Subjective Evaluation of Next-Generation Video Compression Algorithms: A Case Study

Article *in* Proceedings of SPIE - The International Society for Optical Engineering · August 2010

DOI: 10.1117/12.863485 · Source: OAI

Francesca De Simone
Centrum Wiskunde & Informatica
67 PUBLICATIONS **2,043** CITATIONS
SEE PROFILE

Lutz Goldmann
École Polytechnique Fédérale de Lausanne
58 PUBLICATIONS **761** CITATIONS
SEE PROFILE

Jong-Seok Lee
Yonsei University
224 PUBLICATIONS **7,160** CITATIONS
SEE PROFILE

Touradj Ebrahimi
École Polytechnique Fédérale de Lausanne
717 PUBLICATIONS **27,256** CITATIONS
SEE PROFILE

Subjective Evaluation of Next-Generation Video Compression Algorithms: A Case Study

Francesca De Simone[a], Lutz Goldmann[a], Jong-Seok Lee[a],
Touradj Ebrahimi[a] and Vittorio Baroncini[b]

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

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 next-generation video coding standard. The performance of 27 coding technologies have been evaluated with respect to two H.264/MPEG-4 AVC anchors, considering 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. The obtained subjective quality scores show high consistency and support an accurate comparison of the performance of the different coding solutions.

Keywords: subjective quality assessment, High Efficiency Video Coding (HEVC), Joint Collaborative Team on Video Coding (JCT-VC), H.264/MPEG-4 AVC

1. INTRODUCTION

The current trend in video consumption clearly shows that the already large quantity of video material distributed over broadcast channels, digital networks and packaged media is going to increase in the coming years. As an effect of the growing popularity, the users' demand for increased resolution and higher quality is driving the efforts of the technological development. From this point of view, the evolution of video acquisition and display technologies is much faster than that of network capabilities. Thus, a clear need for a new video coding standard with higher efficiency when compared to the current state-of-the-art H.264/MPEG-4 AVC[1] has been identified.

In order to develop the next-generation video coding standard, a group of video coding experts from ITU-T Study Group 16 (VCEG) and ISO/IEC JTC 1/SC 29/WG 11 (MPEG), called the Joint Collaborative Team on Video Coding (JCT-VC)*,[2] has been created. The JCT-VC standardization effort is being referred to as High Efficiency Video Coding (HEVC). The new standard targets a wide variety of applications such as mobile TV, home cinema and UHDTV. It will support next-generation acquisition and display devices featuring progressive scanned video with higher frame rates and resolutions (from WVGA to HDTV and UHDTV), as well as improved picture quality in terms of noise level, color gamut and dynamic range. HEVC aims at a substantially improved coding efficiency compared to the AVC High Profile, i.e. reducing the bit rate requirements by half while keeping comparable image quality, probably at the expense of increased computational complexity. Depending on the application scenario, a trade-off between computational complexity, compression ratio, robustness to errors and processing delay should be supported.

Further author information:
Francesca De Simone: E-mail: francesca.desimone@epfl.ch
Lutz Goldmann: E-mail: lutz.goldmann@epfl.ch
Jong-Seok Lee: E-mail: jong-seok.lee@epfl.ch
Touradj Ebrahimi: E-mail: touradj.ebrahimi@epfl.ch
Vittorio Baroncini: E-mail: vittorio@fub.it

*`http://www.itu.int/ITU-T/studygroups/com16/jct-vc/`

The activities for the standardization of HEVC started in February 2009, when a joint MPEG and VCEG Call for Test Material (CfT)[3] was issued, in order to select suitable video content for evaluating the performance of the coding solution under development. A Call for Evidence (CfE)[4] followed in April 2009, in order to gather potential technologies able to fulfill the defined requirements. Finally, the Call for Proposals (CfP)[5] was published in January 2010. 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

Following the CfP, twenty-seven complete proposals were received. All proposals used a coding architecture conceptually similar to AVC, containing the following basic elements: (a) Block-based coding (b) Variable block sizes (c) Block motion compensation (d) Fractional-pel motion vectors (e) Spatial intra prediction (f) Spatial transform of residual difference (g) Integer-based transform designs (h) Arithmetic or VLC-based entropy coding (i) In-loop filtering. However, the individual coding tools differed a lot between the individual proposals.

