Multimedia Group
TEST PLAN

Draft Version 1.21
March 9, 2008

Editors Note: unresolved issues or missing data are annotated by the string <<XXX>>

Contacts:

D. Hands	Tel: +44 (0)1473 648184	Email: david.2.hands@bt.com
K. Brunnstrom	Tel: +46 708 419105	Email: kjell.brunnstrom@acreo.se

Editorial History

Version	Date	Nature of the modification
1.0	July 25, 2001	Initial Draft, edited by H. Myler
1.1	28 January, 2004	Revised First Draft, edited by David Hands
1.2	19 March, 2004	Text revised following VQEG Boulder 2004 meeting, edited by David Hands
1.3	18 June 2004	Text revised during VQEG meeting, Rome 16-18 June 2004
1.4	22October 2004	Text revised during VQEG meeting, Seoul meeting October 18-22, 2004
1.5	18 March 2005	Text revised during MM Ad Hoc Web Meeting, March 10-18, 2005
1.5a	22 April 2005	Text revised to include input from GC, IC and CL
1.5b	29 April 2005	Text revised during VQEG meeting, Scottsdale 25-29 April 2005
1.5e	30 September 2005	Text revised during VQEG meeting, Stockholm 26-30 September 2005
1.6	20 November 2005	Text updated following audio calls held on 12 October 2005 and 2 November 2005.
1.7	29 November 2005	Text updated following audio call held on 29 November 2005.
1.8	8 December 2005	Text updated following audio call held on 8 December 2005.
1.9	22 December 2005	Text updated following final call for comments (comments to be received by 15 December 2005).
1.10	25 January 2006	Text updated following comments received during review period (comments to be received by 5 January 2006).
1.11	14 February 2006	Final core text as agreed by VQEG. Only outstanding work is completing all annexes. Current status of the Annexes is as follows. Annexes I, II, VIII and IX are agreed. Annexes IV, VI and VII where tracked changes remain have yet to be agreed. Text still to be provided for Annexes III and V.
1.12	23 April 2006	Editorial revisions made following audio call of 2 March 2006.
1.13	16 June 2006	Text updated following VQEG meeting, Boston, April 2006.
1.14	7 September 2006	Test updated following audio call of 7 September 2006.
1.15	29 September 2006	Updated based on Tokyo meeting agreements
1.16	7 February 2007	Updated following audio calls held on 17 and 26 January and on 7 February 2007.
1.17	9 May 2007	Updated following the VQEG meeting in Paris, 7-11 May 2007
1.18	18 May 2007	Minor correction of Section 8.4.1 and update text of the experimental test design in.
1.19	19 Sept 2007	Updated following the VQEG meeting in Ottawa, 2008
1.20	5 Dec 2007	Updated following audio call 5 Dec 2007
1.21	9 March 2008	Updated following the VQEG meeting in Kyoto, 2008

Summary

1. Introduction

This document defines the procedure for evaluating the performance of objective perceptual quality models submitted to the Video Quality Experts Group (VQEG) formed from experts of ITU-T Study Groups 9 and 12 and ITU-R Study Group 6. It is based on discussions from various meetings of the VQEG Multimedia working group (MM) recorded in the Editorial History section at the beginning of this document.

The goal of the MM group is to evaluate perceptual quality models suitable for digital video quality measurement in multimedia applications. Multimedia in this context is defined as being of or relating to an application that can combine text, graphics, full-motion video, and sound into an integrated package that is digitally transmitted over a communications channel. Common applications of multimedia that are appropriate to this study include video teleconferencing, video on demand and Internet streaming media. The measurement tools evaluated by the MM group may be used to measure quality both in laboratory conditions using a FR method and in operational conditions using RRNR methods.

In the first stage of testing, it is proposed that video only test conditions will be employed. Subsequent tests will involve audio-video test sequences, and eventually true multimedia material will be evaluated. It should be noted that presently there is a lack of both audio-video and multimedia test material for use in testing. Video sequences used in VQEG Phase I remain the primary source of freely available (open source) test material for use in subjective testing. The VQEG does desire to have copyright free (or at least free for research purposes) material for testing. The capability of the group to perform adequate audio-video and multimedia testing is dependent on access to a bank of potential test sequences.

The performance of objective models will be based on the comparison of the MOS obtained from controlled subjective tests and the MOSp predicted by the submitted models. This testplan defines the test method or methods, selection of test material and conditions, and evaluation metrics to examine the predictive performance of competing objective multimedia quality models.

The goal of the testing is to examine the performance of proposed video quality metrics across representative transmission and display conditions. To this end, the tests will enable assessment of models for mobile/PDA and broadband communications services. It is considered that FR-TV and RRNR-TV VQEG testing will adequately address the higher quality range (4 Mbit/s and above) delivered to a standard definition monitor. Thus, the Recommendation(s) resulting from the VQEG MM testing will be deemed appropriate for services delivered at 4 Mbit/s or less presented on mobile/PDA and computer desktop monitors.

It is expected that subjective tests will be performed separately for different display conditions (e.g. one specific test for mobile/PDA; another test for desktop computer monitor). The performance of submitted models will be evaluated for each type of display condition. Therefore it may be possible for one model to be recommended for one display type (e.g., mobile) and another model for another display format (e.g., desktop monitor).

The objective models will be tested using a set of digital video sequences selected by the VQEG MM group. The test sequences will be processed through a number of hypothetical reference circuits (HRCs). The quality predictions of the submitted models will be compared with subjective ratings from human viewers of the test sequences as defined by this testplan.

A final report will be produced after the analysis of test results.

2. List of Definitions

Intended frame rate is defined as the number of video frames per second physically stored for some representation of a video sequence. The intended frame rate may be constant or may change with time. Two examples of *constant* intended frame rates are a BetacamSP tape containing 25 fps and a VQEG FR-TV Phase I compliant 625-line YUV file containing 25 fps; these both have an absolute frame rate of 25 fps. One example of a *variable* absolute frame rate is a computer file containing only new frames; in this case the intended frame rate exactly matches the effective frame rate. The content of video frames is not considered when determining intended frame rate.

Anomalous frame repetition is defined as an event where the HRC outputs a single frame repeatedly in response to an unusual or out of the ordinary event. Anomalous frame repetition includes but is not limited to the following types of events: an error in the transmission channel, a change in the delay through the transmission channel, limited computer resources impacting the decoder's performance, and limited computer resources impacting the display of the video signal.

Constant frame skipping is defined as an event where the HRC outputs frames with updated content at an effective frame rate that is fixed and less than the source frame rate.

Effective frame rate is defined as the number of unique frames (i.e., total frames – repeated frames) per second.

Frame rate is the number of (progressive) frames displayed per second (fps).

Live Network Conditions are defined as errors imposed upon the digital video bit stream as a result of live network conditions. Examples of error sources include packet loss due to heavy network traffic, increased delay due to transmission route changes, multi-path on a broadcast signal, and fingerprints on a DVD. Live network conditions tend to be unpredictable and unrepeatable.

Pausing with skipping (formerly frame skipping) is defined as events where the video pauses for some period of time and then restarts with some loss of video information. In pausing with skipping, the temporal delay through the system will vary about an average system delay, sometimes increasing and sometimes decreasing. One example of pausing with skipping is a pair of IP Videophones, where heavy network traffic causes the IP Videophone display to freeze briefly; when the IP Videophone display continues, some content has been lost. Another example is a videoconferencing system that performs constant frame skipping or variable frame skipping. Constant frame skipping and variable frame skipping are subsets of pausing with skipping. A processed video sequence containing pausing with skipping will be approximately the same duration as the associated original video sequence.

