

Subjective testing methodology in MPEG video verification

Charles Fenimore^{*a}, Vittorio Baroncini^b, Tobias Oelbaum^c, Thiw Keng Tan^d

^aNational Institute of Standards and Technology, Gaithersburg MD, USA 20899-8940

^bFondazione Ugo Bordonini, Via B. Castiglione, 59, 00142 – Rome, Italy

^cInstitute for Data Processing, Munich University of Technology, D-80290 Munich, Germany

^dNTT DoCoMo Inc., Multimedia Laboratories, 3-5 Hikarinooka, Yokosuka, Kanagawa, 239-8536
Japan

ABSTRACT

The development of new video processing, new displays, and new modes of dissemination and usage enables a variety of moving picture applications intended for mobile and desktop devices as well as the more conventional platforms. These applications include multimedia as well as traditional video and require novel lighting environments and bit rates previously unplumbed in Moving Picture Experts Group (MPEG) video compression. The migration to new environments poses a methodological challenge to testers of video quality. Both the viewing environment and the display characteristics differ dramatically from those used in well-established subjective testing methods for television. The MPEG Test Committee has adapted the television-centric methodology to the new testing environments. The adaptations that are examined here include:

- The display of progressive scan pictures in the Common Intermediate Format (CIF at 352x288 pixel/frame) and Quarter CIF (QCIF at 176x144 pixel/frame) as well as other, larger moving pictures requires new ways of testing the subjects including different viewing distances and altered ambient lighting.
- The advent of new varieties of display technologies suggests there is a need for methods of characterizing them to assure the results of the testing do not depend strongly on the display.
- The use of non-parametric statistical tests in test data analysis. In MPEG testing these appear to provide rigorous confidence statements more in line with testing experience than those provided by classical parametric tests.

These issues have been addressed in a recent MPEG subjective test. Some of the test results are reviewed; they suggest that these adaptations of long-established subjective testing methodology for TV are capable of providing practical and reliable measures of subjective video quality for a new generation of technology.

KEYWORDS: compression; display measurement; methodology; MPEG; multimedia; non-parametric statistics; progressive-scan video; subjective assessment; video quality

1. INTRODUCTION

The MPEG Testing Group confronts a great variety of video formats in verifying the performance of modern compression tools. For example, in recent MPEG Verification Tests of the Advanced Video Codec (AVC/H.264) the image formats ranged from 176x144 to 1920x1080 pixels per frame, having frame rates from 8 to 60 frames per second. These tests confirmed that when compared to the compression tools in the older MPEG-2 and MPEG-4 standards the advanced coding tools of AVC/H.264 provide a doubling of compression efficiency in about 3/5 of the test conditions that were examined [1]. A brief summary of the results is given in Section 7, while the detailed reports of these results are to be published elsewhere [2]. The focus of this report is on the adaptation of standard television test methodology to newer multimedia and high definition formats and viewing configurations.

The basis for television quality testing has long been the International Telecommunication Union's Recommendation for Subjective Testing of Televisions Systems, BT-500 [3]. The Recommendation represents many years of development of TV picture quality testing, largely tuned to the demands of interlaced television, both at 25 F/s (Frames/second), having 576 lines per picture height, and at 30 F/s, having 486 lines per picture height. Yet, in the last decade a new generation of

* fenimore@nist.gov; 1 301 975 2428; <http://www.nist.gov/>.

moving picture platforms has emerged. These are predominantly progressively scanned with various picture sizes and frame rates, largely dictated by the demands of information technology and mobile applications. The new technologies for display and video processing imply that the test methods must change to reflect the novel elements of the display platforms and the viewing environments in which they are used. Two notable changes are those of picture size and the change to progressive display of video. The implications of these changes as they relate to multimedia and high definition testing have not been fully developed. Section 2 describes the formats (ranging from QCIF to HD video) and the materials used in those tests.

The large outstanding issue with respect to testing design is the formulation of reliable procedures that are well-adapted to the new video formats and displays. In the AVC/H.264 Verification Tests, Liquid Crystal Displays (LCDs) were used for the QCIF and CIF testing conducted at Istituto Superiore delle Comunicazioni e delle Tecnologie dell'Informazione (ISCTI) in Rome. In both the portable platform and computer environment, the customary use is to have one viewer per display. Moreover, the viewing distance may differ from the four to six picture heights specified for standard definition by Rec. 500-10 [3]. The third and fourth sections address new aspects of the methodology of multimedia definition testing as it was performed in the AVC/H.264 Verification Tests (VT). The methodology is not specific to testing compression performance and should apply broadly to a range of multimedia assessments.

