PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

JPEG XS call for proposals subjective evaluations

David McNally, Tim Bruylants, Alexandre Willème, Touradj Ebrahimi, Peter Schelkens, et al.

David McNally, Tim Bruylants, Alexandre Willème, Touradj Ebrahimi, Peter Schelkens, Benoit Macq, "JPEG XS call for proposals subjective evaluations," Proc. SPIE 10396, Applications of Digital Image Processing XL, 103960P (19 September 2017); doi: 10.1117/12.2275137

SPIE.

Event: SPIE Optical Engineering + Applications, 2017, San Diego, California, United States

JPEG XS call for proposals subjective evaluations

David McNally[1], Tim Bruylants[2,3], Alexandre Willème[4], Touradj Ebrahimi[1], Peter Schelkens[2,3], Benoit Macq[4]

[1]École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland [2]Vrije Universiteit Brussel, Belgium [3]imec, Belgium [4]ICTEAM, Université Catholique de Louvain, Louvain-la-Neuve, Belgium

ABSTRACT

In March 2016 the Joint Photographic Experts Group (JPEG), formally known as ISO/IEC SC29 WG1, issued a call for proposals soliciting compression technologies for a low-latency, lightweight and visually transparent video compression scheme. Within the JPEG family of standards, this scheme was denominated JPEG XS. The subjective evaluation of visually lossless compressed video sequences at high resolutions and bit depths poses particular challenges. This paper describes the adopted procedures, the subjective evaluation setup, the evaluation process and summarizes the obtained results which were achieved in the context of the JPEG XS standardization process.

Keywords: JPEG, JPEG XS, Subjective evaluations, Image compression, Video compression, low latency, low delay, Visually lossless coding, Transparent coding

1. INTRODUCTION

The need to store, transmit and process uncompressed video sequences arrises at many stages of a content production work flow. With the emergence of new video formats such as Ultra-High Definition (UHD), High Dynamic Range (HDR), High Frame Rate (HFR) and omnidirectional (360), both the storage and bandwidth requirements are dramatically increasing. As a consequence, ever higher demands are placed on current video processing infrastructure. Given the substantial costs of upgrading or replacing deployed infrastructure to handle new video formats, an attractive alternative would be to apply transparent compression to reduce video stream sizes. Such compression needs to be visually lossless, low latency and, for cost and integration reasons, should be of low complexity. Addressing this challenge, a range of industrial and standardization initiatives have been launched.[1,2,3]

In July 2015, the JPEG committee (joint working group between ISO/IEC JTC1/SC29/WG1 ITU-T/SG16) launched a new standardization work item named JPEG XS[4] in view of creating a standard for a frame based, visually transparent, low complexity codec. In March 2016, the JPEG committee issued the JPEG XS Call for Proposals[5] ("CfP") and complemented this in June 2016 with detailed submission guidelines and associated evaluation procedures.[6] This document defined a schedule both for proponents answering to the CfP as well as for the evaluation of the submitted proposals. This evaluation process was structured into three activities: Objective evaluations of the proponent submitted content,[7] subjective assessments of the same and a compliance analysis of the proposed technologies in terms of CfP specified latency and complexity requirements. The results of these assessment steps were collected and evaluated[8,9] during the JPEG committee meeting in October 2016. Here the decision was taken to merge two of the proposed technologies, launch a range of core experiments[10] and initiate the drafting process for the JPEG XS standard (ISO/IEC 21122).[11] The current status of the JPEG XS technical development is outlined in.[12]

This paper summarizes the subjective evaluations performed on proponent submissions in response to the JPEG XS CfP. Section 2.1 discusses the test methodology and outlines how the visually (near) lossless content was subjectively assessed. Sections 2.2, 2.4 and 2.6 provide details on the selected content on which subjective evaluations were performed, the chosen setup for these evaluations and the participating test labs respectively. Section

For further information contact: Touradj Ebrahimi or David Mc Nally. EPFL STI IEL GR-EB, ELD 234, Station 11, CH-1015 Lausanne, Switzerland - E-mail: touradj.ebrahimi@epfl.ch or david.mcnally@epfl.ch

Applications of Digital Image Processing XL, edited by Andrew G. Tescher, Proc. of SPIE
Vol. 10396, 103960P · © 2017 SPIE · CCC code: 0277-786X/17/$18 · doi: 10.1117/12.2275137

2.3 describes the processing performed both by proponents and by the test labs to generate the content which was subjectively assessed. Section 2.7 then outlines validation and analysis of the subjective test data collected and summarizes the evaluation results presented to the JPEG committee. Section 3 then draws conclusions of this effort.