Pausing without skipping (formerly frame freeze) is defined as any event where the video pauses for some period of time and then restarts without losing any video information. Hence, the temporal delay through the system must increase. One example of pausing without skipping is a computer simultaneously downloading and playing an AVI file, where heavy network traffic causes the player to pause briefly and then continue playing. A processed video sequence containing pausing without skipping events will always be longer in duration than the associated original video sequence.

Refresh rate is defined as the rate at which the computer monitor is updated.

Simulated transmission errors are defined as errors imposed upon the digital video bit stream in a highly controlled environment. Examples include simulated packet loss rates and simulated bit errors. Parameters used to control simulated transmission errors are well defined.

Source frame rate (SFR) is the intended frame rate of the original source video sequences. The source frame rate is constant. For the MM testplan the SFR may be either 25 fps or 30 fps.

Transmission errors are defined as any error imposed on the video transmission. Example types of errors include simulated transmission errors and live network conditions.

Variable frame skipping is defined as an event where the HRC outputs frames with updated content at an effective frame rate that changes with time. The temporal delay through the system will increase and decrease with time, varying about an average system delay. A processed video sequence containing variable frame skipping will be approximately the same duration as the associated original video sequence.

3. List of Acronyms

ACR-HRR	Absolute Category Rating with Hidden Reference Removal
ANOVA	ANalysis Of VAriance
ASCII	ANSI Standard Code for Information Interchange
CCIR	Comite Consultatif International des Radiocommunications
CIF	Common Intermediate Format (352 x 288 pixels)
CODEC	COder-DECoder
CRC	Communications Research Centre (Canada)
DVB-C	Digital Video Broadcasting-Cable
DMOS	Difference Mean Opinion Score
FR	Full Reference
GOP	Group Of Pictures
HRC	Hypothetical Reference Circuit
ILG	Independent Laboratory Group
ITU	International Telecommunication Union
LSB	Least Significant Bit
MM	MultiMedia
MOS	Mean Opinion Score
MOSp	Mean Opinion Score, predicted
MPEG	Moving Picture Experts Group
NR	No (or Zero) Reference
NTSC	National Television Standard Code (60 Hz TV)
PAL	Phase Alternating Line standard (50 Hz TV)
PS	Program Segment
PVS	Processed Video Sequence
QAM	Quadrature Amplitude Modulation
QCIF	Quarter Common Intermediate Format (176 x 144 pixels)
QPSK	Quadrature Phase Shift Keying
VQR	Video Quality Rating (as predicted by an objective model)
RR	Reduced Reference
SMPTE	Society of Motion Picture and Television Engineers
SRC	Source Reference Channel or Circuit
VGA	Video Graphics Array (640 x 480 pixels)
VQEG	Video Quality Experts Group
VTR	Video Tape Recorder

4. Subjective Evaluation Procedure

4.1. The ACR Method with Hidden Reference Removal

This section describes the test method according to which the VQEG multimedia (MM) subjective tests will be performed. We will use the absolute category scale (ACR) [Rec. P.910] for collecting subjective judgments of video samples. ACR is a single-stimulus method in which a processed video segment is presented alone, without being paired with its unprocessed ("reference") version. The present test procedure includes a reference version of each video segment, not as part of a pair, but as a freestanding stimulus for rating like any other. During the data analysis the ACR scores will be subtracted from the corresponding reference scores to obtain a DMOS. This procedure is known as "hidden reference removal."

4.1.1. General Description

The selected test methodology is the single stimulus Absolute Category Rating method with hidden reference removal (henceforth referred to as ACR-HRR). This choice has been selected due to the fact that ACR provides a reliable and standardized method (ITU-R Rec. 500-11, ITU-T P.910) that allows a large number of test conditions to be assessed in any single test session.

In the ACR test method, each test condition is presented singly for subjective assessment. The test presentation order is randomized according to standard procedures (e.g. Latin or Graeco-Latin square, or via random number generator). The test format is shown in Figure 1. At the end of each test presentation, human judges ("subjects") provide a quality rating using the ACR rating scale below. Note that the numerical values attached to each category are only used for data analysis and are not shown to subjects.

5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

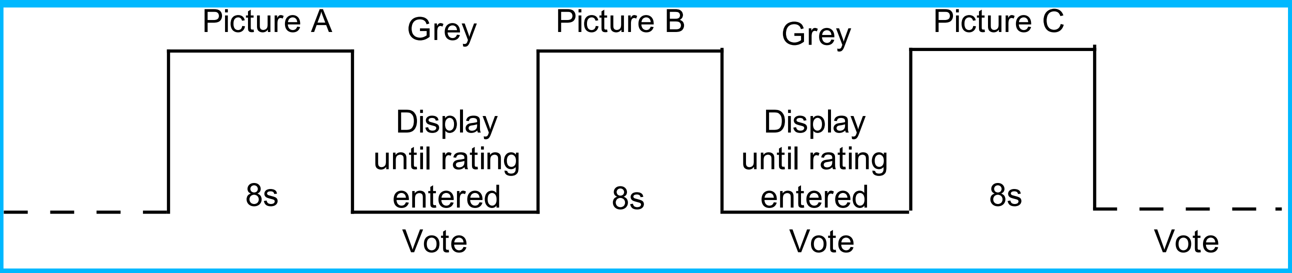


Figure 1 – ACR basic test cell.

The length of the SRC and PVS should be 8 s.

Instructions to the subjects provide a more detailed description of the ACR procedure. The instruction script appears in Annex I.

4.1.2. Application across Different Video Formats and Displays

The proposed MM test will examine the performance of objective perceptual quality models for different video formats (VGA, CIF and QCIF). Section 4.1.3 defines format and display types in detail. Video applications targeted in this test include internet video, mobile video, video telephony, and streaming video.

Presently, VQEG MM assumes a rolling programme of tests. The audio-video tests are expected to involve three separate stages. It is expected that Stage 1 will assess video quality only; the current Test Plan covers Stage 1. It is expected that Stage 2 will assess audio quality only. It is expected that Stage 3 will assess overall audio-video quality.

The test instructions request subjects to maintain a specified viewing distance from the display device. The viewing distance has been agreed as:

- QCIF: nominally 6-10 picture heights (H), and let the viewer choose within physical limits (natural for PDAs).
- CIF: 6-8H and let the viewer choose within physical limits.
- VGA: 4-6H and let the viewer choose within physical limits

H=Picture Heights (picture is defined as the size of the video window)

We note regarding the Stage 2 and Stage 3 audio and audio-video tests, that the room must be acoustically isolated and conform to relevant international standards (e.g. ITU-T Rec. P.800. and ITU-R Rec. BS.1116). Use of headphones will be investigated and perhaps included or mandated in the test (e.g., Stax diffused field equalized Headphones). The specification and selection of audio cards is to be decided.

4.1.3. Display Specification and Set-up

Given that the subjective tests will use LCD displays, it is necessary to ensure that each test laboratory selects appropriate display specification and common set-up techniques are employed. This Test Plan requires that LCD displays meet the following specifications:

Monitor Feature	Specification
Diagonal Size	17-24 inches
Dot pitch	< 0.30
Resolution	Native resolution (no scaling allowed)
Gray to Gray Response Time (if specified by manufacturer, otherwise assume response time reported is white-black)	< 30 ms (<10 ms if based on white-black)
Color Temperature	6500K
Calibration	Yes
Calibration Method	Eye One / Video Essentials DVD
Bit Depth	8 bits/colour
Refresh Rate	>= 60 Hz
Standalone/laptop	Standalone
Label	TCO ´03 or TCO ‘06 (TCO ’06 preferred)

The LCD shall be set-up using the following procedure:

- Use the autosetting to set the default values for luminance, contrast and colour shade of white.
- Adjust the brightness according to Rec. ITU-T P.910, but do not adjust the contrast (it might change balance of the colour temperature).
- Set the gamma to 2.2.
- Set the colour temperature to 6500 K (default value on most LCDs).

