Subjective evaluation of JPEG XR image compression

Francesca De Simone[a], Lutz Goldmann[a], Vittorio Baroncini[b], Touradj Ebrahimi[a]

[a]Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[b]Fondazione Ugo Bordoni (FUB), Rome, Italy

ABSTRACT

In this paper a procedure for subjective evaluation of the new JPEG XR codec for compression of still pictures is described in details. The new algorithm has been compared to the existing JPEG and JPEG 2000 standards when considering compression of high resolution 24 bpp pictures, by mean of a campaign of subjective quality assessment tests which followed the guidelines defined by the AIC JPEG ah-hoc group. Sixteen subjects took part in experiments at EPFL and each subject participated in four test sessions, scoring a total of 208 test stimuli. A detailed procedure for statistical analysis of subjective data is also proposed and performed. The obtained results show high consistency and allow an accurate comparison of codec performance.

Keywords: subjective quality assessment, high resolution pictures, codec performance, JPEG XR, JPEG 2000, JPEG.

1. INTRODUCTION

When a new technology for compression of digital still or moving pictures is submitted to the attention of the international standardization community, ad-hoc groups of experts are created to evaluate it. The evaluation usually consists of comparative studies with existing or concurrent technologies to test the compression efficiency achieved by the proposed coding algorithm, computational complexity, and any additional functionalities. The compression efficiency of different coding strategies can be reliably compared only by means of subjective tests, carried out at the premises of several independent institutions and according to common evaluation methodologies defined by experts. Examples of such activities can be found in the standardization of H.264/AVC technology for video compression,[1] JPEG or JPEG 2000 technologies for image compression,[2] as well as for the recent JPEG XR compression technology,[3] which is the latest image compression standard by the JPEG committee.[4]

The need for subjective tests is due to the fact that the compression efficiency of a coding algorithm expresses its ability to maximize the visual quality of a compressed image or video sequence versus the number of bits used to represent it. Even if a considerable effort has been spent by the research community to develop algorithms which can objectively evaluate the quality of digital pictures, i.e. objective quality metrics, the ability of existing metrics to predict human judgement remains limited. Furtheremore, the lack of standardization in the field of objective quality assessment and the lack of extensive and reliable comparisons of state-of-the-art metrics make the results obtained using existing algorithms not very reliable.[5] Thus, the benchmark for any kind of quality assessments remains the subjective results collected by means of experiments which are time consuming and have to be carefully designed.

In subjective tests, a group of subjects is asked to watch a set of still or moving pictures and to rate their quality. The scores assigned by observers are usually averaged in order to obtain a mean opinion scores (MOS), under the assumption that they follow a Gaussian distribution. In order to produce meaningful MOS values, the test material needs to be carefully selected and the subjective evaluation procedure must be rigorously defined. Although many recommendations by international standardization bodies exist on how to perform subjective

Further authors' information:
Francesca De Simone: E-mail: francesca.desimone@epfl.ch
Lutz Goldmann: E-mail: lutz.goldmann@epfl.ch
Vittorio Baroncini: E-mail: vittorio@fub.it
Touradj Ebrahimi: E-mail: touradj.ebrahimi@epfl.ch

Figure 1: MMSPG test room with calibrated Eizo CG301W screen and controlled lighting system.

quality evaluation of moving pictures, only a few standards are available which provide guidelines for performing subjective quality assessment of still pictures.[6]

In this paper, we present a detailed procedure for subjective quality evaluation of high resolution still pictures, by describing its application to performance evaluation of the JPEG XR compression algorithm. The new technology has been compared to existing JPEG and JPEG 2000 algorithms, considering compression of high resolution 24 bpp pictures, by means of a campaign of subjective quality assessment tests which followed the general guidelines provided in the core experiment plan defined by the AIC JPEG XR ad-hoc group.[7] Sixteen naive subjects took part in experiments at EPFL and each subject participated in four test sessions, scoring a total of 208 test stimuli. A detailed procedure for the statistical analysis of subjective data is also proposed, which allows an accurate comparison of codec performance.

The test conditions of our experiment, including description of test environment, dataset and configuration of coding algorithms, are detailed in section 2. The adopted test methodology, in terms of test design and statistical analysis of subjective data, is presented in section 3. Finally, section 4 discusses the results of the experiment, while concluding remarks are drawn in section 5.

2. TEST CONDITIONS

2.1 Laboratory

The experiment for subjective quality evaluation of JPEG XR was conducted at the Multimedia Signal Processing Group (MMSPG) quality test laboratory at EPFL (shown in figure 1), which is compliant with the recommendations for subjective evaluation of visual data issued by ITU-R.[8] The laboratory setup is intended to assure the reproducibility of results by avoiding involuntary influence of external factors.

An Eizo CG301W LCD monitor with native resolution of 2560x1600 pixels was used to display the test stimuli. The monitor was calibrated using an EyeOne Display2 color calibration device according to the following profile: sRGB Gamut, D65 white point, 120 cd/m^2 brightness and minimum black level. The room is further equipped with a controlled lighting system that consists of neon lamps with 6500 K color temperature. The illumination level measured on the screen is 30 lux and the ambient black level is 0.5 cd/m^2.