In order to compare the performance of the proposed technologies in terms of coding efficiency, the encoded video material provided by the proponents was evaluated in an extensive subjective quality assessment test campaign.[5] The subjective quality 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 results of the subjective tests were presented at the first meeting of the JCT-VC in Dresden in April 2010.

In this paper, we present the details and the results of the subjective quality evaluation performed at EPFL, which included all the high definition (HD) test material, i.e. HD1080p video sequences with frame rate up to 60 fps (referred to as class B test material) and HD720p video sequences at 60 fps (referred to as class E test material), and involved a total of 494 test subjects, over a period of four weeks. While similar efforts have been carried out as part of the standardization process of previous video coding technologies,[1,6] the test campaign described in this paper was by far the most extensive in the history of video coding standardization.

The paper is structured as follows: the MMSPG test laboratory, where the test campaign took place, is described in section 2. The set of HD test material and the adopted test methodology are detailed in section 3, while the statistical analysis of the collected subjective data and the results are presented in sections 4 and 5, respectively. Finally, concluding remarks are drawn in section 6.

2. TEST EQUIPMENT AND ENVIRONMENT

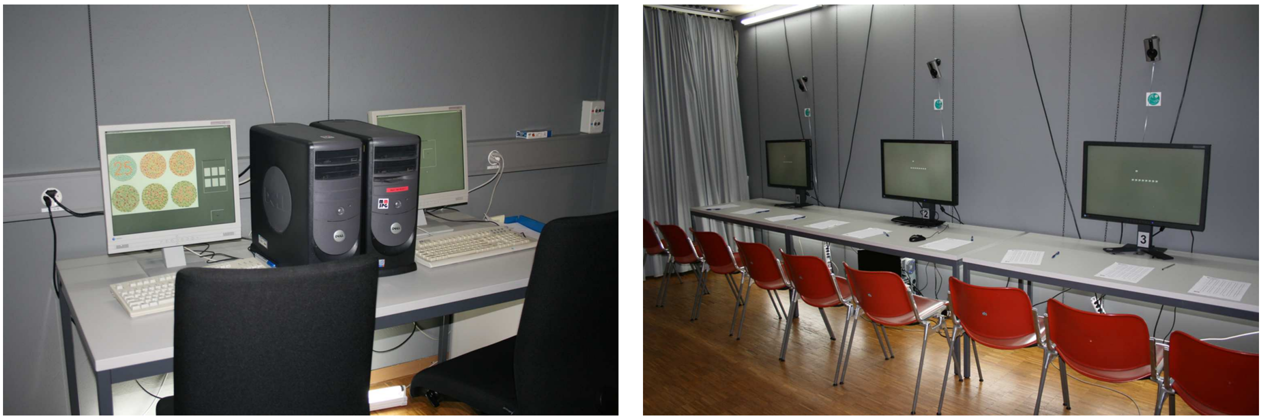
In a subjective quality assessment test, a set of video sequences is presented in a predefined order to a group of subjects, who are asked to rate their visual quality on a particular rating scale. The test has to be carried out according to precise methodologies and in a controlled test environment in order to produce reliable and repeatable results, avoiding involuntary influence of external factors.[7]

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 HD1080p 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 in real time YUV 4:2:0 color subsampled HD1080p video sequences at 60 fps requires a data rate of 237 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 video server and the software used to display the video sequences are listed in Table 1.

Another important element of the hardware facilities 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 application scenarios, high quality LCD monitors were selected, rather than the prohibitively expensive and not very common CRT reference monitors.[8] In order to avoid the ghost effect, which is typical for LCD displays, a small response time is needed.[9] Based on these requirements, an Eizo

Category	Model
Motherboard	Asus Rampage II Extreme X58
Processor	Intel Core i7 975 Extreme
Graphics	ATI Radeon 5870
RAM	OCZ Memory 3x2 GB PC3-12800
SDD (Playback)	OCZ Z-Drive R 512 GB
HDD (Storage)	Western Digital 3x1 TB
Operating system	Windows 7 Enterprise 64 bit
Video player	Media Player Classic 64 bit

Table 1. Server configuration with hardware and software details.



