Assessment of quality of JPEG XL proposals based on subjective methodologies and objective metrics

Pinar Akyazi, Touradj Ebrahimi

Pinar Akyazi, Touradj Ebrahimi, "Assessment of quality of JPEG XL proposals based on subjective methodologies and objective metrics," Proc. SPIE 11137, Applications of Digital Image Processing XLII, 111370N (6 September 2019); doi: 10.1117/12.2530196

SPIE.

Event: SPIE Optical Engineering + Applications, 2019, San Diego, California, United States

Assessment of quality of JPEG XL proposals based on subjective methodologies and objective metrics

Pinar Akyazi and Touradj Ebrahimi

Multimedia Signal Processing Group (MMSPG)
Ecole Polytechnique Fédérale de Lausanne
CH 1015, Lausanne, Switzerland

ABSTRACT

The Joint Photographic Experts Group (JPEG) is currently in the process of standardizing JPEG XL, the next generation image coding standard that offers substantially better compression efficiency than existing image formats. In this paper, the quality assessment framework of proposals submitted to the JPEG XL Call for Proposals is presented in details. The proponents were evaluated using objective metrics and subjective quality experiments in three different laboratories, on a dataset constructed for JPEG XL quality assessment. Subjective results were analyzed using statistical significance tests and presented with correlation measures between the results obtained from different labs. Results indicate that a number of proponents superseded the JPEG standard and performed at least as good as the state-of-the-art anchors in terms of both subjective and objective quality on SDR and HDR contents, at various bitrates.

Keywords: JPEG XL, image compression, standardization, objective quality assessment, subjective quality assessment.

1. INTRODUCTION

Joint Photographic Experts Group (JPEG) is a sub-group of ISO/IEC Joint Technical Committee 1, Subcommittee 29, Working Group 1 (ISO/IEC JTC1/SC29/WG1). The group has created the image coding standard JPEG more than two decades ago, as well as many other standards including JPEG 2000, JPEG XS, JPEG XT and JPEG XR. While maintaining the previous standards, JPEG is active in development of novel coding standards in multimedia. In 2017, JPEG published a Call for Proposals for creating the next generation image coding standard, referred to as JPEG XL.[1] The Call targeted development of a standard for image coding that offers substantially better compression efficiency than existing image coding formats (e.g. $> 60\%$ over JPEG), along with features desirable for web distribution and efficient compression of high-quality images.

The final Call for Proposals was issued in April 2018 with a deadline for expression of interest and registration in August 2018. Submissions were gathered in September 2018 and the performance of proponents was evaluated via subjective and objective quality assessment tests, following the Recommendations in ITU-R BT.2022,[2] ITU-R BT.500-13[3] and ITU-T P.910.[4] A total of seven proponents were compared to four anchors at eight different bitrates during objective and four different bitrates during subjective evaluation. All proponents were evaluated using standard dynamic range (SDR) contents and high dynamic range (HDR) contents at various resolutions with different characteristics. This paper aims to describe the methodology and results of the quality assessment of JPEG XL, which were first published as an output document of the 81st JPEG meeting, Vancouver, Canada, 13-19 October 2018.

2. BACKGROUND

Billions of images are captured, created, uploaded, and shared daily, which creates an immediate need for efficient image compression. Applications are becoming increasingly image-rich, and websites and user interfaces (UIs) rely on images for sharing experiences and stories, visual information and appealing design. On the low end of the spectrum, UIs can target devices with stringent constraints on network connection and/or power consumption. Even though network download speeds are improving globally, in many situations bandwidth is constrained

Further author information: (Send correspondence to authors) E-mail: firstname.lastname@epfl.ch

Applications of Digital Image Processing XLII, edited by Andrew G. Tescher, Touradj Ebrahimi, Proc. of SPIE
Vol. 11137, 111370N · © 2019 SPIE · CCC code: 0277-786X/19/$21 · doi: 10.1117/12.2530196

to speeds that inhibit responsiveness in applications. On the high end, UIs utilize images that have larger resolutions, higher dynamic range and wider color gamut, as well as higher bit depths, which leads to a further explosion of image data.[5]

Standards such as JPEG and PNG are still widely used as the primary coding formats, however, reduced network transmission times are needed for more interactive applications. With websites and UIs containing up to hundreds of images or several high-resolution images, the accumulated data can amount up to several megabytes worth, which could be equivalent to more than a minute of video. While video streams can be buffered before playback, image-based UIs have to be responsive and interactive, without several seconds of loading and stalling when downloading or scrolling.

Recently, evidence has been presented of compression technologies that outperform image coding standards in common use.[6,7] Several metrics showed the HEVC/H.265 HM encoder with SCC extensions[8] to be superior according to most metrics, and for most test images. Subjectively, Daala[9] was competitive, with a limited difference in MOS scores. Although there is evidence of technical advances, there is no widespread standard that has state-of-the-art compression performance, and is widely supported in consumer devices and browsers.

The goal of JPEG XL is to develop a new image coding standard that provides state-of-the-art image compression performance, and that addresses shortcomings in current standards. The activity aims to (i) achieve significant compression efficiency improvement over coding standards in common use at equivalent subjective quality, e.g. $> 60\%$ over JPEG, (ii) offer features for web applications, such as support for alpha channel coding and animated image sequences, and (iii) offer support of high-quality image compression, including higher resolution, higher bit depth, higher dynamic range, very high quality and wider color gamut coding. To encourage widespread adoption, an important goal for this standard is to support a royalty-free baseline.

JPEG XL targets a large variety of use cases including image-rich UIs and web pages on bandwidth-constrained connection such as social media applications, media distribution applications, cloud storage applications, media web sites, animated image applications, mobile applications and games, and high quality imaging applications such as rapid photo viewing, HDR/WCG user interfaces, augmented/virtual reality, image bursts, high-end photography, image mosaics, depth images, and printing. The requirements of the final call for proposals are depicted in Tables 1-3.

Table 1: Necessary attributes of uncompressed images that the targeted image coding standard is expect to support.

Support uncompressed images with attributes
Image resolution: from thumbnail-size images up to at least 40 MP images.
Transfer functions including those listed in BT.709[10] and BT.2100[11]
Bit depth: 8-bit and 10-bit
Color space: at least RGB, YCbCr, ICtCp. • Input type of the encoder shall match output type of the decoder. • Internal color space conversion is permitted (as part of the proposal).
Color primaries including BT.709[10] and BT.2100.[11]
Chrominance subsampling (where applicable): 4:0:0, 4:2:0, 4:2:2, and 4:4:4.
Different types of content, including natural, synthetic, and screen content.

In addition to the requirements in Tables 1-3, the proposals were expected to include a high-level description of the submission including block diagrams of encoder and decoder, as well as arguments on why the proposal is meeting the requirements. Binary encoder/decoder executables and scripts, encoded-decoded materials and results, algorithm and design description, technical documentation and complexity analysis were submitted with the proposals. The submissions were then evaluated by researchers at three different universities, namely, Ecole Polytechqnique Fédérale de Lausanne (EPFL), Vrije Universiteit Brussel (VUB) and Télécom ParisTech (TPT), using the contents and methodologies defined in the Call.

Table 2: Compressed bitstream requirements that submissions are required to cover.

