J. Vis. Commun. Image R. xxx (2011) xxx–xxx

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Towards high efficiency video coding: Subjective evaluation of potential coding technologies

Francesca De Simone *, Lutz Goldmann, Jong-Seok Lee, Touradj Ebrahimi

Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

ARTICLE INFO

Article history:
Received 22 August 2010
Accepted 28 January 2011
Available online xxxx

Keywords:
Subjective quality assessment
Subjective test
Statistical analysis
High definition (HD)
Video coding
High efficiency video coding (HEVC)
Joint collaborative team on video coding (JCT-VC)
H.264/MPEG-4 AVC

ABSTRACT

This paper describes the details and the results of the subjective quality evaluation performed at EPFL, as a contribution to the effort of the joint collaborative team on video coding (JCT-VC) for the definition of the high efficiency video coding (HEVC) standard. The performance of twenty-seven coding technologies has been evaluated with respect to two H.264/MPEG-4 AVC anchors, for high definition (HD) test material. The test campaign involved a total of 494 naive observers and took place over a period of four weeks. While similar tests have been conducted as part of the standardization process of previous video coding technologies, the test campaign described in this paper is by far the most extensive in the history of video coding standardization. A detailed statistical analysis of the subjective results is provided. The results show high consistency and support an accurate comparison of the performance of the different coding technologies.

1. Introduction

Digital pictures and video sequences are captured, processed and finally presented to human observers who directly or indirectly judge their visual quality. The need for objective quality measures, which try to predict observers' opinions, is closely related to the need for optimizing the different processing steps of multimedia data, with the final goal of maximizing user's satisfaction.

Even if a considerable effort has been devoted by the research community to development of metrics which can objectively evaluate the quality of digital pictures and video sequences, the ability of existing metrics to predict human judgment remains limited. The reasons behind this are manifold. The primary problem consists in the high complexity of human visual system, whose fundamental principles are still not completely understood, which thus limits the precision of its modeling efforts. Second, the perception of quality is a very context dependent concept. User's expectations play a fundamental role in the evaluation of multimedia quality, modifying the acceptability threshold and influencing the actual perception of visual distortions. Most of the objective measures designed so far do not take into account this application dependency. Finally, 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. Thus, the benchmark for any kind of formal quality assessment remains subjective opinions, collected by means of experiments which have to be carefully designed.

In subjective quality assessment tests, a group of subjects is asked to watch a set of still or moving pictures, i.e. stimuli, and to rate their visual quality by using a particular rating scale. The scores assigned by the observers to each test stimulus are usually averaged in order to obtain a mean opinion score (MOS). The MOSs obtained for different stimuli can be compared using statistical significance tests in order to analyze the results of the quality evaluation and to understand how the perceived quality of each stimulus is related to that of the others. The tests have to be carried out according to precise methodologies and in a controlled environment in order to produce reliable and repeatable results, avoiding involuntary influence of external factors [1].

Examples of formal subjective quality assessment activities, can be found in the standardization processes of image and video coding technologies, such as MPEG-2 [2], H.264/MPEG-4 AVC [3], JPEG [4], JPEG 2000 [5] and JPEG XR [6]. In particular, when a new compression technology 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

* Corresponding author.
E-mail addresses: francesca.desimone@epfl.ch (F. De Simone), lutz.goldmann@epfl.ch (L. Goldmann), jong-seok.lee@epfl.ch (J.-S. Lee), touradj.ebrahimi@epfl.ch (T. Ebrahimi).

1047-3203/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jvcir.2011.01.008

compression efficiency achieved by the proposed coding algorithm, its computational complexity, and any additional functionalities. The compression efficiency of a coding algorithm describes its ability to maximize the visual quality of a compressed image or video sequence versus the number of bits used to represent it. Thus, for the reasons explained before, 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.

In this paper, the details and the results of the subjective quality evaluation performed as a contribution to the effort of the joint collaborative team on video coding (JCT-VC) [7] for the definition of the next generation video coding standard, are presented. This effort, referred to as high efficiency video coding (HEVC), targets a wide variety of applications, such as mobile TV, home cinema and UHDTV, and aims at a substantially improved coding efficiency compared to the current state-of-the-art H.264/MPEG-4 AVC High Profile [3]. The primary goal is to reduce the bit rate requirements by half while keeping comparable image quality, probably at the expense of increased computational complexity.

In January 2010, a joint ITU-T and ISO/IEC call for proposals (CfP) [8] was issued to gather and evaluate the performance of potential technologies able to serve as starting point for the design of the new standard. According to the CfP, each proponent willing to test its own coding technology was required to:

- Develop and submit a binary executable of the proposed codec
- Encode and decode a predefined set of test material with the proposed codec
- Evaluate the objective quality of the coded material using the Peak Signal to Noise Ratio (PSNR)
- Provide an algorithmic description of the technology

Twenty-seven complete proposals were received. The encoded video material provided by the proponents was evaluated with respect to two H.264/MPEG-4 AVC anchors in the most extensive subjective quality assessment test campaign in the history of video coding standardization. The subjective tests were performed in three laboratories: the Fondazione Ugo Bordoni (FUB), in Rome, Italy, the European Broadcasting Union (EBU) in Geneva, Switzerland, and the Multimedia Signal Processing Group (MMSPG) at Ecole Polytechnique Fédérale de Lausanne (EPFL), in Lausanne, Switzerland.

The test campaign described in this paper was performed at the premises of the MMSPG laboratory at EPFL and involved all the high definition (HD) test material, i.e. HD 1080p video sequences with frame rates up to 60 fps and HD 720p video sequences at 60 fps. A total of 494 naive observers participated in the tests, which took place over a period of four weeks.

The paper is structured as follows. The dataset used in our subjective tests is described in Section 2. Particularly, a description of the original video contents that each proponent had to encode and decode, as defined in [8], as well as a brief overview of the codecs and of the coding conditions under analysis, are included. In Section 3, the MMSPG test laboratory, where the test campaign took place, is described. Our test environment has been previously used to perform many other formal subjective test campaigns, like those described in [9], [10] and [11]. The adopted test methodology is detailed in Section 4, while the statistical analysis of the collected subjective data is presented in Section 5. The detailed analysis of the subjective results, provided in Section 6, extends the results reported in the public report of the subjective test campaign [12]. Finally, a brief overview of the coding tools proposed by each proponent and a discussion regarding the best-performing solutions are presented in Section 7, and concluding remarks are drawn in Section 8.

Table 1
Details of the classes of test material selected to evaluate the performance of the proposals.

Class	Resolution	Framerate	Videos
A	2560 × 1600	30	Traffic (S01), PeopleOnStreet (S02)
B1	1920 × 1080	24	Kimono (S03), ParkScene (S04)
B2	1920 × 1080	50–60	Cactus (S05), BasketballDrive (S06), BQTerrace (S07)
C	832 × 480	30–60	BasketballDrill (S08), BQMall (S09), PartyScene (S10), RaceHorses (S11)
D	416 × 240	30–60	BasketballPass (S12), BQSquare (S13), BlowingBubbles (S14), RaceHorses (S15)
E	1280 × 720	60	Vidyo1 (S16), Vidyo2 (S17), Vidyo3 (S18)

2. Dataset

2.1. Contents

The test material selected for evaluating the performance of the proposals aimed at covering many relevant application scenarios for the next generation video coding standard. The dataset described in the CfP included 5 classes with different spatial and temporal resolutions as shown in Table 1. All the test sequences were progressively scanned, with YUV 4:2:0 color sampling and 8 bits per sample.

As already mentioned above, the test campaign at EPFL involved only the HD content, i.e. the classes B1, B2 and E data, thus, 2, 3 and 3 different contents, respectively. The first frames of these video contents are shown in Fig. 1.

2.2. Codecs

All proponents used a coding architecture conceptually similar to AVC. However, the individual coding tools differed a lot between the individual proposals. Apart from the 27 proposals, two H.264/MPEG-4 AVC anchors were included in the codec set as benchmarks, namely [13]: anchor alpha (A), corresponding to AVC High Profile (HP) with hierarchical B frames (IbBbBbBbP), CABAC and 8 × 8 transform, and anchor beta (B), corresponding to AVC High Profile (HP) with hierarchical P frames (IpPp), CABAC and 8 × 8 transform.

2.3. Coding conditions and bit rates

For each class of contents, a set of combinations of coding conditions and bit rates was specified. Particularly, two coding conditions were considered:

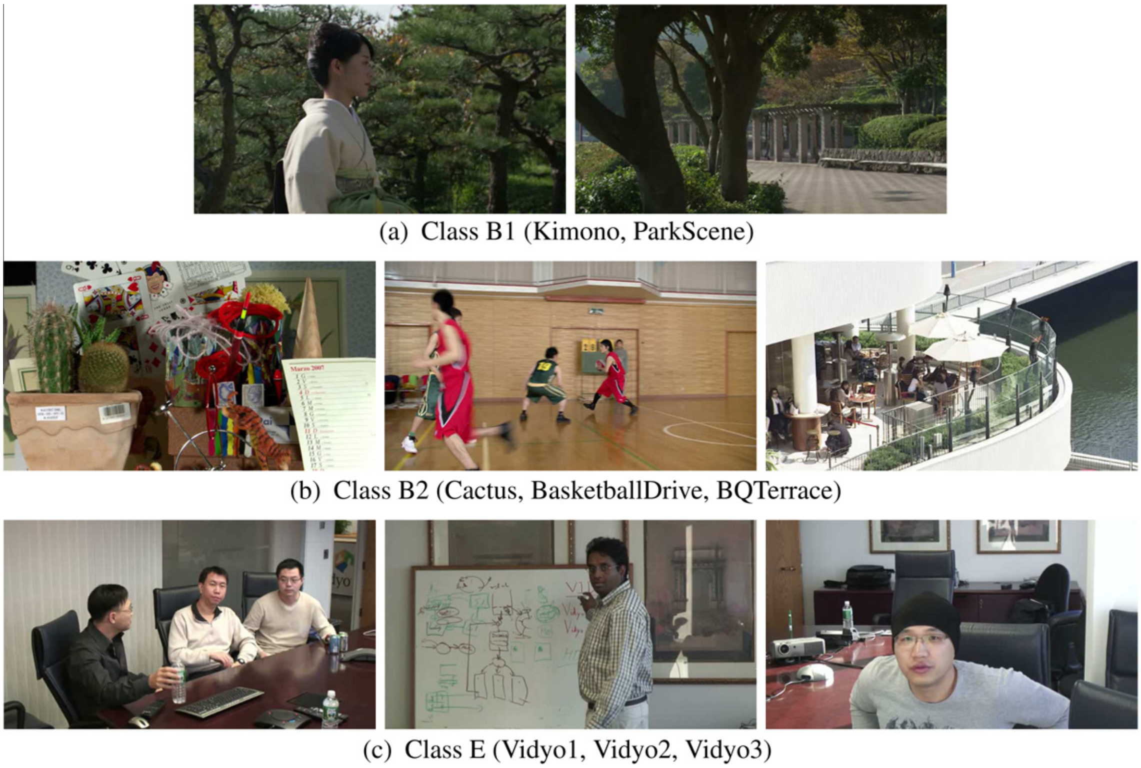
- Random access (RA): group of pictures (GOPs) size is not larger than 8-pictures and the bitstream allows random access intervals of 1.1 seconds or less.
- Low delay (LD): there is no picture reordering between the decoder processing and the output and bit rate fluctuation characteristics and frame-level multi-pass encoding techniques are allowed.