(a) Screening area

(b) Testing area

Figure 1. MMSPG subjective visual quality test laboratory, compliant with ITU recommendation.[7]

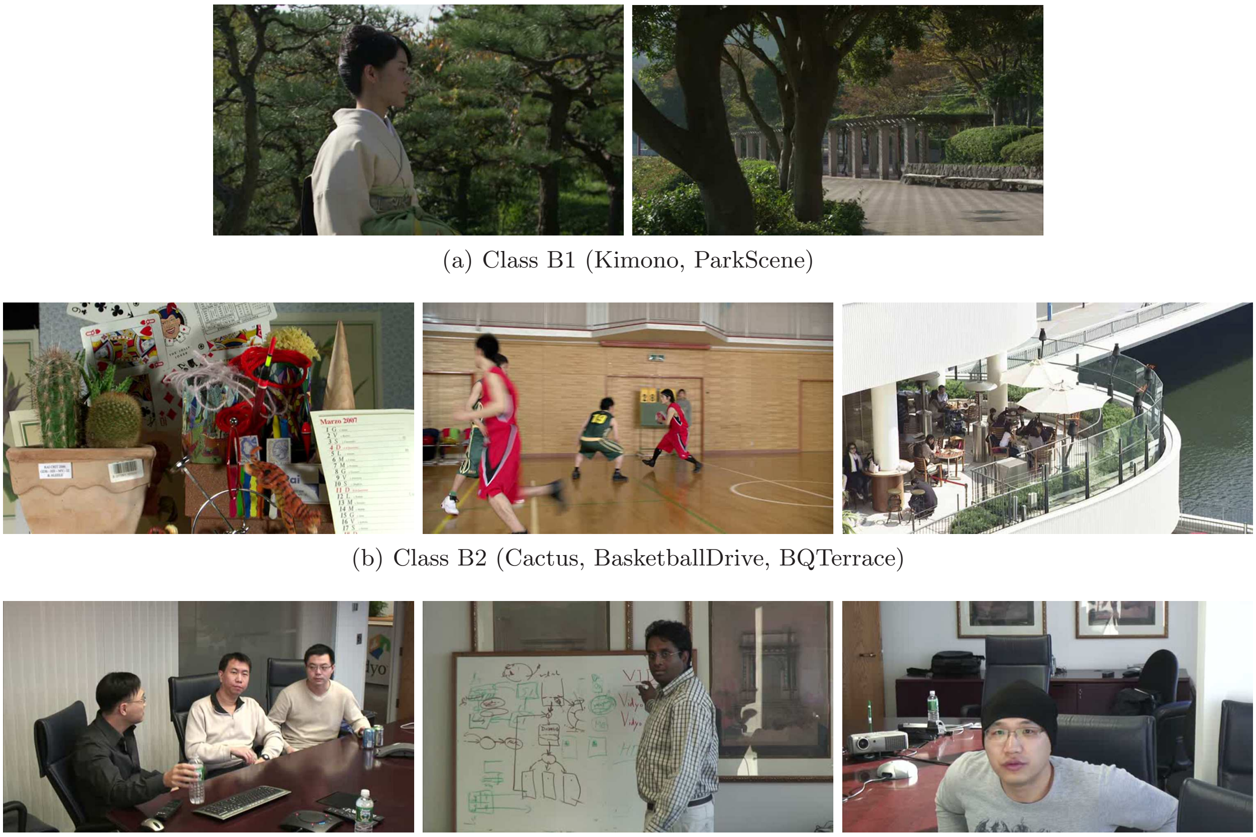
CG301W monitor, with a native resolution of 2560x1600 pixels, a gray-to-gray response time of 6 ms, and a black-white-black response time of 12 ms, was selected for our test. Three of these monitors were connected to the graphic board of the video server, using two DVI and one display port (DP) output of the graphic board by means of a DP to DVI adapter.