Core requirements
Significant compression efficiency improvement over coding standards in common use at equivalent subjective quality.
Hardware/software implementation-friendly encoding and decoding (in terms of parallelization, memory, complexity, power consumption)
Support for alpha channel / transparency coding.
Support for animation image sequences.
Support for 8-bit and 10-bit bit depth.
Support for high dynamic range coding.
Support for wide color gamut coding.
Support for efficient coding of images with text and graphics.

Table 3: Desirable bitstream requirements that submissions are encouraged to cover.

Desirable requirements
Support for higher bit depth (e.g. 12 to 16-bit integer or floating-point HDR) images.
Support for different color representations, including Rec. BT.709, ITU-R BT.2022,[2] Rec. BT.2100, LogC.
Support for embedded preview images.
Support for very low file size image coding (e.g. ¡200 bytes for 6464 pixel images).
Support for lossless alpha channel coding.
Support for a low-complexity profile.
Support for region-of-interest coding.

3. QUALITY ASSESSMENT FRAMEWORK

3.1 Anchor Generation

The proposals were evaluated against four anchors: JPEG,[12] JPEG 2000,[13] HEVC/H.265[14] and WebP[15] (only for 8-bit SDR contents). At the evaluation time, a specific reference implementation was chosen for each anchor, as follows: JPEG XT reference software (v1.53) for JPEG, Kakadu (v7.10.2) for JPEG 2000, HM-16.18+SCM-8.7 for HEVC/H.265 and cwebp 1.0.0 for WebP. Each selected content was encoded by all anchors using 8 target bitrates in the list $[0.06, 0.12, 0.25, 0.50, 0.75, 1.00, 1.50, 2.00]$bpp. The proponents were required to target the same bitrates over all contents, both for SDR and HDR data. The bitrate range was selected extensive enough to investigate all rate-distortion scenarios from very low rates corresponding to very low image quality to very high bitrates corresponding to transparent image quality. Configurations and command lines for anchor generation are given in Table 4. The 12-bit setting was used for HDR images. SDR and HDR command lines were different only for HEVC/H.265.

All input images were in RGB 4:4:4 format. The conversions were handled using HDRConvert[16] with the following command line:

HDRConvert -f HDRConvertBT709PPMToYCbCr420fr.cfg -p SourceFile=<RGB444_input>
-p SourceWidth=<width> -p SourceHeight=<height> -p OutputFile=<YCbCr420_input>
-p OutputWidth=<width> -p OutputHeight=<height> p SourceBitDepthCmp0=<bit_depth>
-p SourceBitDepthCmp1=<bit_depth> -p SourceBitDepthCmp2=<bit_depth>
-p OutputBitDepthCmp0=<bit_depth> -p OutputBitDepthCmp1==<bit_depth>
-p OutputBitDepthCmp2=<bit_depth> -p OutputChromaFormat=1

JPEG XT accepts only 4:4:4 chroma format and the subsampling to 4:2:0 is executed internally. For JPEG XT, the parameter OutputChromaFormat was set to 3 instead of 1. The codebase for anchor generation is available online,[17] including the configuration files used during conversions. To ensure reproducibility of results and ease the objective assessment of different proposals, a Docker container was created to automatically perform the objective assessment of a given set of codecs. The container (i) automatically downloads and configures all

anchor codecs, metrics and dependencies, (ii) allows easy addition of new (proprietary) codecs by placing binaries and Python encoder/decoder scripts in the designated folder, (iii) allows testing new contents, (iv) includes all running encoding, decoding, and objective evaluation scripts, and (v) automatically generates performance curves of objective results.

Table 4: Selected parameters and settings for the anchors.

Anchor	Software	Input format	Command line
JPEG	JPEG XT v1.53	RGB 4:4:4 8-bit	jpeg -qt 3 -h -v -oz -q <qp> -s 1x1,1x1,1x1 <input> <output>
		RGB 4:4:4 10-bit	jpeg -qt 3 -g 1 -h -v -oz -q <qp> -R 2 -s 1x1,1x1,1x1 <input> <output>
		RGB 4:4:4 12-bit	jpeg -qt 3 -g 1 -h -v -oz -q <qp> -R 4 -s 1x1,1x1,1x1 <input> <output>
		YCbCr 4:2:0 8-bit	jpeg -qt 3 -h -v -c -oz -q <qp> -s 1x1,2x2,2x2 <input> <output>
		YCbCr 4:2:0 10-bit	jpeg -qt 3 -g 1 -h -v -c -oz -q <qp> -R 2 -s 1x1,2x2,2x2 <input> <output>
		YCbCr 4:2:0 12-bit	jpeg -qt 3 -g 1 -h -v -c -oz -q <qp> -R 4 -s 1x1,2x2,2x2 <input> <output>
JPEG 2000	Kakadu v7.10.2	RGB 4:4:4 8/10/12-bit	kdu_compress -i <input> -o <output> rate <bpp>
		YCbCr 4:2:0 8/10/12-bit	kdu_v_compress -i <input> -o <output> rate <bpp> -precise -tolerance 0
HEVC	HM-16.18+SCM-8.7	RGB 4:4:4 8/10-bit YCbCr 4:2:0 8/10-bit	TAppEncoderStatic -c encoder_intra_main_scc.cfg -f 1 -fr 1 -q <qp> -wdt <width> -hgt <height> –InputChromaFormat=<chroma_format> –InternalBitDepth=<bit_depth> –InputBitDepth=<bit_depth> –OutputBitDepth=<bit_depth> –ConformanceWindowMode=1 -i <input> -b <output> -o /dev/null
		RGB 4:4:4 12-bit YCbCr 4:2:0 12-bit	TAppEncoderStatic -c encoder_intra_main_rext.cfg -f 1 -fr 1 -q <qp> -wdt <width> -hgt <height>–InputChromaFormat=<chroma_format> –InternalBitDepth=<bit_depth> –InputBitDepth=<bit_depth> –OutputBitDepth=<bit_depth> –ConformanceWindowMode=1 -i <input> -b <output> -o /dev/null
		RGB 4:4:4 12-bit HDR YCbCr 4:2:0 12-bit HDR	TAppEncoderStatic -c encoder_intra_main_rext.cfg -f 1 -fr 1 -q <qp> -wdt <width> -hgt <height>–InputChromaFormat=<chroma_format> –InternalBitDepth=<bit_depth> –InputBitDepth=<bit_depth> –OutputBitDepth=<bit_depth> –ConformanceWindowMode=1 InputColourSpaceConvert=RGBtoGBR -i <input> -b <output> -o /dev/null
WebP	cwebp 1.0.0	YCbCr 4:2:0 8-bit	cwebp -m 6 -q <qp> -s <width> <height> <depth> <input> -o <output>

3.2 Objective Quality Assessment

Objective quality assessment was carried out over all 8 bitrates for all codecs, in YCbCr color space. The RGB 4:4:4 outputs were converted to YCbCr 4:4:4 using HDRConvert. Selected metrics for objective quality assessment of SDR contents were PSNR, SSIM, MS-SSIM for all contents, with the addition of VIF and VMAF only for 8-bit contents. The first three metrics have been computed using HDRMetrics[16] whereas the VMAF FFmpeg plugin is used for the last two. PSNR is computed on Y channel, and by averaging the PSNR over separate channels.