The experiments involved only one subject per display assessing the test material. The subject was seated in line with the center of the monitor, at a distance approximately equal to the height of the screen but was encouraged to vary the viewing distance whenever needed, to better inspect the high resolution image shown on the screen.

2.2 Dataset

In order to support data reproducibility and comparison with results of JPEG committee, the dataset established by the AIC JPEG ad-hoc group for evaluation of JPEG XR was used.[7] It contains 10 high resolution pictures at 24 bit per pixel (bpp) with a variety of characteristics, including different objects, color distributions and textures. Each image was adapted to the screen resolution and test methodology by cropping a representative region of 1280x1600 pixels. This allows display of two images side by side at native resolution on the screen.

Figure 2: Cropped images from the dataset established by the AIC JPEG ad-hoc group for evaluation of JPEG XR. The four top images form the training set while the remaining six are the test set.

The whole image set was split into a training set of four images (referred to as *p04*, *p14*, *p22*, *p30*) and a testing set of six images (referred to as *p01*, *p06*, *p10*, *bike*, *cafe*, *woman*). Figure 2 depicts the dataset.

2.3 Codecs

For the lossy compression of high resolution images three different image codecs were considered, including the widespread JPEG,[9] JPEG 2000[2] and the recently developed JPEG XR codec.[3] To facilitate experiment reproduction, the command line calls used for coding and the entire set of coded and decoded test pictures are available at *http://mmspg.epfl.ch/iqa*.

According to the guidelines provided in the core experiment plan for JPEG XR evaluation,[7] the following six test coding bitrates were selected: 0.25, 0.50, 0.75, 1.00, 1.25 and 1.50 bpp. While JPEG 2000 reference software provides a rate control option to achieve a precise target bit per pixel value, for JPEG and JPEG XR implementations the bitrate can be only controlled by varying an input parameter called "'quality factor"'. Thus, in order to obtain a certain target bitrate, the "'quality factor"' which leads to the closest bitrate lower than or equal to the target bitrate was selected.

2.3.1 JPEG

JPEG[10] is a block-based image compression standard developed in 1992. It exploits presence of redundant and irrelevant information within an image through several steps. First of all, a color space transformation from RGB to YCbCr is applied, to split the image into luminance and chrominance components. Since human visual system (HVS) is less sensitive to color details, chroma components are usually subsampled. The image is then divided into 8x8 non-overlapping blocks on which discrete cosine transform (DCT) is applied. The resulting DCT coefficients for each block are quantized and compressed using entropy coding such as Huffman or arithmetic coding.

For JPEG coding, the IJG implementation* version 6b was used. Images were encoded with the following parameters:

- 4:2:0 subsampling.
- Visually optimized quantization matrix.
- Huffman coding.

*`http://www.ijg.org/`

2.3.2 JPEG 2000

JPEG 2000[11] is a wavelet-based compression standard for still images, also used for compression of image sequences such as those in digital cinema. It has been designed by the JPEG committee who was also behind the original JPEG image compression. Beside outperforming the original DCT-based JPEG standard in terms of compression efficiency in many situations, JPEG 2000 offers a large number of features useful in multimedia applications.

For JPEG 2000 coding, the Kakadu implementation[†] version 6.0 was used. Two different configurations were considered. The first configuration uses chrominance subsampling and requires some external pre- and post-processing steps. The following parameters were used:

- Pre-processing (before encoding): RGB to YCbCr conversion and 4:4:4 to 4:2:0 downsampling.
- 64x64 code block size, 1 layer, no precincts, 9x7 wavelets, 5 decomposition levels and rate control.
- No visual weighting.
- Postprocessing (after decoding): 4:2:0 to 4:4:4 upsampling and YCbCr to RGB color conversion.

In order to estimate the influence of chrominance subsampling on the codec performance, a second configuration was used, encoding 4:4:4 RGB images directly without any subsampling. The parameters were the same as before but pre- and post-processing steps were not needed and visual weighting was applied. The visual weigthing was not used for the 4:2:0 coding since the weighting tables in JPEG 2000 have been designed and optimized for 4:4:4 content and their usage for 4:2:0 coding would be formally incorrect.

2.3.3 JPEG XR

JPEG XR is a recent compression algorithm based on the HD Photo technology developed by Microsoft Corporation for digital imaging applications.[12] JPEG XR has recently become an international standard by JPEG committee. It is characterized by a block-based image compression scheme. The codec design aims at optimizing image quality and compression efficiency while at the same time requiring low-complexity in encoders and decoders implementations. As a result, even if it uses many of the same fundamental building blocks as in other traditional image and video compression schemes, e.g. color conversion, transform, quantization, coefficient scanning and entropy coding, different design objectives and taking advantage of the latest progress in the field of compression, have led to different solutions when compared to other state of the art. In particular, JPEG XR uses a reversible Lapped Bi-orthogonal Transform (LBT) and an alternative coefficient coding approach.

Two implementations of the JPEG XR codec were considered in the tests. The first implementation was provided by Microsoft Corporation (MS) and includes an adaptive quantization steps size control.[12] The second was provided by Pegasus Corporation (PS) and relies on a model of human visual perception for optimazing the quantization strategy.[13] For both implementations the following parameters were used:

- 4:2:0 subsampling.
- One level overlapping filter.

