ISO/IEC JTC 1/SC29/WG1 **N89022**
89th JPEG Meeting, Online, October 2020

ISO/IEC JTC 1/SC 29/WG 1
(ITU-T SG16)

Coding of Still Pictures

JBIG
Joint Bi-level Image Experts Group

JPEG
Joint Photographic Experts Group

TITLE: Report on the JPEG AI Call for Evidence Results

SOURCE: Requirements
EDITORS: João Ascenso, Evgeniy Upenik, Michela Testolina, Elena Alshina, Atanas Boev, Nicola Giuliani

STATUS: Final

REQUESTED ACTION: Distribute

DISTRIBUTION: Public

Contact:
ISO/IEC JTC 1/SC 29/WG 1 Convener – Prof. Touradj Ebrahimi
EPFL/STI/IEL/GR-EB, Station 11, CH-1015 Lausanne, Switzerland
Tel: +41 21 693 2606, Fax: +41 21 693 7600, E-mail: Touradj.Ebrahimi@epfl.ch

TABLE OF CONTENTS

1. Purpose of This Document

The objective of this document is to report the objective and subjective performance evaluation as well as the coding complexity of all Call for Evidence (CfE) proponents submissions, this means the four teams which have made submissions for the call. Additionally, the performance assessment of two teams which have only perform submission for the learning-based image coding challenge organized at IEEE 22nd International Workshop on Multimedia are also included.

2. CfE Proposals

In the JPEG AI Call for Evidence, five deep learning (DL) based image codecs were submitted proposing several state-of-the-art image codecs and proposing different tools and methods, namely training methodologies and loss functions, new convolutional layers and non-local attention modules as well as improved probability models for entropy coding. Moreover, one hybrid Versatile Video Coding (VVC Intra) plus DL based codec of a latent residual signal was proposed. The following proposals are briefly characterized here:

1. Team NJU-VISION [1]:
   - Variational autoencoder structure, with non-local attention modules
   - Masked 3D CNN based prediction is used to obtain more accurate conditional statistics
2. Team Four-leaf Clover [2]:
   - Filtering between high frequency and low frequency components
   - Octave convolution (GoConv) and octave transposed-convolution (GoTConv)
3. Team NCTU [3]:
   - Multi-layer image compression framework with VVC Intra as base layer and learned residual codec as the enhancement layer
   - Enhancement layer includes a local attention block
   - Multi-rate encoding with a single, programmable autoencoder
4. Team Nokia [4]:
   - Learned image compression, based on meta-learning for latent tensor overfitting.
   - Novel probability model and generation of content-specific decoders.

3. CfE Proposals Performance Evaluation Procedure

This Section describes the evaluation procedure of all proponents' submissions to the Call for Evidence, namely the test images, coding conditions and objective quality metrics. Objective quality assessment was carried out at the target bitrates and the decoded images obtained were cross-checked using the submitted bitstreams and corresponding decoders.

3.1. Test Material

The test images used in the CfE are defined in this Section. The test set includes sixteen images depicted in Figure 1 with different characteristics and content. There is a mix of uncompressed and pristine images, with spatial resolutions: 1336×862 (min) to 3680×2456 (max) and all were tested for low-level similarity against JPEG AI training set. Some of the images were down-sampled (spatially and bit-depth) and cropped.

Matterhorn

Racing car

Sardinia festival

Rotunda of Mosta

Las Vegas sign

Train

Windows

Transmission towers

Port

Curiosity rover

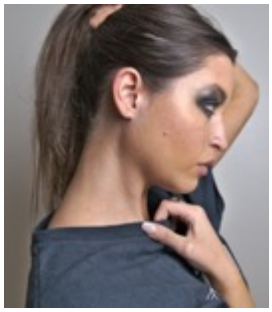
Bell pepper

Woman

Dog

Beach

Harbor



Ponytail

Figure 1. Thumbnails of the JPEG AI test set used in the Call for Evidence.

3.2. Coding Conditions

Target bitrates for the objective evaluations include **0.06, 0.12, 0.25, 0.50, 0.75, 1.00, 1.50, and 2.00 bpp**. The maximum bitrate deviation from the target bitrate should not exceed **15%.** The proponents must declare for every test image which target bitrate their decoder and models can reach. The bitrates specified must account accurately the total number of bits necessary for generating the encoded file (or files) out of which the decoder can reconstruct a lossy version of the entire image. The main rate metric is the number of bits per pixel (bpp) defined as:

BPP = (N_TOT_BITS)/(N_TOT_PIXELS)

where N_TOT_BITS is the number of bits for the compressed representation of the image and N_TOT_PIXELS is the number of pixels in the image.

3.3. Anchor Codecs

Proposals will be compared against the following anchors:

- JPEG (ISO/IEC 10918-1 | ITU-T Rec. T.81)
- JPEG 2000 (ISO/IEC 15444-1 | ITU-T Rec. T.800)
- HEVC Intra (ISO/IEC 23008-2 | ITU-T Rec. H.265)

The configurations detailed below are relevant for the definition of anchors.