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 consisted of neon lamps with 6500 K color temperature, while the color of all the background walls and curtains present in the test area was mid grey. 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 activities and of possible unexpected events which could influence the test results. A picture of the MMSPG test environment is shown in Figure 1(b).

Depending on the resolution of the test material, different viewing conditions (number of subjects, viewing position) were used. For class B test material (see Table 2), 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 (see Table 2), 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.

Class	Resolution	Frame rate	Videos
B1	1920x1080	24	Kimono, ParkScene
B2	1920x1080	50-60	Cactus, BasketballDrive, BQTerrace
C	832x480	30-60	BasketballDrill, BQMall, PartyScene, RaceHorses
D	416x240	30-60	BasketballPass, BQSquare, BlowingBubbles, RaceHorses
E	1280x720	60	Vidyo1, Vidyo2,Vidyo3

Table 2. Details of the classes of test material selected to evaluate the performance of the proponents.



(a) Class B1 (Kimono, ParkScene)

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

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

Figure 2. Sample frames of the individual video sequences from the different classes considered in the subjective test.

3. DATASET AND TEST METHODOLOGY

3.1 Dataset

The test material selected for evaluating the performance of the proponents aimed at covering many relevant application scenarios for the next-generation video coding standard. The dataset described in the CfP included 4 classes with different spatial and temporal resolutions as shown in Table 2. 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 class B and E data, thus, 5 and 3 different contents, respectively. The first frames of these video contents are shown in Figure 2.

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

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

Table 3. Overview of the test conditions defined for the different classes, in terms of combinations of coding bit rates (Mbps) and coding condition.

- Random access (RA): group of picture (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 for each content is shown in Table 3.

Apart from the 27 proponents, two H.264/MPEG-4 AVC anchors were included as benchmarks, namely: anchor alpha (A), corresponding to AVC High Profile (HP) with hierarchical B frames (IbBbBbBbP), CABAC and 8x8 transform, and anchor beta (B), corresponding to AVC High Profile (HP) with hierarchical P frames (IpPp), CABAC and 8x8 transform.

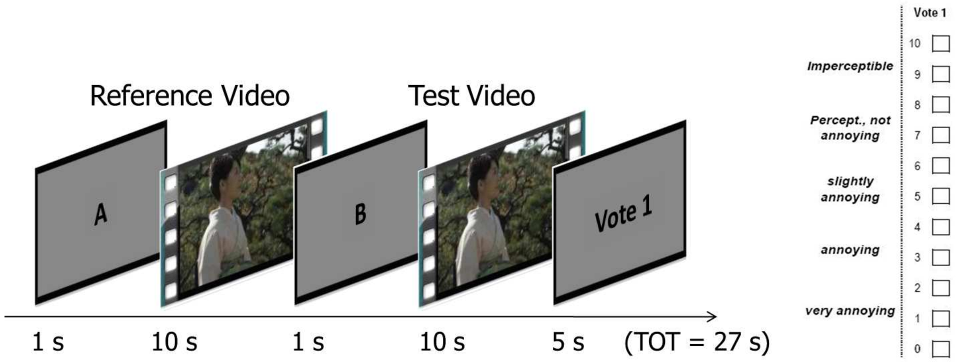
3.2 Test methodology

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.[7]

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 Figure 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 defects that she/he observes in stimulus B. The used rating scale is shown in Figure 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 3.

