INTERNATIONAL ORGANIZATION FOR STANDARDIZATION
ORGANISATION INTERNATIONALE NORMALISATION
ISO/IEC JTC 1/SC 29/WG 11
CODING OF MOVING PICTURES AND AUDIO

ISO/IEC JTC 1/SC 29/WG 11 N2826

July 1999

Source: **Test and Video Group**
Title: **Draft**
Status: **Report Of The Formal Verification Tests On MPEG-4 Coding Efficiency for Low and Medium Bit rates**

SUMMARY

This report illustrates the results of the formal subjective verification test made to evaluate the performance of MPEG-4 Video Coding Efficiency compared with MPEG-1 Video Coding at Low and Medium bit rates.

INTRODUCTION

The visual part of the MPEG-4 standard will provide a toolbox containing tools and algorithms bringing solutions to a number of functionalities and covering a wide range of bit rates.
It was recognised that the verification tests should also address the functionality of coding efficiency compared with other existing standards.
This document describes the test procedures and the results of the coding efficiency test. The test was carried out at the CCETT laboratory.

TEST CONDITIONS

The MPEG-4 Coding Efficiency Test, testing the MPEG-4 VM vs. the MPEG-1 encoder, was performed according to the conditions specified in the Table 4.

MPEG4 VM vs MPEG-1 Coding Efficiency Tests							
	Low bit rate				Medium bit rate		
Sequences	Carphone, Foreman, Coastguard				Dancer, Stefan, Table Tennis		
Resolution	QCIF (176x144)		CIF (352x288)		CIF (352x288)		
Bit rate	40 kbps	64 kbps	128 kbps	256 kbps	384 kbps (440stefan)	512 kbps	768 kbps
Input frame rate	7.5 Hz	7.5 Hz	7.5 Hz	10 Hz	25 Hz	25 Hz	25 Hz
Period of I	1stVOP only	1stVOP only	1stVOP only	1stVOP only	N=24	N=24	N=24
Period of P	M=1	M=1	M=1	M=1	M=1	M=1	M=3
Rate control	TM5	TM5	TM5	TM5	TM5	TM5	TM5

Table 4: **Coding conditions for the Coding Efficiency Test**

In this test only frame based sequences were examined, the content based case is already reported in N2711.
MPEG-1 was used instead of MPEG-2 because for progressive sequences these two are identical, except that MPEG-1 uses less overhead for header information and thus is more efficient. The test uses typical test sequences, encoded with the same rate control for both MPEG-1 and MPEG-4 to compare the coding algorithms without the impact of different rate control schemes.
The tests for the high bit rate case will follow at a later stage.

PARTICIPANTS

Video Encoding	W. Li, G. Heising, U. Benzler	Optivision, HHI, Univ. of Hannover
Video Decoding	U. Benzler	Univ. of Hannover
Preparation of test tapes	V. Baroncini	FUB
Conduct of tests	S. Pefferkorn	CCETT
Statistical analysis	V. Baroncini	FUB

SOURCE MATERIAL (TEST SEQUENCES)

Display format

The decoded video sequences in CIF/QCIF format (352x288 / 176x144 pel) are upsampled to the ITU-R BT.601 frame format (720x576 pel) using an extended version of the mpeg4_filter program by Andreas Hutter, TU Munich (W1552) that is available by anonymous ftp at ftp://ftp.tnt.uni-hannover.de/pub/MPEG/mpeg4-seqs/mpeg4_filter.c. The upsampling is done according to N0322.

TEST METHOD

The "Coding Efficiency" functionality of the MPEG-4 VM compared with MPEG-1 Video Coding was subjectively assessed using the SS (Single Stimulus) test method. The SS test method has been designed in ITU to test video signal in which quality range was broad. In a SS test session the subjects are requested to compare each sequence independently.

Laboratory set-up at CCETT

The subjective assessment environment consists of two different rooms. The first is the «viewing room» where tests are performed, while the second one is the «technical room» containing VTRs and other equipment necessary to perform tests.

