PERFORMANCE ASSESSMENT OF DEEP LEARNING BASED IMAGE CODECS

Filipe M. Ferreira, Fernando Pereira, João Ascenso, Catarina Brites
Instituto Superior Técnico - Instituto de Telecomunicações

ABSTRACT

Nowadays, deep learning (DL) is becoming as one of the hottest topics in image compression. Since image quality assessment is a critical step for any multimedia application these days, subjective quality assessment methods have become crucial to measure the quality performance of multiple processing systems, notably image codecs. As one of the main focus of this paper, some selected DL methods were subjectively assessed through a Double-Stimulus Impairment Scale (DSIS) test. The results show that DL solutions were very competitive against some benchmarks, for example HEVC, achieving in general the best results. As next step, reference objective quality metrics were applied, confirming the obtained results for the perceptual metrics and showing that for MSE based metrics, there is still plenty of room for improvement. In the end, with the main objective to correlate both tests, Spearman and Pearson correlation metrics were computed, showing the MS-SSIM as the closest representation of the subjective results.

Index Terms **–** deep learning, image compression, subjective assessment, objective quality metrics, correlation metrics.

1. INTRODUCTION

This Thesis is focused on digital still images, notably one of its key technologies, specifically image coding. Basically, image coding is a technology able to represent a digital image with a reduced number of bits regarding its sample-based representation while providing the required target quality. Throughout the years some image codecs have been widely used as compression benchmarks. JPEG, the most widely used lossy solution, is based on the Discrete Cosine Transform (DCT) and is mainly used for digital photography. Other coding solutions have derived from the JPEG standard, for instance JPEG XT which includes a JPEG compliant base layer. On the other hand, the more recent JPEG 2000 image coding standard is based on discrete wavelength transforms. Other reference codecs like HEVC Intra or WebP have emerged along the years and offer today more efficient image coding solutions.

On the other hand, deep learning (DL) is the field of study that takes advantage of specific computational models composed of multiple layers, called neural networks, to represent data in many levels of abstraction. There are many types of neural networks being developed and mastered today with the most varied applications. For instance, in Ren et al. [1], deep neural networks are being used to detect specific elements in images with a certain degree of confidence, see Figure 1.

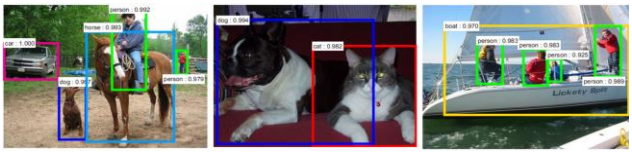

Figure 1: Deep learning application examples for object detection in images [1].

Motivation

As it could be expected, deep neural networks have also arrived to the image coding field, targeting the improvement of conventional image compression results, for example using recurrent (RNN), convolutional and adversarial neural networks or variational auto-encoders (VAE). However, since these techniques are rather recent there are no comparative studies available in the literature, notably in terms of formal subjective assessment. Recognizing the needs from the digital image industry, which is constantly evolving, notably asking for more compression, many companies and universities have started research activities to develop novel image codecs based on deep learning architectures. The need for more efficient image compression motivates this research, ultimately targeting replacing the conventional image codecs

The main objective of this study is first to report the compression efficiency performance of emerging deep learning based image codecs, conducted through subjective and objective assessments, notably in comparison with several available, conventional image coding benchmarks. In the end, this studies are correlated with each other, with the main focus to evaluate which objective quality metric can correlate better with the subjective results. Note that the formal subjective assessment is the second performed after the JPEG AI tests which have not yet been published and considered a smaller number of deep learning based codecs.

To achieve its objectives, this paper is organized as follows: Section 2 reviews some DL-based image codecs. Posteriorly, Section 3 describes the subjective assessment performed and the corresponding results, while Section 4 is dedicated to the objective assessment. Finally, Section 5 covers the correlation between both assessments and Section 6 presents not only the main conclusions taken, but also the plan for future stages of this work.

2. DEEP LEARNING BASED IMAGE CODECS: BRIEF REVIEW

Keeping in mind that in the last few years, NNs became a key technology to improve the performance of several fields of study, this section is dedicated to three relevant selected papers, adopting different types of NNs, more precisely VAEs and RNNs. Note that the solutions used for assessment come from these sources.

