SUBJECTIVE QUALITY ASSESSMENT OF THE EMERGING AVC/H.264 VIDEO CODING STANDARD

Tobias Oelbaum[1] Vittorio Baroncini[2] Thiow Keng Tan[3] Charles Fenimore[4]

[1]Munich University of Technology, Germany [2]FUB, Italy [3]NTT DoCoMo, Inc., Japan [4]NIST, USA

ABSTRACT

The results of a formal subjective test comparing the AVC/H.264 video coding standard with the widely used MPEG-2 video coding standard was recently released by the MPEG working group.

New products based on the recently completed AVC/H.264 video coding standard are being announced every other week. Similarly, there have been numerous reports and claims about the visual quality that can be achieved with this new standard. It is therefore important for the broadcasting community to understand the actual benefits that come with this new video coding standard compared to previous standards.

This paper tries to give further insights and indications to the important question on what the benefits and gains really are. An introduction to the above test and a description of the test conditions and the test environment evaluation is presented. This is followed by results from subjective tests of current standard television resolution as well as ATV up to HDTV resolutions.

This work is the result of the activities of the JVT/MPEG Ad Hoc Group on AVC Verification Test. The authors of this paper acted as chairmen of the Ad Hoc Group during the period it was active from July 2002 to December 2003.

INTRODUCTION

Advance Video Coding (AVC, also know under its ITU-T denotation H.264) is the natural successor to the enormous successful MPEG-2 video coding standard. After close to a decade of video coding using MPEG-2 technology for broadcasting, cable and digital storage media (such as DVD), there is now a very high interest in new technologies that could deliver a better coding efficiency compared to current technologies. New video and TV formats like ATV and HDTV that provide more fidelity than standard definition TV are gaining more and more importance. Customers are asking for higher quality pictures, while broadcasters seek to use the same available bandwidth to deliver more channels and DVD publishers are putting more content onto the same disk. All this requires a video codec capable of more efficient compression for transmitting and storing video.

AVC/H.264 was developed within the Joint Video Team (JVT) in a joint effort of experts from the Motion Picture Experts Group (MPEG) of ISO/IEC and the Video Coding Experts Group of ITU-T (VCEG). This combination has already shown its success in the development of MPEG-2/H.262. After the initial work by VCEG and more than two years of joint development, the final draft was approved by MPEG in March 2003, followed by ISO/IEC ballot approval in October 2003. Similarly, Recommendation H.264 in the ITU-T was completed by a “decision” (final approval) by ITU-T SG 16 in May 2003. In parallel with the

final phase of the standardization process, a formal verification test of the new standard was conducted by MPEG and published in December 2003.

Since then many companies have announced products using AVC/H.264, showing impressive demos and promoting the new standard for the next generation of video broadcasting and DVD distribution. For example, the soccer world cup in 2006 will be broadcasted in 1080i HDTV and the DVD-Forum is considering AVC/H.264 as one possible codec for the blue ray HD-DVD.

Subjective visual tests were carried out in November and December 2003 to analyse the performance of the new standard in comparison to existing video coding technology. This paper presents the results of this formal verification test within MPEG and tries to provide some insight of the current capabilities of AVC/H.264.

TEST CONDITIONS FOR THE COMPARISON OF AVC/H.264 TO MPEG-2

Unlike the MPEG-2 video coding standard, AVC/H.264 targets a wider range of video applications, ranging from video at mobile devices and bit rates as low as below 30Kbit/s to HDTV and bit rates of 20Mbit/s and above. To cover this wide range of resolutions, frame rates and bit rates the test was split into three different cases: one for the low end targeting mainly the mobile market and current internet bit rates, the second one targeting today's standard definition TV, testing both PAL and NTSC resolutions, and a third test targeting the upcoming new high resolution standards. In this contribution only the two latter test cases are reviewed. Hereinafter, we shall refer to these two cases as SD and HD, respectively.