2. SUBJECTIVE QUALITY ASSESSMENTS OF THE JPEG XS PROPOSALS

This section details the subjective quality assessments undertaken to evaluate the proponent submissions received in answer to the JPEG XS Call for Proposals.[5] Together with the objective quality assessments,[8,7] as well as the latency and the complexity analysis of the submitted proposals, the results described here formed the basis upon which the JPEG committee selected proponent technologies to be carried forward through the JPEG XS standardization process.
This section summarizes the chosen test methodology, the test content selection, content processing, test anchors, subjective evaluation setup, participating test labs, test execution, results validation and cross-checking.

2.1 Test methodology

The adopted test methodology followed the recommendations set forth in in ISO/IEC 29170-2 (AIC Part-2) Draft Amendment 2[13] which specifically addresses the challenge of subjective evaluations for near lossless image coding systems. In particular, this draft amendment extends AIC Part-2[14] and normalizes the evaluation and grading of subjective evaluations for transparent coding systems in a manner which is independent of the chosen display technology or system. Subjects are selected from the general population and are to constitute a representative group in terms of gender, age and interests. Each subject is screened for visual acuity and undergoes a training session during which both the objectives and the procedures of the test are explained and demonstrated.
AIC-Part-2 proposes two alternative test methodologies. In the first, subjects are asked to select the one out of two images which most closely looks like a third image which is identified as the reference image. In the second, the reference and test images are shown side-by-side. In this case, the test image is presented as a video which is generated by interleaving the test and the reference image at a given temporal frequency. As a consequence of this presentation, visual discrepancies of the test image with respect to the reference image appear as "flickering". This mode of presentation facilitates recognition of visible differences in the test image and reduces strain on the subjects who are asked to identify the test image within a relatively short time.
For the JPEG XS CfP evaluations the *flicker test* modality of AIC Part-2 was adopted. In particular, the following conditions for the flicker test were adopted: The reference and interleaved reference-test images were shown side-by-side on a 4K monitor (4096 pixels width - see also Section 2.4). Depending on the resolution of the reference sequence, the reference and interleaved reference-test images were cropped to fill approximately 1/2 of the screen width (see Sections 2.2 and 2.3 for details). The subjective tests were performed only with still images. No video sequences were used due to the fact that motion in video sequences masks the flicker to be identified by subjects in the interleaved reference-test sequences. Each stimulus was shown to subjects four times at randomized points during a test session. While mandated by AIC Part-2, the participating test labs decided *not* to use a head rest. Given the size of the displays used and the choice of using most of the screen area to present the reference and test images, it was deemed easier for subjects to identify flicker if they could move their heads slightly. The admissible range of head movement was clearly detailed to subjects during the training session. Finally, AIC Part-2 mandates a forced choice paradigm: Subjects are obliged to choose the image they consider to be the test image (in this case: the image exhibiting flicker). For the JPEG XS CfP subjective evaluations, the decision was made to replace this with a ternary voting system where subjects had the option to cast a *no decision* vote. As a consequence, each subject vote could fall into three categories: The test image was correctly identified, the reference image was wrongly identified as the test image or the subject could not decide which image is the test image. An no decision vote could only be cast once a subject had exhausted the allocated viewing time for a given reference-test image pair (10 seconds). Offering a ternary choice to subjects reduces subject stress and fatigue and was deemed beneficial for the reliability of the subjective evaluation results. This in light of the fact that each subject underwent three test sessions of approximately 20 minutes duration during which only a small number of different stimuli had to be evaluated, potentially leading to fatigue and loss of interest.

2.2 Test content selection

As detailed in[7] and[6] , proponents were asked to code a total of 13 test sequences and six test images at a range of different bitrates and submit this coded content to the test labs. As noted above, no video sequences were used during the subjective evaluations. Given the tight latency and complexity constraints imposed on potential JPEG XS technologies, no inter frame coding compression could be considered and therefore single frame quality assessments were deemed sufficient to judge the overall quality of a proposed coding scheme.[6]
Figure 1 shows the six images selected for subjective evaluation. These images were chosen to cover a wide range of different content types and coding challenges. The three images shown on the top row were stills provided to the proponents. The three images on the bottom row were randomly selected frames taken out of three of the 13 test sequences provided to proponents. The red rectangles on each image indicate the cropping regions which were selected for subjective evaluation.