3.3.1. JPEG (ISO/IEC 10918-1 | ITU-T Rec. T.81)

JPEG does not specify a rate allocation mechanism allowing to target a specific bitrate. Hence, an external rate control loop is required to achieve the targeted bitrate. The following conditions apply:

- Software to be used: JPEG XT reference software, v1.53
  - Available http://jpeg.org/jpegxt/software.html
  - License: GPLv3
- Command-line to use within the rate-control loop:
  - jpeg -q [QUALITY_PARAMETER] [INPUTFILE] [OUTPUTFILE]

3.3.2. JPEG 2000 (ISO/IEC 15444-1 | ITU-T Rec. T.800)

JPEG 2000 anchor generation supports two configurations: 1) PSNR optimized; and 2) Visually optimized. A target rate can be specified using the –rate [bpp] parameter. The following conditions apply:

- Software to be used: Kakadu, v7.10.2
  - Available http://www.kakadusoftware.com
  - License: demo binaries freely available for non-commercial use
- Command-line examples:
  - **MSE weighted:** kdu_compress -i [INPUTFILE] -o [OUTPUTFILE] -rate [BPP] Qstep=0.001 -tolerance 0 -full -precise
  - **Visually weighted:** kdu_compress -i [INPUTFILE] -o [OUTPUTFILE] -rate [BPP] Qstep=0.001 -tolerance 0 -full -precise -no_weights
- Decoding: kdu_expand -i [INPUTFILE .mj2] -o [OUTPUTFILE .yuv] -precise

3.3.3. HEVC (ISO/IEC 23008-2:2018 | ITU-T Rec. H.265 (v5))

For HEVC, an external rate control loop is required to achieve targeted bitrate. The HEVC rate-distortion (RD) performance for the target bitrates are obtained with the following conditions:

- Available software: HEVC Test Model HM-16.20+SCM-8.8
  - Available https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/tags/HM-16.20+SCM-8.8/
  - License: BSD
- FFMPEG will be used to convert the PNG (RGB) to YUV files following the BT.709 primaries according to:
  - ffmpeg -hide_banner -i input.png -pix_fmt yuv444p10le -vf scale=out_color_matrix=bt709 -color_primaries bt709 -color_trc bt709 -colorspace bt709 -y output.yuv
  - The ICC profiles are copied from the original contents to all the respective processed images in order to preserve the representation of colors for subjective evaluation.
- Configuration files to be used are available https://jpegai.github.io/public/encoder_intra_main_scc_10.cfg

4. Objective Quality Metrics

This section defines the objective image quality metrics used for the assessment of the learning-based image coding solutions. The source code of all metrics is available at the CfE + challenge website (jpegai.github.io). Objective quality testing shall be done by computing several quality metrics, including MS-SSIM, VMAF, VIFP, NLPD, FSIM, between compressed and original image sequences, at the target bitrates mentioned in Section 2.2.

4.1. MS-SSIM

Multi-Scale Structural SIMilarity (MS-SSIM) [5] is one of the most well-known image quality evaluation algorithms and computes relative quality scores between the reference and distorted images by comparing details across resolutions, providing high performance for learning-based image codecs [6]. The MS-SSIM [5] is more flexible than single-scale methods such as SSIM by including variations of image resolution and viewing conditions. Also, the MS-SSIM metric introduces an image synthesis-based approach to calibrate the parameters that weight the relative importance between different scales. A high score expresses better image quality.

4.2. VMAF

The Video Multimethod Assessment Fusion (VMAF) metric [7] developed by Netflix is focused on artifacts created by compression and rescaling and estimates the quality score by computing scores from several quality assessment algorithms and fusing them with a support vector machine (SVM). Even if this metric is specific for videos, it can also be used to evaluate the quality of single images and has been proved that performs reasonably well for learning-based image codecs [2]. Since the metric takes as input raw images in the YUV color space format, the PNG (RGB color space) images are converted to the YUV 4:4:4 format using FFMPEG (BT.709 primaries). A higher score of this metric indicates better image quality.

4.3. VIFp

The Visual Information Fidelity (VIF) [8] measures the loss of human perceived information in some degradation process, e.g. image compression. VIF exploits the natural scene statistics to evaluate information fidelity and is related to the Shannon mutual information between the degraded and original pristine image. The VIF metric operates in the wavelet domain and many experiments found that the metric values agree well with the human response, which also occurs for learning-based image codecs. A high score expresses better image quality.

4.4. NLPD

The Normalized Laplacian Pyramid (NLPD) is an image quality metric [9] based on two different aspects associated with the human visual system: local luminance subtraction and local contrast gain control. NLP exploits a Laplacian pyramid decomposition and a local normalization factor. The metric value is computed in the normalized Laplacian domain, this means that the quality of the distorted image relative to its reference is the root mean squared error in some weight-normalized Laplacian domain. A lower score expresses better image quality.

4.5. FSIM

The Feature Similarity (FSIM) metric [10] is based on the computation of two low level features that play complementary roles in the characterization of the image quality and reflects different aspects of the human visual system: 1) the phase congruency (PC), which is a dimensionless feature that accounts for the importance of the local structure and the image gradient magnitude (GM) feature to account for contrast information. Both color and luminance versions of the FSIM metric will be used. A high metric value express better image quality.

4.6. IW-SSIM

Information Content Weighted Structural Similarity Measure (IW-SSIM) [11] is an extension of the structural similarity index based on the idea of information content weighted pooling. This metric assumes that when natural images are viewed, pooling should be made using perceptual weights that are proportional to the local information content. Moreover, advanced statistical models of natural image are employed to derive the optimal weights which are combined with multiscale structural similarity measures to achieve the best correlation performance with subjective scores from well known databases.