The scan rate of the PC monitor must be at least 60 Hz.

The LCD display shall be a high-quality monitor. It is preferred that all subjective tests use the same LCD monitor panel. This will facilitate data analysis using data from different tests. Annex V contains a list of preferred LCD monitors for use in the subjective tests.

Video sequences will be displayed using a black border frame (0) on a grey background (128). The black border frame will be of the following size:

36 lines/pixels VGA

18 lines/pixels CIF

9 lines/pixels QCIF

[Ed. Note: The size of the black border may change after some initial trials of the software change. This will not require 2/3 to change size of black border.]

The black border frame will be on all four sides.

4.1.4. Test Method

All subjective tests will be run using the same software package. The software package will include the following components:

- Entry system for subject details (e.g. name, age, gender)
- Test screens (prompts to users, grey panel, ACR scale, response input, data capture, data storage)
- Timing control
- Correct video play-out check
- Video player

Annex V describes the test method to be used in the VQEG Multimedia testing. Annex V also provides minimum computer specifications (including required OS) required when using this subjective test software package.

4.1.5. Subjects

Different subjective experiments will be conducted by several test laboratories. Exactly 24 valid viewers per experiment will be used for data analysis. A valid viewer means a viewer whose ratings are accepted after post-experiment results screening. Post-experiment results screening is necessary to discard viewers who are suspected to have voted randomly. The rejection criteria verify the level of consistency of the scores of one viewer according to the mean score of all observers over the entire experiment. The method for post-experiment results screening is described in Annex VI. Only scores from valid viewers will be reported in the results spreadsheets as described in Section 4.2[1].

The following procedure is suggested to obtain ratings for 24 valid observers:

1. Conduct the experiment with 24 viewers
2. Apply post-experiment screening to eventually discard viewers who are suspected to have voted randomly
3. If n viewers are rejected, run n additional subjects.
4. Go back to step 2 and step 3 until valid results for 24 viewers are obtained

It is preferred that each viewer be given a different randomized order of video sequences where possible. Otherwise, the viewers will be assigned to sub-groups, which will see the test sessions in different randomized orders. A maximum of 4 viewers may be presented with the same ordering of test sequences per subjective test.

Each viewer can only participate in 1 experiment (i.e. one experiment at one image resolution).

Only non-expert viewers will participate. The term non-expert is used in the sense that the viewers' work does not involve video picture quality and they are not experienced assessors. They must not have participated in a subjective quality test over a period of six months.

Prior to a session, the observers should usually be screened for normal visual acuity or corrected-to-normal acuity and for normal color vision. Acuity will be checked according to the method specified in ITU-T P.910 or ITU-R Rec. 500, which is as follows. Concerning acuity, no errors on the 20/30 line of a standard eye chart [I.1] should be made. The chart should be scaled for the test viewing distance and the acuity test performed at the same location where the video images will be viewed (i.e. lean the eye chart up against the monitor) and have the subjects seated. Ishihara or Pseudo Isochromatic plates may be used for colour screening. When using either colour test please refer to usage guidelines when determining whether subjects have passed (e.g. standard definition of normal colour vision in the Ishihara test is considered to be 17 plates correct out of a 38 plate test; ITU-T Rec. P.910 states that no more than 2 plates may be failed in a 12 plate test.

[I.1] Grahm-Field Catalogue Number 13-1240.

4.1.6. Viewing Conditions

Each test session will involve only one subject per display assessing the test material. Subjects will be seated directly in line with the center of the video display at a specified viewing distance (see Section 4.1.2). The test cabinet will conform to ITU-T Rec. P.910 requirements.

[1] Test laboratories can keep data from invalid viewers if they consider this to be of valuable information to them but they must not include them in the VQEG data.

4.1.7. Experiment design

Each subjective experiment will include the same number of 166 PVSs[2]. The 166 PVSs include both the common set of PVSs inserted in each experiment and the hidden reference (hidden SRCs) sequences, i.e. each hidden SRC is one PVS. The common set of PVSs will include the secret PVSs and secret source.

The randomization will be applied across the 166 PVSs. The 166 PVSs can then be split into 2 sessions of 83 PVSs each. In this scenario, an experiment will include the following steps:

1. Introduction and instructions to viewer
2. Practice clips: these test clips allow the viewer to familiarize with the assessment procedure and software. They must represent the range of distortions in the experiment. A number of 6 practice clips. Each of the practice clip will come from a different test. Ratings given to practice clips are not used for data analysis.
3. Assessment of 83 PVSs
4. Short break
5. Practice clips (this step is optional but advised to regain viewer's concentration after the break)
6. Assessment of 83 PVSs

A full matrix approach will be applied for each experiment: each SRC will be processed through each HRC. The test design should be full matrix of 8 by 17 SRC by HRC combinations. In addition to this the ILG will add 30 common PVSs (6 SRCs and 5 HRCs one of which is the hidden reference).

The SRCS used in each experiment must cover a variety of content categories as defined in Section 6.2. At least 6 categories of content must be included in each experiment.

A similar number of PVSs from each type of error will be tested per image resolution. The image resolutions are defined in Section 4.1.2. The different types of error conditions are defined in Section 6.1.3. However different types of error conditions can be mixed between experiments to ensure a balance in the design of each individual experiment [Ed. Note: if one experiment only includes transmission errors, it will be very difficult to obtain a distribution of MOS across the entire voting scale].

4.1.8. Randomization

For each subjective test, a randomization process will be used to generate orders of presentation (playlists) of video sequences. Playlists can be pre-generated offline (e.g. using separate piece of code or software) or generated by the subjective test software itself. As stated in section 4.1.4, it is preferred that each subject be given a different randomized order of video sequences where possible. Otherwise, the viewers will be assigned to sub-groups, which will see the test sessions in different randomized orders. A maximum of 4 subjects may be presented with the same ordering of test sequences per subjective test.

In generating random presentation order playlists the same scene content may not be presented in two successive trials.

Randomization refers to a random permutation of the set of PVSs used in that test. Shifting is not permitted, e.g.
Subject1 = [PVS4 PVS2 PVS1 PVS3]
Subject2 = [PVS2 PVS1 PVS3 PVS4]
Subject3 = [PVS1 PVS3 PVS4 PVS2]
…

If a random number generator is used (as stated in section 4.1.1), it is necessary to use a different starting seed for different tests.

[2] This will allow conducting an ACR experiment within 1 hour, including practice clips and a comfortable break during the experiment.

Example script in Matlab that performs playlists (i.e. randomized orders of presentation) is given below:

```
rand('state',sum(100*clock));  % generates a random starting seed
Npvs=200; % number of PVSs in the test
Nsubj=24; % number of subjects in the test
playlists=zeros(Npvs,Nsubj);
for i=1:Nsubj
        playlists(:,i)=randperm(Npvs);
end
```

4.1.9. Test Data Collection

The responsibity for the collection and organization of the data files containing the votes will be shared by the ILG Co-Chairs and the proponents. The collection of data will be supervised by the ILG and distributed to test participants for verification.