The new standard was compared to two different implementations of MPEG-2 encoders. To be able to compare AVC/H.264 at a very early stage of optimisation - remember that the standard was finalized only months before these tests were performed - to a MPEG-2 encoder with a similar level of optimisation one reference point for the emerging standard was MPEG-2 TM5. TM5 is the reference implementation of MPEG-2 and still the basis of many low-cost MPEG-2 encoders. The second references were highly optimized state of the art MPEG-2 encoders. This allowed comparison of the new standard to current top-level MPEG-2 video encoding technology (later referred to as MPEG-2 HiQ). To ensure anonymity several companies were asked to encode the video sequences selected for the test and the best results were then chosen to run in the test.

AVC/H.264 bit streams were provided by Fraunhofer-HHI (SD and HD sequences), Sony (HD) and Tut Systems (SD). If multiple encodings were available for the same sequence the visually best version was chosen by the means of expert viewing.

Both MPEG-2 and AVC/H.264 bit streams were decoded using available reference software and had to be compliant to respective profile and level constraints as given in Table 1.

Test Case	Codec	Profile	Level	Bit rates (MBit/s)
SD	AVC/H.264	Main	L3	6; 4; 3; 2.25; 1.5
	MPEG-2	Main	Main	6; 4; 3; 2.25
HD	AVC/H.264	Main	L4	20; 10; 6
	MPEG-2	Main	High	20; 10; 6

Table 1 – Bit rates, profiles and levels for codecs tested

As prefiltering of video material in advance to coding is common industry practice, prefiltering was allowed for all participants. Prefiltering for AVC/H.264 encoding was done by Dolby Laboratories. A list of the test sequences used in the test cases can be found in Table 2, stills of the sequences are presented in Appendix A.

Test Case	Sequence *short description*	Resolution (width x height)	Field rate (fields per second)	Frame rate (frames per second)
SD	Football	720x486	60	
	American football, fast motion, camera motion.			
	Husky	720x576	50	
	Husky running, camera motion, high detailed background.			
	Mobile	720x576	50	
	Test sequence, high contrast, complex regular object motion, saturated colours, pan.			
	Tempete	720x486	60	
	Flowers, falling leaves, stones and some water, zoom out, chaotic object motion.			
HD (720p)	Crew	1280x720		60
	NASA crew leaving a building, flashlights, pan.			
	Harbour	1280x720		60
	Harbour scene, water, small sailing boats passing by, many rigs in the foreground.			
HD (1080i)	New Mobile	1920x1080	60	
	Similar to 'Mobile', high details, text.			
	Stockholm Pan	1920x1080	60	
	Pan over Stockholm city, very high details, water, many regular structures.			
HD (1080p)	Riverbed	1920x1080		25
	Close view of a riverbed, water, transparency.			
	Vintage Car	1920x1080		25
	Vintage car driving at a gravel road in a forest, pan.			

Table 2 – Video sequences used for the test

SUBJECTIVE TESTS

Subjective tests were based on ITU-R BT-500, which describes the test conditions and the test setup for subjective visual tests. The tests were prepared and conducted at Fondazione Ugo Bordoni (FUB), Istituto Superiore delle Comunicazioni e della Tecnologia delle Informazioni (ISCTI) in Rome (Italy), Munich University of Technology (Germany) and at the National Institute of Standards and Technology (NIST) (USA).

All tests used the Double Stimulous Continiuos Quality Scale (DSCQS) test method which is described in ITU-R BT-500-11 [2]. The only differences from what is stated in BT-500-11 concerns the displays for the 720p and 1080p cases where a Digital Light Processing (DLP) projector was used instead of a CRT monitor.

The core of the DSCQS test method is the dislay order of encoded and original sequence and the scale which is used to express the quality of a certain video sequence.

In the DSCQS test method each test cell consits of a coded video and its uncoded (original) version. The order of coded and original video is random. One pair of coded and original video sequences is repeated once before the subjects are asked to rate the quality of the two video sequences.