The complete set of combinations of coding conditions and bit rates defined in the CfP is shown in Table 2.

3. Laboratory

As already mentioned, the visual quality assessment laboratory setup is intended to assure the reproducibility of results by avoiding involuntary influence of external factors [1].

Please cite this article in press as: F. De Simone et al., Towards high efficiency video coding: Subjective evaluation of potential coding technologies, J. Vis. Commun. (2011), doi:10.1016/j.jvcir.2011.01.008



(a) Class B1 (Kimono, ParkScene)

(b) Class B2 (Cactus, BasketballDrive, BQTerrace)

(c) Class E (Vidyo1, Vidyo2, Vidyo3)

Fig. 1. Sample frames of the individual video sequences from the different classes considered in the subjective test.

Table 2
Overview of the test conditions defined for the different classes, in terms of combinations of coding bit rates (Mbps) and coding condition.

Class	Method	Condition	BR1	BR2	BR3	BR4	BR5
B1	DSIS	Random access	1.000	1.600	2.500		
B1	DSCQS	Random access				4.000	6.000
B2	DSIS	Random access	2.000	3.000	4.500		
B2	DSCQS	Random access				7.000	10.000
B1	DSIS	Low delay	1.000	1.600	2.500	4.000	
B1	DSCQS	Low delay					6.000
B2	DSIS	Low delay	2.000	3.000	4.500		
B2	DSCQS	Low delay				7.000	10.000
E	DSIS	Low delay	0.256	0.384	0.512	0.850	1.500

3.1. Test equipment

Since the bitstreams of each proponent required a specific decoder, decoded YUV streams were used for the test. When dealing with raw YUV data up to HD 1080p at 60 fps, the task of displaying the video at its native spatial and temporal resolution requires sufficiently powerful hardware. Particularly, to read and display YUV 4:2:0 color subsampled HD 1080p video sequences at 60 fps in real time requires a data rate of 238 MB/s. Since the typical reading speed of current Hard Disk Drives (HDD) is below 160 MB/s, a hardware solution based on Solid State Drives (SSD) was adopted. The details of the hardware and the software used to display the video sequences are listed in Table 3.

Another important element of the hardware needed to perform subjective video quality tests, which could cause visual artifacts due to an incorrect choice, is the display. In order to use a display technology as realistic as possible for the given applications scenarios, high quality LCD monitors were selected, rather than the prohibitively expensive and not very common CRT reference monitors [14]. In order to avoid the ghosting effect, which is typical for LCD displays, a small response time is needed [15]. Based on these requirements, an Eizo CG301W monitor, with a native resolution of 2560 × 1600 pixels, a gray-to-gray response time of 6 ms, and a black-white-black response time of 12 ms, was selected for our test. Three of these monitors were connected to the graphic board of the video server, using two DVI and one display port (DP) to DVI adapter.

Table 3
Server configuration with hardware and software details.

Category	Model
Motherboard	Asus Rampage II Extreme X58
Processor	Intel Core i7 975 Extreme
Graphics	ATI Radeon 5870
RAM	OCZ Memory 3 × 2 GB PC3-12800
SDD (Playback)	OCZ Z-Drive R 512 GB
HDD (Storage)	Western Digital 3 × 1 TB
Operating system	Windows 7 Enterprise 64 bit
Video player	Media Player Classic 64 bit

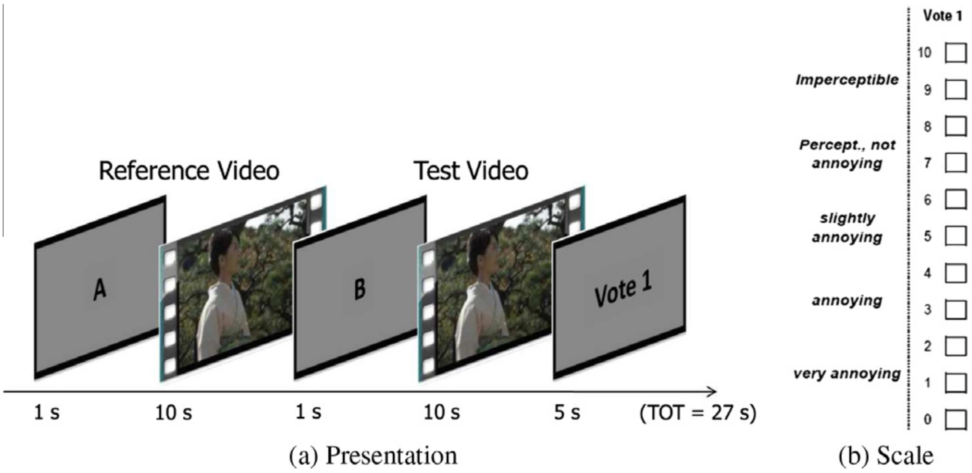
3.2. Test environment and viewing conditions

The monitors were 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 was equipped with a controlled lighting system that consists of neon lamps with 6500 K color temperature. The color of all the background walls and curtains present in the test area was mid-gray. The illumination level measured on the screens was 30 lux and the ambient black level was 0.5 cd/m^2. The test area was controlled by an indoor video security system, with one camera to monitor each screen, in order to keep track of all the test activity

Please cite this article in press as: F. De Simone et al., Towards high efficiency video coding: Subjective evaluation of potential coding technologies, J. Vis. Commun. (2011), doi:10.1016/j.jvcir.2011.01.008

(a) Screening area (b) Testing area

Fig. 2. MMSPG subjective visual quality test laboratory, compliant with ITU recommendation [1].



(a) Presentation (b) Scale

Fig. 3. Double Stimulus Impairment Scale (DSIS) method.

and of possible unexpected events which could influence the test results. The MMSPG test environment included a separate space for screening the participants for correct visual acuity and color vision using Snellen and Ishihara Charts. Pictures of the screening and test area are shown in Fig. 2.

Depending on the resolution of the test material, different viewing conditions (number of subjects, viewing position) were used. For class B, the experiments involved three subjects per display assessing the test material, seated in three different positions (left, center and right) with respect to the center of the monitor, at a distance approximately equal to 2–3 times the height of the test video sequences. For class E data, due to the smaller spatial resolution of the video, the experiments involved two subjects per display, seated in two different positions (left and right) with respect to the center of the monitor, at a distance approximately equal to 2–3 times the height of the test video sequences.

4. Test methodology

Due to the large range of visual qualities present in the test material, two standard test methodologies have been chosen for the experiments, namely the Double Stimulus Impairment Scale (DSIS) method and the Double Stimulus Continuous Quality Scale (DSCQS) method [1].

4.1. DSIS

According to the DSIS methodology, pairs of sequences, i.e. stimuli A and B, are sequentially presented to the subject and she/he is asked to rate the quality of the second stimulus, as shown in Fig. 3(a). The subject is told about the presence of the reference video, having the best expected quality, as stimulus A and she/he is asked to rate the level of annoyance of the visual impairments that she/he observes in stimulus B. The used rating scale is shown in Fig. 3(b). This method is useful for assessing the quality of test material with major impairments. For this reason, the class B test material coded with the lower bit rates and all the class E test material have been assessed using DSIS, as indicated in Table 2.

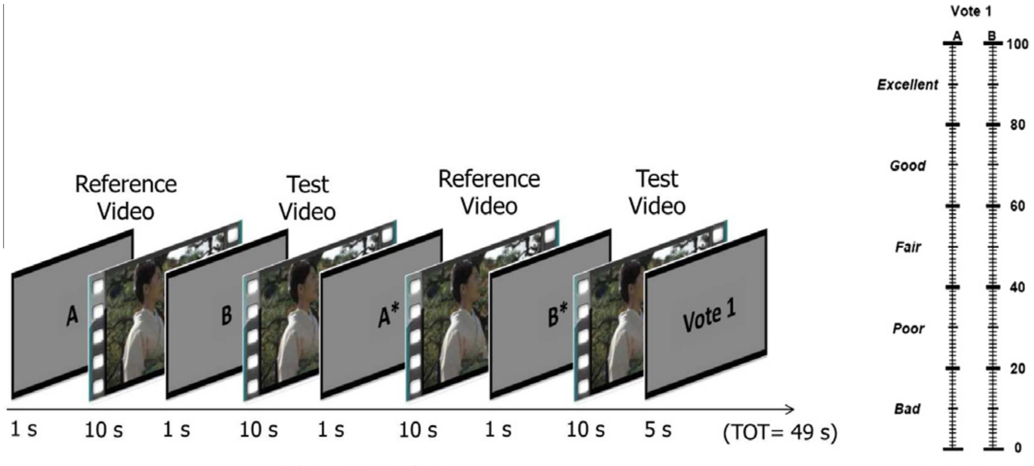
4.2. DSCQS

In the DSCQS methods, pairs of sequences, i.e. stimuli A and B, are presented twice sequentially to the observer and then she/he is asked to rate the quality of both stimuli, as shown in Fig. 4(a). The stimulus A is always the reference video but the subject is not told about it. The selected rating scale is shown in Fig. 4(b). DSCQS is useful for assessing the quality of test material with minor impairments. Thus, it was used for the class B test material coded with the highest bit rates, as listed in Table 2.

4.3. Test sessions

Considering the two classes of contents, i.e. class B and E, and the two test methodologies, three different tests took place: a DSIS test for class E data, a DSIS test for part of the class B data, and a DSCQS test for the remaining class B data. Due to the large number of codecs and the wide range of test conditions, i.e. combinations of coding condition and bit rate, a detailed session planning was necessary.

Please cite this article in press as: F. De Simone et al., Towards high efficiency video coding: Subjective evaluation of potential coding technologies, J. Vis. Commun. (2011), doi:10.1016/j.jvcir.2011.01.008



(a) Presentation

(b) Scale

Fig. 4. Double Stimulus Continuous Quality Scale (DSCQS) method.