5. Objective Performance Evaluation

In this Section, RD performance evaluation results are shown for all quality metrics defined in the previous Section. The results are shown in Figure 2 for the average of all CfE test images. In Annex A, RD performance evaluation results are shown for each image, this means the 16 images under evaluation. As shown, for MS-SSIM, FSIM and IW-SSIM, team05 and team06 show much better performance results compared to HEVC. For VIFp, HEVC is the top contender while team08 and team05 have the leading RD performance which is actually very close to HEVC. For VMAF, team05 and team06 also show very high performance but HEVC and JPEG2000 also achieve a similar level of performance. Finally, NLPD shows that team06 and team08 and HEVC achieve the highest RD performance and very similar.

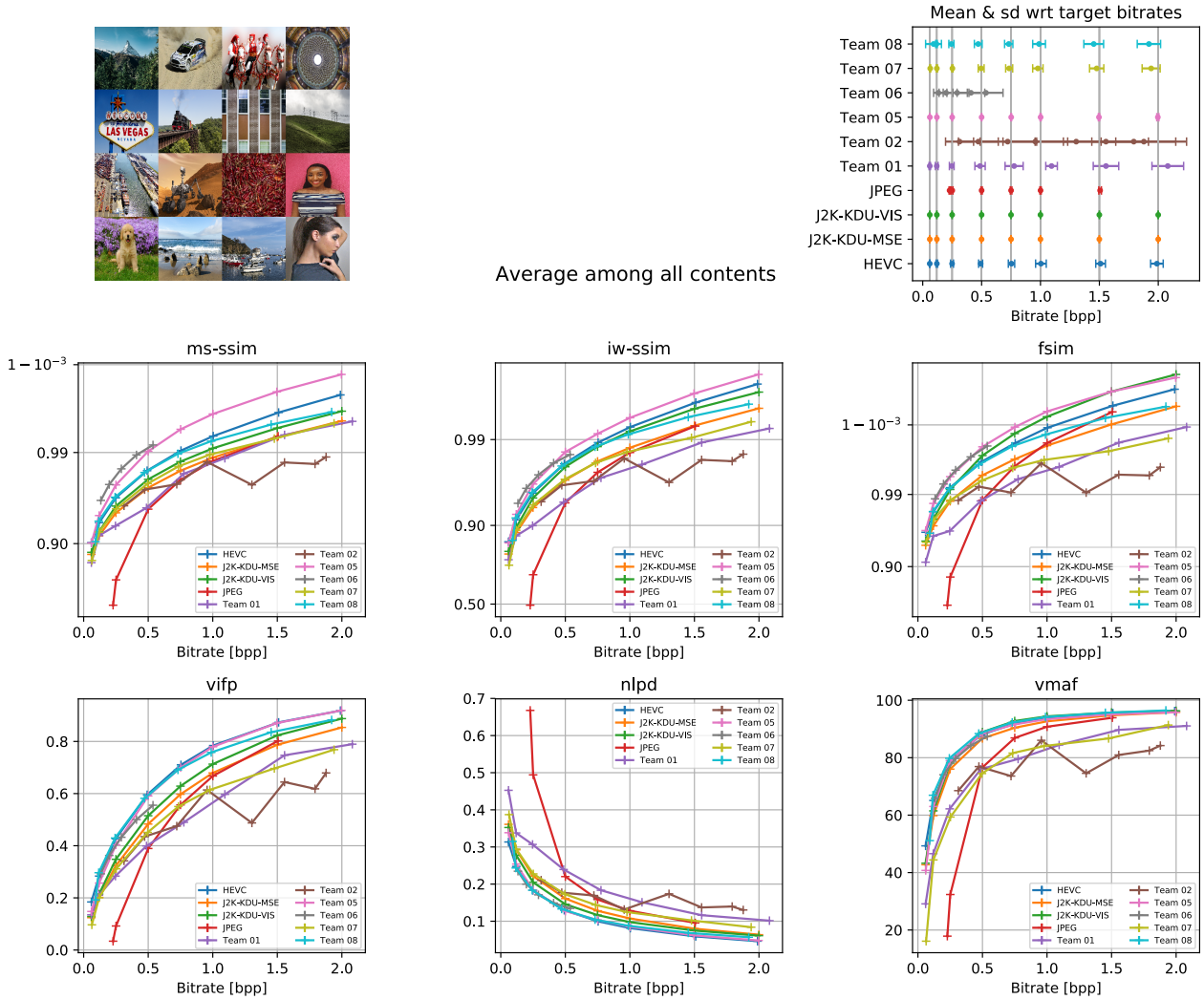


Figure 2. RD performance averaged among all the test images plotted per metric for each codec (in log(1-Q) scale for SSIM based metrics). The upper right-hand corner subplot reports the means with the standard deviations for factual bitrates of stimuli per codec, per target bitrate and their relative positions with respect to the target bitrates (grey vertical lines).

6. Complexity Evaluation

All submissions have been decoded on the same system equipped with a 14 core Intel(R) Core(TM) i9-10940X CPU @ 3.30GHz processor, 64GB RAM and a NVIDIA TITAN RTX GPU with 24GB memory. Execution time was measured with the Unix time command. Figure 3 shows the 'real' and the combined 'user' and 'sys' time for each team and each QP. 'Real' time is all elapsed time from start to finish of the call (wall clock time). 'User' time is the amount of CPU time spent within the process (outside the kernel) while 'Sys' time is the amount of CPU time spent within the kernel. The combined 'sys' and 'user' time is the actual CPU time of the process CPUs and cores. The time measurements, peak GPU memory usage and used frameworks are summarized in Table 1. Numbers are shown in seconds and accumulated over all QPs. The peak GPU memory usage was measured for each team and is shown in megabytes. Measurements for team #08 are missing due to too high GPU memory demand. Also, team #06 didn't submit bitstreams for QPs 100, 150 and 200 and thus, time measurements are incomplete for team #06.