4.2. Data Format

4.2.1. Results Data Format

The following format is designed to facilitate data analysis of the subjective data results file.

The subjective data will be stored in a Microsoft Excel spreadsheet containing the following columns in the following order: lab, test, type, subject #, month, day, year, session, resolution, rate, age, gender, order, scene, HRC, ACR Score. Missing data values will be indicated by the value -9999 to facilitate global search and replacement of missing values. Each Excel spreadsheet cell will contain either a number or a name. All names (e.g., test, lab, scene, hrc) must be ASCI strings containing no white space (e.g., space, tab) and no capital letters. Where exact text strings are to be used, the text strings will be identified below in single quotes (e.g., 'original'). Only data from valid viewers (i.e., viewers who pass the visual acuity and color tests) will be forwarded to the ILG and other proponents.

Below are definitions for the Excel spreadsheet columns:

Lab: Name of laboratory's organization (e.g., CRC, Intel, NTIA, NTT, etc.). This abbreviation must be a single word with no white space (e.g., space, tab).

Test: Name of the test. Each test must have a unique name.

Type: Name of the test category. [Note: exact text strings will be specified after individual test categories have been finalized.]

Distance: Viewing distance (e.g. 6to10H, 6to8H, 4to6H).

Subject #: Integer indicating the subject number. Each laboratory will start numbering viewers at a different point, to ensure that all viewers receive unique numbering. Starting points will be separated by 1000 (e.g., lab1 starts numbering at 1000, lab2 starts numbering at 2000, etc). Subjects' names will *not* be collected or recorded.

Month: Integer indicating month [1..12]

Day: Integer indicating day [1..31]

Year: Integer indicating year [2004..2006]

Session: Integer indicating viewing session

Resolution: One of the following three strings: 'vga', 'cif' or 'qcif'.

Rate: A number indicating the frames per second (fps) of the original video sequence.

Age: Integer number that indicates the subject's age.

Gender: 'f' for female, 'm' for male

Order: An integer indicating the order in which the subject viewed the video sequences [or trial number, if scenes are ordered randomly].

Scene: Name of the scene. All scenes from all tests must have unique names. If a single scene is used in multiple tests (i.e., digitally identical files), then the same scene name must be used. Names shall be eight characters or fewer.

HRC: Name of the HRC. For reference video sequences, the exact text 'reference' must be used. All processed HRCs from all tests must have unique names. If a single HRC is used in multiple tests, then the same HRC name must be used. HRC names shall be eight characters or fewer.

ACR Score: Integer indicating the subject's ACR score (1, 2, 3, 4, or 5).

See Annex II for an example.

4.2.2. Subjective Data Analysis

Difference scores will be calculated for each processed video sequence (PVS). A PVS is defined as a SRCxHRC combination. The difference scores, known as Difference Mean Opinion Scores (DMOS) will be produced for each PVS by subtracting the score from that of the hidden reference score for the SRC used to produce the PVS. Subtraction will be done per subject. Difference scores will be used to assess the performance of each full reference and reduced reference proponent model, applying the metrics defined in Section 8.

For evaluation of no-reference proponent models, the absolute (raw) subjective score will be used. Thus, for each test sequence, only the absolute rating for the SRC and PVS will be calculated. Based on each subject's absolute rating for the test presentations, an absolute mean opinion score will be produced for each test condition. These MOS will then be used to evaluate the performance of NR proponent models using the metrics specified in Section 8.

5. Test Laboratories and Schedule

Given the scope of the MM testing, both independent test laboratories and proponent laboratories will be given subjective test responsibilities. All laboratories will report to VQEG (MMTEST Reflector) the test environment they plan to use prior to conducting the subjective test.

5.1. Independent Laboratory Group (ILG)

The independent laboratory group is composed of IRCCyN (France), T&W (Italy), FT (France), CRC (Canada), INTEL (USA), NTIA (USA), Nortel (Canada), Acreo (Sweden), Ericsson (Sweden) and Verizon (USA).

5.2. Proponent Laboratories

A number of proponents also have significant expertise in and facilities for subjective quality testing. Proponents indicating a willingness to participate as test laboratories are Genista, NTT, Opticom, SwissQual, Psytechnics, KDDI, and Yonsei. It is clearly important to ensure all test data is derived in accordance with this testplan. Critically, proponent testing must be free from charges of advantage to one of their models or disadvantage to competing models.

The maximum number of subjective experiments run by any one proponent laboratory is 3 times the lowest non-zero number run by any other proponent laboratory, per image size.

The maximum number of non-secret PVSs included in overall test by any single proponent laboratory is 30%. The ILG will assign complete tests for each proponent to run, which will ideally not include PVSs generated by the proponent.

See Annex IV for details on fees and conditions for proponents participating in the VQEG Multimedia tests.

5.3. Test procedure and schedule

1.1.

1. Approval of test plan [14 February 2006]

2. Declaration of intent to participate and the number of models to submit [6 weeks prior to model submission date]

3. VQEG compiles a list of HRCs that are of interest to the MM test. Proponents will send details of proposed HRCs and indicate which ones they can create to the points of contacts and example PVSs (HRC point of contacts Quan Huynh Thu and Philip Corriveau). [24 March 2006]

4. Input due from all ILG and Proponents regarding general NDA (Confidentiality agreement) [November 1, 2006]

5. All General NDAs signed. [December 1, 2006].

6. Each test lab will produce a test design for each subjective test they plan to perforom. This test design should be sent to Arthur W for initial review to ensure no duplication. [1 June 2007]. Those MM proponents that have not submitted an MM Subjective Test Design to Arthur Webster (webster@its.bldrdoc.gov) by June

1, 2007 will not be allowed to run any subjective experiments and will not be eligible to have access to source material or processed sequences. This will require them to pay an ILG to design and run a subjective test for them in order to have access to the processed sequences and MM test subjective data.

7. A full and final review of test plans will be performed by the ILG and proponent labs. [16 June 2007]

8. Proponents informed by ILG to whom fee payment is made. [11 May 2007].

9. Content license agreements distributed to proponents and ILG. [31 May 2006]

10. Each proponent must have signed (and content provider received) all NDAs by October 10, 2006. No guarantees of a three month review of the video source will be possible if NDAs are not received by content providers by October 10, 2006.

11. The proposed lists of HRCs for each experiment are examined by VQEG for problems (e.g., one organization creating too many HRCs, overlap between experiments, using NTT guidelines). [9 June 2006]

12. ILG send invoice to proponent. [1 July 2007]

13. All source video sent to content point of contact (Chulhee Lee) [29 September 2006]

14. Source video sent by POC to most ILG. [October 15, 2006]

15. ILG/VQEG will select source video pool files and distribute source pool to regional points of contact (Asia: Chulhee Lee, Europe: Kjell Brunnstrom, North America: Filippo Speranza) [October 15, 2006]

16. Participating organizations obtain copies of source pool, except for the secret SRC. The requesting organization has to pay any costs involved. [October 25, 2006]

17. Receiving organization must send acknowledgement to MM reflector that they have received the source pool files. When all proponents have acknowledged to the MM reflector that they have received all source pool, there will be a 3 month period until the submission of models. Secret content may still be collected by ILG. Proponents are not allowed to provide secret content. [November 1, 2006]

18. Fee payment if applicable. Payment will be made directly from each proponent to the selected testing facility, according to a table agreed on by ILG and distributed to the proponents [1 Aug 2007]

19. By 31 January, 2007, a list of the ILG labs available for model submission & model checking will be made available on the MM reflector (mmtest@its.bldrdoc.gov) and MM proponents (MMProponents@opticom.de). [31 January 2007]

20. Proponents submit their models (executable and, only if desired, encrypted source code). Models should be submitted to Kjell Brunnstrom (Kjell.Brunnstrom@acreo.se) and Filippo Speranza (filippo.speranza@crc.ca). Procedures for making changes after submission will be outlined in a separate document (see Annex VII on storing encrypted version of submitted source code).