The equipment

Test sequences were recorded on a D1 tape and displayed to the observers on a Sony PVM 2054 QM using a D1 Sony VTR controlled by the IQ++ platform including hardware and software and described below. The monitor were calibrated using a Minolta Luminance meter, a Photo Research spectrophotometer and the standard Pluge signal.
Votes were collected using the IQ++ sliders.
Six groups of 3 non-expert observers were participated to the test, organised into 2 sessions. Observers were different for 50 and 60 Hz tests. Prior to the test, subjects were screened for visual acuity by using a Monoyer Optometric Table. Besides, test for normal colour vision were performed using Ishihara's tables. These tables are designed to detect colour blindness or strong colour vision deficiencies; that is to say, to check ability of the observers in discriminating colours. The subject who is not able to pass the test is discarded.
All the viewing conditions were compliant with the ITU-R B.T. Rec. 500.

The software

The software platform, called IQ++, has been developed and commercialised by CCETT in the framework of the RACE/MOSAIC project. It allows processing of the results of subjective evaluations based on the most used methodologies described in ITU-R Rec. B.T. 500. The high level of modularity in the conception of IQ++ allows easy update and build-up of new test methods, which may be an adaptation of existing methods, or the result of brand new research in methodology.
IQ++ is composed of three modules: the 'Test Preparation' module, the 'Test Driver' module and the 'Result Processing' module, all running on Windows'95.

«Test Preparation» module

This module deals with the building-up of a test, from:

- The selected method (if known, otherwise a new one can be created)
- The list of test conditions (algorithms, codecs)
- The list of images or video sequences

From this data the software produces the test, creating in particular the random list of different cells. This list must be compliant with the rules associated to the chosen test method.

This module also produces a «SELO» type file which is able to drive a VTR for the automatic editing of a video tape, according to the test characteristics.
Finally, this programme provides the possibility to manipulate the test frameworks. These frameworks describe the test method and are abstractions of the tests themselves.
Standardised methods are provided in the form of predefined test frameworks, but the user can edit these structures to create his own modifications. This module is capable of managing these new methods in the same way it does the standardised methods. A personal library of methods can easily be built.

«Test Driver» Module

This module allows to drive a test using the information from the previous module.
The test structure, based on key-concepts such as session, presentation and results, gives the user broad flexibility. It is therefore possible to carry out the test several times for different observers, e.g. multi results test. It is also possible to split a test into several sessions and to carry it out, session by session, for different groups of observers, e.g. multi session test, or to repeat the same test for the same observers, e.g. multi-presentation test. The combination of these different functionalities can provide a solution for all possible configurations.
During the test the VTR is set on the position «remote control» and it is fully controlled by the IQ++ software, but the test can also run asynchronously, if the playing system has not a remote control with the appropriate protocol.

«Results Processing» Module

This module deals with the processing of the results. The format of the input data file is compliant with the last developments of the ITU-R Rec. B.T. 500. This obviously means that the output data file format of the previous module «test driver» is also compliant.
The «Results Processing» module allows the processing of raw data and the presentation of the results in the simplest possible manner, notably owing to the use of graphs. This module is very easy to be used because, for example, it is based mainly on contextual menus. It allows different forms of processing to be associated to the same test. Each form of processing is itself composed of a series of operations. Parameters are included in each operation, such as scale range modifications, etc.
From raw data the user is able to choose the «by default» form of processing, based on the type of test, if standardised, or to generate the processing list himself by selecting different operations.
The results are displayed by a piece of external graphic software. The current software uses Microsoft Excel, but it is possible to create a link to other software.

The experimental design of the test

Instructions to the subjects

Before starting the tests the subjects are properly instructed on the task they are supposed to do.
To avoid any possible bias, the instruction are read from a printed paper (see Appendix A).
In this way all subjects receive the same instructions.

The training phase

A subjective assessment test must include a training phase, during which the subjects try a short test session that reproduces the same condition of the real test.
When the training section is finished, the experimenter should check eventual errors and answer to the subjects questions, if any.