In order to retain the subjects' concentration, a subjective video quality test session should not last more than 30 min [1]. For the same reason, it is preferable to alternate as many different contents as possible in the same session. Furthermore, to avoid a possible effect of the presentation order, the stimuli are randomized in a way that the same content is never shown consecutively. Three dummy presentations were included at the beginning of each session, in order to stabilize subject's rating. Additionally, a pair of reference stimuli was included in each session, to check subjects' reliability.

As shown in Fig. 3, one DSIS presentation, i.e. presentation of two stimuli and rating time, takes approximately 27 s. Therefore, test sessions of 33 presentations (i.e. 3 dummies + 29 stimuli + 1 reference pair), corresponding to a duration of 15 minutes, have been designed. For the DSIS class B test, a total of 928 test sequences (29 codecs × 32 combinations of content, coding condition and bit rate) had to be assessed, leading to a total of 32 test sessions. Likewise, for the DSIS class E test, a total of 435 test sequences (29 codecs × 15 combinations of content, coding condition and bit rate) had to be assessed, leading to a total of 15 test sessions.

Considering the DSCQS method, as shown in Fig. 4, one DSCQS presentation, i.e. two consecutive presentations of two stimuli and rating time, takes 49 s. We decided to have test sessions of 22 presentations (i.e. 3 dummies + 18 stimuli + 1 ref vs ref) corresponding to a session duration of 18 min. Since we had to evaluate a total of 522 test sequences (29 codecs × 18 combinations of content, coding condition and bit rate) for the DSCQS class B test, a total of 29 DSCQS test sessions were conducted.

4.4. Observers

In order to have each class B test sequence rated by 27 people and each class E sequence rated by 18 people, this session planning resulted in 4 weeks of test activity with 8 test sessions per half day. Each class B session was attended by three groups of 9 people each (3 subjects in front of each screen), while each class E session was attended by three groups of 6 people each (2 subjects in front of each screen).

A total of 494 non-expert viewers, screened for correct visual acuity and color vision, took part in the test campaign. Thirty percent of the observers were female and the age of the subjects ranged from 21 to 38 years. Each subject was paid 100 CHF for two half days of test activity.

At the beginning of each half a day, the subjects of each group took part in a 15-min training session where oral instructions were provided to explain the task and a viewing session was performed to allow the viewer familiarize with the assessment procedure. The selected training sequences had quality levels representing the different labels used on the rating scales. The experimenter explained the meaning of each label reported on the scale and related them to the presented sample sequences.

In order to collect subjects' scores, the subjects were provided with scoring sheets to enter their quality scores. The scores were then converted offline into electronic data. To identify and correct possible conversion errors, the scores were entered by two operators and inconsistencies were subsequently checked.

5. Statistical analysis of the results

The statistical analysis of subjective results is based on the assumption that the score s_{ij} obtained from subject i for stimulus j, defined by the controlled experimental variables (i.e. video content, codec, coding condition, bit rate), can be modeled as [16]:

$$s_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \tag{1}$$

where μ is the overall mean, i.e. the mean score computed across all subjects and all stimuli, α_i is a factor which accounts for the subject effect, i.e. the score variability across subjects due to their physiological variables (e.g. variations in visual sensitivity) and cognitive variables (e.g. mood and expectation), β_j is the treatment effect, i.e. a factor which accounts for the influence of the controlled experimental variables inherent in 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 $N(0;\sigma^2)$. 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 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 6, are detailed in the following subsections. The results of different groups of subjects were merged before performing the statistical analysis of the data, assuming that no re-alignment procedure was needed across them.

5.1. Distribution analysis

In order to perform the statistical analysis correctly, the assumption of a normal distribution of the data under analysis has been verified. In particular, if the data is normally distributed or it can be transformed to normally distributed, it can then be summarized by the 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 is usually a better descriptor of the central tendency of a distribution and non parametric methods of analysis need to be applied.

The distribution of the collected data can be analyzed for each subject across different test conditions, or for each test condition across different subjects. A Shapiro–Wilk test was used to verify the normality of the distributions [17]. The results of this test showed that the score distributions for each subject across different test conditions, are not normally distributed, while the majority of the score distributions across subjects are normal or close to normal. The results of this test justify the processing applied to the data which is detailed hereafter.

5.2. Outlier detection

In order to detect and remove subjects whose scores appear to deviate strongly from the other scores in a session, outlier detection was performed. It was applied separately for each session, over the set of scores obtained by 27 subjects for class B and 18 subjects for class E.

For each score set, a score s_{ij} was considered as outlier if $s_{ij} > q_3 + 1.5(q_3 - q_1) \vee s_{ij} < q_1 - 1.5(q_3 - q_1)$, where q_1 and q_3 are the 25th and 75th percentiles of the scores distribution, respectively [17]. This range corresponds to approximately ±2.7 the standard deviation or 99.3% coverage if the data is normally distributed. A subject was considered as an outlier, and thus all his/her scores were removed from the results of the session, if more than 20% of his/her scores over the session were outliers. Across all the test sessions, a maximum of two subjects per session were discarded as outliers.

5.3. Mean opinion scores and confidence intervals

After removing the outliers, statistical measures were computed to describe the score distribution across the subjects for each test condition (combination of content, coding condition and bit rate) and codec. For the DSIS methodology, the mean opinion score (MOS) was computed as:

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

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

For the DSCQS methodology, the differential mean opinion score (DMOS) was computed as:

$$DMOS_j = \frac{\sum_{i=1}^{N} \left(s_{ij}^A - s_{ij}^B\right)}{N} \tag{3}$$

where N is the number of valid subjects and s_{ij}^A and s_{ij}^B are the scores for the reference and the test sequence, respectively. In order to facilitate the comparison among the DSIS and DSCQS results for class B, the DMOS values, in the range [100,0], were converted to MOS values in the range [0,10], according to:

$$MOS(DSCQS)_j = \frac{100 - DMOS_j}{10} \tag{4}$$

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 the estimated mean. The $100 \times (1 - \alpha)\%$ confidence intervals (CI) for the MOS values were computed using the student's t-distribution, as follows:

$$CI_j = t(1 - \alpha/2, N) \cdot \frac{\sigma_j}{\sqrt{N}} \tag{5}$$

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). Again N corresponds to the number of subjects after outlier detection and σ_j is the standard deviation of the scores for a single test condition j across all subjects. The confidence intervals were computed for α equal to 0.05, which corresponds to a degree of significance of 95%.

5.4. Multiple comparison procedure

While the ranking of the proponents varied across the different classes, coding conditions and bit rates, some proponents generally performed better than others. In order to accurately analyze the performance of each proponent and evaluate whether the MOS values were significantly different from those obtained by the anchors, a multiple comparison procedure was applied separately to the scores of each test condition [17].

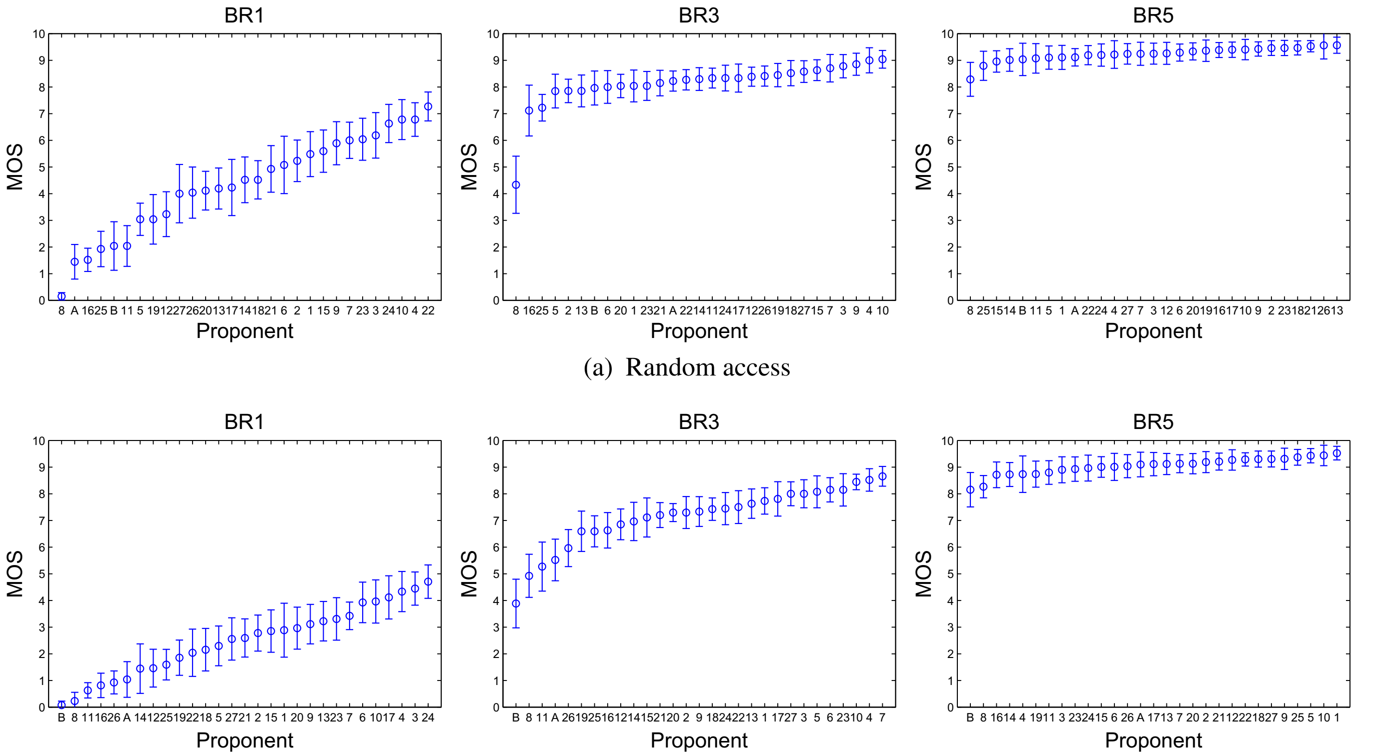
To compare two groups of scores and understand whether their means are statistically significantly different, a simple t-test can be performed by defining a significance level that determines the cutoff value of the t statistic. For example, the value $\alpha = 0.05$ can be specified to insure that when there is no real difference among the two means, a significant difference will be incorrectly detected less than 5% of the time. When there are many group means, a large number of pairs need to be compared. By applying an ordinary t-test in this situation, the $\alpha = 0.05$ value would apply to each comparison, so the chance of incorrectly finding a significant difference would increase with the number of comparisons. Multiple comparison tests are designed to provide an upper bound on the probability that any comparison will be incorrectly found significant [17].