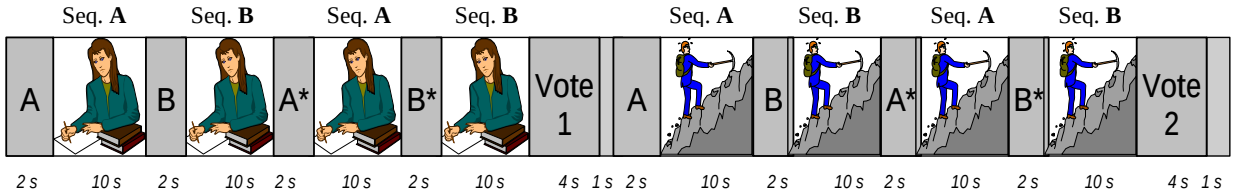

Figure 1 – Sequence ordering for DSCQS test method

For rating a continious scale ranging from 0 (low quality) to 100 (high quality) was used. The subjects were asked to rate both (the coded and the original) video sequences in terms of the overall quality of each clip. Note that the subjects were not informed, that one of the video sequence is the original but were just asked to rate the quality of both test clips.

To receive meaningful and consistent results at least 20 subjects were involved in each test. All subjects passed the Snellen test for visual acuancy and were tested for color blindness using Ishihara test charts.

Before the actual tests start each subject passed a training phase to become familiar with the testing procedure. To adjust the range from a badly encoded video to a perfect reconstruction, the first three sequences of each test are choosen to reflect the whole range of possible quality during this test. The votes for these first three pairs of video were later ignored. They were just used to allow each subject to set their personal range from 'bad' to 'perfect'.

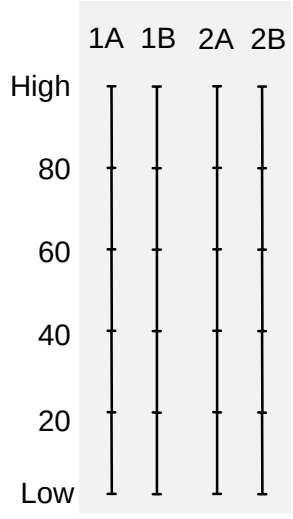

Figure 2 – DSCQS voting scale

Figure 3 - MOS scale with typical interpretation

The vote for each coded sequence was then calculated as $Q_{Loss} = Q_{Orig} - Q_{Coded}$. As the DSCQS method allows the coded video to be rated better than the original (as it may appear in cases where the coded video is visually transparent) Q_{Loss} was clipped to 0. Q_{Loss} was then matched to a standard Opinion Scale (OS) scale by calculating $Q_{OS} = 5 - Q_{Loss}/20$.

All data was then statistically processed to obtain the Mean Opinion Scale (MOS) by averaging the votes of all subjects. In addition the Standard Deviation and the 95% Confidence Intervall (CI) were computed. It is assumed that a lack of overlap with the 95%

CI provides a strong indication of the existence of differences (from the statistical point of view) between adjacent MOS values. A MOS value of 4.5 or above shows that the quality of the respective video is transparent - meaning that coded and original video are statistically indistinguishable.

RESULTS

Standard TV resolution

AVC/H.264 showed an increase in coding efficiency of a factor of at least 1.5 in 8 of 12 statistically conclusive cases when compared to a highly optimized MPEG-2 encoder (MPEG-2 HiQ). For the two low motion sequences 'Mobile' and 'Tempete' AVC/H.264 can achieve the same quality as MPEG-2 with not more but half the bit rate. For these two cases a very good quality can be reached with a bit rate as low as 1.5MBit/s.

The maximum difference between AVC/H.264 and its MPEG-2 competitors can be seen at the 'Mobile' sequence where the quality of AVC/H.264 at 1.5Mbit/s is as good as the one for MPEG-2 at 6MBit/s.

On the other hand the differences for the 'Husky' sequence, which contains a lot of fast object and camera motion and a high textured background, are comparably small, when it comes to the higher bit rates. What can be noticed here is that at bit rates below 3Mbit/s the gain achieved by AVC/H.264 becomes bigger. This effect can also be noticed looking at the 'Football' sequence, which is of similar complexity as the 'Husky' sequence.

As expected the difference in coding efficiency is even bigger when AVC/H.264 is compared to MPEG-2 TM5.

The graphs in Figure 4 show some selected data points. A complete set of results can be found in [3]. Beside the MOS value the graphs also show the 95% Confidence Intervals.