Test organisation

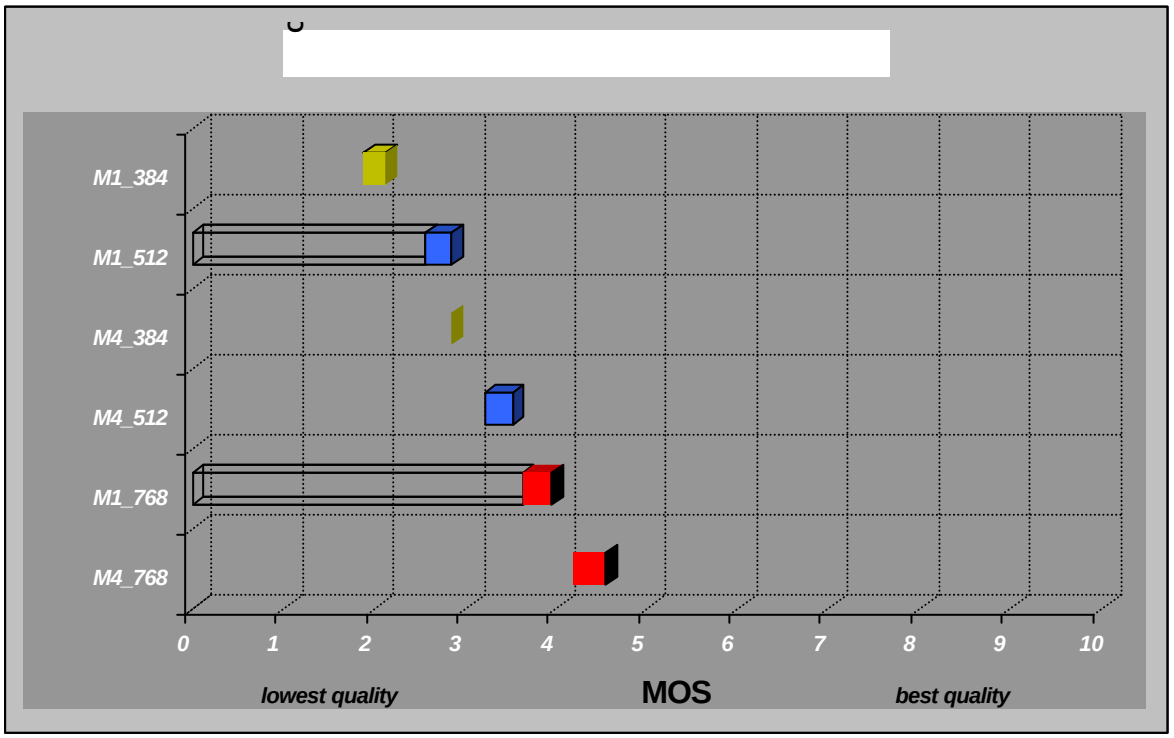
The test has been organised in two different sessions, the Low Bit Rate and the Medium Bit Rate.
Each session was made up of all the possible combinations of sequences and coding conditions; furthermore 5 dummy combinations have been added at the beginning of each session, to allow the stabilisation of the opinions of the subject. The votes collected during the stabilisation phase have been discarded.
To limit the effect of the contextual effect (easily common in a SS test) the sequences under test are presented twice in a session; the subjects are not aware of this situation and repeat their opinion as they see a new set of sequences.

TEST RESULTS

The results are reported in the tables and in the graphs here below. Table 1 and Graph 1 provide the results of the Medium Bit Rate test; table 2 and Graph 2 provide the results of the Low Bit Rate test.