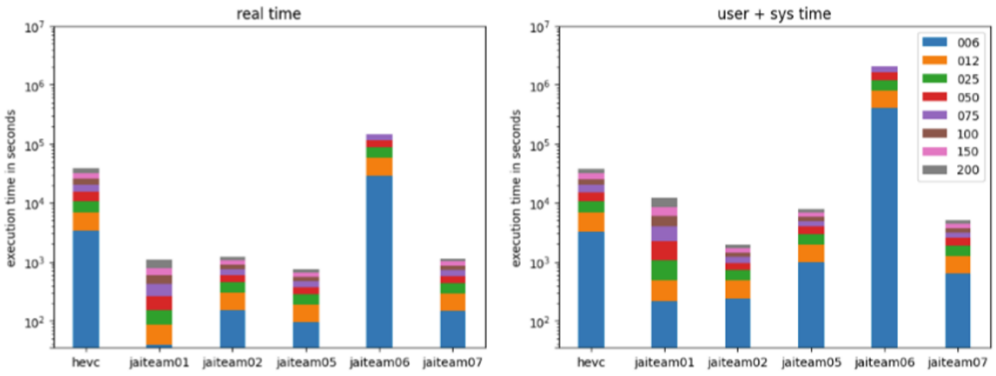


Figure 3. Time measurements using the Unix time command. Each team is shown as a single bar and each bar consists of a stacked time measurement for each QP (8 in total, except for team06 which submitted only 5). In left the real time is presented and in right it is presented the combined user + sys time.

Table 1. Peak GPU memory usage in MB, software framework, real time and combined user + sys time (accumulated over all QPs).

team #	01	02	05	06	07	08	hevc
Σ real	1111	1204	751	146755	1150	NA	38170
Σ (user + sys)	12419	1937	7891	2045271	5025	NA	37924
peak GPU in MB	9979	8941	0	0	0	> 24000	0
framework	pytorch	TF	TF	pytorch	TF	TF	-

7. Subjective Visual Quality Evaluation

During the 88th JPEG Meeting it was decided to perform subjective perceptual visual evaluation of the contributions submitted to JPEG AI Call for Evidence following a crowdsourcing approach, mainly because of the COVID-19 pandemic. Subjective quality evaluation of the compressed images was performed on the test dataset. The semi-controlled crowdsourcing setup has been proven in the past its reliability, i.e. maintains a low variance of the scores [12]. The Amazon Mechanical Turk was used as crowdsourcing platform and to collect the scores the QualityCrowd2 [13] web-based software was used to show the images and obtain the scores.

The Double Stimulus Continuous Quality Scale (DSCQS) methodology was used, where subjects watch side by side the original image and the impaired decoded image and both are scored in a continuous scale. This scale is divided into five equal lengths which correspond to the normal ITU-R five-point quality scale, notably Excellent, Good, Fair, Poor and Bad. This method requires the assessment of both original and impaired versions of each test image. The observers are not told which one is the reference image and the position of the reference image is changed in pseudo-random order. The subjects assess the overall quality of the original and decoded images by inserting a mark on a vertical scale. The vertical scales are printed in pairs to accommodate the double presentation of each test picture.

The subjective test methodology will follow BT500.13 [14] and a randomized presentation order for the stimuli, as described in ITU-T P.910 [15] will be used; the same content is never displayed consecutively. There is no presentation or voting time limit.

After successfully passing the screen size test, subjects see written instructions explaining the purpose of the experiment, the voting interface and the meaning of the quality rating. Each session begins with three training examples to familiarize the participants with the graphical interface and the range of visual quality. First example presents a processed image that is considered to be of “Bad” quality. The second training example shows and image of “Excellent” quality; and the last one is of “Fair” quality. The scores from the training examples are discarded. At the end of the training subjects are informed that the actual evaluation starts and are then presented with 240 image pairs to score.

This Section is organized as follows:

1. Visual quality rating software and the crowdsourcing platform to get subjects for the visual assessment task.
2. Subjective evaluation test characteristics, such as the restrictions imposed to the screen resolution.
3. Selected test images and image cropping
4. Subject population statistics
5. Data analysis and processing
6. Rate-MOS performance analysis

7.1. Visual Quality Rating Software and Crowdsourcing Platform

QualityCrowd2 is the latest version of this software which differs in many points from the previous QualityCrowd software. Instead of providing a web interface for designing and defining a test batch, QualityCrowd2 introduces a text-based definition of tests. Quality tests are now defined through little QualiyCrowd-Scripts (short: QC-Scripts). The QC-Scripting language is easy to learn and gives the operator even more control over his test than before. For the purpose of JPEG AI CfE subjective evaluations QualityCrowd2 was extended to be able to enforce screen size restrictions for participants.

Amazon Mechanical Turk (MTurk) is a crowdsourcing website for businesses (known as Requesters) to hire remotely located "crowdworkers" to perform discrete on-demand tasks that computers are currently unable to do. It is operated under Amazon Web Services, and is owned by Amazon. Employers post jobs known as Human Intelligence Tasks (HITs), such as identifying specific content in an image or video, writing product descriptions, or answering questions, among others. Workers, colloquially known as Turkers or crowdworkers, browse among existing jobs and complete them in exchange for a rate set by the employer. To place jobs, the requesting programs use an open application programming interface (API), or the more limited MTurk Requester site. Figure 4 and 5 show what the subjects see before and after accepting a task.