Figure 1. Test images used for subjective evaluations - clockwise from top-left: *Tools*, *Fly*, *Musik*, *Drums*, *CrowdRun* and *ScreenContent*.

2.3 Proponents coding and submitted content processing

When passing through a typical video production workflow, content is subject to multiple encode-decode cycles or *generations*. For the subjective evaluations of JPEG XS, proponents were required[5] to perform seven encode-decode cycles on all input content and at all target bit rates. This content was to be submitted, at the resolution, color representation and bit depth of the original content, to the test labs. To assure fairness of the evaluation process, spot checks of the submitted content were performed against content processed at the test labs using proponents submitted binary implementations of their codec submissions.
Following the submission deadline, the cropping regions to be used during the subjective evaluations were fixed and disseminated to all interested.

Table 1. Cropping parameters for subjective evaluation including the mandated target bitrates for coding.

Name	Input		Cropped						Target Bitrates (bits per pixel)
	w	h	x_0	x_1	y_0	y_1	w	h	
Tools	1524	1200	0	1520	0	1200	1520	1200	4, 4.8, 6
Fly	1920	1080	0	1040	0	1080	1040	1080	6, 8
Musik	1920	1080	0	1920	0	1080	1920	1080	6, 8
Drums	3840	2160	780	2700	0	2160	1920	2160	5, 6, 7.5
CrowdRun	3840	2160	1920	3840	0	2160	1920	2160	3.3, 4
ScreenContent	4096	2160	420	2340	0	2160	1920	2160	4.8, 6, 8

After cropping, the following post-processing steps were performed for each coded proponent image in order to generate a subjective test stimulus:

1. The reference images after cropping were losslessly formatted as a 24 fps video sequence with a duration of 10 seconds.
2. The cropped proponent images were interleaved with their respective reference images at a rate of 8 Hz and losslessly formatted as a 24 fps, 10 second video sequence.
3. The reference and (interleaved) proponent video sequences were horizontally joined into into a single 24 fps video sequence while adding both horizontal and vertical black zones around and between the reference and test images to fit the 3840 x 2160 resolution of the target display.
4. The above step was performed both with the reference image on the right hand and left hand sides of the display.

In addition, the same process was used to generate both training and control stimuli as detailed in Table 2. As a result of the described process and the mandated target bitrates at which each image was to be processed, each proponent submission resulted in a total of 15 stimuli (or 30 stimuli including step 4 above). With seven submitted proposals to the JPEG XS CfP, a single subjective rating set for all content (proponents and control) required the assessment of 108 stimuli.

Figure 2. The EPFL test lab setup used during JPEG XS subjective evaluations

2.4 Subjective evaluation setup

The viewing setup for subjective evaluations (Figure 2 shows an example) followed the recommendations in ISO/IEC 29170-2 AIC Part-2[14] and its associated Draft Amendment 2[13] - particularly:

- The viewing distance was set to approximately 0.5 meters.
- During the anchor evaluations it was decided not to use a headrest as set forth in.[13]
- The background lighting was set in accordance to ITU-R BT.500-11[15] with a ambient brightness between 12 and 20 cd/m^2 depending on the test lab.
- A monitor with 4K UHD resolution (3840 x 2160 pixels) and 10 bit color depth was used.
- The screen luminance was set to between 115 and 120 cd/m^2 depending on the manufacturer of the screen.

- Screen settings were configured to display the stimuli with minimal additional processing. In particular, the color representation mode was set to ITU-R BT.709-6.[16]
- All test labs used the same testing software[17] which served to orchestrate the subjective evaluation sessions. Ratings were solicited through dialogues shown on the test monitor and were provided by subjects through keyboard input. This data was collected in one tab separated ascii files per subject. The test software, MPV media player,[18] was capable of coping with 10 bit color depth content, for the playback of the individual stimuli. These stimuli were prepared as outlined in Section 2.3 above.

2.5 Test anchors

The evaluations of the proponents content were complemented with anchor evaluations. In subjective evaluations, only a subset of anchors were used, namely JPEG 2000 using Tile-Based Allocation configuration and VC-2.[19] The respective anchor target bitrates for each image were the same as those mandated for the proponent technologies. Details of the software and configurations used to generate subjective evaluation anchor content can be found in Annex A of.[6] The so produced anchors added an additional 30 stimuli to be rated, bringing the total number of stimuli for a complete test set to 138.