The new applications call for display on computer monitors and on portable devices, in addition to traditional cathode ray tubes (CRTs). The potential variety of displays poses two challenges to the test designer: how to assure the test results do not depend critically on the particularities of the display technology and how to modify the existing test methods to assure the results are reliable and repeatable. In characterizing the systems, one sets the available controls to best assure adequate resolution, contrast and brightness, and motion rendition. Particularly in addressing display attributes for high definition testing, we have used a body of metrological methods that has been developed in the last few years. These methods permit the characterization of displays independent of technology, be they cathode ray projection, LCD, plasma, or other technology [4, 5]. This matter is addressed in Section 5, particularly as it relates to the setup of the high definition portions of the AVC/H.264 tests conducted at the National Institute of Standards and Technology.

The last technical adaptation is the use of non-parametric tools in the statistical analysis of the AVC/H.264 testing data. Section 6 describes the results of using a computationally intensive test of significance, the Kruskal-Wallis method [6]. As applied to the high definition portions of the AVC/H.264 tests, it suggests the use of nonparametric methods permits inference more in line with testing experience than the inferences associated with a, possibly unwarranted, assumed normal distribution of the test data.

The recent MPEG subjective tests of video quality indicate that these new test methods are sufficiently powerful to permit the developers and users of video to consistently and reliably determine delivered picture quality on a variety of platforms.

2. IMAGE FORMATS AND TEST MATERIALS

The Advanced Video Codec (AVC/H.264) has profiles and levels corresponding to video formats that range from low resolution, low frame rate multimedia to high definition, high frame rate video. During November and December 2003, the MPEG Testing Group ran Verification Tests of the AVC/H.264 over a wide range of formats. These formats of the uncompressed source materials includes QCIF: 176x144 pixels/frame (Pix/F), progressively scanned; CIF: 352x288 Pix/F, progressive; two standard definition formats: 586 X 720 Pix/F and 486 X 720 Pix/F (both interlaced); and three high definition formats: 1280X720 Pix/F, progressive, and 1920X1080 Pix/F both interlaced and progressive.) Tables 1 - 4 indicate the pixels per frame, frame rates, names of test clips, and the compression rates used in the recent AVC/H.264 tests.

2.1 Multimedia Definition (MD) Test

The MD tests were conducted at ISCTI, Fondazione Hugo Bordoni, in Rome, Italy. Test conditions for the MD Baseline Test are listed in Table 1.

Test	MD Baseline Test	
Codecs	AVC/H.264 Baseline @ L2 compared against MPEG-4 Part 2 SP @ L3	
Resolution	CIF (352x288)	QCIF (176x144)
Sequences	Foreman, Head with Glasses, Paris, PanZoom	Foreman, Head with Glasses, Paris, PanZoom
Input rate	15 frames per second	10 frames per second
Bitrate	768 kbps, 384 kbps, 192 kbps, 96 kbps	192 kbps, 96 kbps, 48 kbps, 24 kbps

Table 1: Test conditions for the MD Baseline test.

The test conditions for the MD Main Test are listed in the table below.

Test	MD Main Test			
Codecs	AVC/H.264 Main @ L2 compared against MPEG-4 Part 2 ASP @ L3			
Resolution	CIF (352x288)		QCIF (176x144)	
Sequences	Mobile & Calendar, Husky,	Tempete, Football	Mobile & Calendar, Husky,	Tempete, Football
Input rate	12 frames per second	15 frames per second	8 frames per second	10 frames per second
Bitrate	768 kbps, 384 kbps, 192 kbps, 96 kbps		192 kbps, 96 kbps, 48 kbps, 24 kbps	

Table 2: Test conditions for the MD Main test.