All proponents must submit the first version all models by 9 February. The ILG will validate that each submitted model runs on their computer, by running the model on the test vectors produced by NTIA, and showing that the model outputs the MOS score expected by the proponent. If necessary, a different ILG may be asked to validate the proponent's model (e.g., if another ILG has a computer that may have an easier time running the model.) Each ILG will try to validate the first submitted version of a model within one week of model submission. [9 February]

All proponents have the option of submitting updated models, between 9 February and 21 February. Such model updates may be either:

(1) Intended to make the model run on the ILG's computer.

(2) Model improvements, intended to replace the previous model submitted. Such improved m odels will be checked as time permits.

If the replacement model runs on the ILG computer, it will replace the previous submission. If the replacement model is not able to run on the ILG computer within one week, the previous submission will be used. ILG checks on models may exceed the model submission deadline. ILG request that proponents try to limit this to one replacement model, so that the ILG are not asked to validate an excessive number of models.

Model Submission Deadline for all proponents and all models is 21 February. Models received after 21 February will not be evaluated in the MM test, no matter what the reason for the late submission. [21 February]

21. NDAs for secret SRCs distributed and signed by proponents and the ILG [1 June 2007].

22. ILG selects and distributes all SRC used for each experiment, secret SRC, backup SRCs, and common set of PVSs to be included in every experiment, as proposed by NTT (e.g., 5 SRC & 5 HRC, which would be 25 of 160 video sequences or 15%). This is the step where the ILG (or other organization) must have deinterlaced and resized the 12 second source video sequences. [3 weeks after model submission March 2, 2007]

23. Proponents and ILG inform VQEG of any problem source content. Problem content must be made available on ftp site and reviewed by proponents / ILG. Majority decision needed to reject suspect source. [1 June 2007]

24. Organizations will generate the PVSs by 14 Sept 2007

25. If a proponent testlab believes that their experiment is unbalanced in terms of qualities or have calibration problems, they may ask the ILG and the proponent group to review the selection of test material. If 2/3rd majority agrees then selection of PVSs will be amended by the ILG. An even distribution of qualities from excellent to bad is desirable. [Sept 2007 VQEG meeting]

26. Proponents check calibration of all PVSs and identify potential problems. They may ask the ILG to review the selection of test material and replace if necessary. If a proponent or ILG does not review the test content, then they lose the right to object to the content composition of that test. [Nov. 8, 2007]. Decision of how to handle the non-conforming PVSs will be taken at an audio Nov 9, 2007. Finished 12 Dec 2007.

27. Proponents run their models on all PVSs and submit raw objective data to the ILG [19 Dec 2007]

28. Each organization runs their test and submits results to the ILG. Any source content used in a subjective test with a MOS of <4 will be evaluated by the ILG. The ILG will determine whether the source and its associated processed files are valid. Any invalid test content will be removed before data analysis is performed. Subjective test finished by [18 Jan 2008]

29. Verification of submitted models [18 Jan 2008]

30. ILG distribute subjective and objective data to the proponents and other ILG [1 Feb 2008]

31. Preliminary statistical analysis by ILG [15 Feb 2008]

32. Updated statistical analysis by proponents and further statistical analysis by the ILG [29 Feb 2008]

33. Draft final report [7 March 2008]

34. Deadline for posting a fitting to other proponent [14 March 2008]

35. Deadline for selection of fitting [21 March 2008]

36. First major revision of draft [27 March 2008]—Audio call

37. Approval of final report [18 April 2008] —Audio call/Face2Face??

The ILG will verify that the submitted models (1) run on the ILG's computers and (2) yield the correct output values when run on the test video sequences. Due to their limited resources, the ILG may encounter difficulties verifying executables submitted too close to the model submission deadline. Therefore, proponents are strongly encouraged to submit a prototype model to the ILG well before the verification deadline, to work out platform compatibility problems well ahead of the final verification date. Proponents are also strongly encouraged to submit their final model executable 14 days prior to the verification deadline date, giving the ILG two weeks to resolve problems arising from the verification procedure.

The ILG requests that proponents kindly estimate the run-speed of their executables on a test video sequence and to provide this information to the ILG.

6. Sequence Processing and Data Formats

Separate subjective tests will be performed for different video sizes. One set of tests will present video in QCIF (176x 144 pixels). One set of tests will present CIF (352x288 pixels) video. One set of tests will present VGA (640x480). In the case of Rec. 601 video source, aspect ratio correction will be performed on the video sequences prior to writing the AVI files (SRC) or processing the PVS.

Note that in all subjective tests 1 pixel of video will be displayed as 1 pixel native display. No upsampling or downsampling of the video is allowed at the player.

Presently, VQEG has access to a set of video test sequences. For audio-video tests this database needs to be extended to include new source material containing both audio and video.

6.1. Sequence Processing Overview

The test material will be selected from a common pool of video sequences. If the test sequences are in interlace format then a standard, agreed de-interlacing method will be applied to transform the video to progressive format. All source material should be 25 or 30 frames per second progressive and there should not be more than one version of each source sequence for each resolution. The de-interlacing algorithm will de-interlace Rec. 601 (or other, e.g., HDTV) formatted video into a progressive format, e.g., VGA, CIF, or QCIF. Algorithms will be proposed on the VQEG reflector and approved before processing takes place. Uncompressed AVI files will be used for subjective and objective tests. Tools are being sought to convert from the various coding schemes to uncompressed AVI (see Annex VIII for a description of the tools used for conversion). The progressive test sequences used in the subjective tests should also be used by the models to produce objective scores.

It is important to minimize the processing of video source sequences. Hence, we will endeavor to find methods that minimize this processing (e.g., to perform de-interlacing and resizing in one step).

6.1.1. Duration of Source Sequences

Source content may be obtained from content stored on tape or on hard drive, provided it meets the quality requirements outlined in Section 6.1.2. Source content must be at least 8 seconds and 20 frames long (220 frames for 625-line source; 260 frames for 525-line source).

All source sequences will be 12 seconds duration (300 frames for 625-line source; 360 frames for 525-line source) for processing through each HRC. Original SRCs will be of 12s duration. Each original 12s SRC will be processed by the relevant HRC. The 12s output obtained by processing the original SRC will then be reduced to produce an 8s PVS. For the original SRC, this will be achieved by removing the first 2s and final 2s. For PVS, this will be achieved by removing the first (2 + N) seconds and final (2 – N) seconds, where N is the temporal registration shift needed for the temporal registration limits in section 7.4. Only the middle 8s sequence will be stored for use in subjective testing and for processing by objective models.

Wherever possible, the original source sequence should be 12 seconds. Where the original source sequence is shorter than 12 seconds, but at least 8 seconds and 20 frames, then it will be edited as follows to create a 12 second version. Pad out the beginning and end of the sequence with additional video content. E.g. concatenate three versions of the source video together, reduce to 12s ensuring that the middle 8s is the correct content for use in the subjective test. The shorter source (e.g., 8s + 20frames) must be centered in the 12s SRC. Furthermore, any 8s PVS created from such a SRC cannot contain the pad (e.g., cascaded content from beginning or end.)