Full Resolution Image Compression with Recurrent Neural Networks (RNN-C)

In [2], *Toderici et al.* presented a full-resolution lossy image coding method based on RNNs. The proposed coding solution represented a landmark in this field since it offered one of the first solutions capable to show competitive results across a wide range of compression rates and could be applied to images of arbitrary sizes. Basically this solution brought a way to encode and decode images by iterations and with memory, taking advantage, additionally, of recurrent entropy coding models.

Variational Image Compression with a Scale Hyperprior (BH)

The solution described by *Ballé et al.* in 2018 takes advantage of VAEs to propose an end-to-end optimized deep image coding solution [3]. This architecture includes side information in the form of a hyperprior to efficiently capture spatial dependencies in the latent representation space. The hyperprior makes it possible

to learn an entropy coding model in a similar way as the core compression model learns the image representation.

Aiming performance testing, this method could be optimized for the Mean Square Error (MSE) based metrics, called as BH-E in this document, and for the MS-SSIM perceptual metric, named BH-M.

Joint Autoregressive and Hierarchical Priors for Learned Image Compression (MM)

The third method includes a new model introduced by *Minnen et al.* in [4]. This new model combines hierarchical entropy with autoregressive priors in an image coding context, complementing each other to achieve great results. As the previous method covered, this one could be optimized for both MSE and MS-SSIM metrics, so two distinct solutions will be referred further ahead as MM-E and MM-M, respectively.

3. DEEP LEARNING BASED IMAGE CODING: SUBJECTIVE QUALITY EVALUATION

In general, it is rather clear the increasing demand for high quality images, not only for mobile phones but also for personal computers and tablets, for example. Since image quality assessment is nowadays a critical step for any multimedia application, subjective quality assessment methods have become fundamental to measure the quality performance of multiple processing systems, notably image codecs. In this context, this section describes the subjective methodologies and conditions used to evaluate some of the DL based coding methods introduced before.

3.1 Test Material and Preparation

Prior to the subjective tests, there are several important elements to select and define like the dataset to use, the type of screen and its resolution and the image resolution to code and assess, as well as all the image preparation process before applying the selected coding methods.

The selected image dataset was the JPEG AI, a PNG format dataset and fixed bit-depth of 8 bits, used for image coding experiments with learning-based image codecs [5]. The spatial resolution of its images varies between 960×642 and 8K. Since the subjective assessment tests can only include a limited number of images, 8 images were picked from the available 40 after careful visual inspection, aiming visual diversity and content that could grant difficulties in terms of image coding. The picks are represented in Figure 2. Note that for better referencing, specific names were given to each of the selected images, since the original labels were not discriminatory enough.

Figure 2: Set of JPEG AI dataset images selected for subjective assessment [5].

When it comes to the display conditions, a Dell P2715Q monitor [6] has been selected to take advantage of the large image spatial resolutions. It was also important to define how the images should be presented to the subjects during the quality assessment process. Since a Double Stimulus Impairment Scale (DSIS) assessment method was selected, where two images are put side-by-side on the screen, it was critical to define an image cropped resolution that could fit two images side-by-side on the screen. In order to fit two images in the screen without changing their original aspect, the adopted strategy was to crop the images to a dimension of 1904×2048 pixels, looking to include the most relevant point of interest in the image. These dimensions were defined also taking into account that some vertical spacing, i.e. a grey vertical bar, should be kept between the two images under relative assessment for better separation. The crop performed on the Caterpillar image is represented in Figure 3.

Figure 3: Crop performed on the Caterpillar original image (left) and final test image (right).

At the end, there were also two special cases: the *Church* was downsampled to 2048×1360 pixels before cropping, since it was not possible to cover the most relevant point of interest in the image; the *Memorial* image had smaller original height than the established crop height, so its original value was maintained while changing the width to 1904 pixels.

3.2 Deep Learning based Test Codecs and Benchmarks

The coding process was applied not only for the selected DL based codecs but also for relevant benchmarks, notably available image coding standards. Referred in Section 2, three DL-based methods, corresponding to a total of 5 used solutions, were chosen:

- **RNN-C** - *Toderici et al.* [2] trained Residual GRU model [7] available in Python and Tensorflow.
- **BH-E and BH-M** - *Ballé et al.* [3] hyperprior model, with two pre-trained models available [8], one optimized for the MSE, other for the MS-SSIM.
- **MM-E and MM-M** - *Minnen et al.* [4] new model combining hierarchical entropy with autoregressive priors [8]. Again, two pre-trained models were available, one optimized for the MSE and another for the MS-SSIM.