The original sequences for the MD tests were CIF or QCIF versions derived by spatially sub-sampling the SD version of the sequence and were deemed to have a frame rate of 30 F/s. For CIF at 15 F/s, the encoder dropped every second frame from the original sequence. Thus, from an original image sequence with frames numbered 1, 2, 3, 4 to 300, frames 1, 3, to 299 were encoded. For QCIF at 10 fps, the encoder dropped every second and third frame from the original sequence. Thus, for the same original sequencing of frames, only frames 1, 4, 7 ... 298 were encoded. For the display during the test, the decoded frames were replicated to achieve the same number of frames as the original.

2.2 Standard Definition (SD) Main Test

The SD tests were conducted at the Institute for Data Processing, Munich University of Technology, in Munich, Germany. Test conditions for the SD Main test are listed in Table 3:

Test	SD Main Test	
Codecs	AVC/H.264 Main @ L3 compared against MPEG-2 <u>MP@ML</u> (MPEG-2 TM5 & HiQ)	
Resolution	SD (586 X 720 pixels/frame)	SD (486 X 720 pixels/frame)
Sequences	Mobile & Calendar, Husky	Tempete, Football
Input rate	50 fields per seconds	60 fields per seconds
Bitrate	6 Mbps, 4 Mbps, 3 Mbps, 2.25 Mbps, 1.5 Mbps (AVC/H.264 only)	

Table 3: Test conditions for the SD Main test.

2.3 High Definition (HD) Main Test

The HD tests were conducted at NIST in Gaithersburg, Maryland, USA. Test conditions for the HD Main test are listed in Table 4:

Test	HD Main Test		
Codecs	AVC/H.264 Main @ L4 compared against MPEG-2 <u>MP@HL</u> (MPEG-2 TM5 & HiQ)		
Resolution	720(60p)	1080(30i)	1080(25p)
Sequences	Harbour, Crew	Stockholm Pan, New Mobile & Calendar	Vintage Car, Riverbed
Input rate	60 frames/second	60 fields/econd	25 frames/second
Bitrate	20Mbps, 10Mbps, 6Mbps	20Mbps, 10Mbps	20Mbps, 10Mbps, 6Mbps

Table 4: Test conditions for the HD Main test.

3. MULTIMEDIA DEFINITION (MD) TESTING

Unlike television picture quality testing, for subjective multimedia definition (MD) the design of the viewing environment, the setup and use of the displays, and the scoring of picture quality have not been standardized. The recent AVC/H.264 Verification Tests employed the methodology described in the next two sections.

The laboratory set-up for the MD double stimulus impairment scale (DSIS) test method at the ISCTI laboratory was based on using 4 LCD monitors from a single manufacturer properly adjusted to have the same contrast and luminance level. The sequences were displayed in their original resolution and the displays were driven in the native resolution of 1280x1024 pixel. These procedures differ from previous test methods for QCIF and CIF resolution video material in that no spatial upsampling was applied for displaying the video. This resulted in a picture height of 8.6 cm for CIF and 4.3 cm for QCIF resolution video.

The ambient light level was close to the level emitted by the displays. The only light source illuminating the test room came from a uniform illumination of the background wall, obtained by means of white fluorescent lamps placed on the floor and at the ceiling of the wall. The colour of the background wall and of the other walls was as close as possible to D65 Pantone grey tone.

The displays were placed on a standard desk one meter from the background wall and about 30 cm behind the front edge of the desk. Each station was separated from the others by a grey curtain hung from the ceiling. The test area was not illuminated by external light during the test. Using four monitors at four stations permitted the testing of up to four subjects simultaneously. The test area was acoustically isolated with attenuation higher than 40dB from sources in the external environment.

All subjects passed a Snellen test for visual acuity, with vision correctible to 20/20, and a color blindness test using Ishihara color charts. Subjects were between 20 and 35 years old. Instructions given to the subjects and other details are described in the MPEG test report [1]. Each subject sat at a station facing a monitor and was asked to keep both arms on the edge of the desk and not to move the head closer or farther from the screen. A scoring sheet was put on each desk in front of the monitor.

4. MULTIMEDIA DOUBLE STIMULUS IMPAIRMENT SCALE TESTING (MM-DSIS) METHOD

Typical multimedia (MM) displays, such as computer CRT monitors, LCDs, and projectors, are progressively scanned. The Multimedia Double Stimulus Impairment Scale (MM-DSIS) Test method is designed to assess the multimedia signals based on the same protocol as the Double Stimulus Impairment Scale (DSIS) test method defined in international

standard ITU-R Rec. 500 [3]. The main difference is the use of progressive displays such as LCD and computer CRT monitors. For the most part, MM viewing is conducted with one viewer per display, implying for efficiency that one conducts these tests with simultaneous viewing at several stations. This was done in the AVC/H.264 Verification Tests.