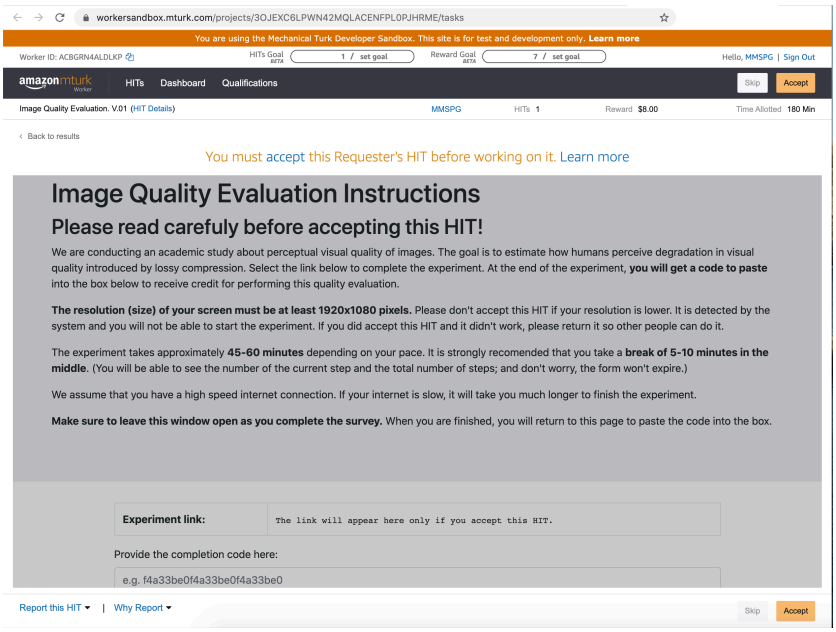


Figure 4. Preview of the task available to workers before accepting. Note that the link is not shown.

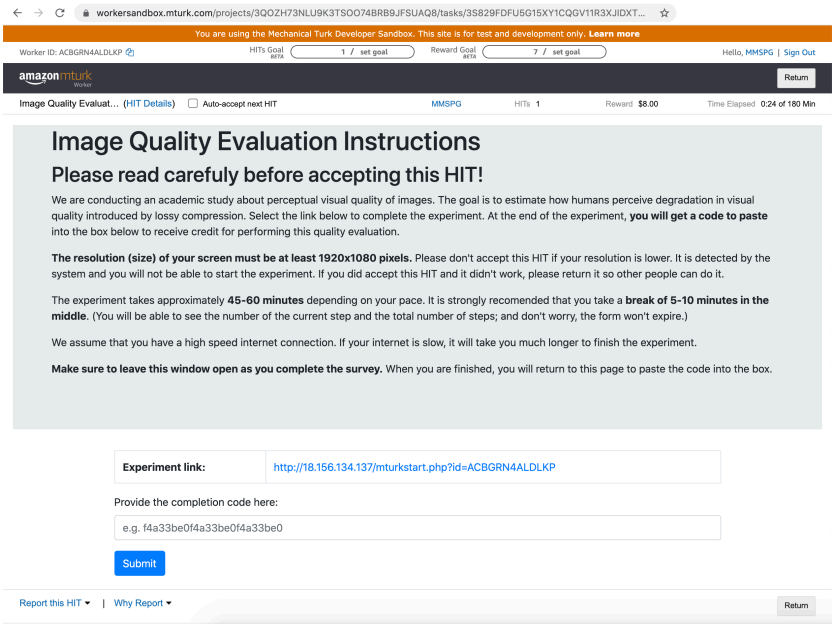


Figure 5. View of the task available to a worker after accepting.

7.2. Subjective Evaluation Test Characteristics

The test starts with a mandatory screen resolution check, the main purpose of which is to make sure that a HiDPI/Retina mode is not enabled and that the resolution is 1920x1080 or higher. Figure 6 and 7 shows the screen resolution checks made for two cases, one failure and one success. After, the subjective test starts and both reference and degraded images are shown side by side along with the scales for scoring both images. The layout of the voting step is show in Figure 8.

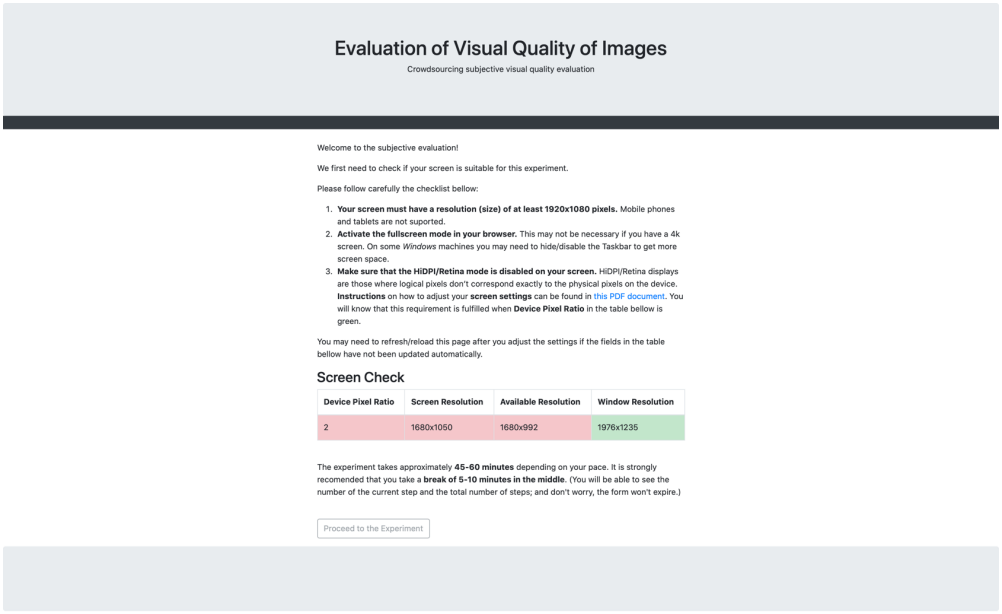


Figure 6. Worker screen validation layout for the case of incompatible screen size.