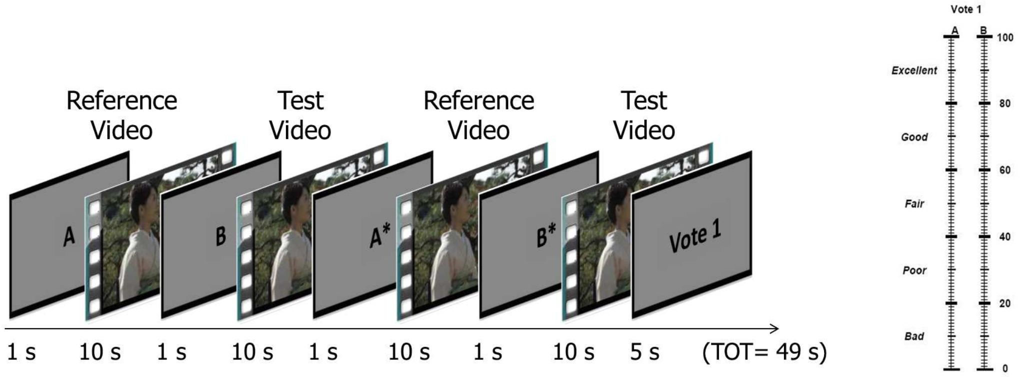
On the other hand, in the DSCQS method, pairs of sequences, i.e. stimuli A and B, are presented twice sequentially to the observer and then she/he is asked to rate the quality of both stimuli, as shown in Figure 4(a). The stimulus A is always the reference video but the subject is not told about it. The selected rating scale is shown in Figure 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 3.

Considering the two content classes, i.e. class B and E, and the two test methodologies, three different tests took place: a DSIS test for part of the class B data, a DSIS test for class E data, and a DSCQS test for the remaining class B data.



(a) Presentation (b) Scale

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



(a) Presentation (b) Scale

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

3.3 Session plan

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.

In order to retain the concentration of the subjects, a subjective video quality test session should not last more than 30 minutes.[7] 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 the reliability of the subjects.

As shown in Figure 3, one DSIS presentation, i.e. presentation of two stimuli and rating time, takes approximately 27 seconds. Therefore, test sessions of 33 presentations (i.e. 3 dummies + 29 stimuli + 1 reference pair), corresponding to a duration of 14.85 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 Figure 4, one DSCQS presentation, i.e. two consecutive presentations of two stimuli and rating time, takes 49 seconds. 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 17.96 minutes. 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.

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 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 old. Each subject was paid 100 CHF for two half days of test activity. All participants were screened for correct visual acuity and color vision using Snellen and Ishihara Charts. A picture of the screening area in shown in Figure 1(a).

The training of the subjects of each group was conducted at the beginning of each half a day, as a 15-minute training session, where oral instructions were provided to explain the task and a viewing session was performed to allow the viewer to familiarize with the assessment procedure. The video sequences used as training samples had quality levels representative of the labels reported on the rating scales: the experimenter explained the meaning of each label reported on the scale and related them to the presented sample sequences.

To collect evaluation scores, the subjects were provided with scoring sheets to enter their quality scores. The scores were then offline converted into electronic version. Furthermore, all the scores were converted by two operators to identify and correct any eventual manual mistake.

4. SCORE PROCESSING

4.1 Normality test

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 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.

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.[10] 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. The results of this test showed that the score distributions for each subject across different test conditions were not normally distributed, while the majority of the score distributions across subjects were normal or close to normal. The results of this test justify the processing applied to the data which is detailed hereafter.

4.2 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 25^{th} and 75^{th} percentiles of the scores distribution, respectively.[10] 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. A maximum of two subjects per session were discarded as ouliers, across all the test sessions.

4.3 Statistical measures

After removing the outliers, statistical measures were computed to describe the score distribution across the subjects for each of the test conditions (combination of content, coding condition and bit rate).

For the DSIS methodology, the mean opinion score (MOS) is computed as

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

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) is computed as

$$DMOS_j = \frac{\sum_{i=1}^{N} (s_{ij}^{A} - s_{ij}^{B})}{N} \tag{2}$$

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

$$MOS(DSCQS)_j = \frac{100 - DMOS_j}{10}. \tag{3}$$

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:

$$CI_j = t(1 - \alpha/2, N) \cdot \frac{\sigma_j}{\sqrt{N}} \tag{4}$$

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 a single test condition j across the subjects i. The confidence intervals were computed for an α equal to 0.05, which corresponds to a degree of significance of 95%.