The basic unit of testing is the cell, consisting of a pair of video clips of original and encoded video materials. The original material is presented first and the encoded material is presented second. This sequencing allows subjects to concentrate on any impairment they may see and to make a fair comparison having the original as a reference. Presentation order of the cells is determined by randomizing on codec and encoding bit rate.

In this method the subjects express their subjective judgement of the level of impairment by putting a mark in a single box chosen from five levels of quality (Fig. 1). The boxes are labelled with quality adjectives, i.e. Excellent, Good, Fair, Poor, or Bad. During data entry and analysis, the discrete marked levels ranging from Excellent to Bad are converted to discrete numerical scores ranging from 5 to 1.

For each basic test cell of original and encoded video material, the subject votes with a single mark in a column. The subject can recognise the correct column in which to put the mark from the number on the top on each column. The column number is the same as the number that is displayed in the last imagery of each basic test cell; thus, the Nth cell is identified when the viewing screen displays “VOTE N”. Fig. 1 displays a sample scoring sheet for 8 basic test cells. Data collection is done using either paper scoring sheets or custom software allowing automatic voting and entry of the data.

	1	2	3	4	5	6	7	8	
Excellent	<input type="checkbox"/>								
Good	<input type="checkbox"/>								
Fair	<input type="checkbox"/>								
Poor	<input type="checkbox"/>								
Bad	<input type="checkbox"/>								

Figure 1: Example of MM-DSIS Test scoring sheet with 8 basic test cells.

From the point of view of the sequencing of the test clips, a basic test cell for the MM-Test is one of two variants:

-Variant I has a single presentation of the stimulus (original plus encoded clips play once.)

A	Clip “A”	B	Clip “B”	Vote
2s	10s	2s	10s	5s

Table 5: Temporal scheme of Variant I

-Variant II has two presentations of the stimulus (original plus encoded clips play twice),

A	Clip “A”	B	Clip “B”	A*	Clip “A”	B*	Clip “B”	Vote
2s	10s	2s	10s	2s	10s	2s	10s	5s

Table 6: Temporal scheme of Variant II

This leads to a basic test cell duration of **29 seconds** for Variant I and **53 seconds** for Variant II.

The selection of Variant I or Variant II is based on the following factors:

- Stress level of the test (too many conditions lead to too much effort for the subjects leading to viewer fatigue)
- High quality of the images (this suggests Variant I; Variant II is suggested when the image quality range is wide.)

The test should not exceed 30 minutes in order to reduce viewer fatigue. Test lengths in excess of 30 minutes require added test sessions with intervening rest pauses for the subjects. A stabilization phase of 5 basic test cells is included at the beginning of each session.

5. HIGH DEFINITION TESTING, DISPLAY AND VIEWING ENVIRONMENT

Those AVC/H.264-VT HD test clips that were progressively scanned, were presented on a 3-chip digital micro-mirror device (DMD). For the 1920x1080 Pix/F imagery the viewed image was the center-cut 1280x1024 pixel window. The 1920x1080 Pix/F interlaced imagery was displayed on a studio CRT monitor. The DMD technology is just one of several actual or proposed technologies for projection of very high quality, very high-resolution imagery. In order to assure that the results from our tests can be properly interpreted, we followed recommendations of the Video Electronics Standards Association (VESA) and the ITU [4] to make some critical display measurements.

The principle measurements are made using color bars to set the brightness and contrast controls and with the peak brightness (luminance) at 12-ft Lamberts. The measured system Gamma was 2.60. The dominant source of ambient illumination was back scattered light from the projector. With the CRT, D-65 fluorescents were used to backlight a translucent photographer's backdrop. This provided a surround for the CRT having a brightness which was 10 - 15% of the peak brightness of the screen. Viewers were provided with LED white light pens for scoring the image clips. The resolution of the DMD display was measured using the contrast modulation method described in a report on display characterization for the previous MPEG Digital Cinema Tests [5]. Measurement of the contrast modulation requires finding the contrast for a number of Grille patterns [Fig. 2]. Using this method, the NIST DMD-device is determined to