		Dancer	*Table Tennis*	*Stefan*	*GENERAL*	*NSSD*
M4_768	**Mean**	**4,86**	**4,47**	**3,25**	**4,19**	**M1_768**
	C.I.	0,64	0,55	0,53	0,35	
M1_768	**Mean**	**3,97**	**4,11**	**2,81**	**3,63**	**M4_384**
	C.I.	0,58	0,56	0,43	0,32	
M4_512	**Mean**	**4,25**	**3,08**	**2,31**	**3,21**	**M4_384**
	C.I.	0,54	0,44	0,42	0,31	
M4_384	**Mean**	**3,14**	**2,56**	**2,19**	**2,63**	**M1_384**
	C.I.	0,40	0,35	0,35	0,22	
M1_512	**Mean**	**3,00**	**2,97**	**1,72**	**2,56**	**M1_384**
	C.I.	0,48	0,48	0,40	0,29	
M1_384	**Mean**	**2,06**	**2,19**	**1,39**	**1,88**	
	C.I.	0,36	0,44	0,39	0,24	
GENERAL	**Mean**	**3,55**	**3,23**	**2,28**	**3,02**	
	C.I.	0,24	0,22	0,19	0,13	

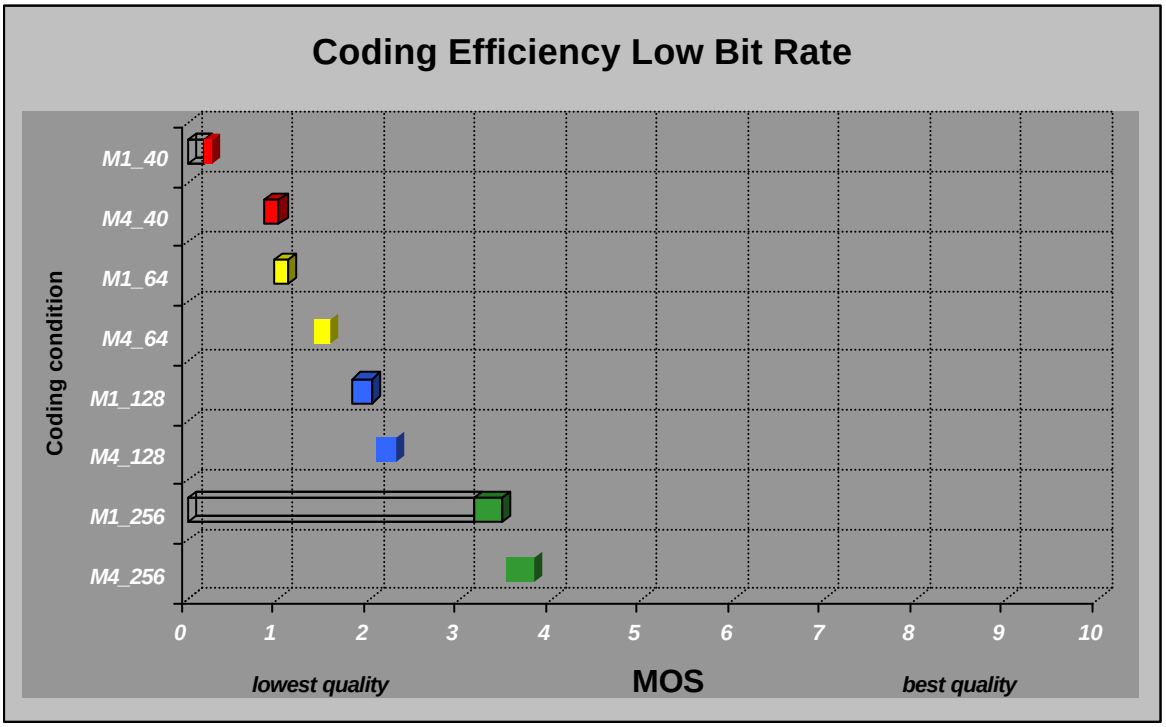
Table 1 – Results of the Medium Bit Rate test