When it comes to the benchmark image codecs, both JPEG XT and JPEG 2000 were selected, as well as the Intra mode of the HEVC video coding standard and the Intra mode of WebP codec. It is important to mention that, before coding the images with JPEG 2000, WebP and HEVC Intra codecs, a RGB to YCbCr conversion was applied, with 4:2:0 chroma subsampling, as this is the format commonly used by these codecs.

The various codecs were controlled using various parameters to obtain different rate-distortion (RD) points. Thus, after encoding each image, its corresponding bitrate was computed in bit per pixel (bpp), obtained dividing the bits for the encoded image by the image resolution. The selected base rates were: 0.06, 0.125, 0.25 and 0.5 bpp, but to conduct the test, some had to be adapted. Annex A contains a table with all the chosen bpp values. Overall, each coding solution was used to code 8 images, each with 4 different rates, thus leading to a total of 32 different decoded images for each codec. Since 9 codecs have been selected, a total of **288** decoded images have been used for subjective assessment.

3.3 Subjective Evaluation Protocol

The Double-Stimulus Impairment Scale (DSIS) method has been selected from a vast list of options and, in this protocol, a subject is presented with a series of image pairs, each pair containing an unimpaired reference image and the same image impaired. The subject is requested to compare them, voting on the impaired image quality (always in comparison to the reference image), according to the following Impairment Scale:

1. Very Annoying.
2. Annoying.
3. Slightly Annoying.
4. Perceptible, but not Annoying.

5. Imperceptible.

The series of image pairs, with different images and several levels of degradation, is presented to the subject in a random order, covering all conditions once and never repeating any condition. In the end, the entire set of opinion scores (for all stimulus) given by the subject are gathered and after the Mean Opinion Score (MOS) is computed for each impaired image.

3.4 Subjective Test Results

First, when it comes to the subjects and according to Recommendation ITU-R BT.500-13 [9], expert or non-expert viewers could have been selected, as long as they were not directly involved in the project. Since at least 15 subjects are needed to obtain statistically meaningful conclusions, the DSIS test was conducted with a total of 18 subjects to take into account the possibility that some could be outliers. In this case, no outliers were detected. After computing the average MOS scores, the 95% confidence interval was computed for each, while following the Annex 2 of [9]. A total of 5184 scores were obtained and their distribution is presented in Figure 4. The results show a trend towards the higher scores as the number of the 4 and 5 scores is much larger than the number of 1 and 2 scores.

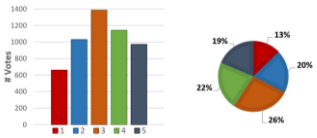

Figure 4: Overall DSIS scores distribution.

For each test image, three MOS-rate charts were constructed; in these charts, the vertical bars around the average MOS represent the respective 95% confidence interval. The first chart represented the MOS-rate performance for benchmark codecs; the second chart included the MOS-rate performance for the DL coding solutions; the third chart gathered the best two solutions from each of the previous two charts to better compare the best benchmark and DL image coding solutions.

MOS-rate charts for *Tiger* and *Texan* are shown in Figure 5. The *Tiger* had some good areas to detect impairments and, overall, all the coding solutions have shown balanced scores, covering the full scale, and maintained a rising trend with rate. For the *Texan* though, it showed to be a bit tricky for subjects to evaluate, because it did not have clear focus points where impairments could be easily detected

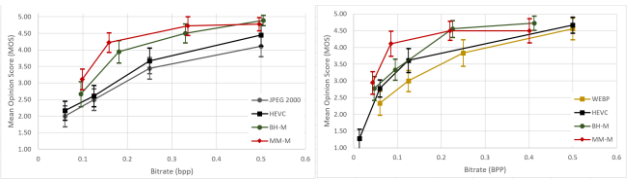

Figure 5: *Tiger* (left) *and Texan* (right) MOS versus rate charts for best solutions.

The image *Girl* did not show to be an easy image to detect impairments, just like *Zip*. This may be a justification for the high scores obtained. Figure 6 shows the MOS charts for both.