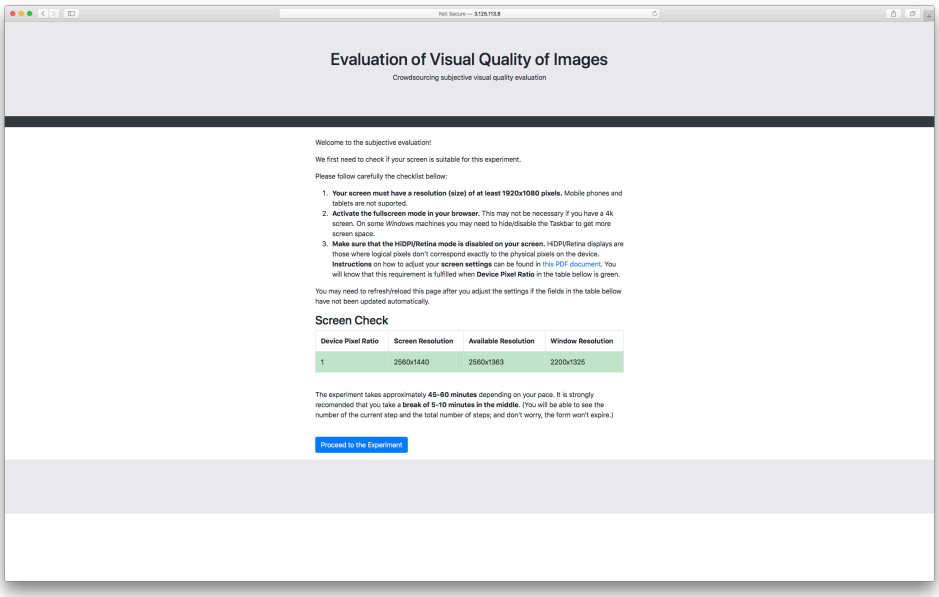


Figure 7. Worker screen validation layout for the case of fulfilled requirements.

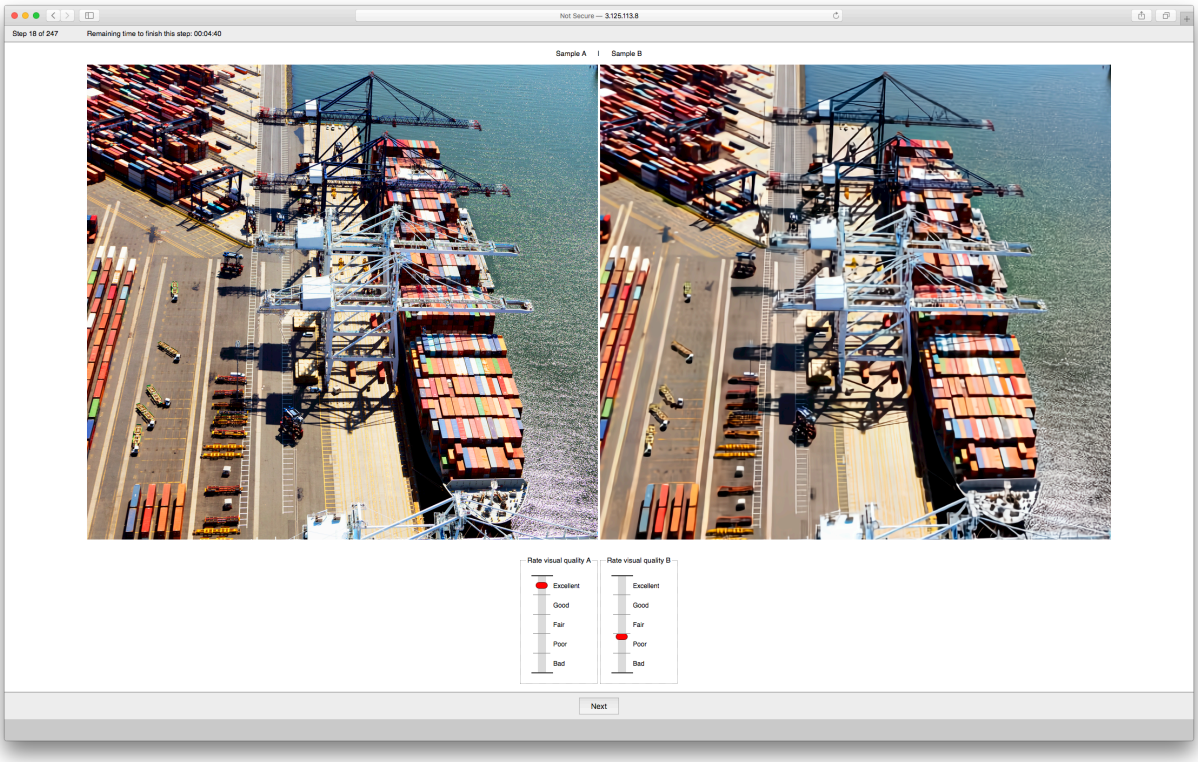


Figure 8. Layout of the voting step in QualityCrowd2 implementing DSCQS methodology.