3. TEST METHODOLOGY

While several methodologies have been proposed by international standardization bodies for subjective quality evaluation of moving images,[14] there is a lack of detailed recommendations for subjective quality evaluation of still images.[14] Although the former methodologies may serve as a good starting point for defining the latter, there are fundamental differences between video sequences and images, that have to be considered in an efficient test design. Evaluation of images is different from evaluation of moving images mainly due to the way eyes explore the visual area. When assessing moving images, the fovea tracks the region of interest in the scene.

[†] `http://www.kakadusoftware.com/`

Figure 3: Graphical user interface for the DSCQS methodology adapted to high resolution images.

On the other hand, when evaluating still images, eyes explore in a more extensive way the whole visual area, attempting to get the sensation of maximum resolution for any portion of the image.

In order to develop a suitable methodology for subjective quality evaluation of high resolution still images, we propose an adaptation of the double-stimulus continuous quality scale (DSCQS) method.[14]

3.1 DSCQS for still images

According to the original DSCQS methodology, two video sequences are displayed sequentially, one being always the reference, i.e. unimpaired video. Subjects are not told about the presence of a reference in every pair and, after visualization, are asked to rate the quality of both stimuli, using for each a continuous quality scale ranging from 0 to 100 with five distinct quality levels.

In order to adapt this evaluation methodology to still images, two main aspects should be considered. First of all, images can be compared more easily when viewed in parallel. Therefore, images are not shown sequentially but simultaneously by splitting the screen horizontally into two parts. Furthermore, a closer exploration of the high resolution content is supported by not limiting the time the user has to judge the quality of each image in the pair.

In order to reduce complexity of subjective tests, the proposed methodology has been simplified following the suggestion proposed in the JPEG XR core experiment plan.[7] The subject is informed about the presence of a reference and an impaired image in every test pair, whose position are randomly switched. Instead of judging the quality of both images, he/she is asked to detect the impaired image in the pair and rate only its quality.

Figure 3 shows the developed graphical user interface (GUI) that implements the adapted DSCQS methodology for high resolution images. Two images are displayed simultaneously on a split screen. After analyzing them, the subject clicks on the screen area and a voting dialog appears: the observer has to rate only the quality of the impaired image choosing the slider corresponding to that stimulus. This way the subject implicitly specifies which image is the reference and which is the impaired.

3.2 Training session

Before the start of tests, oral instructions are provided to subjects to explain their task. Additionally, a training session is performed to allow the viewer to familiarize with the assessment procedure and the GUI. Contents shown in the training session are not used in test sessions and the data gathered during training are not included in the final test results. The four training images, shown in figure 2, were coded at six bitrates of interest, as specified in the previous section. Out of these 24 images, five training samples were manually selected by an expert viewer such that the quality of each sample became representative of one categorical quality level on the

bpp	Session 1 X and 3 -				
	C1	C2	C3	C4	C5
0.25	-	X	-	X	-
0.5	X	-	X	-	X
0.75	-	X	-	X	-
1.0	X	-	X	-	X
1.25	-	X	-	X	-
1.5	X	-	X	-	X

bpp	Session 2 O and 4 •				
	C1	C2	C3	C4	C5
0.25	•	O	•	O	•
0.5	O	•	O	•	O
0.75	•	O	•	O	•
1.0	O	•	O	•	O
1.25	•	O	•	O	•
1.5	O	•	O	•	O

Figure 4: Distribution of different codecs and bitrates across four test sessions. Sessions 1 and 3 used images *p01*, *p06*, *p10* sessions 2 and 4 used images *bike*, *cafe*, *woman*. C1 to C5 indicate the codecs, i.e. JPEG 2000 4:2:0, JPEG 4:2:0, JPEG XR PS 4:2:0, JPEG XR MS 4:2:0, JPEG 2000 4:4:4, respectively.

rating scale. The training material was presented to each subject exactly as the test material, namely, in side by side image pairs where one of the two stimuli is always the reference image.

During the display of each training pair, the experimenter explained the meaning of each label reported on the scale, as summarized below: "In this experiment you will see pairs of high resolution still pictures on the screen that is in front of you. Each time a pair is shown, you should first detect which picture is the impaired one and then judge its quality by choosing one point on the continuous quality scale corresponding to the following categories:

- *Bad*: the detection of the reference image is very easy and strong artifacts (i.e. artifacts which destroy the scene structure or create new patterns) are detected in the entire impaired image.
- *Poor*: the detection of the reference image is very easy and noticeable artifacts (i.e. artifacts which can be detected at a first glance) are detected in a major part of the impaired image.
- *Fair*: the detection of the reference image is easy and several noticeable artifacts are detected in some localized areas of the impaired image.
- *Good*: you have to focus to detect the reference image but still the difference with the impaired image is detectable, since the latter presents some visible artifacts.
- *Excellent*: you are not sure about which image is the reference, because the two images look alike exactly (in this case you can randomly choose which stimuli to rate and assign a score 100) or they present some differences but it is difficult to define these differences as artifacts (in this case you consider the image which you prefer as the reference and rate the other).