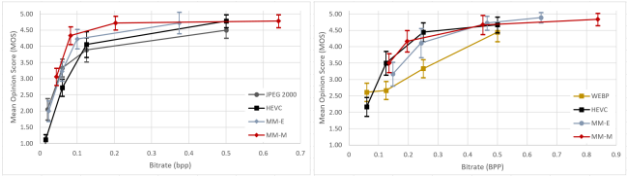

Figure 6: *Girl* (left) and *Zip* (right) MOS versus rate charts for the best solutions.

The *Emperor* is a really detailed image and thus differences between the reference and impaired images were easily spotted. When it comes to the *Caterpillar*, it was a bit trickier. The obtained MOS-rate charts are presented in Figure 7.

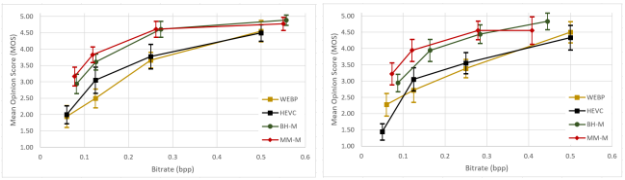

Figure 7: *Emperor* (left) and *Caterpillar* (right) MOS versus rate charts for the best solutions.

The *Memorial* and the *Church* revealed to be a great test images since these showed much detail and it was easy to find impairments. The *Church* even features a rather small airplane trace that has a large impact when detecting if the image was impaired or not. The MOS-rate charts are presented in Figure 8.

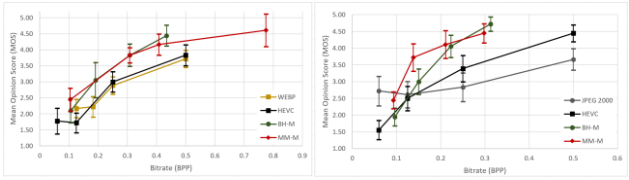

Figure 8: *Memorial* (left) and *Church* (right) MOS versus rate charts for the best solutions.

In general, the obtained MOS scores and MOS-rate charts did not show major incongruences and even allow to achieve pretty conclusive results, notably considering the wide variety of content and detail present in the test images. The major conclusions of this subjective quality assessment study are:

1. JPEG XT was the worst performing benchmark coding solution as it is an older codec; in fact, it scored really poor for some of the images assessed, notably for the lower rate. For the benchmarks, a rather natural order was maintained, with HEVC at the top and WebP and JPEG 2000 competing to come in second place and even doing similar for some specific rates.

2. Among the DL-based coding solutions, naturally, the MS-SSIM optimized solutions were in general on the top since this was a subjective test. Overall, MM-M was clearly the solution achieving the best performance.

3. Overall, DL-based coding solutions clearly outperformed the benchmarks, even noticing the RNN-C have showed a fairly bad performance on the tests.

Although it was not expected due to their recent emergence, the DL-based coding solutions, notably those MS-SSIM optimized, clearly outperformed the best conventional image codecs such as JPEG 2000 and HEVC Intra. The MS-SSIM optimization may have played a major role, like predicted initially, since the benchmarks are optimized for the MSE. These conclusions bring interesting hints for the next section where an objective assessment study was performed using several objective quality metrics and involving a larger number of test images.

4. DEEP LEARNING BASED IMAGE CODING: OBJECTIVE QUALITY EVALUATION

After conducting a subjective quality evaluation test, a very natural next step is to perform an objective quality study with some widely used objective quality metrics, notably to assess how reliable these metrics are. Considering the images used for the subjective assessment and all the other images present in the JPEG AI dataset, some renowned quality metrics were computed for each of these images taking into account a reference. The five chosen objective quality metrics were: $PSNR_Y$, $PSNR_{YCbCr}$, MS-SSIM, SSIM and VIF. All the selected quality metrics were computed with the *Sewar* python package [10]. Figure 9

represents average RD charts for the $PSNR_{YCbCr}$ (left) and MS-SSIM (right) on the JPEG AI dataset.

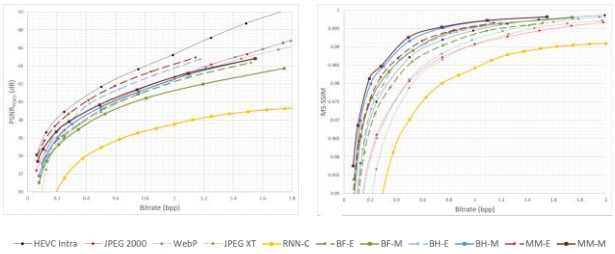

Figure 9: *PSNRYCbCr* (left) and *MS-SSIM* (right) RD performance on the JPEG AI dataset.