6. Results

6.1. Performance for each test condition

Some representative plots for one content of class B1 are shown in Fig. 5. The plots show the MOS and CI results for the 27 proponents (labeled 1 to 27) and two anchors (labeled A and B), sorted with increasing MOS values. From the 5 test bit rates, only the results for the lowest (BR1), the middle (BR3), and the highest (BR5) are shown. The same plots for all the other contents are shown in Figs. A.11, A.12, A.13, A.14, A.15, A.16, A.17 included in the Appendix at the end of the paper. As it can be seen from the plots, the confidence intervals usually cover not more than roughly a two units interval on the MOS scale, thus indicating that the variations between the subjects are rather small and the obtained results are reliable. The results show that, especially for lower bit rates, the performance of the individual proponents differs considerably and that some of them clearly outperform the anchors, as the quality of the same coded video for a fixed bit rate is significantly better. The differences among the proponents decrease towards the highest bit rates, since the quality of the coded data is anyway very high. Furthermore, it is interesting to see that the MOS scores for the same bit rates are generally higher for the random access (RA) when compared to the low delay (LD) scenario, since the latter constraint generally leads to a lower efficiency.

Please cite this article in press as: F. De Simone et al., Towards high efficiency video coding: Subjective evaluation of potential coding technologies, J. Vis. Commun. (2011), doi:10.1016/j.jvcir.2011.01.008



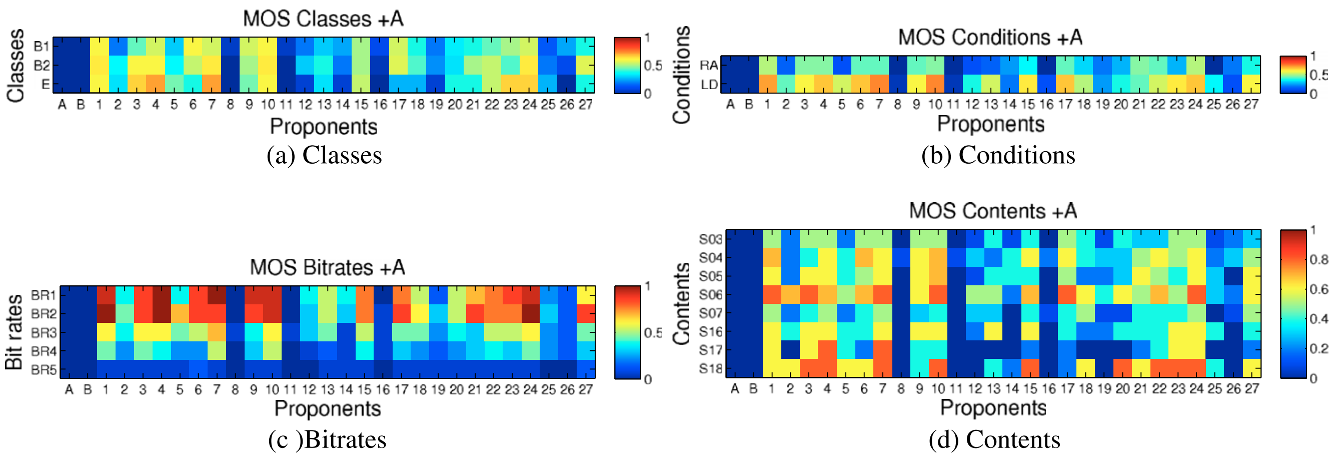
(a) Random access

(b) Low delay

Fig. 5. MOS/CI results for class B1 content *Kimono (S03)* for low (BR1), middle (BR3) and high (BR5) bit rates. The proponents are ordered individually for each bit rate with increasing MOS value.

Table 4
Results of the multiple comparison test expressed in terms of number of times that each proponent performs better, equal or worse than each anchor (A or B) or all the other proponents (P), expressed in % over the entire set of test conditions.

Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
Proponent	7	1	4	10	24	3	6	9	15	22	23	17	27	21	18	5	13	20	2	12	25	14	19	26	16	8	11	A	B
P > *A*	63	62	62	62	62	57	52	51	51	51	51	49	48	45	38	37	35	35	31	22	20	18	18	12	5	2	2	0	0
P = *A*	37	38	38	38	38	43	48	49	49	49	49	51	52	55	62	63	65	65	69	78	72	82	82	88	94	89	98	100	91
P < *A*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	2	9	0	0	9
Proponent	4	7	10	3	9	22	24	1	6	17	23	21	15	13	5	27	18	20	2	19	25	14	12	26	16	A	11	8	B
P > *B*	72	69	66	65	65	65	63	62	60	60	60	58	57	52	51	51	49	49	48	42	37	35	32	17	15	9	6	3	0
P = *B*	28	31	34	35	35	35	37	38	40	40	40	42	43	48	49	49	51	51	52	58	58	65	68	83	85	91	94	89	100
P < *B*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	8	0
Proponent	4	3	7	10	24	22	23	9	1	6	15	17	21	5	27	13	2	20	18	19	14	12	25	26	16	11	A	B	8
P > *P*	18	15	15	14	14	11	11	10	10	8	8	8	8	7	7	6	6	5	5	4	4	3	3	2	1	1	1	0	0
P = *P*	82	84	85	86	86	88	89	89	90	91	91	91	91	89	92	91	91	91	90	89	88	89	71	81	69	64	59	49	54
P < *P*	0	0	0	0	0	1	0	1	0	1	1	1	1	5	2	3	3	3	5	7	9	8	26	17	30	35	40	50	46



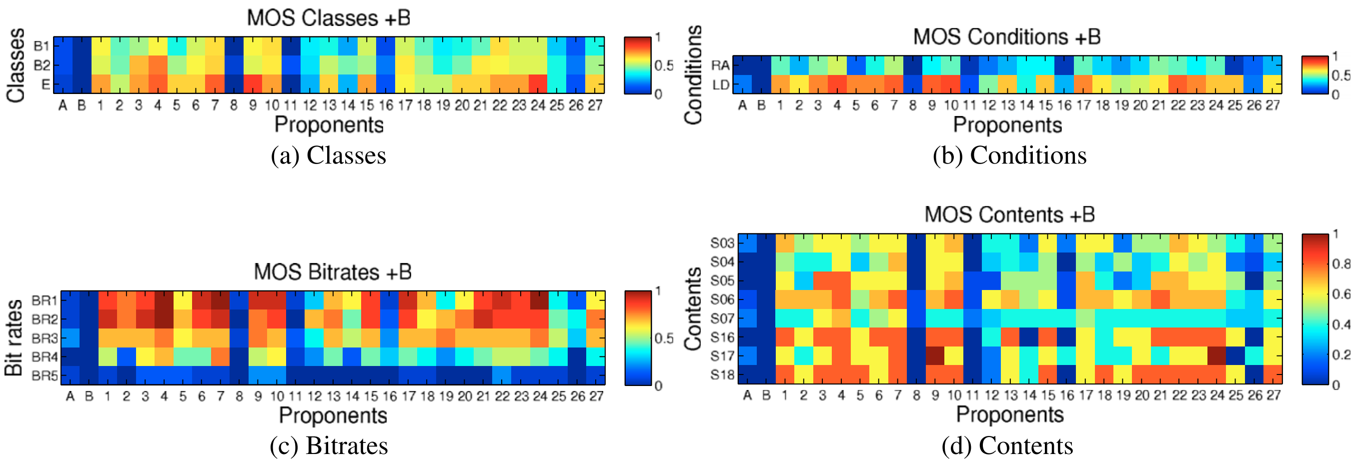
(a) Classes

(b) Conditions

(c)Bitrates

(d) Contents

Fig. 6. Detailed analysis of the percentage where each proponent significantly outperforms *anchor A* according to the subjective MOS values.



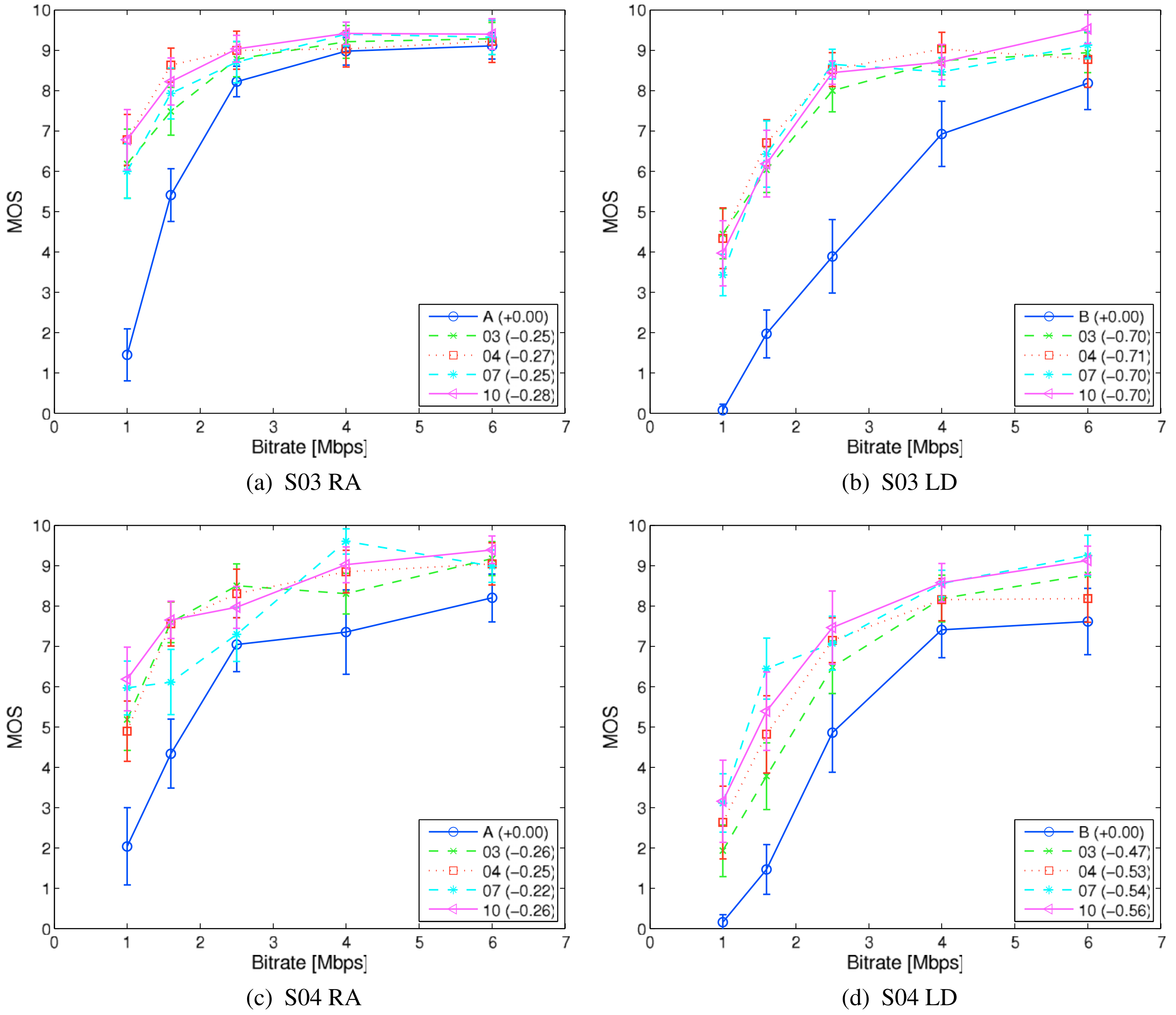
(a) Classes

(b) Conditions

(c) Bitrates

(d) Contents

Fig. 7. Detailed analysis of the percentage where each proponent significantly outperforms *anchor B* according to the subjective MOS values.