3.3 Test sessions

The six test images shown in figure 2 were coded with the five codecs under analysis (JPEG 4:2:0, JPEG 2000 4:2:0, JPEG 2000 4:4:4, JPEG XR MS 4:2:0, and JPEG XR PS 4:2:0, respectively) at the six bitrates of interest (0.25, 0.50, 0.75, 1.00, 1.25, 1.50 bpp), defined in section 2.3, leading to an overall volume of different 180 test conditions.

Since this number of test conditions was too large for a single session, the experiment was split into four sessions. Each session included test material corresponding to three images (*p01*, *p06*, *p10* in sessions 1 and 3, *bike*, *cafe*, *woman* in sessions 2 and 4), all the five codecs, and only a subset of bitrates distributed across sessions according to the scheme shown in figure 4. Therefore, each session contained 45 test pairs. Additionally, four dummy pairs were included at the beginning of each session to stabilize viewer's judgment and three reference versus reference pairs were also included. This led to a total volume of 52 pairs per session.

3.4 Statistical analysis of the subjective results

The statistical analysis of subjective results is based on the assumption that the score m_{ij} obtained from subject i after he/she scores stimulus j defined by the controlled experimental variables (i.e. image content, bit per pixel value, codec) can be modeled as:

$$m_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \quad (1)$$

where μ is the overall mean, i.e. the mean score computed across all the subjects and the stimuli, α_i is the subject effect corresponding to the specific subject i, β_j is the treatment effect corresponding to the specific stimulus j, and ϵ_{ij} is a random variation caused by a range of uncontrolled variables, called experimental error and assumed to be normally distributed $N(0, \sigma^2)$.[15] The statistical analysis aims at answering two questions:

1. Is the variation in the subjective scores a results of the intended variation of controlled experimental variables (i.e. image content, bit per pixel value, codec) or is it more likely to be a random variation?
2. Since experiments are based on a limited sample of subjects, is it possible to draw general conclusions which are valid for the entire population?

Additionally, we want to understand whether the difference between estimated means for different codecs are statistically significant.

The different steps of the statistical analysis, applied in order to answer these questions and to obtain the final results discussed in section 4, are detailed in the following subsections. The results of different sessions have been merged together before performing the statistical analysis of the data, assuming that no re-alignment procedure was needed across them.

3.4.1 Distribution of the data

In order to perform the statistical analysis correctly, it is important to know what is the distribution of the data under analysis. In particular, if the data is normally distributed or it can be transformed to normally distributed, it can then be summarized by arithmetic mean value and variance or standard deviation and can be analyzed using parametric statistics. However, if the assumption of normality is not verified, the median value could be a better descriptor of the central tendency of a distribution and non parametric methods of analysis need to be applied.

Distribution of the collected data can be analyzed for each subject, across different test conditions, or for each test condition, across different subjects. We used Shapiro-Wilk test to verify the normality of distributions.[16] The results of this test showed that scores distributions for each subject across different test conditions, as expected, are not normally distributed, while the majority of scores distributions across subjects (54%) are normal or close to normal (mean p-value equal to 0.25). The results of this test justify the processing applied to the data which is detailed in the next subsections.

3.4.2 Outlier detection and removal

The screening of subjects was performed according to the guidelines described in section 2.3.1 of annex 2 of ITU-R BT. 500-11 recommendation.[14] First, for each stimulus, it is tested whether the distribution of scores across subjects is normal or not. This is done by calculating the Kurtosis coefficient of the distribution: if the coefficient is between 2 and 4, the distribution is assumed to be normal. Then, the score of each observer is compared with an upper and a lower threshold computed as the mean value plus and minus the standard deviation associated to that stimulus (times two, if normal, or times 20, if non-normal). For each subject, every time his/her score is found above the upper threshold a counter P_i is incremented. Similarly, every time his/her score is found below the lower threshold, a counter Q_i is incremented. Finally, the following two ratios are calculated: $P_i + Q_i$ divided by the total number of scores from each subject for the whole session, and $P_i - Q_i$ divided by $P_i + Q_i$ as an absolute value. If the first ratio is greater than 5% and the second ratio is less than 30%, then subject i is an outlier and all his/her results are removed from the experiments.

3.4.3 Mean opinion scores

After the outlier removal, the mean opinion score was computed for each test condition j (i.e. combination of image content, bit-per-pixel value and codec) as:

$$MOS_j = \frac{\sum_{i=1}^{N} m_{ij}}{N} \tag{2}$$

where N is the number of valid subjects and m_{ij} is the score by subject i for the test condition j.

3.4.4 Relationship between the true and the estimated mean values

The relationship between the estimated mean values based on a sample of the population (i.e. the subjects who took part in our experiments) and the true mean values of the entire population is given by the confidence interval of estimated mean. Due to the small number of subjects, the $100 \times (1-\alpha)\%$ confidence intervals (CI) for mean opinion scores were computed using the Students t-distribution, as follows:

$$CI_j = t(1-\alpha/2, N) \cdot \frac{\sigma_j}{\sqrt{N}} \tag{3}$$

where $t(1-\alpha/2, N)$ is the t-value corresponding to a two-tailed t-Student distribution with $N-1$ degrees of freedom and a desired significance level α (equal to 1-degree of confidence). N corresponds to the number of subjects after outliers detection, and σ_j is the standard deviation of a single test condition across subjects. The interpretation of a confidence interval is that if the same test is repeated for a large number of times, using each time a random sample of the population, and a confidence interval is constructed every time, then $100 \times (1-\alpha)\%$ of these intervals will contain the true value. We computed our confidence intervals for an α equal to 0.05, which corresponds to a degree of significance of 95%.