In general, the obtained RD performance results matched what was expected, with HEVC being the best for pixel-wise metrics such as PSNR and the most recent DL codec being the best for more perceptual metrics, such as MS-SSIM. Overall, for the image coding benchmarks, a rather natural order was maintained, notably considering their 'age'. This also confirms some results present in the literature depicted in Section 2. Among the DL-based coding solutions, naturally, the MS-SSIM optimized coding solutions come on top for SSIM and MS-SSIM; the MSE optimized solutions come on top for $PSNR_Y$, $PSNR_{YCbCr}$; and for VIF the results are rather close, with a slight advantage for the MSE optimized coding solutions, showing that although being a perceptual metric, MS-SSIM optimization acts negatively in terms of performance for VIF.

5. SUBJECTIVE AND OBJECTIVE QUALITY ASSESSMENT CORRELATION

After the subjective assessment of a selected set of images and objective assessment for a set of quality metrics, it is possible to obtain the correlation between objective and quality scores and from that evaluate the accuracy of the objective quality metrics.

To have the same scale and behaviour for objective metrics and MOS values, before the performance evaluation, it is necessary to apply a non-linear regression with a logistic function to the objective values, obtaining all the data points into the interval of MOS scores (minimum 1 and maximum 5). In this case, the logistic function was taken from Farid et al. [11] and the predefined parameters were used. The nonlinear regression is obtained using Matlab's *nlinfit* function and then, using the logistic function with new parameters obtained, Matlab's *nlpredci* function is applied, obtaining MOS_p predictions.

After obtaining the MOS_p scores, two important coefficients could be computed to study which objective metric is more correlated with the subjective assessment results, i.e. the objective metrics performance: Pearson (PLCC) and Spearman's Rank (SROCC) Correlation Coefficients. Naturally, the data used for the performance assessment of quality metrics were all the objective quality scores computed for the images assessed subjectively, as well as all the MOS computed for the same images. To present these results, three distinct correlation tests were performed for each of the 5 objective quality metrics: one for all the benchmarks, another for all the DL based solutions and finally one for all the codecs assessed. The PLCC and SROCC results allow to perform a quantitative analysis and these can be observed in Table 1 and Table 2.

Table 1: Pearson Correlation Coefficient results.

METRIC	BENCHMARKS	DL SOLUTIONS	ALL CODECS
$PSNR_Y$	0.7634	0.6388	0.6898
$PSNR_{YCbCr}$	0.7285	0.6524	0.7016
MS-SSIM	**0.8515**	**0.776**	**0.8292**
SSIM	0.7913	0.697	0.7584
VIF	0.7919	0.6898	0.7581

Table 2: Spearman Rank Correlation Coefficient results.

METRIC	BENCHMARKS	DL SOLUTIONS	ALL CODECS
$PSNR_Y$	0.7569	0.5829	0.6646
$PSNR_{YCbCr}$	0.7191	0.5967	0.6794
MS-SSIM	**0.8466**	**0.7375**	**0.8127**
SSIM	0.7797	0.6474	0.7288
VIF	0.7836	0.6601	0.738

It can be denoted that perceptual metrics like the MS-SSIM, SSIM and VIF obtain the best correlation results for PLCC and SROCC, which was expected since these are considered more correlated with human perception. For the benchmarks, all the metrics achieve higher results when compared to the DL solutions. This allows to conclude that these metrics are more reliable for the benchmarks, which was expected since most of these metrics were designed considering the available codecs at that time. Overall, the best metric was clearly the MS-SSIM for both PLCC and SROCC. Charts with the MOS scores as function of MOSp were plotted to see which metric came closer to the subjective results. The charts obtained for the MS-SSIM are shown in Figure 10. The red line shows how strong the correlation was, where the closer points where more correlated than the others.

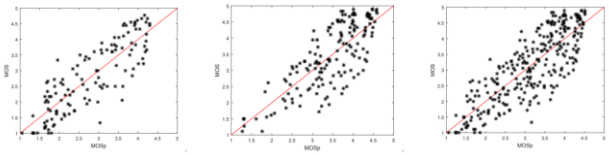

Figure 10: MOS as function of MOS_p for the benchmarks (left), DL solutions (middle) and all codecs (right).