For the 'Tempete' sequence the quality reached with AVC/H.264 at 1.5MBit/s is next to transparent and clearly outperforms the MPEG-2 TM5 at 4MBit/s, while it is slightly better as MPEG-2 HiQ at 3MBit/s. For 'Football' the quality gain is much less, but what can be seen is that AVC/H.264 still provides acceptable quality at 2.25MBit/s, which is not true any more for the MPEG-2 cases.

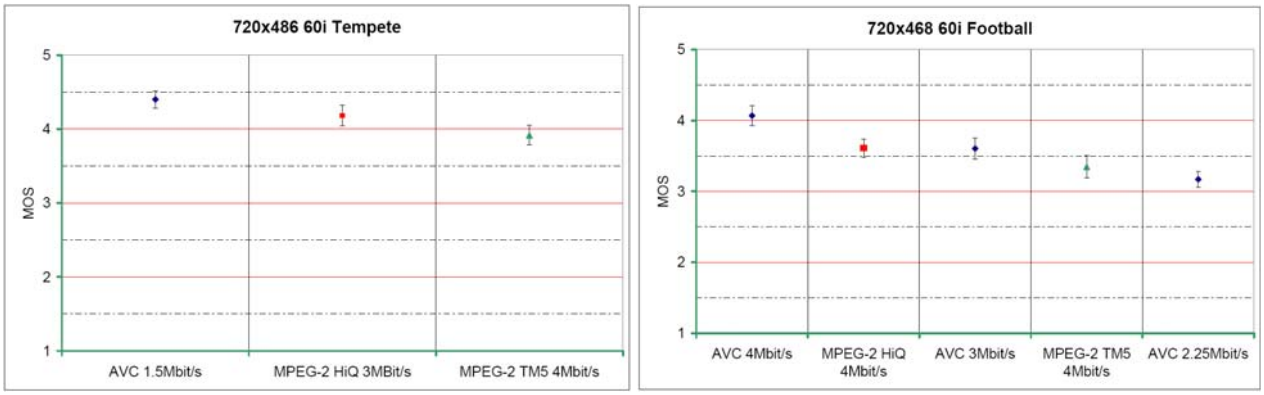

Figure 4 – Selected results for standard TV resolution

ATV and HDTV

Compared to MPEG-2 HiQ AVC/H.264 reveals an increase of coding efficiency by a factor of 1.7 and more in 7 of 9 cases. Except for the very challenging 'Riverbed' sequence AVC/H.264 provides a very good quality even at the lowest bit rate that was tested and reaches transparency for 'Harbour' (720p) and 'Vintage Car' at 6Mbit/s. For all cases the 6Mbit/s AVC/H.264 was at least as good as the 10MBit/s MPEG-2 HiQ.

Selected results for the two sequences 'Riverbed' (1080p) and 'Crew' (720p), which are the two most challenging sequences out of the ATV/HDTV test set, are presented in the graphs shown in Figure 5. AVC/H.264 at 6MBit/s clearly outperforms MPEG-2 at 10 MBit/s for 'Riverbed' and delivers at least equal quality for the 'Crew' sequence.