2.6 Participating test labs

Table 3 lists the test labs which participated in the JPEG XS subjective evaluations and includes the number of subjects as well as the number of test scores per stimuli contributed to this evaluation.

2.7 Data validation, analysis and results

To simplify and unify the verification and reporting of test results across the different test labs, a set of Python scripts and Excel worksheet templates[25] were prepared. This tool set supported parsing of all subject test results and subsequent population of the Excel worksheet in order to generate a range of standardized graphs. The analysis process proceeded in four steps:

Step 1 - validation: We identified (and excluded ratings from) subjects who obviously did not understand or follow the training guidelines or who did not make the required effort to identify flickering as instructed. Subjects were shown three stimuli, randomly placed throughout the test sessions, which should have given rise to an obvious rating choice. Figures 3 shows the ratings, per subject, for three test images and indicates that all subjects were attentive. Three control images, which were more challenging to classify, were shown twice each to

Table 2. Coding of training and control images.

Type	Name	Codec	Bitrate
Training	*ScreenContent*	VC-2[19]	4 bpp
Training	*Drums*	VC-2	5 bpp
Control	*Fly*	VC-2	4 bpp
Control	*Musik*	VC-2	4 bpp
Control	*Tools*	VC-2	4 bpp

Table 3. Participating test labs and sample sizes.

Lab name	Observers	Scores	Monitor used
CVUT[20]	42	84	Eizo CG318-4K
EPFL[21]	40	80	Eizo CG318-4K
Nantes[22]	32	64	Asus PQ321
VUB[23]	33	64	Eizo CG318-4K
Yonsei[24]	40	68	Samsung U28E590D

all subjects at random times during the test session. Figure 4 shows the ratings, by subject, for these six stimuli. To discriminate between "reliable" and "unreliable" subjects a threshold (black line in Figure 4) was computed as the average minus one standard deviation of the number of "ok" votes cast, taken over all subjects:

$$\text{Threshold} = \overline{\text{OK Decisions}} - 1 \times \sigma(\text{OK Decisions})$$

Based on the the data shown on Figure 4 subjects number 19 and 21 appear to have chosen to cast a no-decision vote in lieu of making a correct decision. Also subject number 7 cast wrong or no-decision votes instead of making correct decisions. The scores cast by these three subjects were consequently removed from the analysis. Figure 5 compiles all votes cast by subject. Similar to above, a reliability threshold is defined as the average plus two sigma of all "Wrong" votes cast, computed over all subjects. This graph shows that subjects number 14, 30, 33 and 36 cast "Wrong" votes in lieu of casting a "No Decision" vote suggesting that they either did not fully understand the instructions or that they pre-emptively cast a vote before the viewing time was up and did not wait to cast a "No Decision" vote. Again, these subjects were removed from the analysis.

Step 2 - cross-lab verification: To verify inter-lab consistency, Figure 6 shows a summary of all votes cast for a given test image, plotted by lab and organized by proponent and target bitrate. This figure indicates excellent inter-lab consistency of the subjective test ratings.

Step 3 - analysis: Given the choice of a ternary voting system, the following metric was defined:

$$Score = 2 \times \left(1 - \frac{Ok + 0.5 \times No\ Decision}{Votes}\right)$$

For a given stimulus, *Ok* are the number of times a correct score was given (i.e. the flickering image was correctly identified), *No Decision* is the number of times subjects chose to cast a "No Decision" vote and *Votes* is the total number of scores cast for this stimulus. The lower bound of the so defined *Score* metric is 0 and is achieved if all votes cast are *Ok*; the upper bound is 2 and achieved if all scores cast are *Wrong* (that is, the reference image was systematically mis-identified as the flickering test image) - this case never occurs thanks to the validation of test scores and removal of unreliable subjects outlined above. A *Score* of 1 is achieved if all scores cast are *No Decision*, indicating transparent coding as no subject was able to identify the flickering image. Figure 7 summarizes the data collected by all test labs in terms of the *Score* metric defined above. For each image tested, the scores are arranged by proponent and ordered by target bitrate.