3.4.5 Relationship between estimated mean values

In order to understand whether the difference between two MOS values corresponding to two different codecs is statistically significant, a hypothesis test was used. Particularly, for each bit per pixel value and test image, a two-sided Welch's t-test was conducted to compare all the codecs pairwise.[16] We will refer to the two codecs in the pairwise comparison as codec A and codec B. The Null hypothesis under test is that the results obtained for codec A and for codec B are independent random samples from a normal distributions with equal means, against the alternative hypothesis that means are not equal:

$$H_0 : MOS_A = MOS_B \tag{4}$$

$$H_a : MOS_A \neq MOS_B \tag{5}$$

where MOS_A is the mean opinion score for codec A and MOS_B is the mean opinion score for codec B. The consequence of accepting H_0 is that the difference between means will be zero and that distribution of difference between mean values follows a t-distribution. On the basis of the observed mean values, standard deviations and number of observations for the two codecs, a t-value can be calculated. Since each subject has evaluated both codecs, the t-value is computed as:

$$t_{obs} = \frac{MOS_A - MOS_B}{\sqrt{\frac{\sigma_A^2}{N} + \frac{\sigma_B^2}{N}}} \tag{6}$$

where σ_A^2 and σ_B^2 are the estimated variances and N is the number of observations for each codec, which in our case is equal to the number of valid subjects. The decision rule which allows to reject H_0 at a certain probability level is:

$$t_{obs} < t(\alpha/2, df) \; or \; t_{obs} > t(1-\alpha/2, df) \tag{7}$$

where $t(\alpha/2, df)$ and $t(1-\alpha/2, df)$ are calculated using the table for the t-Student distribution, df is the number of degrees of freedom given by Welch-Satterthwaite's approximation, and α is considered equal to 0.05. If the observed t-value is outside the critical range, the null hypothesis can be rejected and the conclusion is that the two MOS are significantly different at the defined significance level α.

4. SUBJECTIVE RESULTS

Sixteen naive subjects, screened for visual acuity and color blindness, took part to the experiment at EPFL. Five of them were female and 11 males, with ages ranging from 17 to 40 years old. Three outliers were detected and discarded. The subjective results, including both raw and processed scores, are available at *http://mmspg.epfl.ch/iqa*.

Rate distortion plots Figure 5 shows, for each image, the rate distortion plots with MOS and CI values obtained after processing the results. These rate distortion plots show that the confidence intervals are usually small, thus indicating a high level of agreement among different subjects and the effectiveness of training session and data processing. Also, for each image, the MOS values span over the entire range of quality levels, as expected. The only exception to this overall behaviour is on the *cafe* image, where there are almost no values within the *Excellent* quality range. This can be explained considering that in the centre of this picture a striped red curtain is depicted and this area is particulary senstive to artifacts, showing clear quality degradation even at the highest bitrates.

Finally, from the rate distortion curves it is possible to have an overall impression of the performance of different codecs. The JPEG 2000 4:4:4, JPEG XR MS 4:2:0 and JPEG XR PS 4:2:0 curves show a quite stable behaviour across different contents and bitrate values, with JPEG 2000 4:4:4 and JPEG XR MS 4:2:0 outperforming the other codecs in most cases, and JPEG XR PS 4:2:0 having an intermediate behaviour, at times close to the JPEG XR MS 4:2:0 performance. On the other hand, the performance of JPEG 2000 4:2:0 and JPEG 4:2:0 may vary significantly from content to content and also depending on the bitrate value under analysis. Particularly, for the *woman* image, for bitrate values larger than 0.5 bpp, while JPEG 4:2:0 achieves very good performance, JPEG 2000 4:2:0 results as the worse performing codec. This is explained by the fact that no visual weighting was used in JPEG 2000 4:2.0. Lack of visual weighting creates strong distortions on the skin texture at lower bitrates, as reported during development of JPEG 2000 standard. On the contrary, the use of visual weighting in JPEG 4:2:0 prevents a strong deterioration of the skin texture in this image. However, results for images *p01*, *p06* and *p10*, show that performance of JPEG 2000 4:2:0 is similar to JPEG 2000 4:4:4 and JPEG XR codecs, while JPEG 4:2:0 is clearly the less performing codec.

Figure 6 depicts examples of impaired images: a 300x300 pixels portion of reference image *cafe* original image is shown together with the corresponding crops extracted from 0.5 bpp samples using the five codecs under evaluation.

Results of the hypothesis test The above codec performance comparison, based on the analysis of rate distortion curves is confirmed by the results of the hypothesis test which are shown in figure 7 for each bit-per-pixel value separately, in terms of number of rejection of the hypothesis computed over the entire set of six pictures under analysis. Since the null hypotesis is that the two MOS related to two different codecs are statistically the same, when the rejection number is equal to zero (dark blue in the figure) it means that, for the bit per pixel under analysis, the two codecs always have the same performance. As opposite example, when the rejection rate is maximum, i.e. equal to six (dark red in the figure), it means that the two codecs never have the same performance.