6.1.2. Camera and Source Test Material Requirements

The standard definition source test material should be in Rec. 601, DigiBeta, Betacam SP, or DV25 (3-chip camera) format or better. Note that this requirement does not apply to Categories 4 and 8 (Section 6.2) where the best available quality reference will be used. HD source test material should be taken from a professional grade HD camera (e.g., Sony HDR-FX1) or better. Original HD video sequences that have been compressed should show no impairments after being re-sampled to VGA, CIF, and QCIF.

VQEG MM expresses a preference for all test material to be open source. At a minimum, source material must be available for use within VQEG MM proponents and ILG for testing (e.g., under non-disclosure agreement if necessary).

6.1.3. Software Tools

Transformation of the source test sequences (e.g., from Rec. 601 525-line to CIF) shall be performed using Avisynth 2.5.5, VirtualDub 1.6.11, and ffdshow 20050303. Within VirtualDub, video sequences will be saved to AVI files using Video Compression option (Video->Compressor) "ffdshow Video Codec", configured with the "Uncompressed" decoder and the UYVY color space. For the Colour Depth (Video->Color Depth), the setting “4:2:2 YcbCr (UYVY)” is used as output format. The processing mode (Video->) is set to “Full processing mode”.

6.1.4. Colour Space Conversion

In the absence of known color transformation matrices (e.g., such as what might be used by a video display adapter), the following algorithms will be used to transform between ITU-R Recommendation BT.601 $Y'C_B'C_R'$ video and R'G'B' video that is in the range [0, 255]. The reference for these color transformation equations is pages 15-16 of ColorFAQ.pdf, which can be downloaded from:

http://www.poynton.com/PDFs/ColorFAQ.pdf

Transforming R'G'B' to $Y'C_B'C_R'$

1. Compute the matrix transformation:

$$\begin{bmatrix} Y' \\ C_B' \\ C_R' \end{bmatrix} = \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} + \frac{1}{256} \begin{bmatrix} 65.738 & 129.057 & 25.064 \\ -37.945 & -74.494 & 112.439 \\ 112.439 & -94.154 & -18.285 \end{bmatrix} \bullet \begin{bmatrix} R' \\ G' \\ B' \end{bmatrix}$$

2. Round to the nearest integer.

3. Clamp all three components to the range 1 through 254 inclusive (0 and 255 are reserved for synchronization signals in ITU-R Recommendation BT.601).

Transforming $Y'C_B'C_R'$ to R'G'B'

1. Compute the matrix transformation:

$$\begin{bmatrix} R' \\ G' \\ B' \end{bmatrix} = \frac{1}{256} \begin{bmatrix} 298.082 & 0 & 408.583 \\ 298.082 & -100.291 & -208.120 \\ 298.082 & 516.411 & 0 \end{bmatrix} \bullet \left(\begin{bmatrix} Y' \\ C_B' \\ C_R' \end{bmatrix} - \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} \right)$$

2. Round to the nearest integer.

3. Clamp all three components to the range 0 through 255 inclusive.

6.1.5. De-Interlacing

De-interlacing will be performed when original material is interlaced, using the de-interlacing function “KernelDeint” in Avisynth. If the deinterlacing using KernelDeint results in source sequence that has serious artifacts, the Blendfield or Autodeint may be used as alternative methods for deinterlacing. Proprietary algorithms and/or hardware deinterlacing may be used if the above three methods prove unsatisfactory.

To check for de-interlacing problems (e.g. serious artifacts introduced by the de-interlacing process), the ILG will examine source content played back at normal speed, with the option to inspect possible problems at reduced speed.

6.1.6. Cropping & Rescaling

Table 2 lists recommend values for region of interests to be used for transforming images. These source regions should be centered vertically and horizontally. These source regions are intended to be applied *prior* to rescaling and avoid use of over scan video in most cases. These regions are known to correctly produce square pixels in the target video sequence. Other regions may be used, provided that the target video sequence contains the correct aspect ratio.

The source region selection must not include overscan—i.e. black borders from the overscan are not allowed.

TABLE 2. Recommended Source Regions for Video Transformation

From	**To**	**Source Region**
525-line: 720x486 Rec. 601	VGA: 640x480 square pixel	704x480
525-line: 720x486 Rec. 601	CIF: 352x288 square pixel	646x480
525-line: 720x486 Rec. 601	QCIF: 176x144 square pixel	646x480
625-line: 720x576 Rec. 601	VGA: 640x480 square pixel	702x576
625-line: 720x576 Rec. 601	CIF: 352x288 square pixel	644x576
625-line: 720x576 Rec. 601	QCIF: 176x144 square pixel	644x576
1080i: 1920x1080	VGA: 640x480 square pixel	1440x1080
1080i: 1920x1080	CIF: 352x288 square pixel	1320x1080
1080i: 1920x1080	QCIF: 176x144 square pixel	1320x1080
720p: 1280x720	VGA: 640x480 square pixel	960x720
720p: 1280x720	CIF: 352x288 square pixel	880x720
720p: 1280x720	QCIF: 176x144 square pixel	880x720

6.1.7. Rescaling

Video sequences will be resized using Avisynth's 'LanczosResize' function.

6.1.8. File Format

All source and processed video sequences will be stored in Uncompressed AVI in UyVy..

Source material with a source frame rate of 29.97 fps will be manually assigned a source frame rate of 30 fps prior to being inserted into the common pool of video sequences.

AVI is essentially a container format that consists of hierarchical chunks – which have their equivalent in C data structures – which are all preceded by a so called fourcc, a "four character code", which indicates the type of chunk following. Some of the chunks are compulsory and describe the structure of the file, while some are optional and others contain the real video or audio data. The AVI container format which is used for the exchange of files in VQEG MM is originally defined by Microsoft as part of the RIFF file specification in:
"http://msdn.microsoft.com/library/default.asp?url=/library/en-us/wcedshow/html/_dxce_dshow_avi_riff_file_reference.asp"

Other descriptions can be found in:
http://www.opennet.ru/docs/formats/avi.txt
http://www.the-labs.com/Video/odmlff2-avidef.pdf

These last two can be found on the mmpretest ftp server. All these links describe the AVI format in details as far as the container itself is concerned. Since the multitude of chunks is quite confusing, an example C code that reads and writes AVI files down to this level is also included in the archive on the mmpretest reflector (files avilib.c and avilib.h). Please note, that the provided C code falls under the GNU Public License. Please refer to the license statements in the files themselves. The provided C code is very simple to use and should serve all needs of VQEG. Please note that the C code allows opening the data chunk with the UVVY data, but it does not decode this data. In fact, avilib does not know how to interpret these data. All it returns is a pointer to the data and some additional information like image sizes and frame rate. Interpretation of these data is up to the user and described in the following paragraphs.

A description of the UYVY chunk format which is to be used inside the AVI container can be found in http://www.fourcc.org/index.php?http%3A//www.fourcc.org/fccyvrgb.php and below.

UYVY is a YUV 4:2:2 format. The effective bits per pixel are 16. In the AVI main header (after the fourcc "avih"), a positive height parameter implies a top-down image (top line first).Two image pixels form one macro pixel and are stored in one 32bit word with the following byte ordering:

(lowest byte) U_0 Y_0 V_0 Y_1 (highest byte)

6.1.9. Source Test Video Sequence Documentation

Preferably, each source video sequence should be documented. The exact process used to create each source video sequence should be documented, listing the following information:

- Camera specifications
- Source region of interest (if the default values were not used)

- Use restrictions (e.g., “open source”)
- Deinterlacing method

This documentation is desirable but not required.