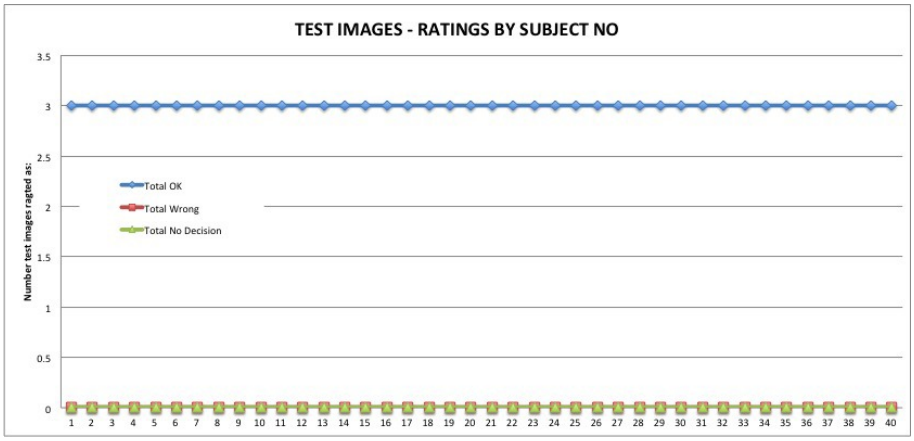

Figure 3. Rating of test images by subject number (EPFL data)

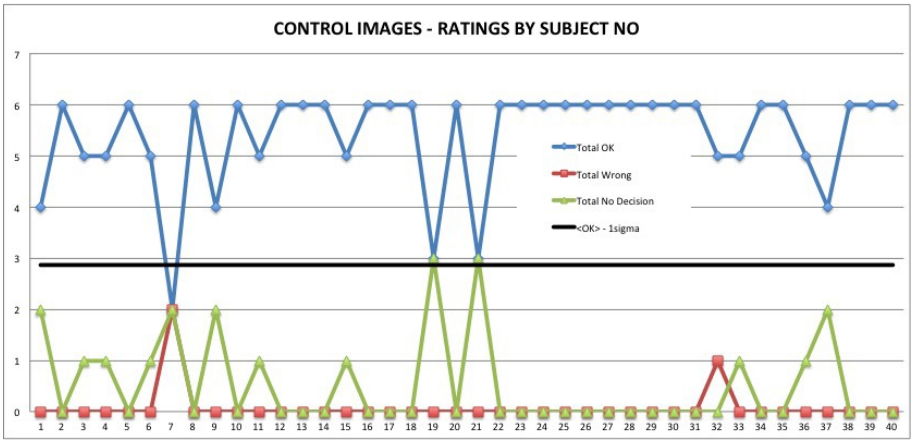

Figure 4. Rating of control images by subject number (EPFL data)

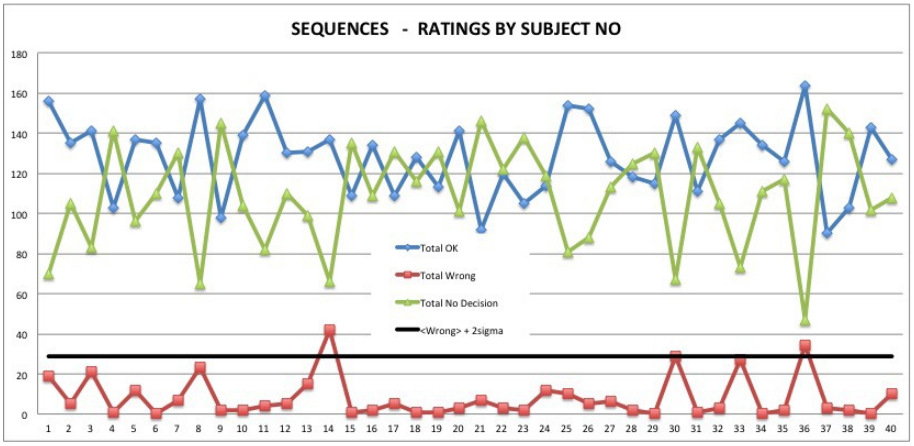

Figure 5. Rating of control images by subject number (EPFL data)

Step 4 - summary: The results shown in Figure 7 are summarized in Table 4. For each proponent and test image a pass mark("$\surd$") is attributed if the subjective evaluation score is greater than 50% in the case of the lowest and greater than 75% in the case of the highest assessed bitrate respectively.