7.3. Selected Test Images and Image Cropping

The images used for subjective evaluation are a subset of the test dataset images and its number was selected based on the number of proposals that were subjectively evaluated. These images were selected by experts based on their characteristics. A fixed set of 8 images was made from an initial selection of images. The images to be assessed during the subjective evaluation will correspond to crops of the decoded images such that relevant coding artifacts are included The size of the cropped images was selected in order to fit the images in the 1920x1080 display still leaving space for the scale at the bottom of each image. The subjective test images characteristics are shown in Table 2 and the cropped images used in the subjective evaluation are shown in Figure 9.

Table 2. Subjective test image cropping characteristics.

Input image	Original size	Cropped size	Top left corner point (x,y)
jpegai02 - Racing car	2144x1424	945×880	(323,164)
jpegai05 - Las Vegas sign	1336x872	945×872	(190,0)
jpegai06 - Train	1544x1120	945×880	(235,248)
jpegai07 - Windows	1472x976	945×880	(258,58)
jpegai09 - Port	1976x1312	945×880	(259,299)
jpegai10 - Curiosity Rover	2000x1128	945×880	(547, 185)
jpegai11 - Bell pepper	1744x1160	945×880	(394,174)
jpegai12 - Woman	1512x2016	945×880	(286,288)

Racing Car | Las Vegas Sign | Train | Windows

Port | Curiosity Rover | Bell pepper | Woman

Figure 9. Cropped images used for the subjective evaluation

7.4. Subject Population Statistics

The number of subjects after screening is 116, with 32 females and 84 males. The age goes from 18 to 70, with the age mean 34.72 and the age median is 32.50. The statistics on the screen resolution and countries are shown in Table 3.

Table 3. Statistics on screen resolution and countries.

ScreenSize	Subject
1920x1080	95
1920x1200	15
2560x1440	3
3440x1440	3
2048x1280	2
2560x1080	2
2560x1600	2
1920x1440	1
2736x1824	1
2880x1800	1
3840x2160	1

Country	Subject
United States	88
India	17
Brazil	8
United Kingdom	3
Honduras	2
Italy	2
Canada	1
Estonia	1
France	1
Greece	1
Not found	1

7.5. Subjective Scores Processing

The subjective scores were processed according to the following steps:

- Outlier detection: Outlier detection was performed according to ITU Rec. BT-500. In this case, 118 naïve subjects properly completed the experiments and 2 outliers were detected and their data have been excluded.
- MOS calculation: The MOS is computed independently for each test condition as:

$$\mathrm{MOS}_i = \frac{1}{N}\sum_{j=1}^{N} s_{ij}$$

  where N is the number of valid subjects and s_{ij} is the score by subject *j* for the test condition *i*.
- DMOS calculation: Because in DSCQS the source reference must also be graded by the subjects, instead of reporting the MOS for both the source reference (SRC) and processed stimuli (PS), the DMOS is calculated according to:

$$\mathrm{DMOS(PS)} = \mathrm{MOS(PS)} - \mathrm{MOS(SRC)} + \max(\text{rating scale})$$

7.6. Rate-MOS Experimental Results

The experimental results obtained in the subjective evaluation test are shown in Figure 10. In this case, rate-DMOS curves were plotted, including the 95% confidence intervals. As shown, team08, team06 has better performance compared to HEVC Intra, which is the most relevant benchmark included in these experiments. The gains of the learning-based image coding solutions compared to HEVC Intra depend on which image is considered but can be rather high as shown for image 07. Moreover, team06 and team08 have similar performance and for some images, one has clear better performance than the other.

