estimation is based on the confidence intervals, which are computed assuming a standard distribution of the subjective scores. In this paper, the Thurstone Case V model was used to convert the paired comparison ratings to equivalent MOS values [7]. For each content, the quality score values were converted to the range $[1, 5]$ by mapping the lowest and highest quality score values to 1 and 5, respectively, as the lower and upper bitrates were selected to be representative of the lowest and best quality (see Section 2.1), respectively. Intermediate values were scaled proportionally. Confidence intervals can be estimated from the paired comparison ratings, but their nature is different from that of confidence intervals computed directly on a discrete or continuous ratings scale. Therefore, only accuracy and monotonicity were considered. First, a regression was fitted to each [VQR, DMOS] data set using logistic fitting. Then, the Pearson linear correlation coefficient (PCC) and the root-mean-square error (RMSE) were computed to estimate accuracy of the VQR. To estimate monotonicity, the Spearman rank order correlation coefficient (SROCC) was computed. Detailed description of this process can be found in [9].

Metric	YCbCr 420 10b			YCbCr 444 12b		
	PLCC	SROCC	RMSE	PLCC	ROCC	RMSE
wPSNR-R	0.4602	0.4626	1.4042	0.4701	0.4993	1.3992
wPSNR-G	0.5330	0.5237	1.3381	0.5289	0.5405	1.3452
wPSNR-B	0.4872	0.4397	1.3811	0.5107	0.4642	1.3628
$wPSNR_{DE0100}$	0.4211	0.4062	1.4344	0.4386	0.4168	1.4247
$wPSNR_{L0100}$	0.5269	0.5085	1.3441	0.4956	0.4642	1.3768
$wPSNR_{DE1000}$	0.3559	0.3481	1.4779	0.3664	0.3191	1.4749
$wPSNR_{L1000}$	0.4957	0.4520	1.3735	0.4677	0.4184	1.4011
$wPSNR_{DE5000}$	0.3766	0.3649	1.4650	0.3958	0.2754	1.4556
$wPSNR_{L5000}$	0.5165	0.4596	1.3542	0.4930	0.4291	1.3791
HDR-VDP-2 Q_{MOS}	**0.9094**	**0.8887**	**0.6579**	**0.9099**	**0.9009**	**0.6576**
HDR-VDP-2 Q	**0.9128**	**0.8993**	**0.6458**	**0.9117**	**0.9070**	**0.6513**
wtPSNR-Y_{yuv}	**0.7659**	**0.7665**	**1.0168**	**0.7315**	**0.7482**	**1.0809**
wtPSNR-U	0.1478	0.2596	1.5641	0.4996	0.4194	1.3747
wtPSNR-V	0.6356	0.6169	1.2209	0.6095	0.6123	1.2576
wtPSNR-YUV	**0.7463**	**0.7222**	**1.0526**	**0.7219**	**0.7482**	**1.0971**
wtPSNR-R	0.6492	0.6398	1.2034	0.6247	0.6062	1.2383
wtPSNR-G	0.7873	0.8016	0.9750	0.7570	0.7802	1.0358
wtPSNR-B	0.7236	0.7070	1.0916	0.7610	0.765	1.0284
wtPSNR-RGB	**0.7305**	**0.7176**	**1.0800**	**0.7214**	**0.7070**	**1.0979**
wtPSNR-X	0.4771	0.4978	1.3899	0.4848	0.5115	1.3864
wtPSNR-Y_{xyz}	0.5331	0.5588	1.3379	0.5232	0.5436	1.3509
wtPSNR-Z	0.4892	0.4397	1.3793	0.5121	0.4642	1.3615
wtPSNR-XYZ	0.4656	0.4657	1.3996	0.4847	0.4840	1.3865
MSE	**0.7267**	**0.7237**	**1.0866**	**0.6912**	**0.6932**	**1.1460**
SNR	0.6392	0.6489	1.2162	0.6363	0.6611	1.2229
SSIM	0.6629	0.6276	1.1840	0.6117	0.6016	1.2540
MS-SSIM	**0.8392**	**0.7711**	**0.8600**	**0.8131**	**0.7619**	**0.9227**
VIFP	**0.7370**	**0.7070**	**1.0689**	**0.7240**	**0.7314**	**1.0934**
wmfPSNRY0	**0.7411**	**0.7405**	**1.0629**	**0.7252**	**0.7451**	**1.0929**
wmfPSNRU0	0.5141	0.5085	1.3566	0.4647	0.5008	1.4036
wmfPSNRV0	0.5011	0.3451	1.3685	0.4854	0.3279	1.3858
wmfPSNR0	0.6735	0.7054	1.1708	0.6579	0.6673	1.1964

Table 2: Accuracy and monotonicity indexes for different metrics.