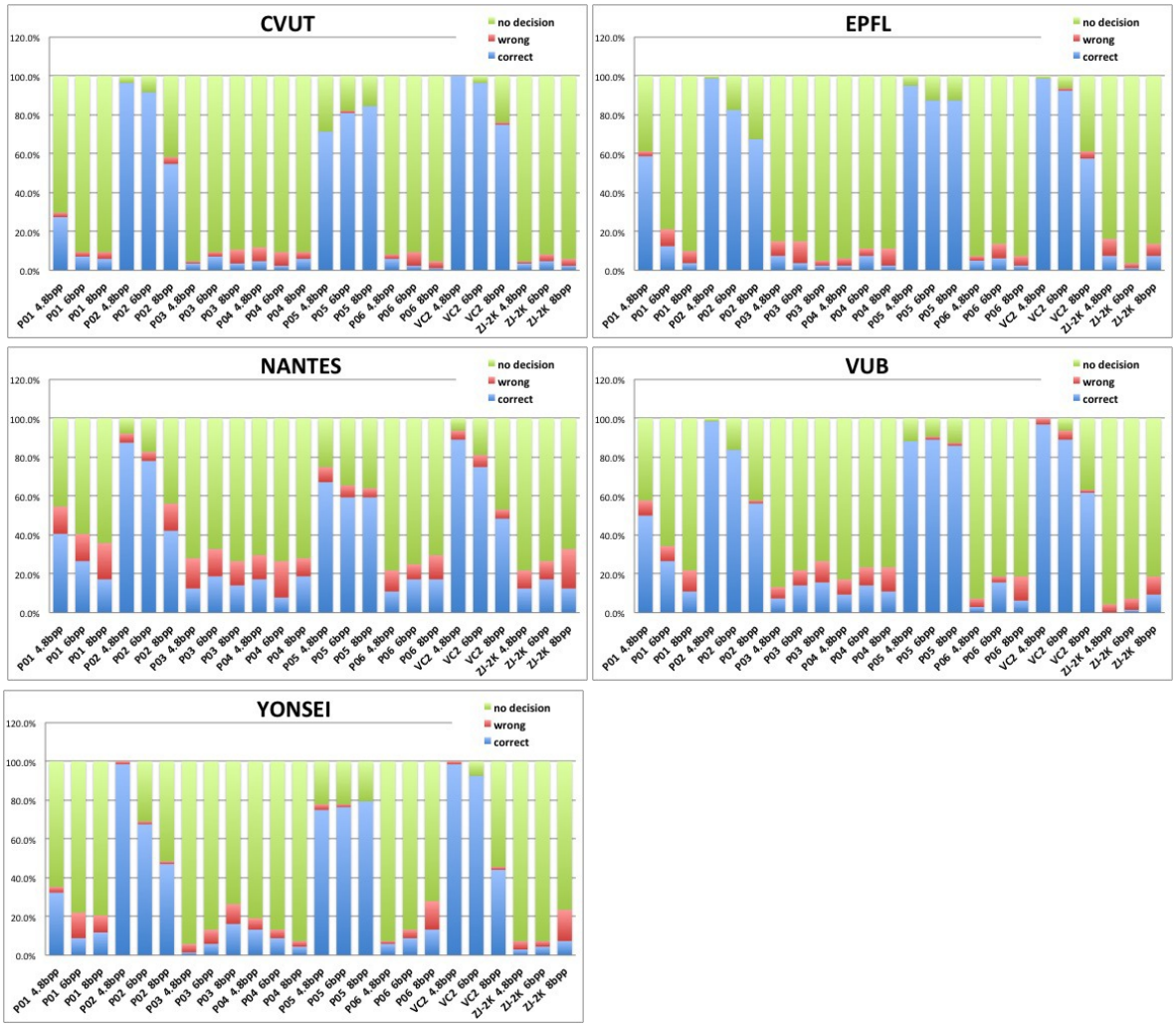

Figure 6. Stacked histograms of all votes cast for sequence *ScreenContent*, by proponent and target bitrate for each test lab - in clockwise order from top left: CVUT, EPFL, Nantes, VUB and Yonsei

Table 4. Pass-fail summary of subjective evaluation results. Proponents in red color were selected by the JPEG committee. The proponent in blue color was disqualified due to latency and complexity compliance issues.