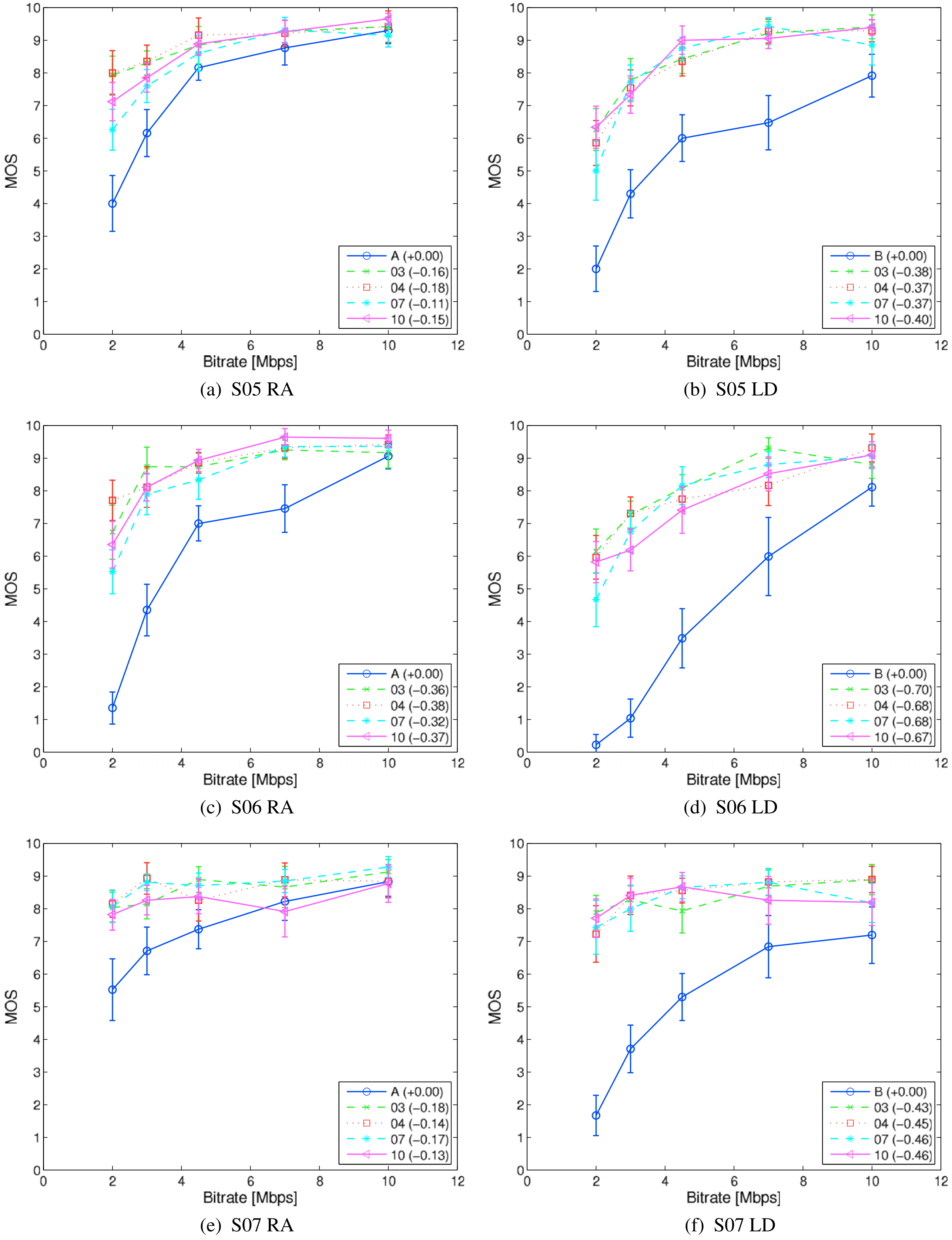
(a) S03 RA

(b) S03 LD

(c) S04 RA

(d) S04 LD

Fig. 8. Rate distortion curves for class B1 contents and corresponding mean bit rate saving in percent for the first four best performing proponents, with respect to anchor A for RA coding condition and to anchor B for LD coding condition.



(a) S05 RA

(b) S05 LD

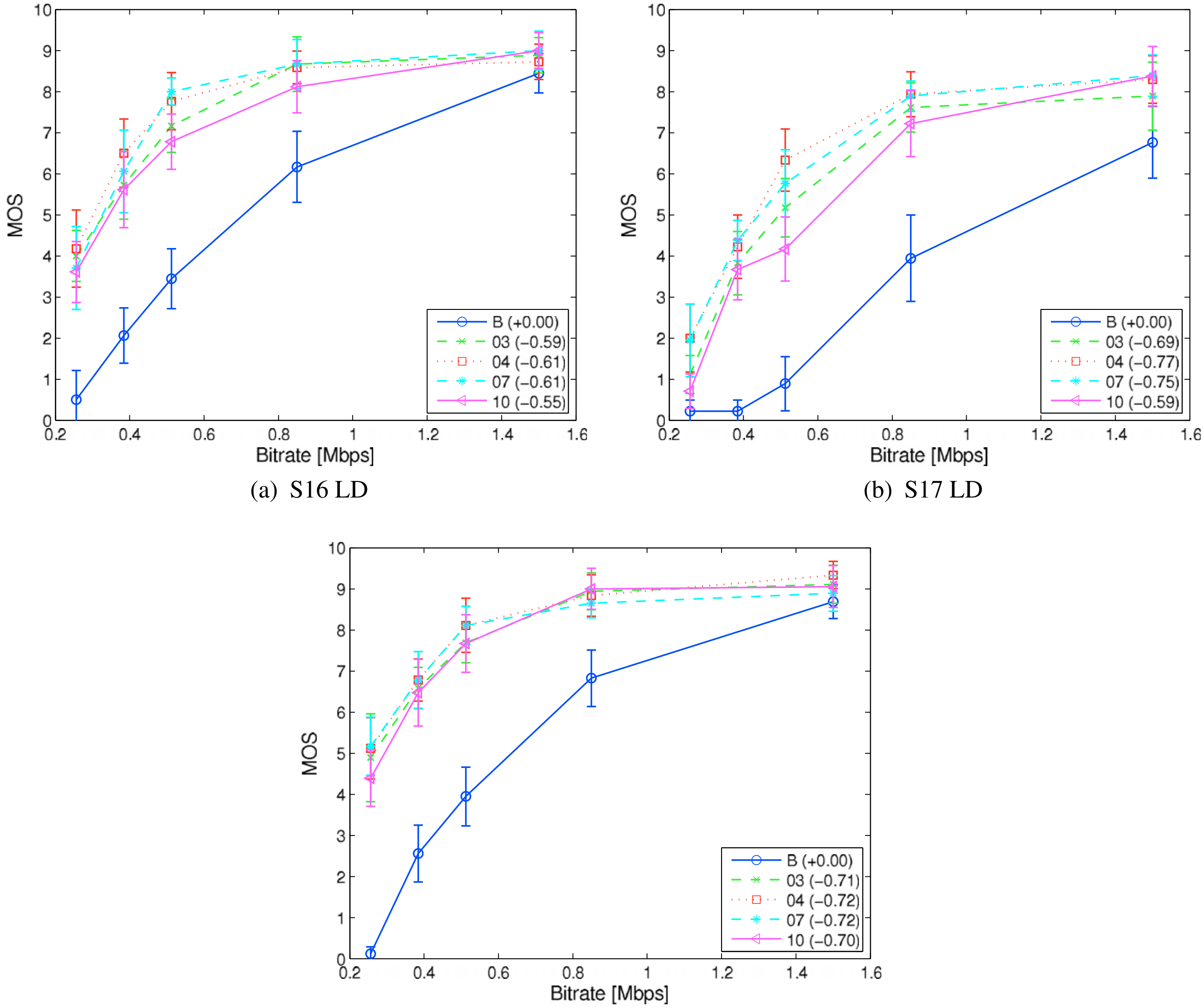
(c) S06 RA

(d) S06 LD

(e) S07 RA

(f) S07 LD

Fig. 9. Rate distortion curves for class B2 contents and corresponding mean bit rate saving in percent for the first four best performing proponents, with respect to anchor A for RA coding condition and to anchor B for LD coding condition.



(a) S16 LD

(b) S17 LD

(c) S18 LD

Fig. 10. Rate distortion curves for class E contents and corresponding mean bit rate saving in percent for the first four best performing proponents, with respect to anchor B for LD coding condition.

6.2. Overall performance across test conditions

In order to get a better overview of the overall performance of the individual proponents across the 65 test conditions, the results of the multiple comparison test for each proponent can be summarized by counting how many times each proponent performed significantly better, equal or worse than each anchor (A and B) and the other proponents (P). These values, expressed in percentages over the entire set of test conditions, are reported in Table 4.

For each of the references, i.e. A, B or P, the proponents have been ranked according to the percentage of performing better than the reference. Apart from some slight variations between similar performing proponents, the ranking is quite consistent across the individual references. The best proponents (3, 4, 7, 10) significantly outperform the anchors and the other proponents in more than 57%, 65%, and 14% of the test conditions. The worst proponents (8, 11, 16, 26) are better than the anchors and the other proponents in less than 12%, 17%, 2% of the cases. Their performance is mostly equal to the anchors and in a few cases even worse.