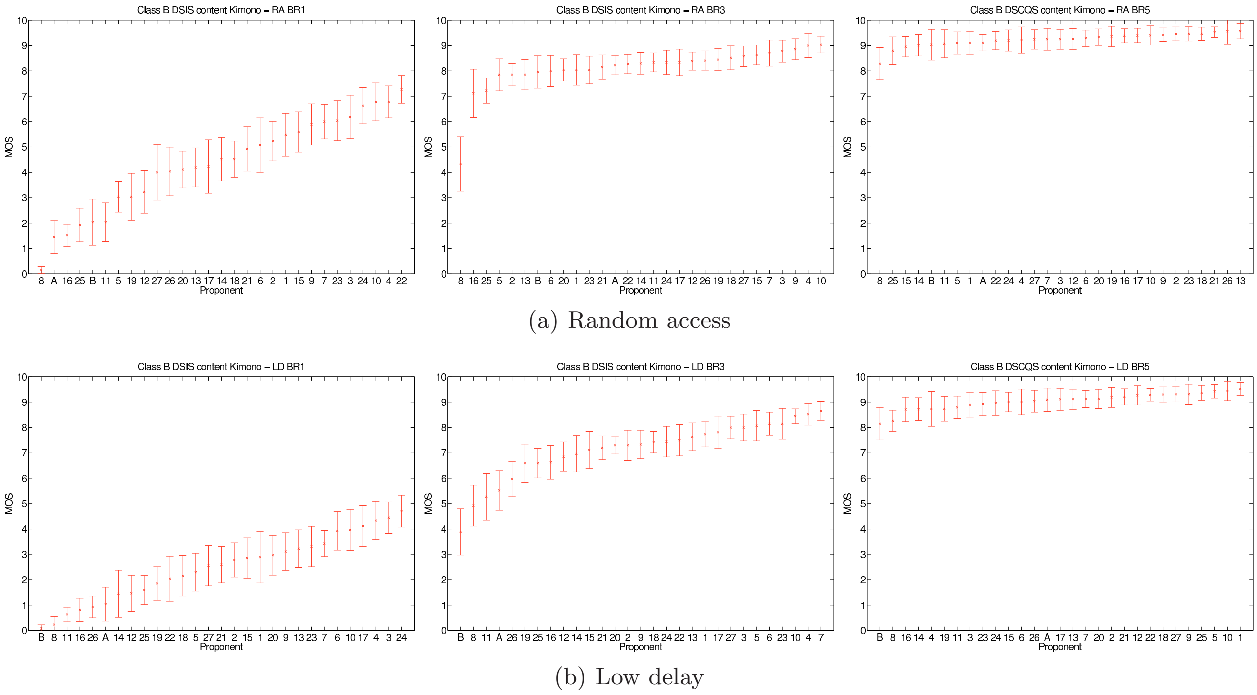
5. RESULTS AND DISCUSSION

5.1 Individual contents and bit rates

Due to page limitations, it is not possible to show the results for all the test conditions of all the classes. Therefore, a few representative results for each of the classes are shown in Figures 5-7. The plots show the MOS and CI results for the 27 proponents (labels 1 to 27) and two anchors (labels A and B), sorted with increasing MOS values. From the considered bit rates, only the lowest (BR1), the middle (BR3), and the highest (BR5) are shown. As it can be seen from the small confidence intervals, the results are reliable and the variations between the subjects are rather small. The results show that, especially for lower bit rates, the performance of the individual proponents differs considerably and some lead to a substantial quality improvement compared to the anchors.