Considering the results of hypothesis testing for 0.25 bpp, it can be noticed that: JPEG 4:2:0 almost never matches the performance of any other codec, apart from JPEG XR PS 4:2:0 which has the same performance on more than 50% of the cases; JPEG 2000 4:4:4 is also most of the times different from all other codecs, having usually the best performance as it is clear from figure 5; the two JPEG XR implementations are equivalent except in one case; the performance of JPEG 2000 4:2:0 is equivalent to JPEG XR codecs on 50% of the cases.

At 0.50 bpp the results slightly change: still, JPEG 4:2:0 almost never reaches the performance of other codecs except for JPEG XR PS 4:2:0; in the majority of cases, JPEG 2000 4:4:4 has the same performance as JPEG 2000 4:2:0 and JPEG XR MS 4:2:0.

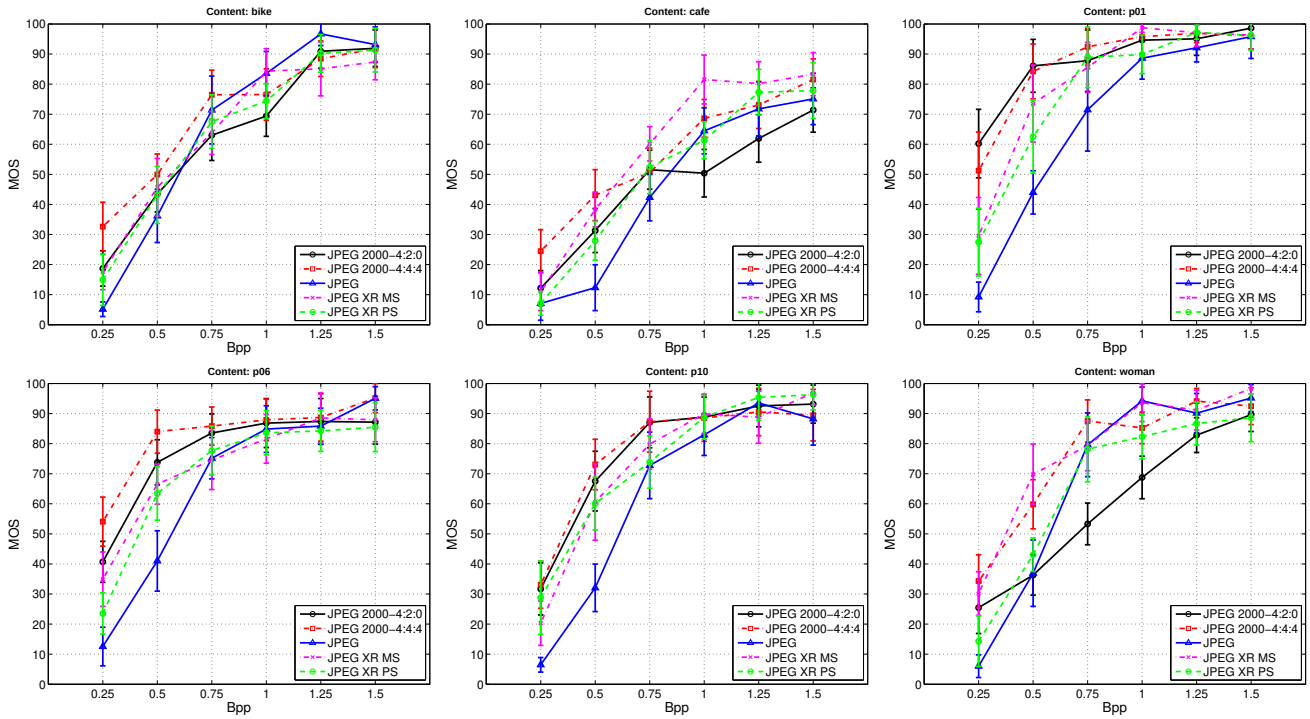

Figure 5: Mean opinion score vs. bitrate results for the different codecs across test images. From top left to bottom right: *bike*, *cafe*, *p01*, *p06*, *p10* and *woman*.