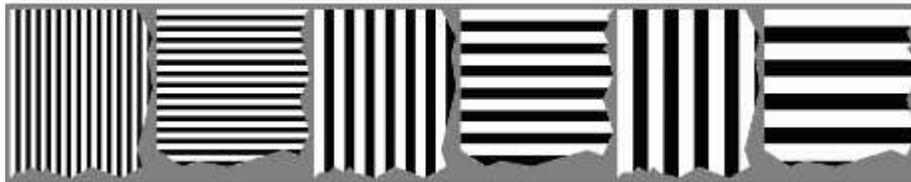


Figure 2: Patches of N-by-N pixel Grille Patterns of Alternating Black and White Stripes.

have full resolution, that is, the measured resolution is equal to the nominal resolution. The narrow line width of the CRT prevented us from determining its resolution.

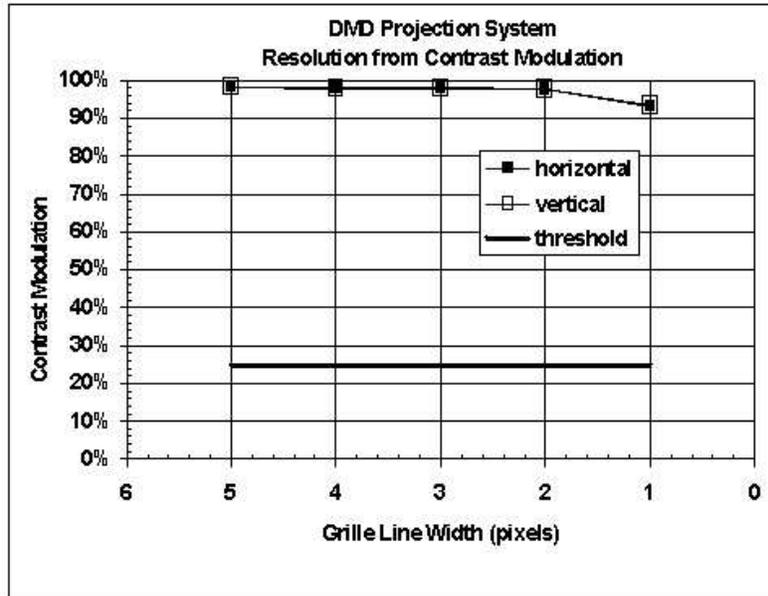


Figure 3: Resolution determined from Contrast Modulation. The contrast is measured on several grille patterns. The system is deemed to resolve patterns for which the contrast modulation exceeds the 25% threshold value represented by the horizontal line.

To accommodate the differing set ups, the full HD tests were broken into 3 sessions. The sessions had 3 or fewer viewers with the exception of one with 4. Typically, viewers who were NIST employees participated for about ½ hour on each of 3 days. Visitors to NIST spent about 4 hours at the Lab, with one-hour rests between ½ hour sessions. Data collection was paper-based using a 100-point (0 to 100) quality scale.

6. STATISTICAL METHODS FOR ANALYSIS - HD DATA

Data was produced at three test sites: multimedia definition at ISCTI, standard definition at the Technical University of Munich, and high definition at NIST. The data were used to obtain the mean opinion scores (MOS), the sample standard deviation (S) and the 95% Confidence Interval (CI). In comparing codecs at various bit rates, a lack of overlap of the 95% CI can provide a strong indication of a significant difference (from the statistical point of view) between MOS values. Where CI values do overlap, the MOS values are considered equivalent (that is the null hypothesis of no difference is accepted) even if the numerical values of the MOS are different. The sample graph (Fig. 4) in this document is based on these computations and assumptions.

This statistical analysis is an adaptation of classical parametric methods. For some time, it has been noted by one of the authors (V.B.) that the classical estimates for a 95% CI are too stringent for the data sets encountered in video quality measurements. For the AVC/H.264-VT the 95% CIs are set as the interval [MOV – S, MOV + S]. This interval is justified by experience with repeated testing.

This observation does not necessarily conflict with statistical theory. The classical analysis depends on satisfying an assumption that the data are normally distributed. Yet, the assumption may fail. In fact, deviations from normality are to be expected. The typical numerical measurement scale used for subjective assessment constrains the scores to lie in a finite range. At the extremes of the scale (corresponding to very high and very low quality) the distribution of measured scores tends to be quite skewed.