For HDR contents, PQ-PSNR-Y and PQ-MS-SSIM-Y metrics were computed using HDRMetrics. In order to carry out objective evaluation on HDR images, inverse PQ transfer function was applied first, leading to 12-bit PQ-RGB 4:4:4 images to obtain linear RGB images. Then, following color space conversion from linear RGB to XYZ, PQ transfer function was applied to Y component and PSNR and MS-SSIM metrics were computed on the Y component only. All command lines are provided in Table 5 and the configuration files are available online.[17]

3.3 Subjective Quality Assessment

The Double Stimulus Impairment Scale (DSIS) Variant I^2 was the test methodology selected for subjective quality assessment. In this test, the stimulus under assessment and the reference are presented simultaneously to the subject. The subject is asked to rate the degree of annoyance of the visual distortions in the stimulus under

Table 5: Command lines for objective metric computations.

DR	Metric	Software	Command line
SDR	PSNR, SSIM, MS-SSIM	HDRMetrics	HDRMetrics -f HDRMetrics.cfg -p Input0File=<reference> -p Input1File=<decoded> -p LogFile=<log_file> -p NumberOfFrames=1 -p Input0Width=<width> -p Input0Height=<height> -p Input1Width=<width> -p Input1Height=<height> -p TFPSNRDistortion=0 -p EnablePSNR=1 -p EnableSSIM=1 -p EnableMSSSIM=1
	VMAF, VIF	FFmpeg	ffmpeg -s:v <width>,<height> -i <decoded> -s:v <width>,<height> -i <reference> -lavfi libvmaf=log_fmt=json:log_path=<log_file> -f null -
HDR	PQ-PSNR-Y PQ-MSSSIM-Y	HDRMetrics	HDRMetrics -f HDRMetrics.cfg -p Input0File=<reference> -p Input1File=<decoded> -p LogFile=<log_file> -p NumberOfFrames=1 -p Input0Width=<width> -p Input0Height=<height> -p Input1Width=<width> -p Input1Height=<height> -p TFPSNRDistortion=1 -p EnablePSNR=1 -p EnableMSSSIM=1

assessment with respect to the reference. The degree of annoyance is divided into five different levels labeled as Very annoying, Annoying, Slightly annoying, Perceptible but not annoying and Imperceptible, corresponding to a quality scale ranging from 1 to 5, respectively.

3.3.1 Content and rate point selection

Subjective tests are costly in terms of time and effort. In order to be able to generalize the results, a minimum of 15 subjects are needed to participate in the experiment. The duration of each experiment depends on the number of contents to be evaluated. To balance this trade-off, selection of contents and rate points has to be handled meticulously.

The content and rate selection was carried out during expert viewing sessions prior to setting up the experiments, both for SDR and HDR tests. All contents in the dataset were encoded using the anchor software and the decoded images were viewed by experts. In order to obtain meaningful results from the experiments, the selected rate points needed to span a range that covers very low to high bitrates, corresponding to very low to transparent visual quality. The anchor with the best performance, i.e. HEVC/H.265, was used to select such rate points and the selection was verified using other anchors. The contents that were too difficult to examine by naive subjects were excluded from the selection.

3.3.2 Data preparation

Selected contents were then processed according to the DSIS framework. For SDR and HDR tests, a 30 inch Eizo 10bit ColorEdge CG301W monitor with a resolution of 4096×2160 and a Sim2 HDR47ES4MB display with 1920×1080 resolution were used by all participating labs, respectively. SDR and HDR stimuli were cropped using FFmpeg[18] to fit their respective screen resolutions. The region to be cropped for each stimulus was determined during expert viewing. Each decoded stimulus was placed side by side with its reference, with a 20 pixel mid-gray colored separation in between. The side-by-side stimuli were then displayed in front of the same mid-gray colored background, and were randomized such that the same content was never presented consecutively.[2] Two dummy sequences were included in each test, about which the subjects were not informed. A training session was conducted for each subject prior to the experiment, during which three stimuli were presented as examples for the two extremes of the voting scale, i.e. Very annoying (1) and Imperceptible (5), along with an example in the middle, i.e. Slightly annoying (3). For half of the subjects the reference was placed at the right side of the screen, whereas for the other half it was placed on the left to avoid position bias. Each experiment was conducted in two sessions to prevent subject fatigue. The monitors were calibrated using an i1 DisplayPro color calibration device according to the guidelines described in.[2,11] Same guidelines were followed to set up the controlled environment for viewing with a mid gray level background inside the test rooms.

During both SDR and HDR tests, viewing time was not restricted. Subjects, however, were instructed to vote within reasonable time for the experiments to proceed smoothly. No viewing distance or position was specified for the SDR tests. On the other hand, HDR tests were conducted with a fixed distance from the screen as instructed by ITU-R BT.2100. Figure 1 depicts the test environment for SDR and HDR experiments.

(a)

(b)

Figure 1: Consenting subjects during SDR (a) and HDR (b) subjective quality assessment tests conducted at EPFL.

4. EXPERIMENTS AND RESULTS

Seven proposals were submitted that fit the requirements of the Call, of which 4 supported bit depths larger than 8, and HDR images. 18 and 20 subjects participated to the subjective quality assessment experiments of the SDR contents in EPFL and VUB. 18, 20 and 17 subjects participated to the subjective quality assessment experiments of the HDR contents in EPFL, VUB and TPT, respectively. A standard outlier detection was performed on all sets of raw scores to remove subjects whose ratings deviated strongly from others.[3] None of the subjects were identified as outliers in our experiments.

The mean opinion score (MOS) and 95% confidence intervals (CIs) assuming a Student's t-distribution of the scores were computed for each test condition.[19] To determine and compare the differences among MOS obtained for different codecs and bitrates, a one-sided Welch test at 5% significance level was performed on the scores. Bitrates that deviated more than 10% from the target bitrates were excluded from statistical significance tests.

4.1 Dataset

The specifications of the full dataset are given in Table 6. Resolutions of SDR images varied from SD to UHD whereas HDR images had HD resolution. SDR color images had BT.709 primaries whereas HDR images had BT.2020 primaries. Contents selected for subjective quality assessment experiments for SDR and HDR tests are presented in Figures 2 and 3, respectively.

Table 6: Distribution of full set of contents.

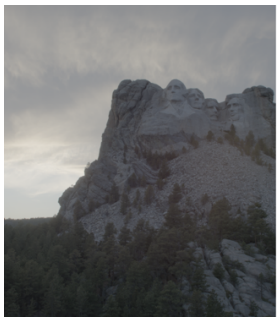
Class	Description	Bit depth	Number of contents
A	Natural images (RGB 4:4:4)	8-bit 10-bit	23 10
B	Grayscale images (4:0:0)	8-bit	4
C	Computer generated images (RGB 4:4:4)	8-bit 10-bit 12-bit	1 1 1
D	Screen content images (RGB 4:4:4)	8-bit	3
E	HDR/WCG images (RGB 4:4:4)	12-bit	24
All			**67**

(a) Training (b) Arri (c) Apple

(d) Bike (e) Cafe (f) Fly

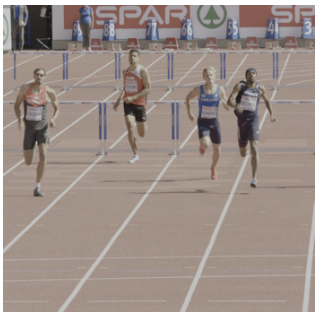
(g) p06 (h) Blender (10-bit) (i) Woman

Figure 2: Thumbnails of SDR contents selected for subjective quality assessment, after cropping for DSIS experiments. All contents have 8-bit depth except for Blender.