6.2. Test Materials

The test material will be representative of a range of content and applications. The list below identifies the type of test material that forms the basis for selection of sequences.

1) video conferencing: (available for research purposes only, NTIA (Rec 601 60Hz); BT (Rec 601 50Hz), Yonsei (CIF and QCIF), FT (Rec 601 50Hz, D1)), NTT (Rec 601 60Hz, D1)

   Currently available: NTIA, NTT, FT

2) movies, movie trailers:(VQEG Phase II), Opticom, IRCCyN, (trailer equivalent, restricted within VQEG)

   Currently available: Psytechnics, SVT, Opticom,

3) sports: (available, 15-20 mins from Yonsei, Comcast), KDDI (7 min D1 and D2, other scenes also available), NTIA (Comcast), IRCCyN

   Currently available: Yonsei, SVT, Psytechnics, Opticom

4) music video: (Intel)), IRCCyN

   Currently available: NTIA,

5) advertisement:

   Currently available: Psytechnics, Opticom

6) animation: (graphics Phase I, cartoon Phase II; Opticom will send material to Yonsei), IRCCyN

   Currently available: Opticom, NTIA

7) broadcasting news: (head and shoulders and outside broadcasting). (available – Yonsei;, possible Comcast), IRCCyN

   Currently available: KBS, Opticom

8) home video: (FUB possibly, BT possibly, INTEL, NTIA). Must be captured with DV camera or better.

   Currently available: NTIA, SwissQual, Yonsei

   There will be no completely still video scenes in the test.

All test material should be sent to the content point of contact (Chulhee Lee, Yonsei) first and then it will be put on the ftp server by NTIA. Ideally the material should be converted before being sent to Chulhee Lee.

The source video will only be used in the testing if an expert in the field considers the quality to be good or excellent on an ACR-scale.

6.2.1. Selection of Test Material (SRC)

The ILG is responsible for selecting SRC material to be used in each subjective quality test. The VQEG MM group will be responsible for deciding upon precise HRCs to be used in the testing. Section 5.3 provides basic guidelines on the process for selecting SRCs and HRCs together with a procedure for the distribution of test content.

6.3. Hypothetical Reference Circuits (HRC)

The subjective tests will be performed to investigate a range of HRC error conditions. These error conditions may include, but will not be limited to, the following:

- Compression errors (such as those introduced by varying bit-rate, codec type, frame rate and so on)
- Transmission errors
- Post-processing effects
- Live network conditions
- Interlacing problems

The overall selection of the HRCs will be done such that most, but not necessarily all, of the following conditions are represented.

6.3.1. Video Bit-rates

- PDA/Mobile (QCIF): 16 kbit/s to 320 kbs (e.g., 16, 32, 64, 128, 192, 320)
- PC1 (CIF): 64 kbit/s to 704 kbit/s (e.g. 64, 128, 192, 320, 448, 704)
- PC2 (VGA): 128kbit/s to 4Mbit/s (e.g. 128, 256, 320, 448, 704, ~1M, ~1.5M, ~2M, 3M,~4M)

6.3.2. Simulated Transmission Errors

A set of test conditions (HRC) will include error profiles and levels representative of video transmission over different types of transport bearers:

- Packet-switched transport (e.g., 2G or 3G mobile video streaming, PC-based wireline video streaming)
- Circuit-switched transport (e.g., mobile video-telephony)

It is important that when creating HRCs using a simulator, documentation is produced detailing simulator settings (for circuit switched HRCs the error pattern for each PVS should also be produced).

Annex III provides guidelines on the procedures for creating and documenting transmission error conditions.

Packet-switched transmission

HRCs will include packet loss with a range of packet loss ratios (PLR) representative of typical real-life scenarios.

In **mobile video streaming**, we consider the following scenarios:

1. Arrival of packets is delayed due to re-transmission over the air. Re-transmission is requested either because packets are corrupted when being transmitted over the air, or because of network congestion on the fixed IP part. Video will play until the buffer empties if no new (error-checked/corrected) packet is received. If the video buffer empties, the video will pause until a sufficient number of packets is buffered again. This means that in the case of heavy network congestion or bad radio conditions, video will pause without skipping during re-buffering, and no video frames will be lost. This case is not implemented in the current test plan as stated in Section 6.3.4.
2. Arrival of packets is delayed, and the delay is too large: These packets are discarded by the video client.

Note: A radio link normally has *in-order delivery*, which means that if one packet is delayed the following packets will also be delayed.

Note: If the packet delay is too long, the radio network might drop the packet.

3. Very bad radio conditions: Massive packet loss occurs.
4. Handovers: Packet loss can be caused by handovers. Packets are lost in bursts and cause image artifacts.

   Note: This is valid only for certain radio networks and radio links, like GSM or HSDPA in WCDMA. A dedicated radio channel in WCDMA uses soft handover, which not will cause any packet loss.

Typical radio network error conditions are:

- Packet delays between 100 ms and 5 seconds.

In **PC-based wireline video streaming**, network congestion causes packet loss during IP transmission.

In order to cover different scenarios, we consider the following models of packet loss:

- Bursty packet loss. The packet loss pattern can be generated by a link simulator or by a bit or block error model, such as the Gilbert-Elliott model.
- Random packet loss
- Periodic packet loss.

Note: The bursty loss model is probably the most common scenario in a 'normal' network operation. However, periodic or random packet loss can be caused by a faulty piece of equipment in the network. Bursty, random, and periodic packet loss models are available in commercially-available packet network emulators.

Choice of a specific PLR is not sufficient to characterize packet loss effects, as perceived quality will also be dependent on codecs, content, packet loss distribution (profiles) and which types of video frames were hit by the loss of packets. For our tests, we will select different levels of loss ratio with different distribution profiles in order to produce test material that spreads over a wide range of video quality. To confirm that test files do cover a wide range of quality, the generated test files (i.e., decoded video after simulation of transmission error) will be:

1. Viewed by video experts to ensure that the visual degradations resulting from the simulated transmission error spread over a range of video quality over different content;
2. Checked to ensure that degradations remain within the limits stated by the test plan (e.g., in the case where packet loss causes loss of complete frames, we will check that temporal misalignment remains with the limits stated by the test plan).

Circuit-switched transmission

HRCs will include bit errors and/or block errors with a range of bit error rates (BER) or/and block[3] error rates (BLER) representative of typical real-world scenarios. In circuit-switched transmission, e.g., video-telephony, no re-transmission is used. Bit or block errors occur in bursts.

In order to cover different scenarios, the following error levels can be considered:

[3] Note that the term 'block' does not refer to a visual degradation such as blocking errors (or blockiness) but refers to errors in the transport stream (transport blocks).

Air interface block error rates: Normal uplink and downlink: 0.3%, normally not lower. High value uplink: 0.5%, high downlink: 1.0%. To make sure the proponents' algorithms will handle really bad conditions up to 2%-3% block errors on the downlink can be used.

Bit stream errors: Block errors over the air will cause bits to not be received correctly over the air. A video telephony (H.223) bit stream will experience CRC errors and chunks of the bit stream will be lost.

Tools are currently being sought to simulate the types of error transmission described in this section.

Proponents are asked to provide examples of level of error conditions and profiles that are relevant to the industry. These examples will be viewed and/or examined after electronic distribution (only open source video is allowed for this).