5.2 Overall performance

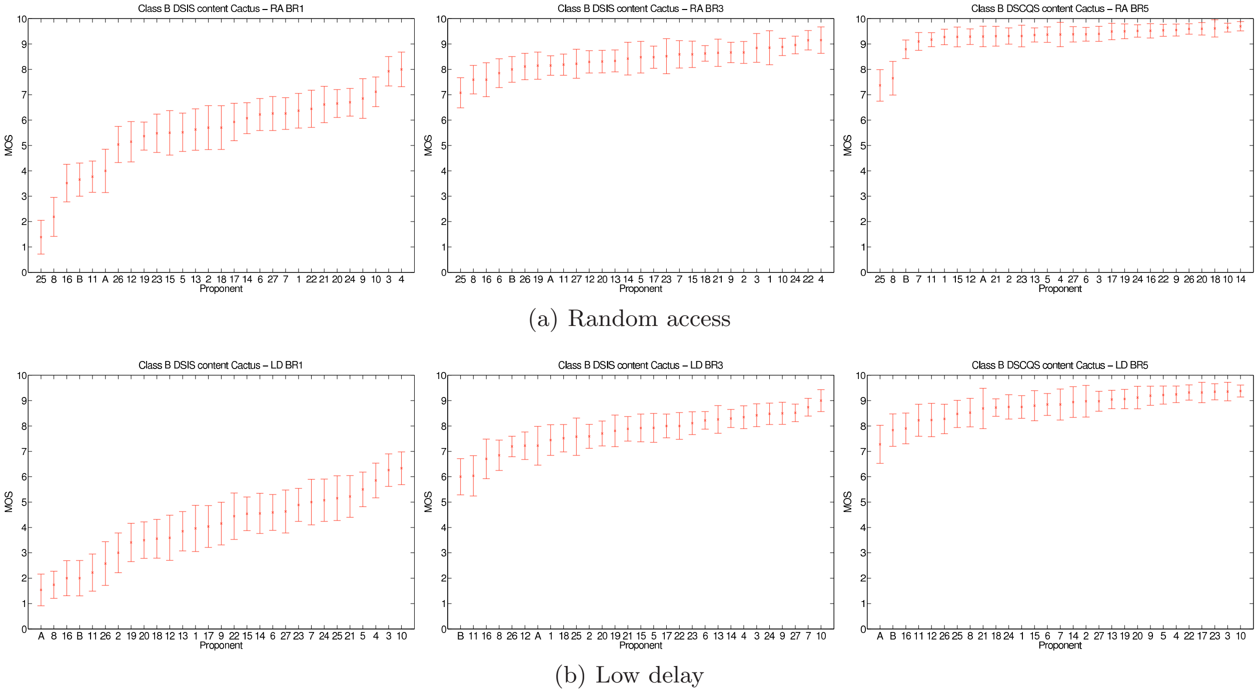
While the ranking of the proponents varies across the different classes, coding conditions and bit rates, some proponents generally perform better than others. In order to accurately analyze the performance of each proponent and evaluate whether the obtained results are significantly different from those obtained with the anchors, a multiple comparison procedure has been applied separately to the scores obtained for each test condition (combination of content, coding condition and bit rate).[10]



(a) Random access

(b) Low delay

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



(a) Random access

(b) Low delay

Figure 6. MOS/CI results for class B2 content Cactus for low (BR1), middle (BR3) and high (BR5) bit rates. The proponents are ordered for each bit rate individually with increasing MOS value.

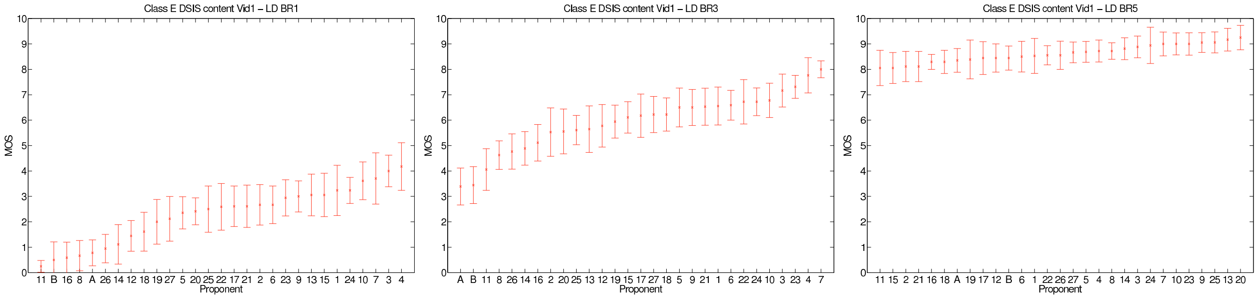


Figure 7. MOS/CI results for class E content Vidyo1 for low (BR1), middle (BR3) and high (BR5) bit rates. The proponents are ordered for each bit rate individually with increasing MOS value.