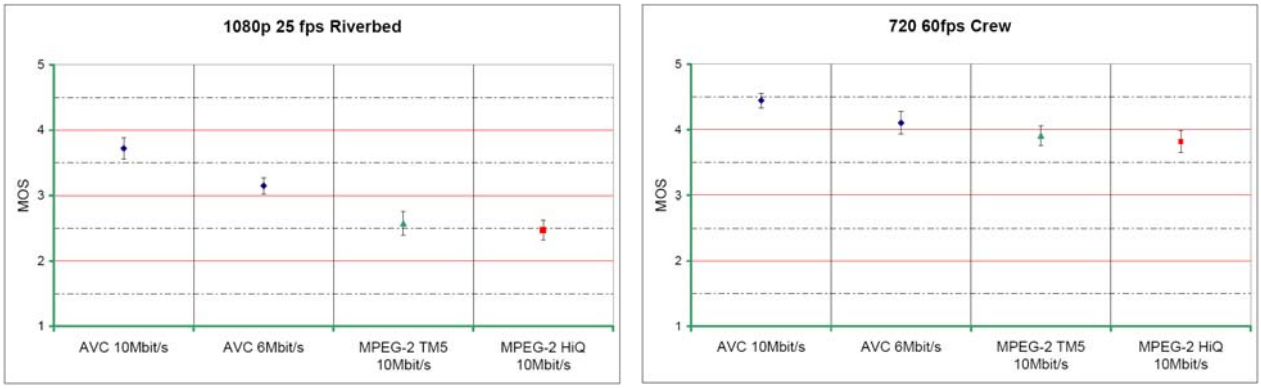

Figure 5 – Selected results for ATV and HDTV resolution

Summary of the results

The tests showed a noticeable superiority of AVC/H.264 compared to state of the art MPEG-2 video encoders in almost all test cases. This is especially true for sequences with comparably slow but complex motion such as 'Mobile' or 'Tempete'.

Subjective comparison reveals much less colour distortions for sequences encoded with AVC/H.264 encoders where saturated colours are present. Furthermore, sequences encoded with AVC/H.264 encoders have less noise-like impairments and appear to have much less blocking artefacts compared to sequences encoded with MPEG-2 encoders.

Sequences like 'Husky' and 'Football', where the differences between sequences encoded with AVC/H.264 and MPEG-2 encoders at middle and high bit rates were comparably small, exposed that current AVC/H.264 implementations tend to smoothen out too many details in sequences with high motion.

When looking at the results, one should take into account two main boundary conditions of this quality evaluation. First of all, the tests took place only a few months after the standard was finalized. This put the companies delivering the AVC/H.264 encoded bit streams at a disadvantage. The AVC/H.264 encoders, which were at a very early stage of optimization, competed with MPEG-2 encoders that have been optimized for nearly a decade.

Secondly, this disadvantage for AVC/H.264 was partly balanced by the fact, that AVC/H.264 offered many more options during encoding: sub pixel interpolation up to 1/4 pixel, several reference frames, several options for bidirectional prediction, sub blocks with only 4x4 pixel – just to name some features. Together with other key features, such as CABAC, these options result in a significant increase in coding efficiency. The test setup, that did not require real time encoding, allowed AVC/H.264 to explore all possible combinations and features to reach the best coding efficiency. For real time AVC/H.264 encoders it will

become hardly feasible to systematically explore that many available encoding options.

So on the one hand one could expect AVC/H.264 encoders having more sophisticated and fine tuned rate control and rate/distortion optimisation strategies as the ones used to produce the test bit streams. On the other hand first real time (or next to real time) implementations will most probably not reach the very high quality provided for this test.

CONCLUSION

Even at a very early stage of optimization AVC/H.264 has shown a remarkable superiority in coding efficiency compared to current MPEG-2 encoding technology.
AVC/H.264 can deliver acceptable or even good quality at bit rates as low as 1.5MBit/s and 6MBit/s for SD and HD sequences, respectively. These are bit rates where MPEG-2 could not deliver acceptable picture quality any more.

As the standard specifies the bit stream syntax and the decoding procedure only but does not restrict encoder strategies, there is much space for further optimization and future AVC/H.264 encoders probably will provide broadcast quality at bit rates even lower than what was selected for this test.

REFERENCES

[1] ISO/IEC 14496-10:2003 Information technology -- Coding of audio-visual objects -- Part 10: Advanced Video Coding

[2] ITU-R BT-500-11 Methodology for the subjective assessment of the quality of television pictures

[3] “Report of the formal Verification Tests on AVC/H.264”. ISO/IEC JTC1/SC29/WG11 MPEG2003/N6231, December 2003 Waikoloa Hawaii, USA (public available through http://www.chiariglione.org/mpeg/quality_tests.htm)

APPENDIX A: TEST SEQUENCES

Figure A1 - Husky 720x576 50i

Figure A2 - Mobile 720x576 50i

Figure A3 - Football 720x486 60i

Figure A4 - Tempete 720x576 60i

Figure A5 - Crew 1280x720 60p

Figure A6 - Harbour 1280x720 60p

Figure A7 – New Mobile 1920x1080 60i

Figure A8 – Stockholm 1920x1080 60i

Figure A9 – Riverbed 1920x1080 25p

Figure A10 – Vintage Car 1920x1080 25p