In order to understand whether the performance of each proponent varied depending on a particular class, content, coding condition or bit rate, the results of the multiple comparison test have been analyzed by grouping them according to these different criteria.

Figs. 6 and 7 show the results of this grouping as pseudo color plots, where the number of times that each proponent performs better than A and B, respectively, over the particular set of test conditions under analysis, has been scaled to the range 0 (the proponent never outperforms the reference) to 1 (the proponent always outperforms the reference).

Again, the results with respect to the two anchors are qualitatively and quantitatively similar. Also, in general, the performance of the proponents are quite consistent across the different groupings and corresponds to the overall ranking derived in Table 4.

With respect to classes (Figs. 6(a) and 7(a)), the differences between the proponents and the anchors are slightly larger for class E than for class B1 and B2. This is confirmed by the results shown in Figs. 6(d) and 7(d). Additionally, it can be noticed that, even within each class, the performance may differ considerably depending on the video content. For content S06 of class B2, the performance of the individual proponents are usually better than for the other two contents (S05, S07). The same is true for content S17 of class E, in comparison to the other two contents (S16, S18). Interestingly, these two contents, S06 and S17, are those having the fastest

Table 5
Set of coding elements optimized by each proponent and its corresponding ranking in terms of overall performance, according to the results described in Section 6.

Id	Architectural elements								Ranks			
	PP	MC	IP	TR	QU	LF	EC	FB	$P > A$	$P > B$	$P > P$	Average
1	1		1	1		1			2	8	9	6
2			1						19	19	17	18
3		1	1		1	1			6	4	2	4
4	1	1	1	1	1	1	1		3	1	1	2
5		1							16	15	14	15
6	1	1	1	1		1			7	9	10	9
7	1	1	1			1			1	2	3	2
8					1		1		26	28	29	28
9	1	1	1	1	1	1	1	1	8	5	8	7
10	1	1		1	1	1	1		4	3	4	4
11							1		27	27	26	27
12	1	1	1	1		1			20	23	22	22
13		1							17	14	16	16
14		1	1	1	1				22	22	21	22
15	1	1	1	1		1			9	13	11	11
16		1	1			1			25	25	25	25
17	1	1	1	1		1			12	10	12	11
18	1	1	1	1		1	1		15	17	19	17
19	1	1		1		1			23	20	20	21
20		1		1		1			18	18	18	18
21				1			1	1	14	12	13	13
22	1	1	1	1		1	1		10	6	6	7
23	1	1	1	1			1		11	11	7	10
24	1	1	1	1	1	1	1		5	7	5	6
25	1	1							21	21	23	22
26		1				1			24	24	24	24
27			1			1			13	16	15	15

motion compared to others in the same class. Thus, this results would suggest that the new coding tools adopted by some of the proponents achieve better performance with respect to AVC particularly when the content is more challenging.

Considering the coding constraints (Figs. 6(b) and 7(b)), the improvement in performance of the proposed techniques with respect to AVC is generally larger for the low delay (LD) when compared to the random access (RA) scenario. This would support the conclusion drawn above with respect to the complexity of the content, since apparently the new coding tools adopted by some of the proponents achieve better performance when the coding constraints are more challenging. It is also interesting to see that some proponents (10, 12) perform much better for one scenario than the other, while others (14, 21) have very similar performance.

Finally, the improvement in performance of the proposed techniques decreases with increasing bit rate (Figs. 6(c) and 7(c)). While for low bit rates (BR1) the best proponents outperform both anchors in more than 90% of the cases, for medium bit rates (BR3) this percentage drops to 60%–70% and is below 10% for high bit rates (BR5). Similarly, the differences in performance between the proponents are larger for low bitrates and become very small for high bitrates.

6.3. Rate distortion curves

In order to visualize and quantify the improvement of the proponents with respect to the AVC anchors in terms of bitrate savings for the same quality level, or in terms of quality improvement for the same bitrates, rate distortion plots for each content and coding condition are provided in Figs. 8–10. In order to keep the figures readable, only the 4 best performing proponents and the anchor corresponding to the actual coding condition have been considered.

The mean percent bit rate saving for each of the proponents has been computed with respect to the anchor A for the random access coding constraint and to the anchor B for the low delay coding constraint, following the guidelines provided in [18].

It can be noticed that, in a number of cases (Figs. 8(b) and (d), 9(d), 10(a) to (c)), the performance of the best proposals can be roughly characterized as achieving similar quality when using only half of the bit rate, or less. Also, the four best proponents have very similar performance and they all reach transparent quality of the coded test material much faster than AVC. Particularly, some contents are easier to code, such as for example content S07, for which all proponents already reach very high quality levels for the lowest bit rate.

Finally, the general conclusions drawn after analyzing the results of the multiple comparison test are confirmed by the analysis of rate distortion curves. For all contents, the major bit rate savings are obtained when the coding constraint and the content is more challenging.

7. Proponents and technologies

As previously mentioned, all the 27 algorithms submitted for evaluation used a hybrid block-transform motion-compensated coding architecture similar to AVC [19]. Most proponents employed improved techniques in the following coding elements [20,21]:

- Picture partitioning (PP)
- Motion prediction, compensation and encoding (MC)
- Intra prediction (IP)
- Transforms (TR)
- Quantization (QU)
- Loop filtering (LF)
- Entropy coding (EC)
- Frame buffering (FB)

A detailed discussion of the different coding tools used by the proponents is out of the scope of this paper and interested readers are referred to [20,21]. Nevertheless, in order to support a better analysis of the technologies proposed by the individual proponents in relation to their performance, Table 5 links the set of coding elements optimized by each proponent to its ranking in terms of overall performance, according to the results described in Section 6.

Considering the rank of each proponent with respect to the individual references (A, B, P) and its average rank, proponents 3, 4, 7 and 10 are clearly those with the best performance (green), while proponents 8, 11, 16 and 26 performed worse than the others (red). Having a look at the individual architectural elements, it is observed that the best proponents employ new features in all or most of the 7 coding elements, which in turn results in better performance. Although the information of the computational complexity of each proposals is not completely available, and thus an accurate comparison of their complexity is not possible, it can be inferred that the complexity increases the more architectural elements have been optimized.

In order to create an optimized unified architecture, it is crucial to understand the relative importance of the individual elements and how they interact with each other. Therefore, after analyzing the results of the subjective test campaign, a Test Model under Consideration (TMuC) [19] has been developed by JCT-VC, which combines the key elements of 7 well-performing proposals [7]. The TMuC will become the basis of a first software implementation which will be used for the investigation and assessment of the selected coding tools.

8. Conclusion

In this paper a detailed description of the EPFL test campaign for the performance evaluation of potential video coding technologies for HEVC has been presented. In the most extensive subjective test campaign in the history of video coding standardization, 27 proposals have been evaluated with respect to each other and two AVC anchors. The evaluation performed at EPFL focused on HD video sequences and involved 494 observers. Subjective quality scores related to a total of 1885 test stimuli have been collected. The obtained results show high consistency and allow an accurate comparison of the performance of the different proposals. The test results clearly indicate that some proposals exhibit a substantial improvement in compression performance, as compared to the corresponding AVC anchors. In a number of cases, the performance of the best proposals can be roughly characterized as achieving similar quality when using only half of the bit rate.

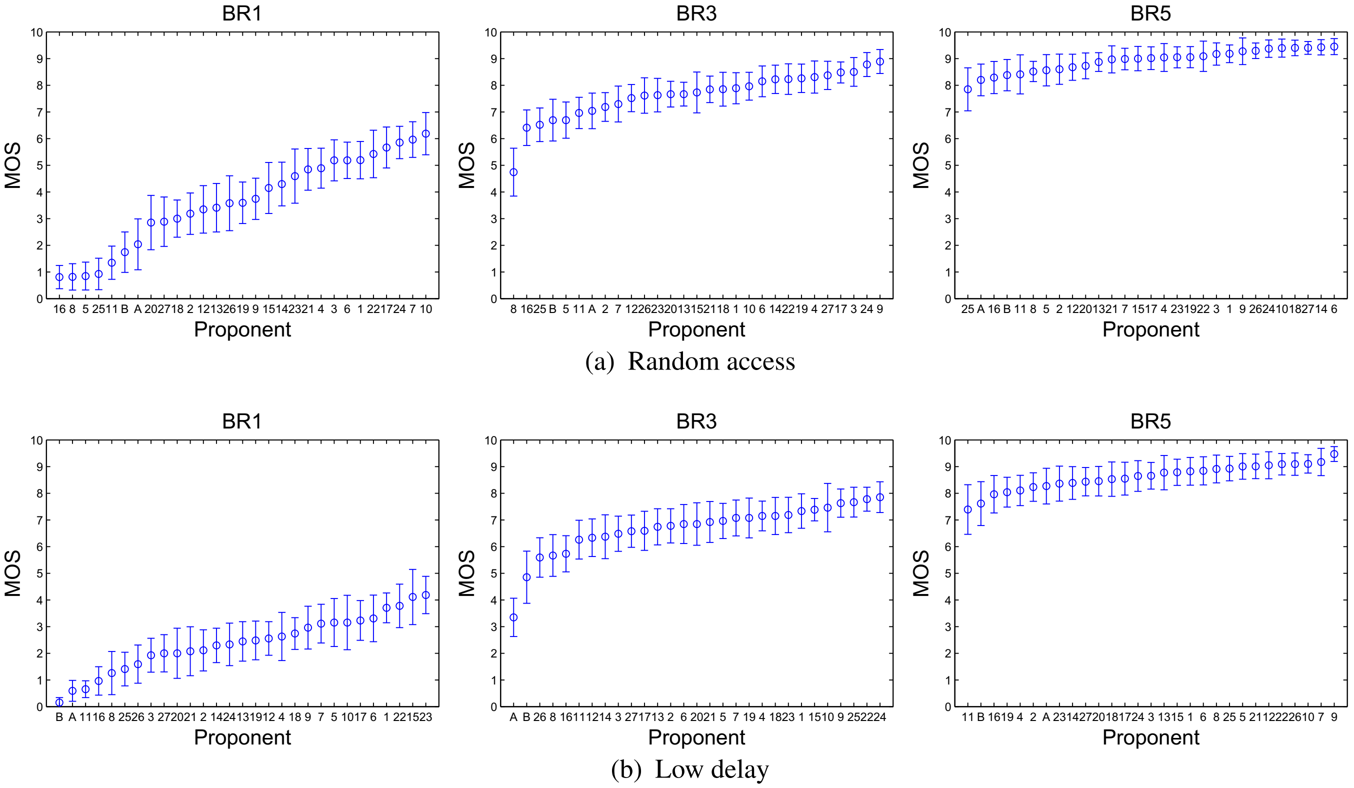
As a result of the analysis of the data collected at EPFL and the other two test laboratories, several elements from the best proposals have been combined to develop an initial test model, called Test Model under Consideration (TMuC), which serves as a starting point for definition of the new video coding standard. The TMuC has similarities to the H.264/MPEG-4 AVC standard, including block-based intra/inter prediction, block transform and entropy coding. New features include increased prediction flexibility, more sophisticated interpolation filters, a wider range of block sizes and new entropy coding schemes. Twice the compression efficiency of H.264/MPEG-4 AVC is expected to be achieved, at the expense of a considerable increase in computational complexity. The performance of the coding algorithm resulting from this integration step will be analyzed by means of formal subjective quality assessment in a next subjective test campaign.