	Tools		Fly		Musik		ScreenContent		CrowdRun		Drums	
	L	H	L	H	L	H	L	H	L	H	L	H
P01							√	√	√	√	√	
P02									√	√		
P03						√	√	√	√	√	√	√
P04			√	√		√	√	√	√	√	√	√
P05									√	√	√	√
P06			√	√			√	√	√	√	√	√

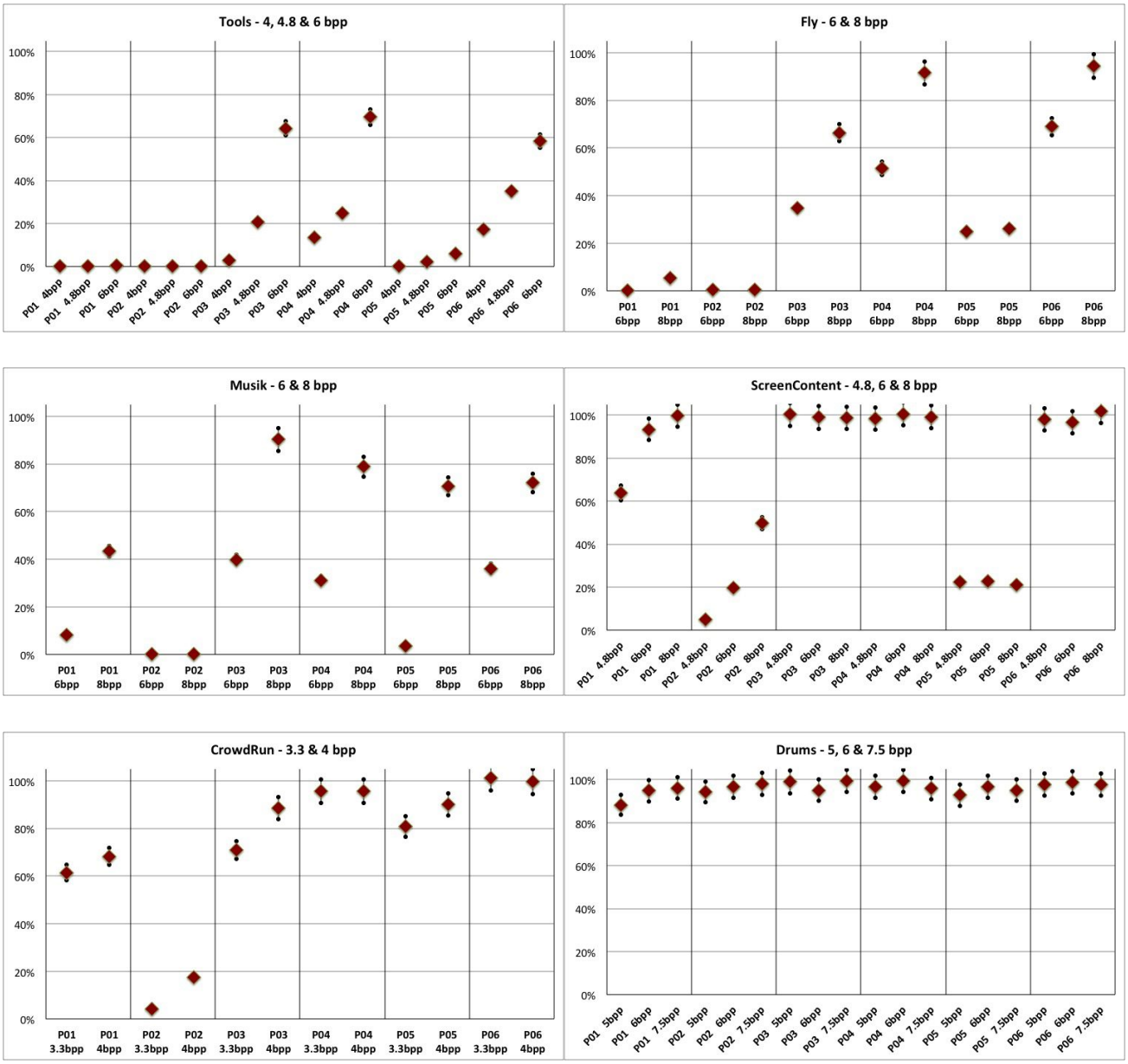

Figure 7. Subjective results for each stimulus and target bitrate ordered by proponents.

3. CONCLUSION

This paper summarizes the subjective assessment procedures performed on content provided by proponents responding to the JPEG XS Call for Proposals. Given the JPEG XS objective of providing near lossless compression, the methodology for these subjective evaluations had three particularities: Reference and processed images were interleaved at a rate of 8 Hz in order to guide subjects and facilitate identification of the processed image. Ternary voting was used instead of traditional binary voting in order to alleviate subject fatigue. As a consequence of this choice, a modified quality score was introduced in order to accommodate ternary votes.
Following a validation and removal process to eliminate votes cast by unreliable subjects, the data of all five participating test labs showed excellent correlation and was combined to compare the subjective performance of the participating proponent technologies. Based on these results, objective evaluations and taking into consideration compliance with the mandated complexity and latency requirements of the submitted proposals, the JPEG committee was able to select two proponent technologies to proceed to the next step in the JPEG XS standardization process.