Using the subjective test score data from the HD portions of the AVC/H.264-VT, we applied a Kruskal-Wallis non-parametric statistical test [6] to determine a 95% confidence statement for the difference between the subjective scores

for video compressed with AVC/H.264 and with one of the two MPEG-2 codecs at various bit rates. The differences between the populations were compared to the confidence statements for the difference of population means, as determined using a “1-sigma” CI [5]. The HD portion of the AVC/H.264-VT required 30 comparisons to determine which, if any, of the three encoders was significantly better than the others and the bit rates at which they became equivalent. In each of the 30 cases considered, inferences based on the “1-sigma” 95% confidence interval agreed with those based on the non-parametric analysis.

7. DISCUSSION AND SUMMARY

New methods for determining the quality of digital moving pictures in multimedia and on very high quality picture systems are described here in sufficient detail that they can be produced in a well-equipped video test facility. The new methods are developed to address the testing needs for progressive displays in new viewing environments for multimedia and high definition video. These methods have been employed in several rounds of video quality tests. The recent AVC/H.264-VT provides a rich opportunity to judge the effectiveness of these methods.

In the original conception, the AVC/H.264-VT was intended to provide viewing data to confirm (or deny) the evidence of signal-to-noise (PSNR) measurements, that suggested a 2:1 gain in compression efficiency was possible when AVC/H.264 was compared to MPEG-2 and MPEG-4 codecs. As seen in Fig. 4, the results appear to confirm such gains in over 60% of the test conditions.

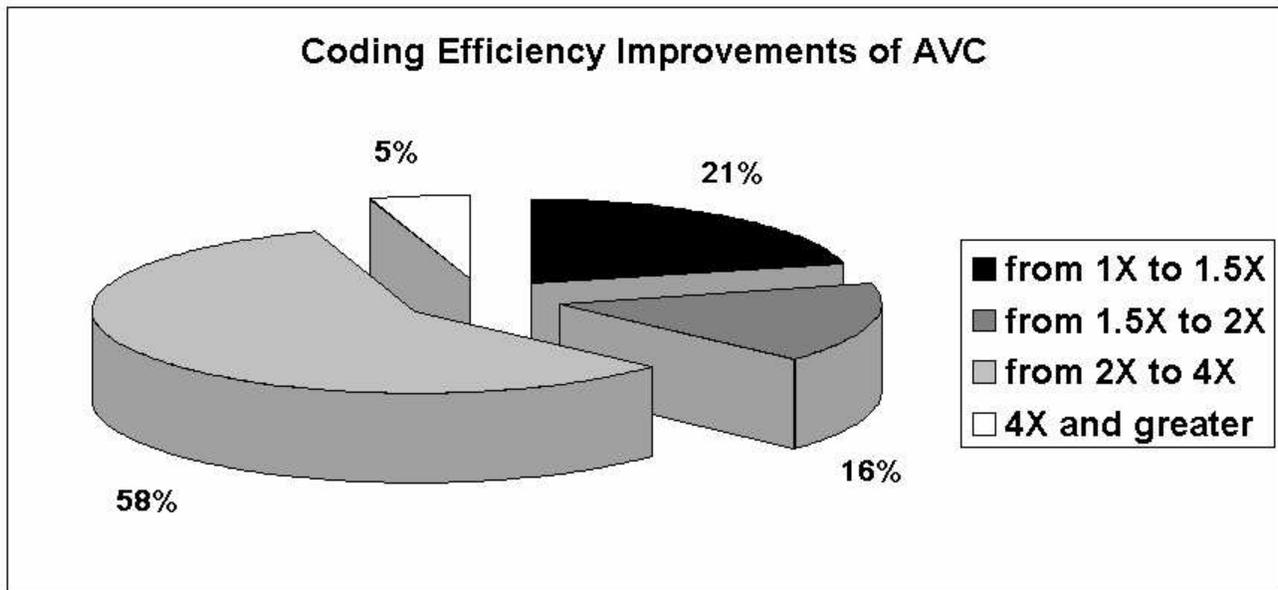


Figure 4: Coding Efficiency Gains in AVC/H.264 relative to MPEG-2 and MPEG-4. The percentage is based on a count of conditions of test clip and comparison codec (both MPEG-2 and MPEG-4). The percent is the fraction of the statistically conclusive test conditions for which the indicated improvement was measured.