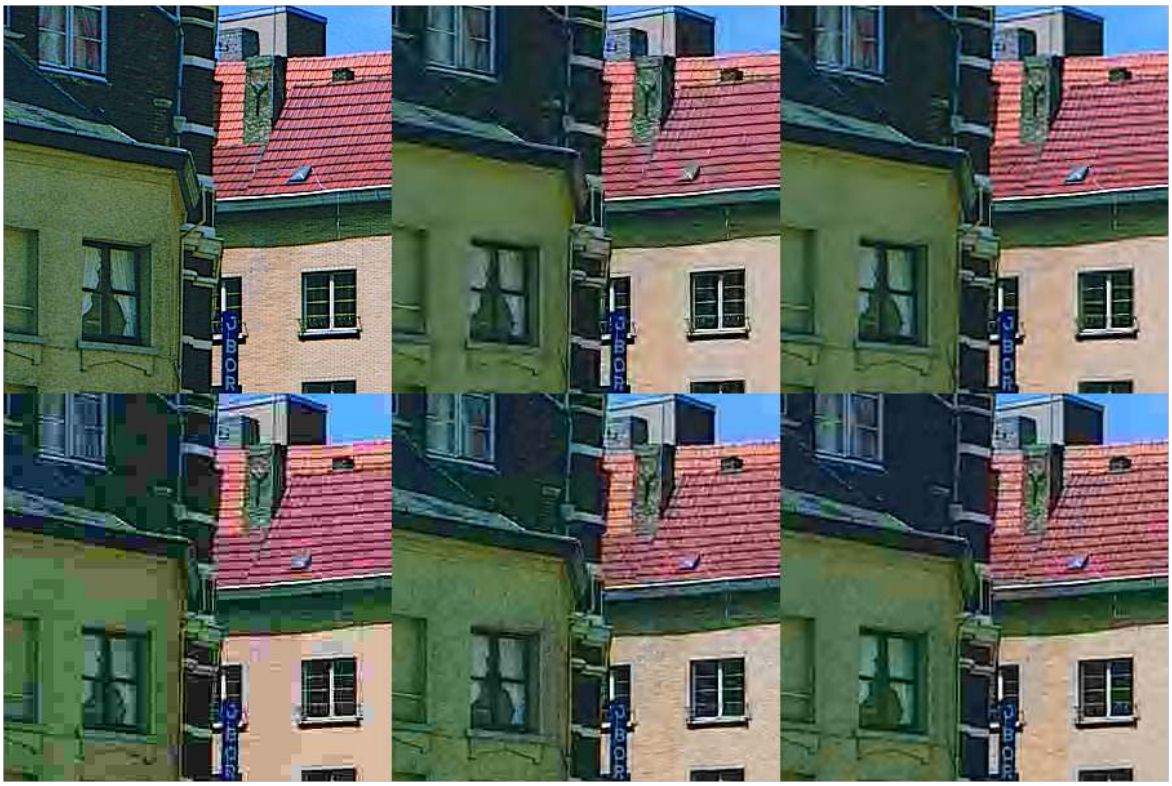

Figure 6: Sample crop of the image *cafe* for different codecs. From top left to bottom right: reference, JPEG 2000 4:2:0, JPEG 2000 4:4:4, JPEG 4:2:0, JPEG XR MS 4:2:0, JPEG XR PS 4:2:0.

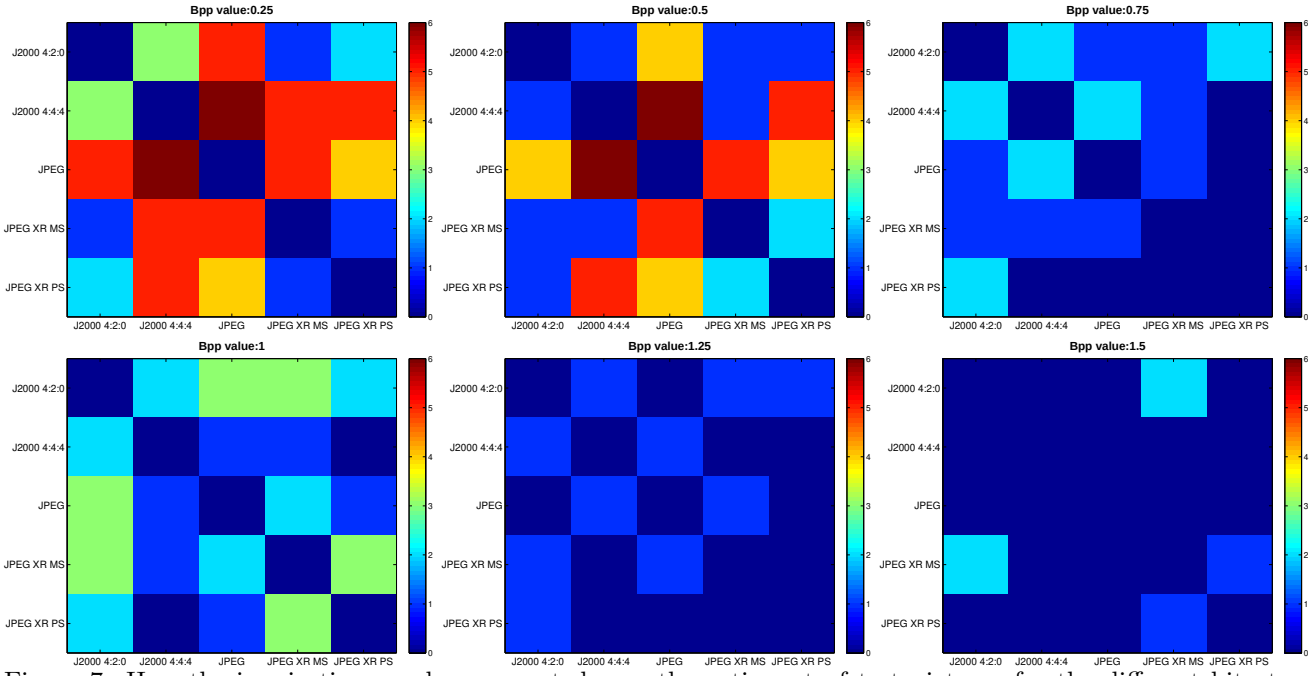

Figure 7: Hypothesis rejection number computed over the entire set of test pictures for the different bitrates (note: the J2000 acronym stands for JPEG 2000). From top left to bottom right: 0.25, 0.5, 0.75, 1, 1.5, 1.75 bpp. The corresponding mean rejection numbers across all the codec pairs are: 2.5, 2, 0.7, 1.2, 0.3, 0.2.