(a) Training

(b) 507



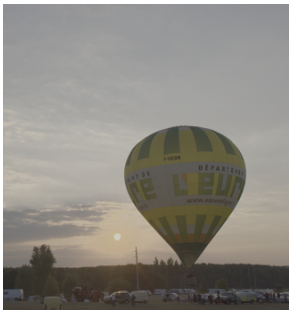
(c) Hurdles

(d) Kitchen

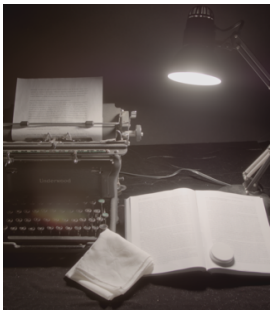
(e) Market

(f) Showgirl

(g) Sintel



(h) Sunrise

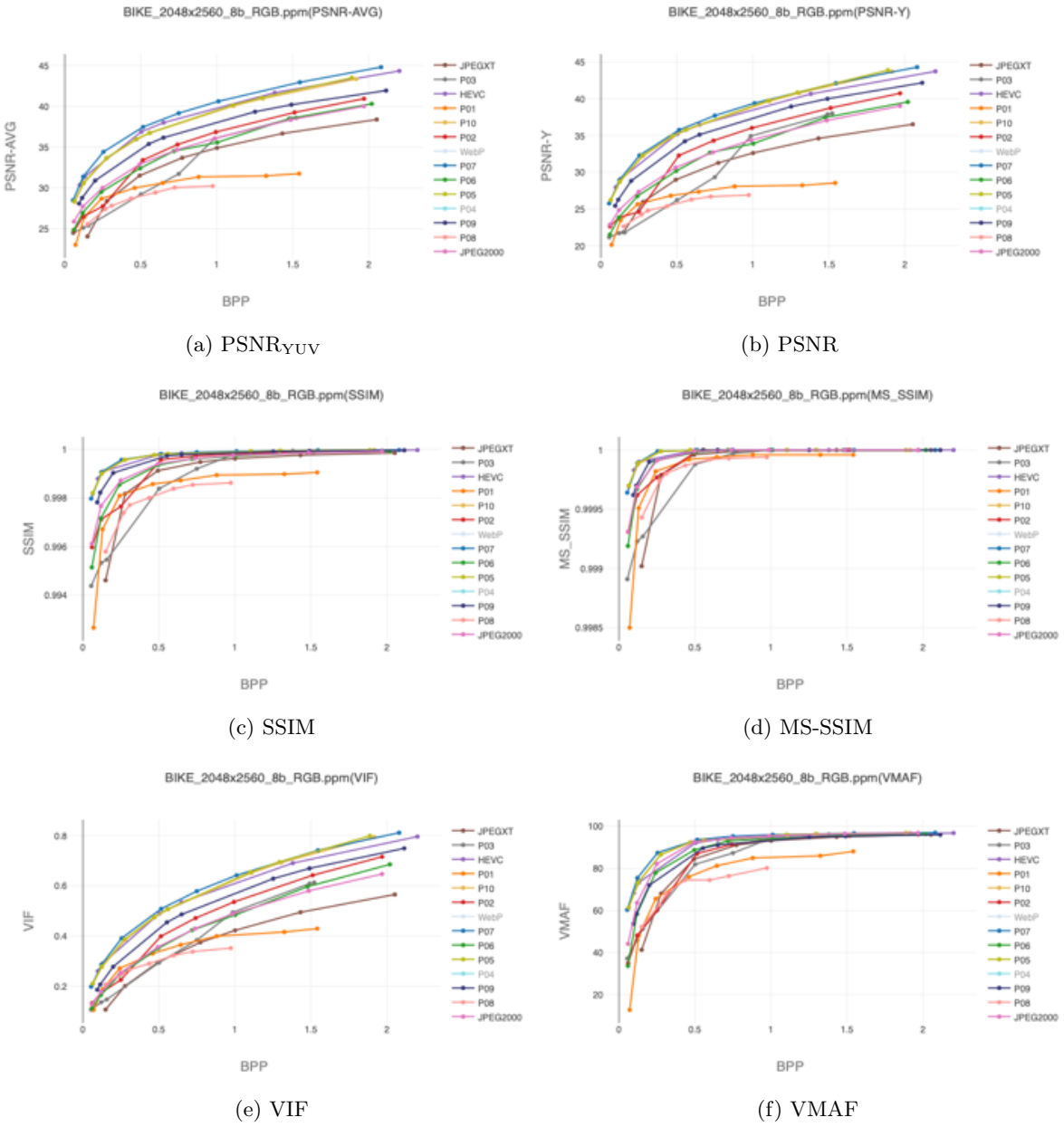


(i) Typewriter

Figure 3: Thumbnails of HDR contents selected for subjective quality assessment, after cropping for DSIS experiments. Linear RGB thumbnails are included here only for demonstration.

4.2 Objective Quality Assessment Results

Objective quality assessment was performed on all 67 contents listed in Table 6, at all 8 bitrates, for all proponents and anchors. Interactive plots were generated using the scripts online. For demonstrative purposes, the objective quality assessment results of the contents Bike in Figure 2d and 507 in Figure 3b are presented in Figures 4 and 5. All objective quality assessment results are stored in .json format along with interactive plots. The data is available online and can be accessed by contacting the authors.

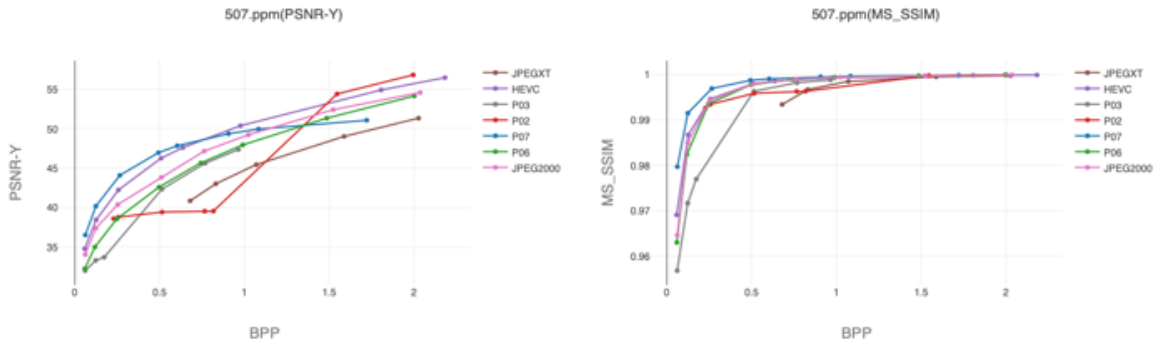


(a) PSNR_{YUV} (b) PSNR

(c) SSIM (d) MS-SSIM

(e) VIF (f) VMAF

Figure 4: Objective results for the SDR content Bike (Figure 2d. Results for codecs accepting RGB 4:4:4 as native format are included in the objective comparison.



(a) PSNR (b) MS-SSIM

Figure 5: Objective results for the HDR content 507 (Figure 3b).

4.3 Subjective Quality Assessment Results

Subjective quality assessment was performed on the selected contents at the screened out bitrates given in Table 7, for all proponents and anchors. The MOS vs. bitrate plots and comparisons between pairwise conditions are presented in Figures 6-9.