The question remains, does the new methodology provide reliable, reproducible measures of MM quality? The AVC/H.264 test provides a partial answer to this question. This answer is based on two observations. First, we note that in all three testing modes (that is for MM, SD, and HD imagery) the subjective measurements confirm the PSNR measurements. In fact, with the new methodology, the multimedia tests measured the most dramatic gains in compression efficiency.

- When compared to MPEG-4 SP, AVC/H.264 Baseline Profile achieved a coding efficiency improvement of 2 times or greater in 14 out of 18 (78%) statistically conclusive cases.

- When compared to MPEG-4 ASP, AVC/H.264 Main Profile achieved a coding efficiency improvement of 2 times or greater in 18 out of 25 (68%) statistically conclusive cases.

These rates compare to about 70% for the SD tests and about 44% for the HD tests.

The second observation is that although the structure of the HD tests (particularly, the compression bit rates used) makes a detailed comparison difficult, the consistency of the MM tests results with those for SD suggests that the new methodology is similar in its power to quantify the quality differences between compressed and uncompressed clips.

The MM methods make use of readily available technology in the form of high performance personal computers with LCD monitors. The testing adaptation to this new platform appears to be successful for Multimedia as described in Section 2. This success may depend on the low frame rate (no higher than 15 F/s) of the QCIF and CIF materials. In particular, the use of LCD displays for the faster frame rates of SD and HD video cannot be taken for granted. Motion rendition quality in contemporary LCD display technology is currently the focus of manufacturing improvement efforts [7]. Execution of higher frame rate testing on LCD platforms may require rigorous validation of test quality.

In the case of HD tests, the results do not provide strong support for the effectiveness of the DMD test platform for subjective testing; neither do they indicate a failure of the test methodology. Confirming the expected doubling of compression efficiency was made difficult by the small number and limited range of bit rates (BRs 20, 10, and 6 Mb/s) spanning fewer than 2 octaves (BR doublings) excepting 1080I video which had only two BRs. This contrasts with the SD testing which had 5 BRs ranging over two octaves and MD which had 4 BRs over three octaves.

ACKNOWLEDGEMENTS

The authors are leaders of the MPEG Test Group. The results reported here are based on the contributions of a large team of investigators, whose assistance has made this work possible. We particularly wish to cite Angelo Ciavardini, Giancarlo Gaudino; Florian Obermeier, Michael Pilz, Testronic Laboratories; Walt Husak, Eric Gsell and Tom McMahon; and John Roberts, Stefan Leigh and Alan Heckert. In addition, the many companies that contributed to the AVC/H.264 Testing are listed in the formal Test Report [1].

REFERENCES

- [1] "Report of the Formal Verification Tests on IS 14496-10 AVC / H.264", ISO/IEC JTC1/SC29/WG11 MPEG2003/N6231, December 2003, Waikoloa, Hawaii, USA. Available at http://www.chiariglione.org/mpeg/working_documents/mpeg-04/avc/avc_vt.zip
- [2] T. Oelbaum, V. Baroncini, T.K. Tan, C. Fenimore, *Subjective quality assessment of the emerging AVC/H.264 video coding standard*, to appear Proceedings 2004 International Broadcast Conference, Amsterdam, Netherlands [2004].
- [3] Recommendation ITU-T BT500.10, *Methodology for the Subjective Assessment of the Quality of Televisions Pictures* [2000].
- [4] ANSI standards, IEC 61947-1 Ed. 1.0 and IEC 61947-2 Ed. 1.0.
- [5] P. A. Boynton and C. Fenimore, *Characterization of Projection Systems for the MPEG-4 Digital Cinema Compression Scheme Evaluation*, NISTIR 6792, July 2001.
- [6] J. Filliben, *DATAPLOT Statistical Software Package* contains the Kruskal-Wallis Test. The software is available for downloading at <http://www.itl.nist.gov/div898/software/dataplot/>.
- [7] J. Ohwada, *Improving the Moving-Image Quality of LCDs by Using Impulse Driving*, Information Display Magazine, June 2004.