At 0.75 bpp the differences among codecs reduce significantly, with the two JPEG XR implementations having excatly the same performance, and JPEG 4:2:0 improving with respect to the previous bpp values.

At 1 bpp, JPEG XR PS 4:2:0 and JPEG 2000 4:4:4 show exactly the same performance over the entire set of images, while the two JPEG XR implementations present again different performance on 50% of the pictures.

Finally, it can be noticed that for bit per pixel values greater than 1 bpp, as both figures 5 and 7 clearly show, the quality of compressed images is either transparent or similar for most of the codecs.

5. CONCLUSION

In this paper a procedure for subjective evaluation of the new JPEG XR codec for compression of still pictures is described in details. Two implementations of JPEG XR are compared to the existing JPEG and JPEG 2000 compression algorithms when considering the compression of high resolution 24 bpp pictures, by means of a campaign of subjective quality assessment tests which followed the guidelines defined by the AIC JPEG ah-hoc group. Sixteen naive subjects took part in experiments at EPFL and each subject participated in four test sessions, scoring a total of 208 test stimuli. A detailed procedure for statistical analysis of subjective data is also proposed and performed. The obtained results show high consistency and allow an accurate comparison of the performance of the different codecs, which is discussed in the paper. The entire set of test and training pictures, as well as the subjective results, are available for download at *http://mmspg.epfl.ch/iqa*.

ACKNOWLEDGMENTS

The work presented here was partially supported by the European Network of Excellence PetaMedia (FP7/2007-2011), and the Swiss National Foundation for Scientific Research in the framework of NCCR Interactive Multimodal Information Management (IM2). The authors thank Dr. Gary Sullivan and Dr. Thomas Richter for providing the JPEG XR MS software and the JPEG XR PS compressed pictures, respectively. Also, a special thank goes to the subjects who participated with high dedication to the subjective test reported in this paper.

REFERENCES

[1] Oelbaum, T., Baroncini, V., Tan, T., and Fenimore, C., "Subjective quality assessment of the emerging AVC/H.264 video coding standard," *Proc. Intern. Broadc. Conf. 2004 (IBC04)* (2004).

[2] Skodras, A., Christopoulos, C., and Ebrahimi, T., "The JPEG 2000 still image compression standard," *IEEE Signal Process. Mag.,* **18**, 36–58 (Sep 2001).

[3] Srinivasan, S., Tu, C., Regunathan, S. L., and Sullivan, G. J., "HD Photo: a new image coding technology for digital photography," *Proc. SPIE* **6696** (2007).

[4] De Simone, F., Goldmann, L., Baroncini, V., and Ebrahimi, T., "Subjective quality assessment of JPEG XR," Tech. Rep. WG1N4995, ISO/IEC JTC1/SC29/WG1 (JPEG) (Apr 2009).

[5] Sheikh, H., Sabir, M., and Bovik, A., "A statistical evaluation of recent full reference image quality assessment algorithm," *IEEE Trans. Image Process.,* **15**, 3440–51 (Nov 2006).

[6] ISO, "Photography – Psychophysical experimental methods for estimating image quality – Part 1: Overview of psychophysical elements," Tech. Rep. ISO 20462-1:2005, ISO (2005).

[7] AIC AhG, "JPEG XR subjective assessment: Core Experiments Description 4.1," Tech. Rep. WG1N5001, ISO/IEC JTC1/SC29/WG1 (JPEG) (2009).

[8] ITU-R, "Methodology for the subjective assessment of the quality of television pictures," Tech. Rep. Rec. BT.500-11, ITU-R (2002).

[9] Wallace, G., "The JPEG still picture compression standard," *IEEE Trans. on Consumer Electron.,* **38**, 18–34 (Feb 1992).

[10] ITU, "Information technology - Digital compression and coding of continuous-tone still images - Requirements and guidelines," Tech. Rep. T.81, ITU (1992).

[11] ITU, "Information technology - JPEG 2000 image coding system: Core coding system," Tech. Rep. T.800, ITU (2002).

[12] Schonberg, D., Sullivan, G. J., Sun, S., and Zhou, Z., "Perceptual encoding optimization for JPEG XR image coding using spatially adaptive quantization step size control," *to appear in Proc. SPIE 7443* (2009).

[13] Richter, T., "Visual quality improvement techniques of HD Photo/JPEG XR," *Proc. Intern. Conf. on Image Process. 2008 (ICIP08)* , 2888–91 (Oct 2008).

[14] ITU-R, "Methodology for the subjective assessment of the quality of television pictures," Tech. Rep. ITU-R BT.500-11, ITU-R (2002).

[15] Keppel, G. and Wickens, T., [*Design and Analysis: A Researcher's Handbook*], Prentice Hall (2004).

[16] Snedecor, G. W. and Cochran, W. G., [*Statistical Methods*], Iowa State University Press (1989).