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. Vittorio Baroncini and the proponents who provided the test material and made this subjective test campaign possible. Also, a special thank goes to Dr. Ulrich Engelke, Dr. Ulrich Reiter, Dr. Junyong You, Mrs. Liyuan Xing, Mrs. Fitri N. Rahayu, and Mr. Christoph Steindl who helped in assisting the test sessions and the subjects who participated with high dedication to the subjective tests.

Appendix A. Additional MOS/CI plots

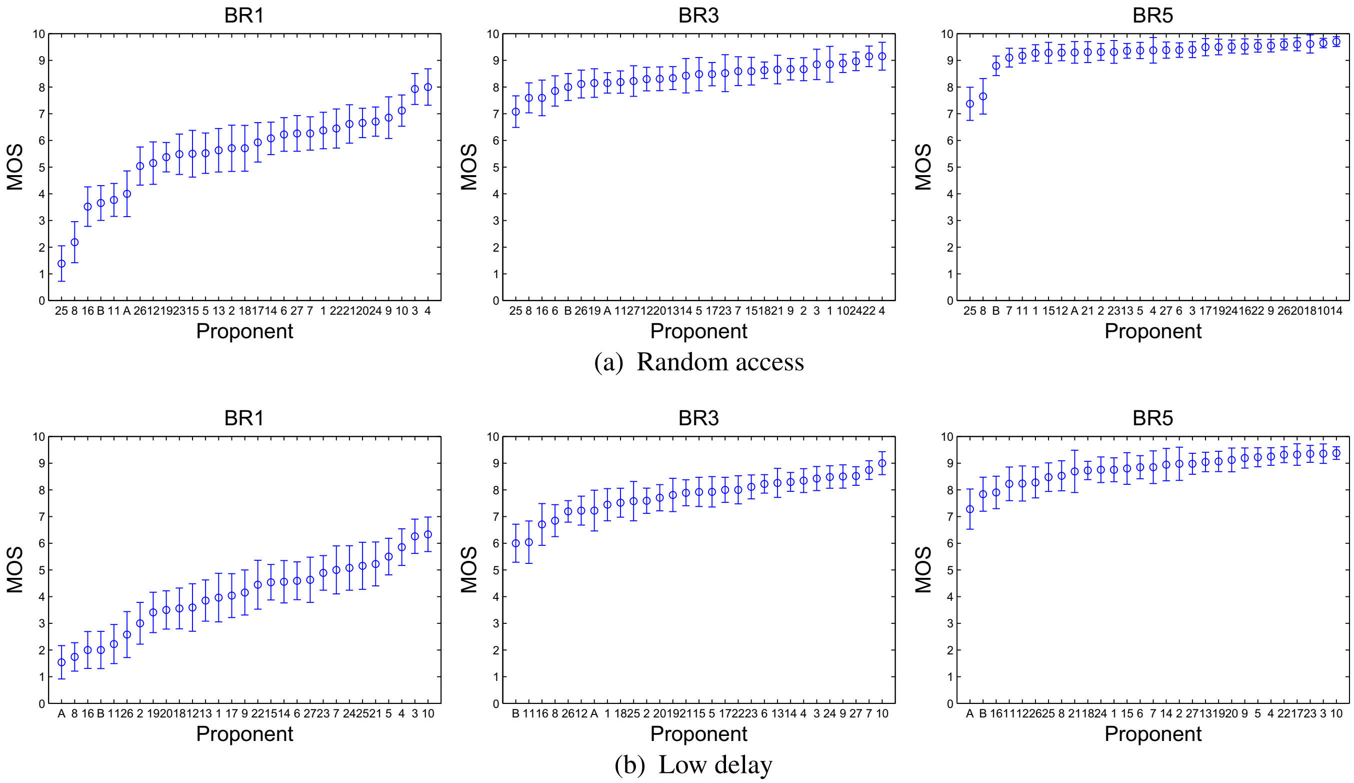
Figs. A.11–A.17



(a) Random access

(b) Low delay

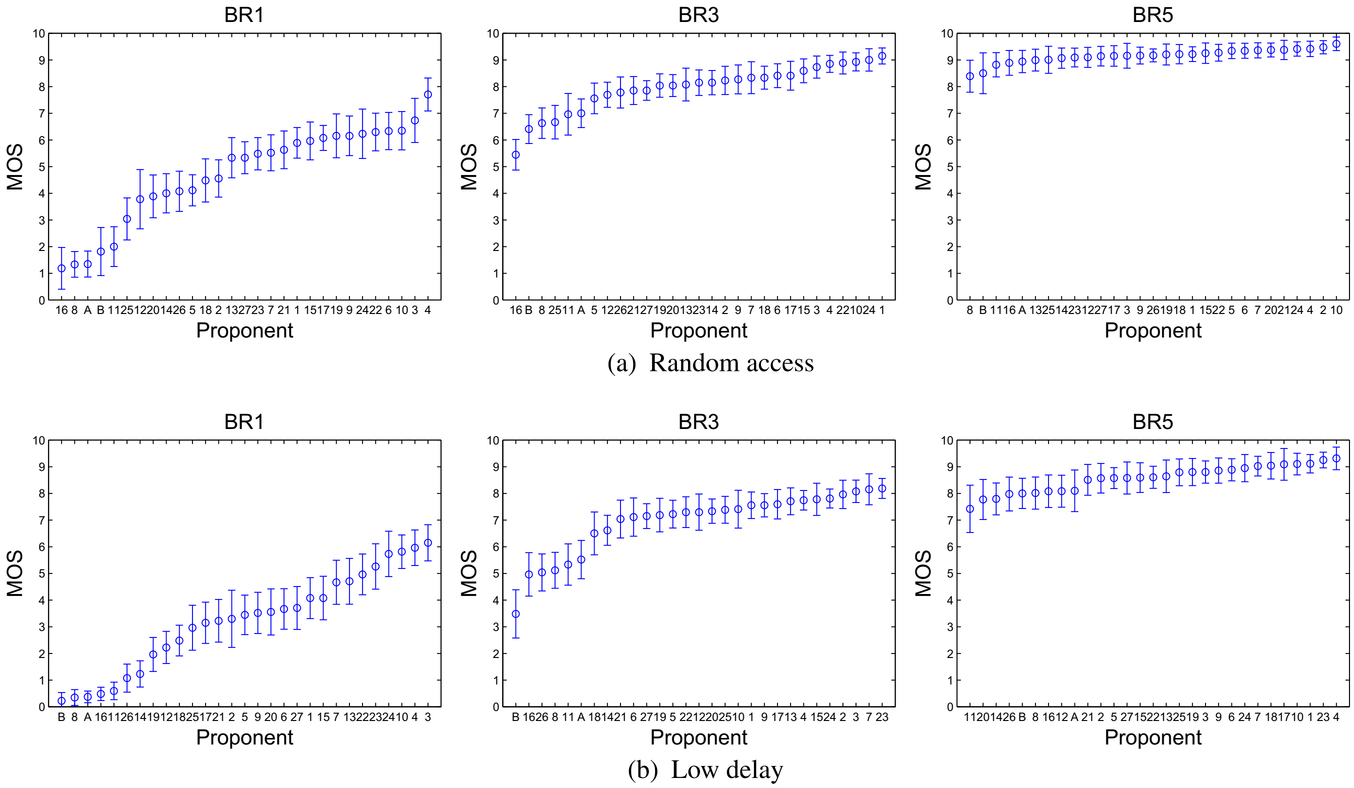
Fig. A.11. MOS/CI results for class B2 content *ParkScene (S04)* for low (BR1), middle (BR3) and high (BR5) bit rates.



(a) Random access

(b) Low delay

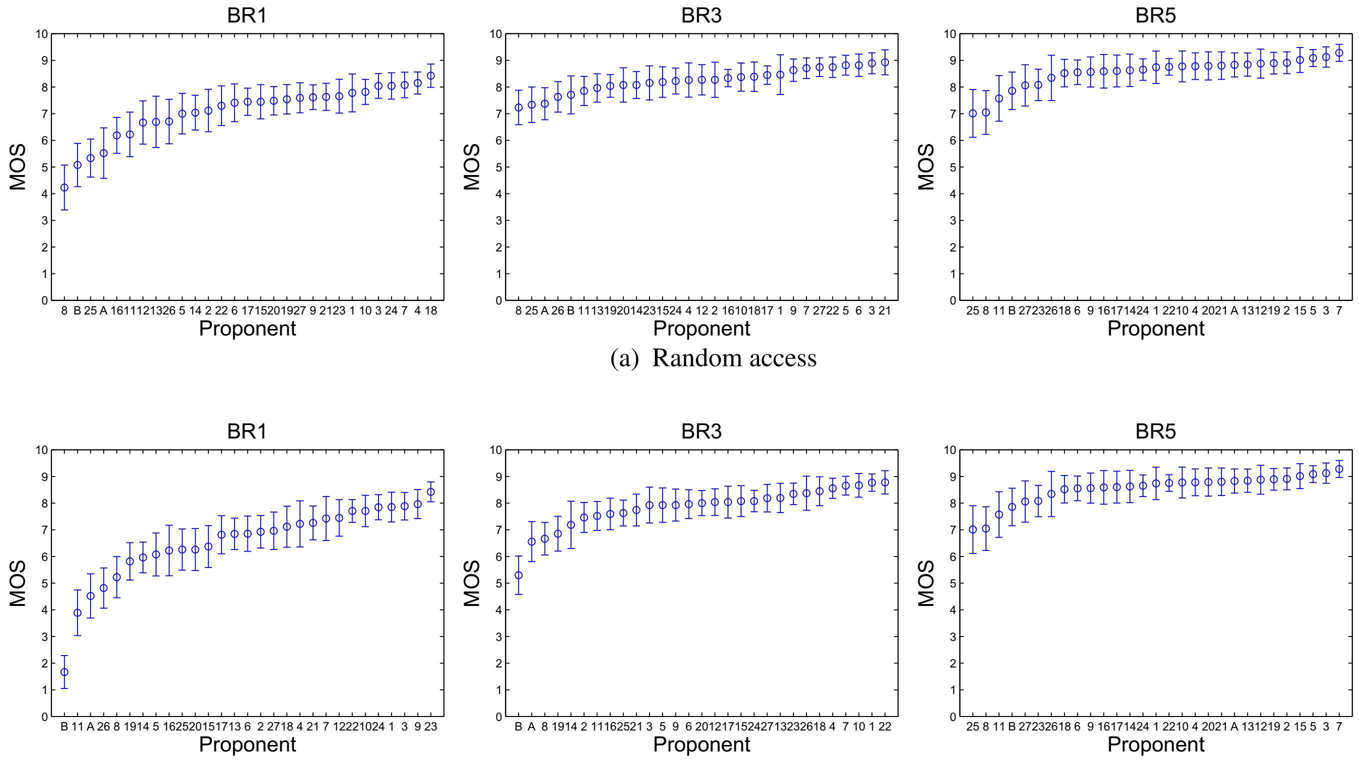
Fig. A.12. MOS/CI results for class B2 content *Cactus (S05)* for low (BR1), middle (BR3) and high (BR5) bit rates.