Table 7: Original resolutions, classes and selected bitrates for subjective quality assessment of SDR contents.

Name	Class	Resolution	Bitrates
Arri	A	2880 × 1620	[0.06, 0.12, 0.25, 0.50]
Apple	D	2560 × 1440	[0.06, 0.12, 0.25, 0.50]
Bike	A	2048 × 2160	[0.06, 0.12, 0.25, 0.50]
Cafe	A	1280 × 1600	[0.06, 0.12, 1.00, 2.00]
Fly	A	1920 × 1080	[0.06, 0.12, 0.25, 0.50]
p06	A	4064 × 2704	[0.06, 0.12, 0.25, 0.50]
Blender	C	4096 × 1744	[0.06, 0.12, 0.25, 0.50]
Woman	A	2048 × 2560	[0.06, 0.12, 0.25, 0.50]
507	E	944 × 1080	[0.06, 0.12, 0.50, 1.00]
Hurdles	E	1920 × 1080	[0.50, 0.75, 1.00, 2.00]
Kitchen	E	944 × 1080	[0.06, 0.12, 0.25, 0.75]
Market	E	1920 × 1080	[0.75, 1.00, 1.50, 2.00]
Showgirl	E	944 × 1080	[0.75, 1.00, 1.50, 2.00]
Sintel	E	944 × 1080	[0.75, 1.00, 1.50, 2.00]
Sunrise	E	1920 × 1080	[0.50, 0.75, 1.00, 2.00]
Typewriter	E	944 × 1080	[0.75, 1.00, 1.50, 2.00]

Throughout SDR and HDR contents, the proponents P03, P06 and P07 were performing as good as, and even better, than the state-of-the-art codecs. The performance of P01 and P05 were also ample, however, these codecs did not support images with bit depths higher than 8. P07 reached transparent quality at the highest bitrate tested for all contens. P03 also reached transparent quality at the highest bitrate except for the screen content Apple. P06, on the other hand, performed below transparent quality for content Arri and Blender. P03 was the best performer at the highest bitrate of 10-bit computer generated image Blender. It must be noted that the confidence intervals of the competing codecs at selected bitrates were usually overlapping. Examining statistically significant differences between codec performances per bitrate per codec on the left side of Figures 6-9 show that P03 was indeed performing better than all other proponents for Blender image at the highest bitrate. P03 also performed better than all other proponents except P01 on the Woman image at the highest bitrate, followed by P03 and P01. At intermediate bitrates, performances of P06 and P03 decreased especially for complex contents such as Bike. Statistically significant differences were not observed at the lowest bitrate in general.

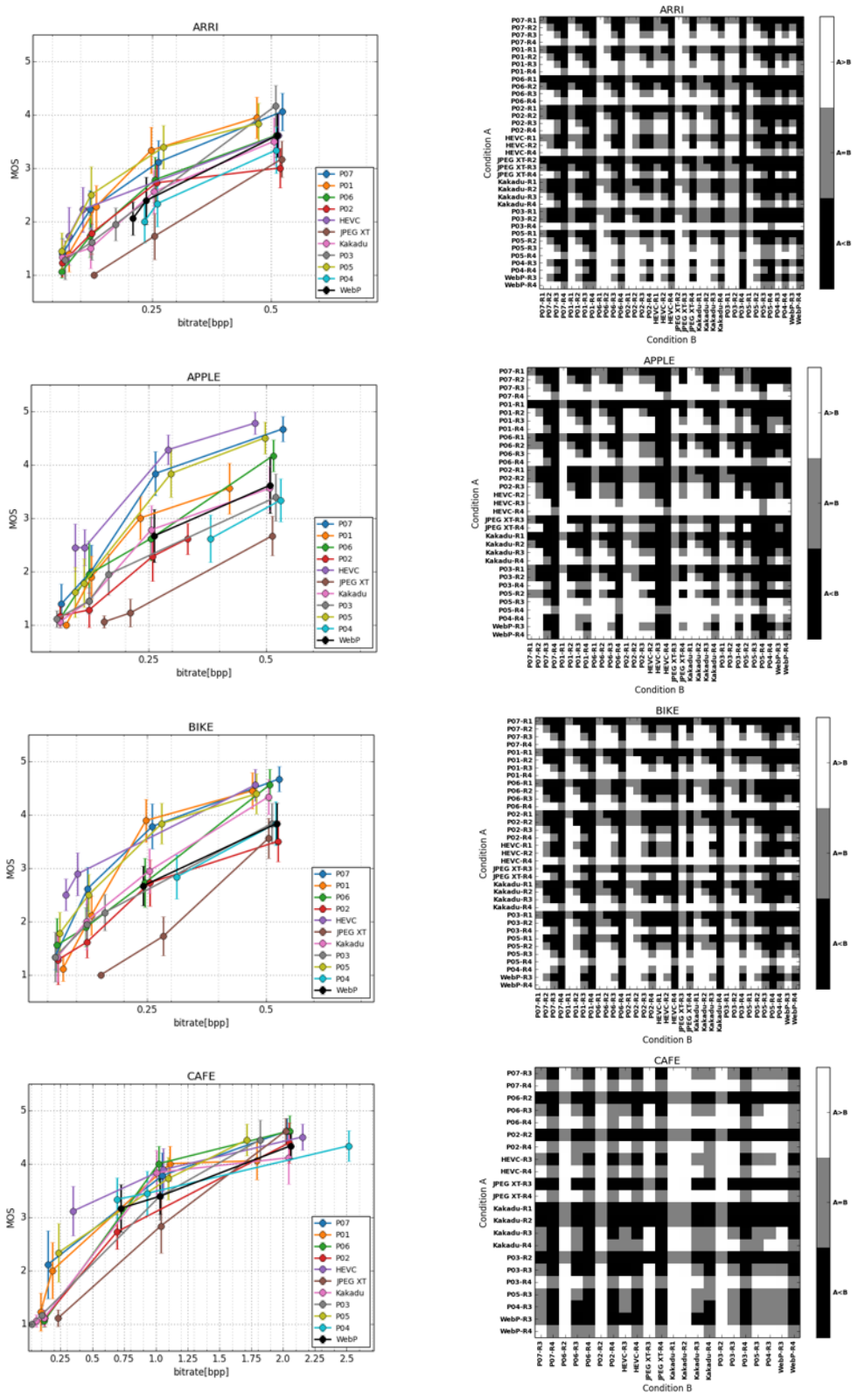


Figure 6: Subjective results for the SDR contents Arri, Apple, Bike and Cafe from top to bottom. MOS vs. bitrates are presented on the right. Comparisons between pairwise conditions are presented on the left.