3.2. Results

Table 2 reports the accuracy and monotonicity indexes, as defined in Section 3.1, for all evaluated metrics. The fitting was applied on all contents at once. Results show that HDR-VDP-2 outperforms all other metrics, with a PCC above 0.9 and a RMSE below 0.66. On the other hand, PSNR and deltaE2000 computed in linear domain are among the worst performing metrics, with a PCC and SROCC below 0.53. However, when PSNR is computed in PQ-TF domain, the correlation with subjective scores is higher, especially when considering the luma channel. These results show that computing PSNR on PQ values rather than linear values improves the correlation with subjective scores, as the perceptual quantizer developed by Miller *et al.* [14] considers luminance sensitivity of the human visual system.

To compute metrics designed for low dynamic range (LDR) on HDR content, previous works [8, 10] have shown that considering a perceptual space rather than the linear domain yields better results. Results show that the metrics computed in perceptually uniform space outperform the PSNR-based metrics computed in linear domain, especially for MS-SSIM, which shows the best performance in this case. Another alternative to compute LDR metrics is to consider multiple-exposure versions or different tone-mapped versions of the HDR content. The first approach is used by the mPSNR metric [16] and the performance is comparable to that of the metrics computed using a perceptually uniform space.

In general, results show that HDR-VDP-2, wtPSNR-Y_{yuv}, wtPSNR-YUV, wmfPSNRY0, wtPSNR-RGB, VIFP, MSE, and MS-SSIM are among the best metrics, with a PCC and SROCC above 0.7. Nevertheless, it is worth mentioning that RMSE is above 1 for most metrics, which indicates that the prediction error is quite high, even though correlation might be high. Performance is usually quite similar between YCbCr420 10bit and YCbCr444 12bit.

Figure 4 depicts the examples of scatter plots of subjective versus objective results. As it can be observed, most objective metric show quite strong content dependency, which explains the low performance when considering all contents at once. To determine whether the difference between two performance index values corresponding to two different metrics is statistically significant, a statistical test was performed as in [9]. Results show that HDR-VDP-2 is statistically better than most metrics, but there are few other significant differences, as there are few (20) data points.

4. CONCLUSION

This paper presented the results of subjective evaluations and their correlation with objective metrics of five HDR video sequences in YCbCr420 10bit and YCbCr444 12bit formats compressed with HEVC encoder at four different encoding operational points. Subjective results show that the perceptual quality increases monotonously for all contents with the increasing bitrate. Generally, for YCbCr444 format, the perceptual quality increases more rapidly. For Market3 and Typewriter sequences, the increase in perceptual quality (with the increasing bitrate) is more aggressive compared to other contents.

Correlation with objective metrics shows that HDR-VDP-2 metric has the highest correlation with subjective scores with PLCC and SROCC coefficients near 0.9. A few variants of PNSR, as well as VIFP, MSE, and MS-SSIM metrics show acceptable correlation results with PLCC and SROCC coefficients close to 0.8.

For the future work, a more extensive subjective experiments and a larger number of objective metrics are needed for a clearer understanding which metrics are the most suitable for the quality evaluation of HDR video encoding.