The results obtained for SROCC are reasonably close to the PLCC, which means that the ranking order is maintained considering the limitations in terms of PLCC. In terms of metric rankings, after the MS-SSIM comes VIF and SSIM, which showed close competitive results for PLCC, while the VIF overcame the SSIM for the SROCC. In the last place come the MSE based metrics: PSNRY and $PSNR_{YCbCr}$, which was expected due to their well-known limitations. Specifically, for the benchmarks, PSNRY achieves better results than the $PSNR_{YCbCr}$ but this order changes for the deep learning solutions. Overall, it can be seen that both correlation measures are not very high and there is a significant room for improvement. Even MS-SSIM which represents the best metric evaluated, has plenty of room to improve and reach higher values, especially for deep learning solutions. Thus, it can be said that although the results were expected, new metrics that target the artifacts introduced by deep learning codecs are needed.

6. SUMMARY AND FUTURE WORK PLAN

Overall, it can be denoted that for perceptual evaluation, taking into account the human visual system and metrics that can simulate our visual judgement, the latest DL-based compression solutions achieve great results. In the future, becoming more competitive for MSE and MS-SSIM based metrics jointly could be a point of focus. In this context, future research work should be focused on even more advanced DL-based coding methods. A vast selection of NN types can be taken into account as, for example, no GAN-based coding solutions were considered.

Finally, to complete this study, the Absolute Category Rating with Hidden Reference (ACR-HR) protocol test could be used. This type of subjective test would allow assessing the decoded images, e.g. in terms of their realism, artificial looking ,etc., and would complement the results and conclusions obtained with the DSIS protocol, targeting a more complete understanding of the coding behavior and impacts of the DL-based coding solutions.

7. BIBLIOGRAPHY

[1] S. Ren, K. He, R. Girshick, J. Sun, "Faster R-CNN: Towards Real Time Object Detection with Region Proposal Networks", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 39, nº 6, pp. 1137-1149, June 2017

[2] G. Toderici et al., "Full Resolution Image Compression with Recurrent Neural Networks", IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, July 2017.

[3] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, N. Johnston, "Variational Image Compression with a Scale Hyperprior", International Conference on Learning Representations (ICLR), Vancouver, Canada, May 2018.

[4] D. Minnen, J. Ballé, G. Toderici, "Joint Autoregressive and Hierarchical Priors for Learned Image Compression", arXiv preprint arXiv:1809.02736, September 2018.

[5] J. Ascenso and P. Akayzi, "JPEG AI Image Coding Common Test Conditions", ISO/IEC JTC 1/SC 29/WG 1 N84035, 84th Meeting, Brussels, Belgium, July 2019

[6] "Dell P2715Q/P2415Q User's Guide", Rev. A00, October 2014

[7] [Online] Available: https://github.com/tensorflow/models/tree/master/research/compression/image_encoder [Accessed 24 09 2019]

[8] [Online] Available: https://tensorflow.github.io/compression/ [Accessed 24 09 2019]

[9] International Telecommunication Union, "Methodology for the Subjective Assessment of the Quality of Television Pictures", ITU-R BT.500-11, January 2012

[10] [Online] Available: https://pypi.org/project/sewar/ [Accesed 29 09 2019]

[11] M.S. Farid, M. Lucenteforte, M. Grangetto, "Perceptual Quality Assessment of 3D Synthesized Images", Proc. IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, July 2017

8. ANNEXES

8.1 Chosen Bitrates for Subjective Assessment

In summary, the coding rates selected for each image and codec are presented in Table 3.

Table 3: Bitrates used for subjective evaluation and correlation analysis for each image and coding method.