Graph 1 – Graphic representation of the results of the Medium Bit Rate test

		Carphone	*Coastguard*	*Foreman*	*GENERAL*	*NSSD*
M4_256	**Mean**	**4,19**	**3,22**	**3,06**	**3,49**	**M4_128**
	C.I.	0,53	0,49	0,52	0,31	
M1_256	**Mean**	**3,42**	**3,14**	**2,89**	**3,15**	**M4_128**
	C.I.	0,58	0,58	0,43	0,31	
M4_128	**Mean**	**2,72**	**1,75**	**1,72**	**2,06**	**M4_64**
	C.I.	0,42	0,33	0,35	0,23	
M1_128	**Mean**	**1,94**	**1,97**	**1,53**	**1,81**	**M4_64**
	C.I.	0,35	0,42	0,30	0,21	
M4_64	**Mean**	**1,58**	**1,25**	**1,36**	**1,40**	**M1_64**
	C.I.	0,33	0,24	0,28	0,17	
M1_64	**Mean**	**1,19**	**0,97**	**0,67**	**0,94**	**M1_40**
	C.I.	0,28	0,29	0,21	0,15	
M4_40	**Mean**	**0,89**	**1,06**	**0,56**	**0,83**	**M1_40**
	C.I.	0,34	0,30	0,21	0,17	
M1_40	**Mean**	**0,08**	**0,22**	**0,25**	**0,19**	
	C.I.	0,09	0,16	0,14	0,08	
GENERAL	**Mean**	**2,00**	**1,70**	**1,50**	**1,73**	
	C.I.	0,20	0,17	0,16	0,10	

Table 2 – Results of the Low Bit Rate test



Graph 2 – Graphic representation of the results of the Low Bit Rate test

In Graph 1 and Graph 2, the comparison among the bars associated with the same bit rate shows the clear superiority in quality of MPEG-4 compared with MPEG-1.

CONCLUSIONS

The tests of the Coding Efficiency functionality show a clear superiority of MPEG-4 toward MPEG-1 at the medium bit rate coding conditions (384 to 768 kbps) whatever the criticality of the scene.
For the low bit rate case, if a statistical analysis based on the C.I. is applied, there is evidence too of superiority of MPEG-4 towards MPEG-1 (to be deleted).

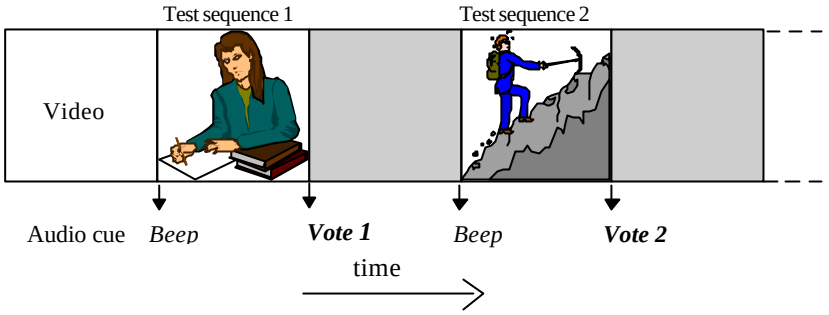
Appendix A: Single Stimulus (SS) Method Instructions

Dear observers,

Thank you for participating in this test.

In the SS tests, a series of test sequences will be displayed on the monitor.

This figure describes what you will see and hear during a SS test session:



Your task is to evaluate the quality of each sequence, by marking one and only one box in the following rating scale:

	☐	10
EXCELLENT	☐	9
	☐	8
GOOD	☐	7
	☐	6
FAIR	☐	5
	☐	4
POOR	☐	3
	☐	2
BAD	☐	1
	☐	0

Your evaluation must reflect your opinion of the global degradation of the whole test sequence. Therefore, vote only after the end of the sequence, and base your evaluation on the entire sequence.

Do not hesitate to rate a sequence either at the top or bottom of the scale, if that is how you believe it should be rated.

A voting form will be distributed before each session. On this form will be a series of rating scales like the one above, one scale for each sequence in the test session. All the scales are numbered. Use scale 1 for the first test sequence, scale 2 for the second one and so on.

During the voting interval for test sequence N, you will hear the audio comment «VOTE N». This will help you know which rating scale to use.

During these tests do not comment on the sequences you have seen or talk with other assessors.

Finally, it is important that you keep your concentration throughout the test session.

Now try this evaluation procedure in a practice session. You will see a series of sequences using the exact same timing as will be used during an actual test session. This will allow you to become familiar with the timing of the test and to practice using the rating scales.

If you have any questions, please ask them now.