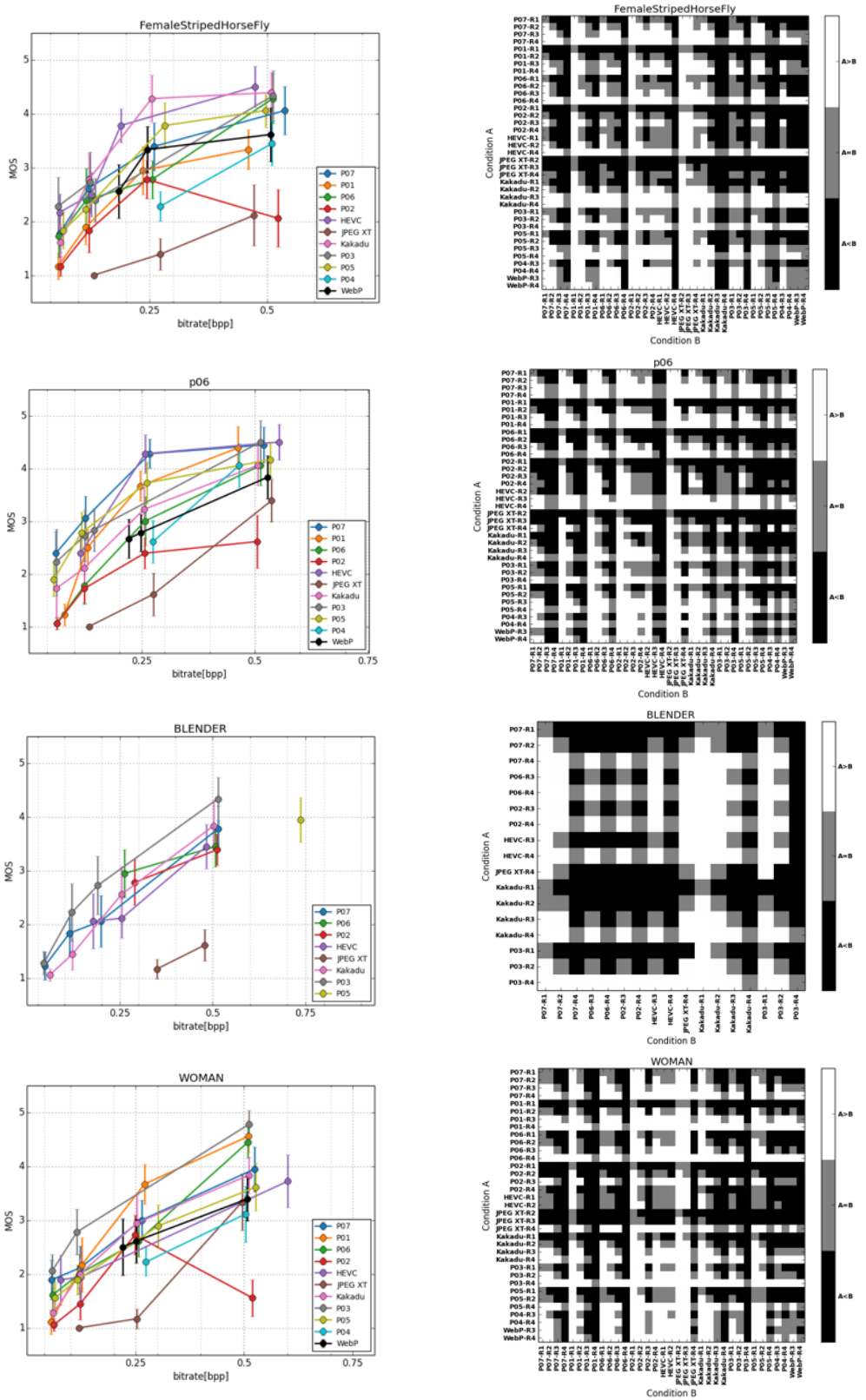


Figure 7: Subjective results for the SDR contents Fly, p06, Blender and Woman from top to bottom. MOS vs. bitrates are presented on the right. Comparisons between pairwise conditions are presented on the left.

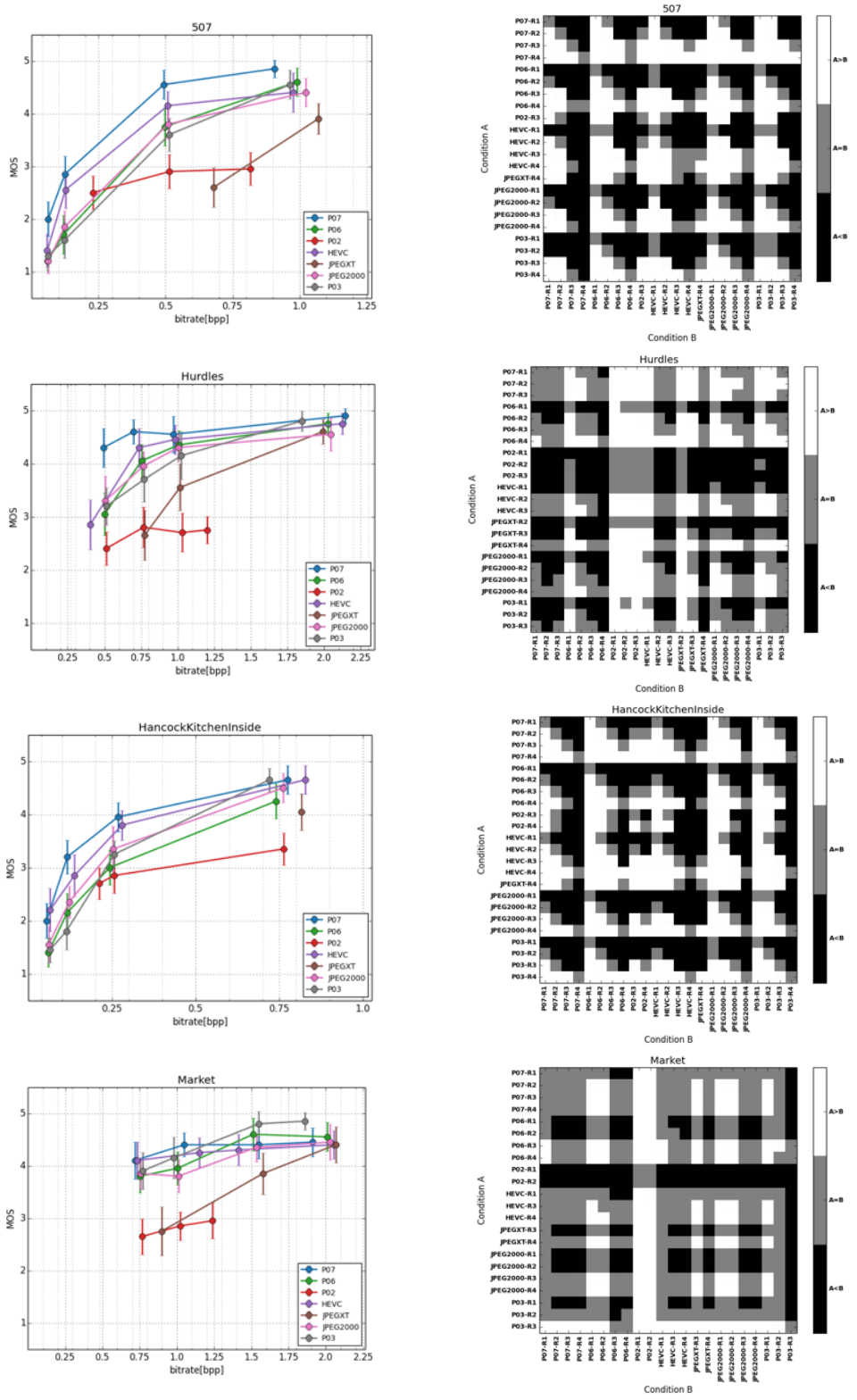


Figure 8: Subjective results for the HDR contents 507, Hurdles, Kitchen and Market from top to bottom. MOS vs. bitrates are presented on the right. Comparisons between pairwise conditions are presented on the left.