	Tiger	*Girl*	*Emperor*	*Caterpillar*	*Memorial*	*Zip*	*Church*	*Texan*
JPEG XT	0.06 0.125 0.25 0.5							
JPEG 2000	0.06 0.125 0.25 0.5	0.02 0.06 0.125 0.5	0.06 0.125 0.25 0.5	0.03 0.06 0.125 0.5	0.06 0.125 0.25 0.5			0.03 0.06 0.125 0.5
WebP	0.06 0.125 0.25 0.5	0.045 0.06 0.125 0.5	0.06 0.125 0.25 0.5		0.125 0.182 0.25 0.5	0.06 0.125 0.25 0.5		
HEVC INTRA	0.06 0.125 0.25 0.5	0.015 0.06 0.125 0.5	0.06 0.125 0.25 0.5	0.05 0.125 0.25 0.5	0.06 0.125 0.25 0.5			0.013 0.06 0.125 0.5
RNN-C	0.125 0.25 0.5 0.75							
BH-E	0.09 0.147 0.24 0.61	0.048 0.065 0.133 0.242	0.089 0.123 0.238 0.643	0.097 0.137 0.217 0.615	0.14 0.231 0.543 0.775	0.177 0.245 0.479 0.673	0.125 0.177 0.345 0.468	0.048 0.065 0.133 0.495
BH-M	0.096 0.181 0.33 0.505	0.06 0.106 0.219 0.365	0.082 0.126 0.273 0.557	0.087 0.164 0.284 0.445	0.106 0.191 0.308 0.434	0.134 0.208 0.314 0.453	0.096 0.15 0.223 0.313	0.047 0.095 0.226 0.412
MM-E	0.075 0.138 0.223 0.567	0.023 0.059 0.1 0.374	0.074 0.11 0.217 0.436	0.067 0.114 0.185 0.576	0.093 0.324 0.53 0.807	0.148 0.243 0.465 0.647	0.09 0.154 0.315 0.438	0.035 0.052 0.114 0.447
MM-M	0.099 0.159 0.336 0.496	0.044 0.082 0.203 0.641	0.078 0.119 0.262 0.55	0.073 0.121 0.279 0.408	0.104 0.308 0.408 0.775	0.135 0.196 0.451 0.838	0.093 0.138 0.211 0.298	0.043 0.085 0.22 0.401

8. ANNEXES

8.1 Chosen Bitrates for Subjective Assessment

In summary, the coding rates selected for each image and codec are presented in Table 3.

Table 3: Bitrates used for subjective evaluation and correlation analysis for each image and coding method.

	Tiger	*Girl*	*Emperor*	*Caterpillar*	*Memorial*	*Zip*	*Church*	*Texan*
JPEG XT	0.06 0.125 0.25 0.5							
JPEG 2000	0.06 0.125 0.25 0.5	0.02 0.06 0.125 0.5	0.06 0.125 0.25 0.5	0.03 0.06 0.125 0.5	0.06 0.125 0.25 0.5			0.03 0.06 0.125 0.5
WebP	0.06 0.125 0.25 0.5	0.045 0.06 0.125 0.5	0.06 0.125 0.25 0.5		0.125 0.182 0.25 0.5	0.06 0.125 0.25 0.5		
HEVC INTRA	0.06 0.125 0.25 0.5	0.015 0.06 0.125 0.5	0.06 0.125 0.25 0.5	0.05 0.125 0.25 0.5	0.06 0.125 0.25 0.5			0.013 0.06 0.125 0.5
RNN-C	0.125 0.25 0.5 0.75							
BH-E	0.09 0.147 0.24 0.61	0.048 0.065 0.133 0.242	0.089 0.123 0.238 0.643	0.097 0.137 0.217 0.615	0.14 0.231 0.543 0.775	0.177 0.245 0.479 0.673	0.125 0.177 0.345 0.468	0.048 0.065 0.133 0.495
BH-M	0.096 0.181 0.33 0.505	0.06 0.106 0.219 0.365	0.082 0.126 0.273 0.557	0.087 0.164 0.284 0.445	0.106 0.191 0.308 0.434	0.134 0.208 0.314 0.453	0.096 0.15 0.223 0.313	0.047 0.095 0.226 0.412
MM-E	0.075 0.138 0.223 0.567	0.023 0.059 0.1 0.374	0.074 0.11 0.217 0.436	0.067 0.114 0.185 0.576	0.093 0.324 0.53 0.807	0.148 0.243 0.465 0.647	0.09 0.154 0.315 0.438	0.035 0.052 0.114 0.447
MM-M	0.099 0.159 0.336 0.496	0.044 0.082 0.203 0.641	0.078 0.119 0.262 0.55	0.073 0.121 0.279 0.408	0.104 0.308 0.408 0.775	0.135 0.196 0.451 0.838	0.093 0.138 0.211 0.298	0.043 0.085 0.22 0.401