(a) Random access

(b) Low delay

Fig. A.13. MOS/CI results for class B2 content *BasketballDrive (S06)* for low (BR1), middle (BR3) and high (BR5) bit rates.



(a) Random access

(b) Low delay

Fig. A.14. MOS/CI results for class B2 content *BQTerrace* (*S*07) for low (BR1), middle (BR3) and high (BR5) bit rates.

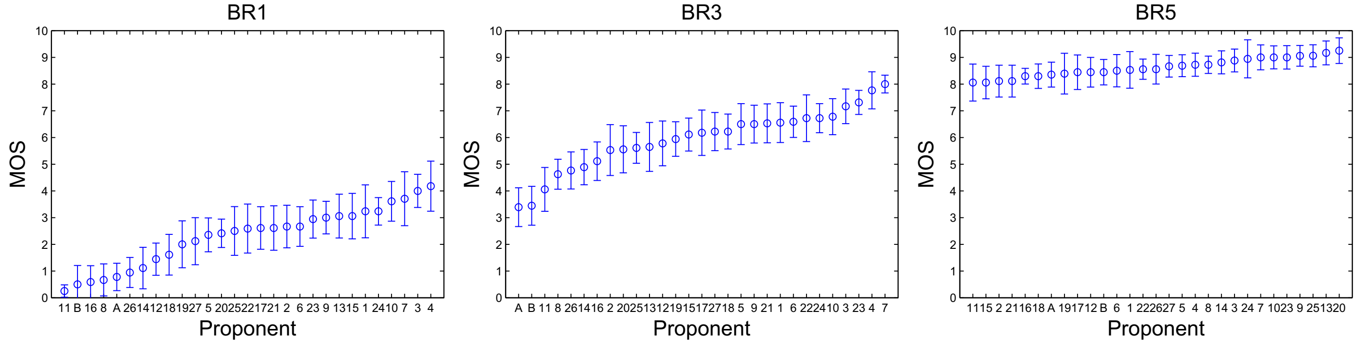


Fig. A.15. MOS/CI results for class E content *Vidyo1 (S16)* for low (BR1), middle (BR3) and high (BR5) bit rates.

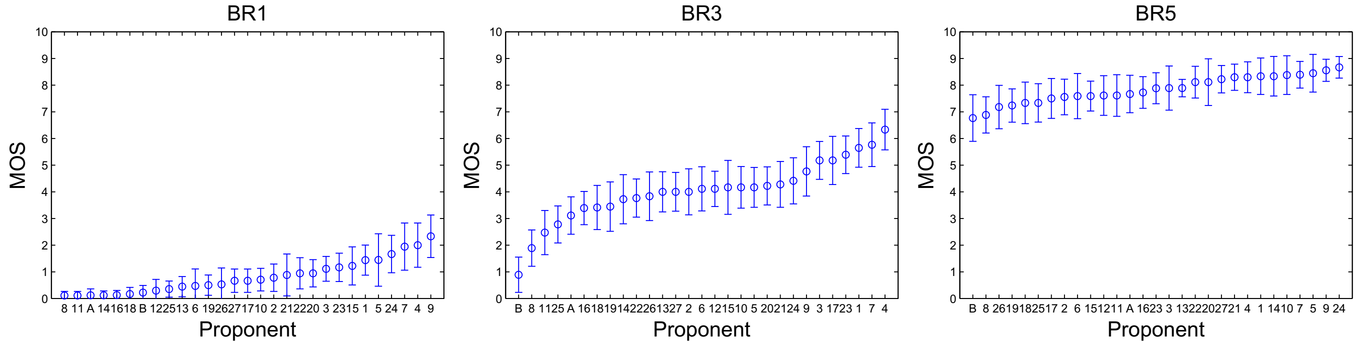


Fig. A.16. MOS/CI results for class E content *Vidyo2 (S17)* for low (BR1), middle (BR3) and high (BR5) bit rates.

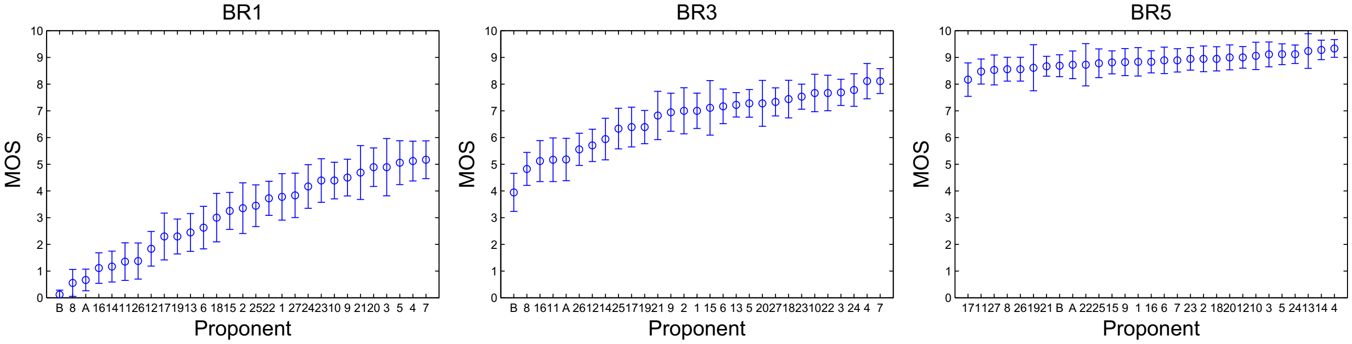


Fig. A.17. MOS/CI results for class E content *Vidyo3 (S18)* for low (BR1), middle (BR3) and high (BR5) bit rates.

References

[1] ITU-R, BT.500-11: Methodology for the subjective assessment of the quality of television pictures, Technical Report BT.500-11, ITU-R, 2002.

[2] ISO/IEC, Information technology – Generic coding of moving pictures and associated audio information: Video, Technical Report 13818-2:2000, ISO/IEC, 2000.

[3] ISO, Information technology – Coding of audio-visual objects – Part 10: Advanced Video Coding, Technical Report ISO/IEC 14496-10:2005, ISO/IEC, 2005.

[4] ITU, Information technology – Digital compression and coding of continuous-tone still images – Requirements and guidelines, Technical Report T.81, ITU, 1992.

[5] ITU, Information technology – JPEG 2000 image coding system: Core coding system, Technical Report T.800, ITU, 2002.

[6] ITU, Information technology – JPEG XR image coding system – Image coding specification, Technical Report T.832, ITU, 2010.

[7] G.J. Sullivan, J.-R. Ohm, Recent developments toward standardization of next-generation video coding technology.

[8] JCT-VC, Joint Call for Proposals on Video Compression Technology, Technical Report VCEG-AM91, MPEG-N11113, ITU-T VCEG and ISO/IEC MPEG, Kyoto, 2010.

[9] J.-S. Lee, F. De Simone, N. Ramzan, Z. Zhao, E. Kurutepe, T. Sikora, J. Ostermann, E. Izquierdo, T. Ebrahimi, Subjective evaluation of scalable video coding for content distribution, in: Proceedings of the International Conference on Multimedia MM'10, 2010, pp. 65–72.

[10] F. De Simone, M. Tagliasacchi, M. Naccari, S. Tubaro, T. Ebrahimi, H.264/AVC video database for the evaluation of quality metrics, in: Proceedings of 35th International Conference on Acoustics, Speech, and Signal Processing.

[11] F. De Simone, L. Goldmann, V. Baroncini, F. Dufaux, T. Ebrahimi, Subjective quality assessment of JPEG XR, Technical Report wg1n4995, JPEG, 2009.

[12] JCT-VC, Report of subjective testing of responses to Joint Call for Proposals (CfP) on video coding technology for High Efficiency Video Coding (HEVC), Technical Report A204, JCT-VC, Dresden, 2010.

[13] ISO, Joint Call for Proposals on Video Compression Technology, Technical Report, ISO/IEC JTC1/SC29/WG11 ITU-T Q6/16 Visual Coding, Kyoto, JP, 2010.

[14] VQEG, Report on the validation of video quality models for high definition video content, Technical Report, VQEG, June, 2010.

[15] S. Tourancheau, P.L. Callet, K. Brunnstrom, B. AndrTn, Psychophysical study of LCD motion-blur perception, in: Proceedings of the Human Vision and Electronic Imaging XIV, vol. 7240.

[16] S. Bech, N. Zacharov, Perceptual Audio Evaluation – Theory, Method and Application, John Wiley & Sons, Ltd, 2006.

[17] G.W. Snedecor, W.G. Cochran, Statistical Methods, Iowa State University Press, 1989.

[18] VCEG, Calculation of average PSNR differences between RD-curves, Technical Report M33, Video Coding Experts Group (VCEG), Austin, Texas, USA, 2001.

[19] JCT-VC, Meeting report of the first meeting of the Joint Collaborative Team on Video Coding (JCT-VC), Technical Report A200, JCT-VC, Dresden, 2010.

[20] JCT-VC, Architectural outline of proposed high efficiency video coding design elements, Technical Report A202, JCTVC, Dresden, 2010.

[21] JCT-VC, Table of proposal design elements for high efficiency video coding (HEVC), Technical Report A203, JCTVC, Dresden, 2010.