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
Proponent	1	10	24	4	7	6	17	3	9	22	15	23	27	21	18	13	5	20	2	12	14	19	25	26	16	11	8	B
$P > A$	70	70	68	67	67	63	63	61	59	57	56	54	54	52	48	41	40	39	33	32	26	23	21	20	7	4	2	0
$P = A$	30	30	32	33	33	37	37	39	41	43	44	46	46	48	52	59	60	61	67	68	74	77	71	80	90	96	87	90
$P < A$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	2	0	11	10
Proponent	4	7	10	3	9	17	22	1	6	24	23	15	21	2	13	18	27	20	5	14	19	12	25	36	16	A	11	8
$P > B$	77	72	71	68	67	67	67	66	66	65	63	61	61	56	54	54	54	52	51	45	44	39	39	22	17	10	9	5
$P = B$	23	28	29	32	33	33	33	34	34	35	37	39	39	44	46	46	46	48	49	55	56	61	56	78	83	90	91	85
$P < B$	0	0	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	10

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), expressed in % over the entire set of test conditions.

To compare two groups of scores and understand whether their means are statistically 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, there are also many pairs to compare. 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.

The overall performance of each proponent can be summarized by counting how many times each proponent performed better, equal or worse than each anchor, as resulting from the multiple comparison test. These values, expressed in percentages over the entire dataset of different contents, coding conditions and bit rates, are reported in Table 4.

On one hand, from these overall performance indexes, it can be easily noticed that at least 8 proponents showed better performance, with respect to both anchors, in more than 60% of the test cases. Furthermore, an accurate analysis of the MOS and CI plots reveals that, 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. On the other hand, only in very rare cases, few proponents performed worse than the anchors.

6. 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. Twenty-seven proponents have been evaluated, in comparison to two AVC anchors, in the most extensive subjective test campaign in the history of video coding standardization. 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 codecs. Particulyrly, the test results clearly indicate that some proposals exhibite 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 an eventual 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 the proponents who provided the test material and made this subjective test campaign possible. Also, a special thank goes to the subjects who participated with high dedication to the subjective evaluations reported in this paper.

REFERENCES

[1] ISO, "Information technology – coding of audio-visual objects – part 10: Advanced video coding," International Standard ISO/IEC 14496-10:2005, ISO/IEC (2005).

[2] Sullivan, G. J. and Ohm, J., "Recent developments in standardization of high efficiency video coding (HEVC)," *to appear in Proceeding of Applications of Digital Image Processing XXXIII* (2010).

[3] ISO, "Call for test materials for future high-performance video coding standardization," Tech. Rep. N10362, ISO/IEC JTC1/SC29/WG11, Lausanne, Switzerland (2009).

[4] ISO, "Call for evidence on high-performance video coding (hvc)," Tech. Rep., ISO/IEC JTC1/SC29/WG11, Maui, US (2009).

[5] ISO, "Joint call for proposals on video compression technology," Tech. Rep., ISO/IEC JTC1/SC29/WG11 ITU-T Q6/16 Visual Coding, Kyoto, JP (2010).

[6] ISO/IEC, "Information technology – generic coding of moving pictures and associated audio information: Video," International Standard 13818-2:2000, ISO/IEC (2000).

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

[8] VQEG, "Report on the validation of video quality models for high definition video content," Tech. Rep., VQEG (June, 2010).

[9] Tourancheau, S., Le Callet, P., Brunnstroem, K., and Andren, B., "Psychophysical study of LCD motion-blur perception," *Proceedings Vol. 7240, Human Vision and Electronic Imaging XIV* (2009).

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