ACKNOWLEDGMENTS

Touradj Ebrahimi and David McNally contributions to this work was possible thanks to funding provided by the Swiss SERI (State Secretariat for Education, Research and Innovation) in the framework of H2020-ICT-2015 ImmersiaTV.

Alexandre Willème's contribution to this work was funded by the Walloon Region (Belgium) Eurostars Project ACHEF E! 9845.

Tim Bruylant's contribution to this work was funded by iMinds (ICON project HD2R)

REFERENCES

[1] VESA, "Display Stream Compression (DSC) Standard v1.1." August 2014.
[2] VESA DSC TG, "Call for Technology: Advanced Display Stream Compression." 15 January 2015.
[3] SMPTE, "VC-2 Video Compression, document SMPTE 2042-1:2012." August 2012.
[4] "Overview of JPEG XS." https://jpeg.org/jpegxs/. Accessed: 2017-07-10.
[5] "Jpeg xs call for proposals for a low-latency lightweight image coding system." 71th SC29/WG1 Meeting La Jolla, 11 March 2016, Doc. wg1n71031.
[6] "JPEG XS proposals evaluation process." 73th SC29/WG1 Meeting Chengdu China, October 2016, Doc. wg1m73000.
[7] Willème, A. and Richter, T and Rosewarne C. and Benoit, M., "Overview of the JPEG XS objective evaluation procedures," *Proc. of Appl. of Digital Image Proc. XL, SPIE* (2017).
[8] "JPEG XS: Overview of proposals objective evaluation." 73th SC29/WG1 Meeting Chengdu China, October 2016, Doc. wg1n73069.
[9] "Jpeg xs subjective test results." 73th SC29/WG1 Meeting Chengdu China, October 2016, Doc. wg1n73068.
[10] "Jpeg xs core experiments #1." 73th SC29/WG1 Meeting Chengdu China, October 2016, Doc. wg1n73017.
[11] "JPEG XS Working Draft 1." 73th SC29/WG1 Meeting Chengdu China, October 2016, Doc. wg1n73016.
[12] Descampe, A., Keinert, J., Richter, T., Fel, S., and Rouvroy, G., "Jpeg xs: a new standard for visually lossless low-latency lightweight image compression,"
[13] "ISO/IEC 29170-2 (AIC Part-2) Draft Amendment 2." 72nd SC29/WG1 Meeting Geneva Switzerland, June 2016, Doc. wg1n72029.
[14] "Advanced image coding and evaluation – Part 2: Evaluation procedure for nearly lossless coding," (2015).
[15] "RECOMMENDATION ITU-R BT.500-11." Methodology for the subjective assessment of the quality of television pictures.
[16] "RECOMMENDATION ITU-R BT.709-6." Parameter values for the HDTV standards for production and international programme exchange.

[17] "JPEG XS Evaluation software package." 73th SC29/WG1 Meeting Chengdu China, October 2016, Doc. wg1m73003.
[18] "A free, opensource, and cross-platform media player." https://mpv.io/. Accessed: 2017-07-10.
[19] "A reference encoder and decoder for SMPTE ST 2042-1 VC-2 Video Compression." https://github.com/bbc/vc2-reference. Accessed: 2017-06-09.
[20] "Multimedia Technology Group - FEE, CTU in Prague, Czech Republic." http://mmtg.fel.cvut.cz.
[21] "Multimedia Signal Processing Group - EPFL, Switzerland." http://mmspg.epfl.ch.
[22] "Image and Video Communications Group - University of Nantes, France." http://ivc.univ-nantes.fr/en.
[23] "ETRO VUB Department of electronics and informatics - Vrije Universiteit Brussel, Belgium." http://www.etrovub.be.
[24] "Multimedia Computing and Machine Learning (MCML) Group - School of Integrated Technology, Yonsei University, Korea." http://mcml.yonsei.ac.kr.
[25] "JPEG XS Proponent Subjective Evaluation Data Processing." 73th SC29/WG1 Meeting Chengdu China, October 2016, Doc. wg1m73023.