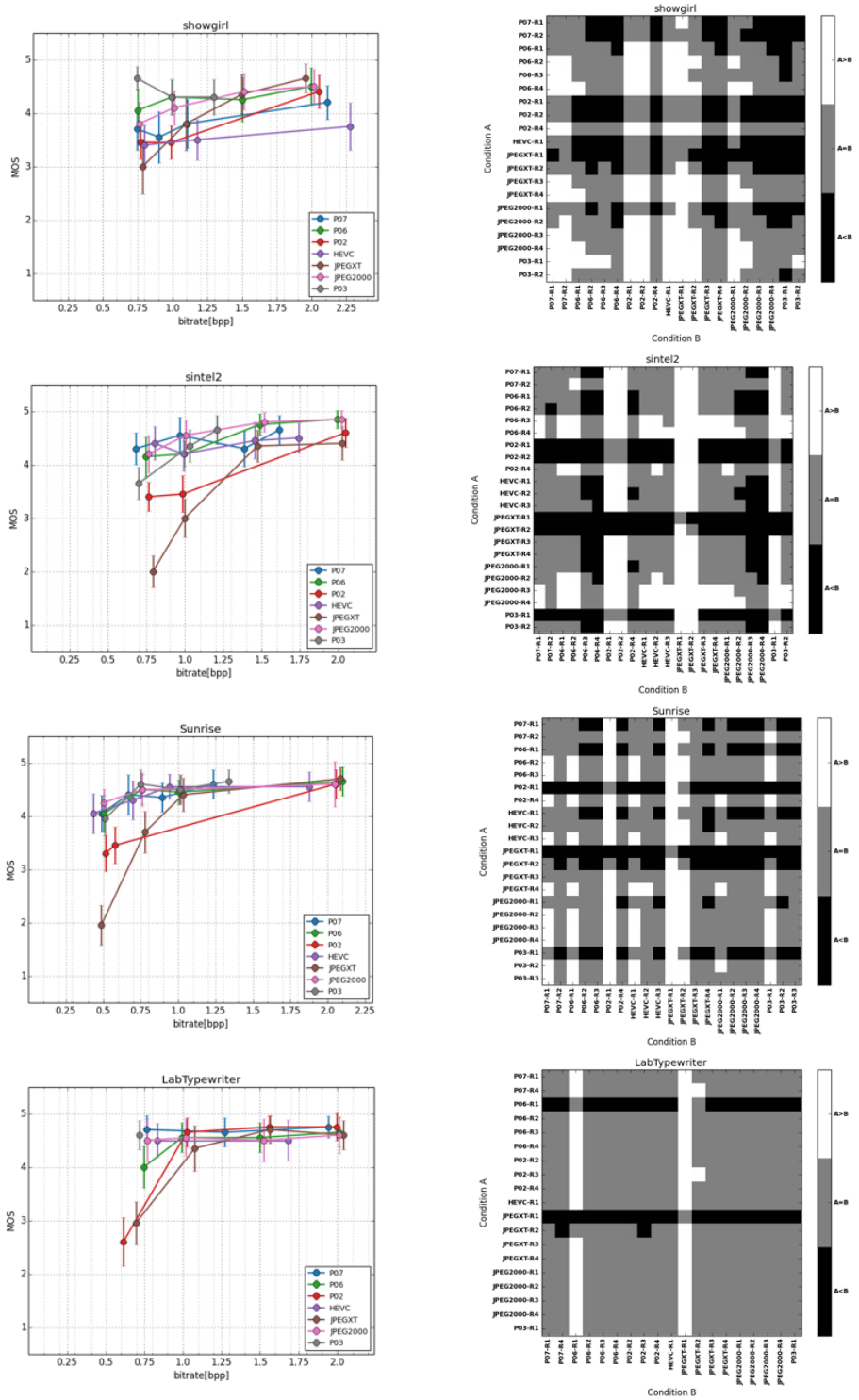


Figure 9: Subjective results for the HDR contents Showgirl, Sintel, Sunrise and Typewriter from top to bottom. MOS vs. bitrates are presented on the right. Comparisons between pairwise conditions are presented on the left.

JPEG and P02 were inferior to other codecs at the lower end of the rate spectrum for most HDR images. Contents Market, Showgirl, Sintel2, Sunrise and Typewriter did not provide statistically significant differences between the performances of other codecs. MOS for 507, Hurdles and Kitchen had more variances along the rate spectrum. P03, P06 and P07 reached transparent quality at the highest bitrate, with P07 having transparent quality at the next lowest bitrate for contents 507 and Kitchen, and remaining at transparent quality at all bitrates for Hurdles. Interestingly, P03's performance at the lowest bitrate for content Showgirl was never inferior to any other codec at any bitrate. These different behaviors indicate the strengths and weaknesses of codecs at certain types of images and regions.

4.4 Correlation between results from different labs

It is important to establish the accuracy of the results of the experiments. Objective and subjective tests were therefore conducted in different labs and results were cross checked. Correlation between the results of subjective quality assessment tests of SDR data ran at EPFL and VUB are given in Table 9 and Figure 10a. Correlation between the results of subjective quality assessment tests of HDR data ran at EPFL, VUB and TPT are given in Table 8 and Figures 10b-10d. The results were compared using multiple metrics such as Pearson linear correlation (PLC), Spearman rank order correlation (SROC), root mean square error (RMSE), outlier ratio, correct estimation rate, under and over estimation rates, correct decision rate, false ranking, false differentiation and false tie rates. The minimum correct decision rate was 88.15% with PLC and SROC going up to 97.68% and 97.75%, respectively.

Table 8: Comparison of subjective quality assessment results for SDR data, gathered by EPFL and VUB.

Key	**Value (%)**	
	Linear fitting	Cubic fitting
Pearson	97.31	97.75
Spearman	97.75	97.75
RMSE	25.25	23.44
OutlierRatio	3.53	2.94
Correct estimation	100.0	100.0
Under estimation	0.00	0.00
Over estimation	0.00	0.00
Correct decision	89.05	90.16
False ranking	0.00	0.00
False differentiation	5.66	5.86
False tie	5.30	3.99

Table 9: Comparison of HDR subjective quality assessment results for SDR data, gathered by EPFL, VUB and TPT.

Key	**Value (%)**					
	Linear fitting			**Cubic fitting**		
	EPFL vs. VUB	EPFL vs. TPT	VUB vs. TPT	EPFL vs. VUB	EPFL vs. TPT	VUB vs. TPT
Pearson	97.28	96.43	97.68	98.12	97.15	97.72
Spearman	92.61	91.44	89.46	92.61	91.44	89.46
RMSE	28.92	33.06	23.91	24.08	29.60	23.69
OutlierRatio	5.80	6.25	4.02	5.80	5.80	4.02
Correct estimation	100.0	100.0	100.0	100.0	100.0	100.0
Under estimation	0.00	0.00	0.00	0.00	0.00	0.00
Over estimation	0.00	0.00	0.00	0.00	0.00	0.00
Correct decision	92.10	88.15	90.93	93.17	90.71	90.94
False ranking	0.00	0.00	0.00	0.00	0.00	0.00
False differentiation	2.62	1.74	1.69	3.46	2.18	1.65
False tie	5.29	10.11	7.38	3.38	7.11	7.42