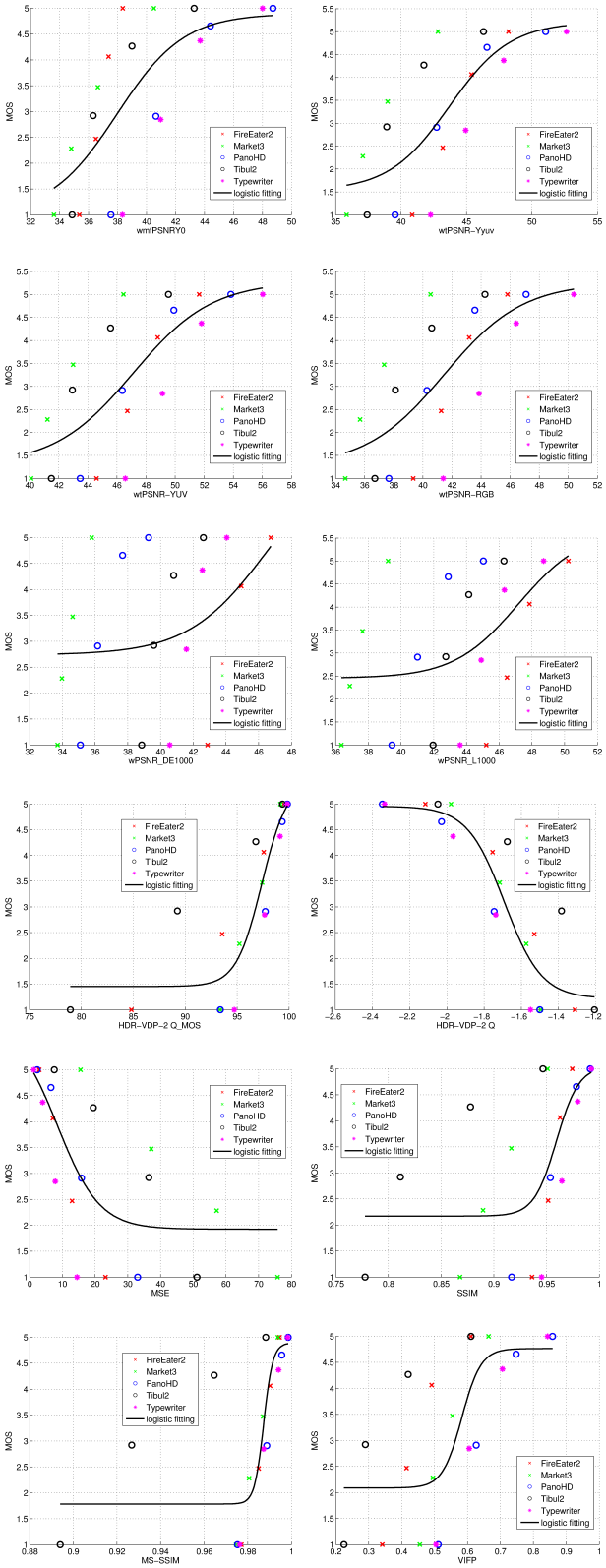


Fig. 4: Subjective results versus objective measure for YCbCr420 10bit sequences.

5. REFERENCES

[1] ISO/IEC JTC1/SC29/WG11 MPEG, "Exploration experiments for HDR and Wide Gamut Content Distribution," Doc. N14549, Saporo, Japan, July 2014.

[2] M. Čadík, M. Wimmer, L. Neumann, and A. Artusi, "Evaluation of HDR tone mapping methods using essential perceptual attributes," *Computers & Graphics*, vol. 32, no. 3, pp. 330–349, 2008.

[3] B. Annighöfer, T. Tajbakhsh, and R.-R. Grigat, "Prediction of results from subjective evaluation of real-time-capable tone-mapping operators applied to limited high-dynamic-range images," *Journal of Electronic Imaging*, vol. 19, no. 1, January 2010.

[4] P. Korshunov and T. Ebrahimi, "Context-dependent JPEG backward-compatible high-dynamic range image compression," *Optical Engineering*, vol. 52, no. 10, September 2013.

[5] R. Mantiuk, G. Krawczyk, K. Myszkowski, and H.-P. Seidel, "High Dynamic Range Image and Video Compression - Fidelity Matching Human Visual Performance." International Conference on Image Processing (ICIP), September 2007.

[6] P. Ledda, A. Chalmers, T. Troscianko, and H. Seetzen, "Evaluation of tone mapping operators using a High Dynamic Range display," in *SIGGRAPH*, August 2005.

[7] P. Hanhart, P. Korshunov, and T. Ebrahimi, "Subjective evaluation of higher dynamic range video," in *Proc. SPIE 9217*. Applications of Digital Image Processing XXXVII, August 2014.

[8] G. Valenzise, F. De Simone, P. Lauga, and F. Dufaux, "Performance evaluation of objective quality metrics for HDR image compression," in *Proc. SPIE 9217*. Applications of Digital Image Processing XXXVII, August 2014.

[9] P. Hanhart, M. Bernardo, P. Korshunov, M. Pereira, A. Pinheiro, and T. Ebrahimi, "HDR image compression: a new challenge for objective quality metrics," in *Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*, September 2014.

[10] C. Mantel, S. Ferchiu, and S. Forchhammer, "Comparing subjective and objective quality assessment of HDR images compressed with JPEG-XT," in *International Workshop on Multimedia Signal Processing (MMSP)*, September 2014.

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

[12] 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.

[13] ISO/IEC JTC1/SC29/WG11 MPEG, "Test sequences and anchor generation for HDR and Wide Gamut Content Distribution," Doc. N14548, Saporo, Japan, July 2014.

[14] S. Miller, M. Nezamabadi, and S. Daly, "Perceptual Signal Coding for More Efficient Usage of Bit Codes," in *SMPTE Conferences*, vol. 2012, no. 10, 2012, pp. 1–9.

[15] T. O. Aydın, R. Mantiuk, and H.-P. Seidel, "Extending quality metrics to full luminance range images," in *Proc. SPIE 6806*. Human Vision and Electronic Imaging XIII, January 2008.

[16] J. Munkberg, P. Clarberg, J. Hasselgren, and T. Akenine-Möller, "High dynamic range texture compression for graphics hardware," *Transactions on Graphics (TOG)*, vol. 25, no. 3, pp. 698–706, 2006.

[17] ITU-T P.1401, "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," International Telecommunication Union, July 2012.