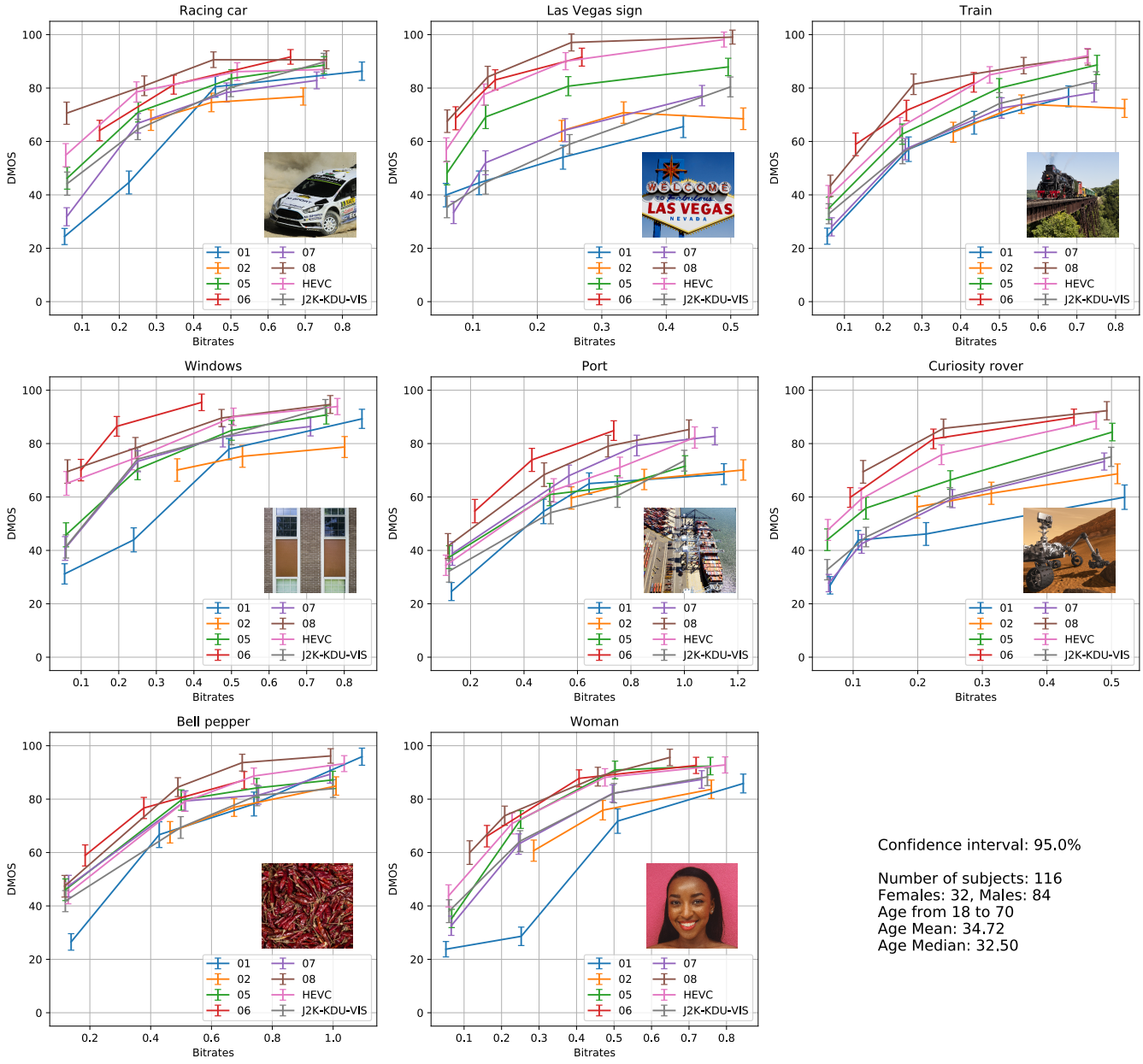


Figure 10. Differential mean opinion scores (DMOS) with 95% confidence intervals (CI) plotted against the bitrate for each codec grouped by contents.

References

[1] ISO/IEC JTC 1/SC29/WG1 M89086, "Four-Leaf Clover Response to JPEG AI Call for Evidence", 88th JPEG Meeting, Online, July 2020.

[2] ISO/IEC JTC 1/SC29/WG1 M89087, "NJU-VISION Response to JPEG AI Call for Evidence", 88th JPEG Meeting, Online, July 2020.

[3] ISO/IEC JTC 1/SC29/WG1 M89088, "NCTU Response to JPEG AI Call for Evidence", 88th JPEG Meeting, Online, July 2020

[4] ISO/IEC JTC 1/SC29/WG1 M89109, "Nokia Response to JPEG AI Call for Evidence", 88th JPEG Meeting, Online, July 2020

[5] Z. Wang, E. P. Simoncelli and A. C. Bovik, "Multi-scale Structural Similarity for Image Quality Assessment", 37th IEEE Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, November 2003.

[6] J. Ascenso, P. Akayzi, M. Testolina, A. Boev, E. Alshina "Performance Evaluation of Learning based Image Coding Solutions and Quality Metrics", ISO/IEC JTC 1/SC29/WG1 N85013, 85th JPEG Meeting, San Jose, USA, November 2019. Available here.

[7] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy and M. Manohara, "Toward A Practical Perceptual Video Quality Metric", [Online], Available here.

[8] H.R. Sheikh and A. C. Bovik, "Image Information and Visual Quality," IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, Canada, August 2004.

[9] L. Zhang, L. Zhang, X. Mou, D. Zhang, "FSIM: a Feature Similarity Index for Image Quality Assessment," IEEE Transactions on Image Processing, vol. 20, no. 8, pp. 2378-2386, August 2011.

[10] V. Laparra, J. Ballé, A. Berardino, and E. P. Simoncelli, "Perceptual Image Quality Assessment using a Normalized Laplacian Pyramid", S&T Symposium on Electronic Imaging: Conf. on Human Vision and Electronic Imaging, San Francisco, CA, USA, February 2016.

[11] Z- Wang, Q. Li, "Information Content Weighting for Perceptual Image Quality Assessment". IEEE Transactions on Image Processing, Vol. 20, No. 5, May 2011.

[12] D. Saupe, F. Hahn, V. Hosu, I. Zingman, M. Rana, and S. Li, "Crowdworkers Proven Useful: a Comparative Study of Subjective Video Quality Assessment," International Conference on Quality of Multimedia Experience (QoMEX), Lisbon, Portugal, June 2016.

[13] C. Keimel, J. Habigt, C. Horch, and K. Diepold, "Qualitycrowd: a Framework for Crowd-based Quality Evaluation," Picture Coding Symposium, Krakow, Poland, May 2012.

[14] ITU-R Recommendation BT.500-13, "Methodology for the Subjective Assessment of the Quality of Television Pictures," International Telecommunications Union, Geneva, Switzerland, 2012.

[15] ITU-T Recommendation P. 910, "Subjective Video Quality Assessment Methods for Multimedia Applications," International Telecommunication Union, Geneva, 2008.