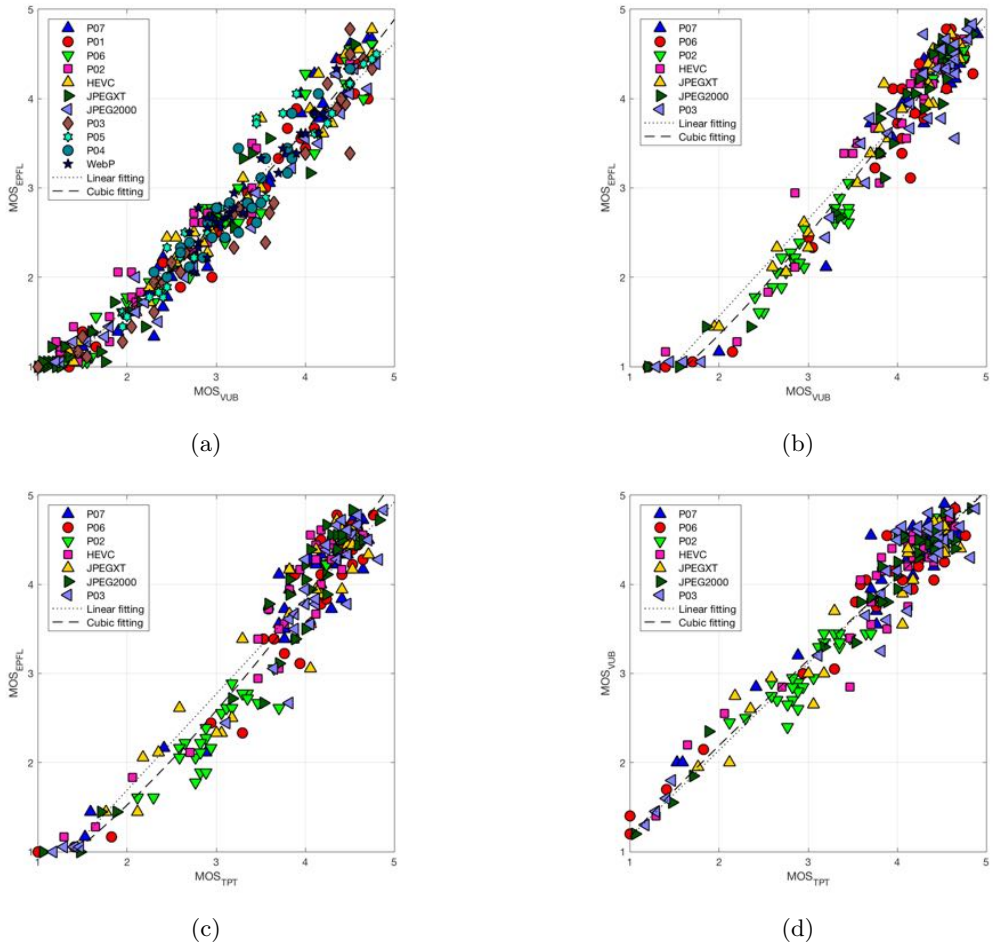


Figure 10: Comparison of subjective quality assessment results gathered by EPFL, VUB and TPT. (a) SDR results from EPFL and VUB, (b) HDR results from EPFL and VUB, (c) EPFL and TPT and (d) VUB and TPT.

5. CONCLUSION

This paper presented the framework and results of the quality assessment of proponents submitted to the JPEG XL Call for Proposals for creating the next generation image coding standard. A total of seven proponents were compared, also with four anchors, at eight different bitrates during objective and four different bitrates during subjective evaluation. Subjective tests were run at three different labs to cross-check the accuracy of the experiments. Objective and subjective test results showed that the Call was able to gather solutions superior to the current JPEG standard. The perfomance of some proponents were as good as or exceeding state-of-the-art codecs for numerous SDR and HDR contents. The quality assessment tests have lead to the selection of two proponents, which were then combined to generate the current version of JPEG XL codec. A verification model was recently developed, which is expected to be finalized and published in October 2019.

ACKNOWLEDGMENTS

The authors would like to thank Peter Schelkens, Saeed Mahmoudpour and Giuseppe Valenzise for carrying out the subjective experiments at VUB and TPT. This paper reports a research performed under the framework of project Digital Eye: Deep Learning Video Quality Assessment Technology, funded by The Swiss Commission for Technology and Innovation (InnoSuisse) under the grant 27403.1 PFES-ES.

REFERENCES

1. "Overview of jpeg xl." https://jpeg.org/jpegxl/index.html. Accessed: 2019-07-29.
2. ITU-R BT.2022, "General viewing conditions for subjective assessment of quality of sdtv and hdtv television pictures on flat panel displays," August 2012.
3. ITU-R BT.500-13, "Methodology for the subjective assessment of the quality of television pictures," January 2012.
4. ITU-T P.910, "Subjective video quality assessment methods for multimedia applications," April 2008.
5. "Final call for proposals for a next-generation image coding standard (jpeg xl)." https://jpeg.org/downloads/jpegxl/jpegxl-cfp.pdf. Accessed: 2019-08-05.
6. M. Bernando, T. Bruylants, T. Ebrahimi, K. Fliegel, P. Hanhart, L. Krasula, A. Pinheiro, M. Rerabek, P. Schelkens, and H. Xu, "Objective and subjective evaluations of some recent image compression algorithms," in *Picture Coding Symposium (PCS)*, 2015.
7. E. Alexiou, I. Viola, L. Krasula, T. Richter, T. Bruylants, A. Pinheiro, K. Fliegel, M. Rerabek, A. Skodras, P. Schelkens, *et al.*, "Overview and benchmarking summary for the icip 2016 compression challenge," in *23rd International Conference on Image Processing*, 2016.
8. J. Lainema, M. M. Hannuksela, V. K. M. Vadakital, and E. B. Aksu, "Hevc still image coding and high efficiency image file format," in *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 71–75, IEEE, 2016.
9. J.-M. Valin, N. E. Egge, T. Daede, T. B. Terriberry, and C. Montgomery, "Daala: A perceptually-driven still picture codec," in *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 76–80, IEEE, 2016.
10. ITU-R BT.709, "Parameter values for the hdtv standards for production and international programme exchange," June 2015.
11. ITU-R BT.2100, "Image parameter values for high dynamic range television for use in production and international programme exchange," July 2018.
12. G. K. Wallace, "The jpeg still picture compression standard," *IEEE transactions on consumer electronics* **38**(1), pp. xviii–xxxiv, 1992.
13. D. Taubman and M. Marcellin, *JPEG2000 image compression fundamentals, standards and practice: image compression fundamentals, standards and practice*, vol. 642, Springer Science & Business Media, 2012.
14. G. J. Sullivan, J. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (hevc) standard," *IEEE Transactions on circuits and systems for video technology* **22**(12), pp. 1649–1668, 2012.
15. "Webp compression study." https://developers.google.com/speed/webp/docs/webp_study. Accessed: 2019-07-29.
16. "Hdrtools." https://gitlab.com/standards/HDRTools/tree/master. Accessed: 2019-08-05.
17. "Jpeg xl quality assessment." https://github.com/pinarakyazi/codec_compare. Accessed: 2019-07-29.
18. "Ffmpeg." http://ffmpeg.org. Accessed: 2019-07-29.
19. F. De Simone, L. Goldmann, J.-S. Lee, and T. Ebrahimi, "Towards high efficiency video coding: Subjective evaluation of potential coding technologies," *Journal of Visual Communication and Image Representation* **22**(8), pp. 734–748, 2011.