6.3.3. Live Network Conditions

Simulated errors are an excellent means to test the behavior of a system under well defined conditions and to observe the effects of isolated distortions. In real live networks however usually a multitude of effects happen simultaneously when signals are transmitted, especially when radio interfaces are involved. Some effects like e.g. handovers, can only be observed in live networks.

The term "live network" specifies conditions which make use of a real network for the signal transmission. This network is not exclusively used by the test setup. It does not mean that the recorded data themselves are taken from live traffic in the sense of passive network monitoring. The recordings may be generated by traditional intrusive test tools, but the network itself must not be simulated.

Live network conditions of interest include radio transmission (e.g., mobile applications) and fixed IP transmission (e.g., PC-based video streaming, PC to PC video-conferencing, best-effort IP-network with ADSL-access). Live network testing conditions are of particular value for conditions that cannot confidently be generated by network simulated transmission errors (see section 6.3.4). Live network conditions should exhibit distortions representative of real-world situations that remain within the limits stated elsewhere in this test plan.

Normally most live network samples are of very good or best quality. To get a good proportion of sample quality levels, an even distribution of samples from high to low quality should be saved after a live network session.

Note: Keep in mind the characteristics of the radio network used in the test. Some networks will be able to keep a very good radio link quality until it suddenly drops. Other will make the quality to slowly degrade.

Samples with perfect quality do not need to be taken from live network conditions. They can instead be recorded from simulation tests.

Live network conditions as opposed to simulated errors are typically very uncontrolled by their nature. The distortion types that may appear are generally very unpredictable. However, they represent the most realistic conditions as observed by users of e.g. 3G networks.

Recording PVSs under live network conditions is generally a challenging task since a real hardware test setup is required. Ideally, the capture method should not introduce any further degradation. The only requirement on capture method is that the captured sequences conform to the file requirements in section 6.1.7 and 7.2.

For applications including radio transmissions, one possibility is to use a laptop with e.g. a built-in 3G network card and to download streams from a server through a radio network. Another possibility is the use of drive test tools and to simulate a video phone call while the car is driving. In order to simulate very bad radio coverage, the antenna may be wrapped with some aluminum foil (Editors note: This strictly a simulation again, but for the sake of simplicity it can be accepted since the simulated bad coverage is overlayed with the effects from the live network).

In order to prepare the PVSs the same rules apply as for simulated network conditions. The only difference is the network used for the transmission.

6.3.4. Pausing with Skipping and Pausing without Skipping

Pausing without skipping events will not be included in the current testing.

Pausing with skipping events will be included in the current testing. Anomalous frame repetition is not allowed during the first 1s or the final 1s of a video sequence. Note that where pausing with skipping and anomalous frame repetition is included in a test then source material containing still sections should form part of the testing.

If it is difficult or impossible to determine whether a video sequence contains pausing without skipping or pausing with skipping, the video sequence will be given the benefit of doubt and considered to contain pausing with skipping. The same applies to anomalous frame repetition in the first 1s or final 1s of video sequence.

Other types of anomalous behavior are allowed provided they meet the following restrictions. The delay through the system before, after, and between anomalous behavior segments must vary around an average delay and must meet the temporal registration limits in section 7.4. The first 1s and final 1s of each video sequence cannot contain any anomalous behavior. At most 25% of any individual PVS's duration may exceed the temporal registration limits in section 7.4. These 25% must have at most maximum temporal registration error of +3 seconds (added delay).

(See section 2 for definitions of “pausing with skipping”, “pausing without skipping” and “anomalous frame repetition”.)

The following is another example of a PVS containing pausing with skipping events, provided that the constraints in this section (6.3.4) and the temporal registration limits in section 7.4 are respected: a PVS with no frame repeats that changes between several different constant delays when anomalous events occurs, for example the delay is 0s at the start, then +0.1 sec for awhile, then -0.25s for awhile, then +0.25s for awhile, etc. Different constant delays occur in typical error-prone network scenarios because it cannot be assured that a pause of X frames is followed by a skip of exactly X frames.

6.3.5. Frame Rates

For those codecs that only offer automatically set frame rate, this rate will be decided by the codec. Some codecs will have options to set the frame rate either automatically or manually. For those codecs that have options for manually setting the frame rate (and we choose to set it for the particular case), 5 fps will be considered the minimum frame rate for VGA and CIF, and 2.5 fps for PDA/Mobile..

Manually set frame rates (constant frame rate) may include:

- PDA/Mobile: 30, 25, 15, 12.5, 10, 8, 5, 2.5 fps

• PC1 (CIF):	30, 25, 15, 12.5, 10, 8, 5 fps
• PC2 (VGA):	30, 25, 15, 12.5, 10,8, 5 fps

Variable frame rates are acceptable for the HRCs. The first 1s and last 1s of each QCIF PVS must contain at least two unique frames, provided the source content is not still for those two seconds. The first 1s and last 1s of each CIF and VGA PVS must contain at least four unique frames, provided the source content is not still for those two seconds.

Care must be taken when creating test sequences for display on a PC monitor. The refresh rate can influence the reproduction quality of the video and VQEG MM requires that the sampling rate and display output rate are compatible. For example,

Given a source frame rate of video is 30fps, the sampling rate is 30/X (e.g. 30/2 = sampling rate of 15fps). This is called frame rate. Then we upsample and repeat frames from the sampling rate of 15fps to obtain 30 fps for display output.

The intended frame rate of the source and the PVS must be identical.

6.3.6. Pre-Processing

The HRC processing may include, typically prior to the encoding, one or more of the following:

- Filtering
- Simulation of non-ideal cameras (e.g. mobile)
- Colour space conversion (e.g. from 4:2:2 to 4:2:0)
- Interlacing of previously deinterlaced source.

This processing will be considered part of the HRC.

6.3.7. Post-Processing

The following post-processing effects may be used in the preparation of test material:

- Colour space conversion
- De-blocking
- Decoder jitter

Deinterlacing of codec output including when it has been interlaced prior to codec input.

6.3.8. Coding Schemes

Coding Schemes that will be used may include, but are not limited to:

- Windows Media Player 9
- H.263
- H.264 (MPEG-4 Part 10)

- Real Video (e.g. RV 10)
- MPEG 4
- JPEG 2000 Part 3
- VC1

6.3.9. Processing and Editing Sequences

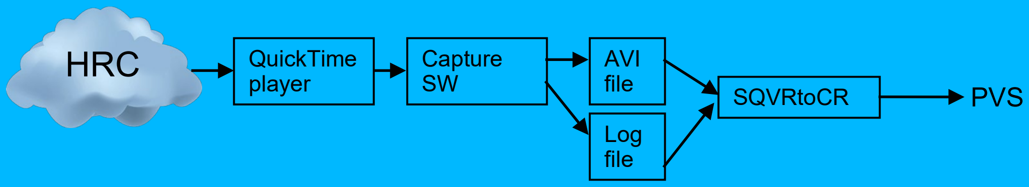
Test sequences will be captured from the decoded video in uncompressed format. The two capture methods below have been identified, but others may be used as well. Strict documentation of how PVSs have been produced should be forwarded to the ILG.

SwissQual method

Video capture is done using proprietary software developed at SwissQual. The software captures an uncompressed video signal directly from QuickTime player v7.0 generating two files. The first file contains video data in AVI format whereas a second file contains a list of the time-stamps of the received frames. The input signal can also contain a variable frame rate.
QuickTime 7.0 supports most known video encoding formats like: MPEG-4, H.261, H.263, H.264, Cinepak, DV-PAL/NTSC, Intel Indeo etc.

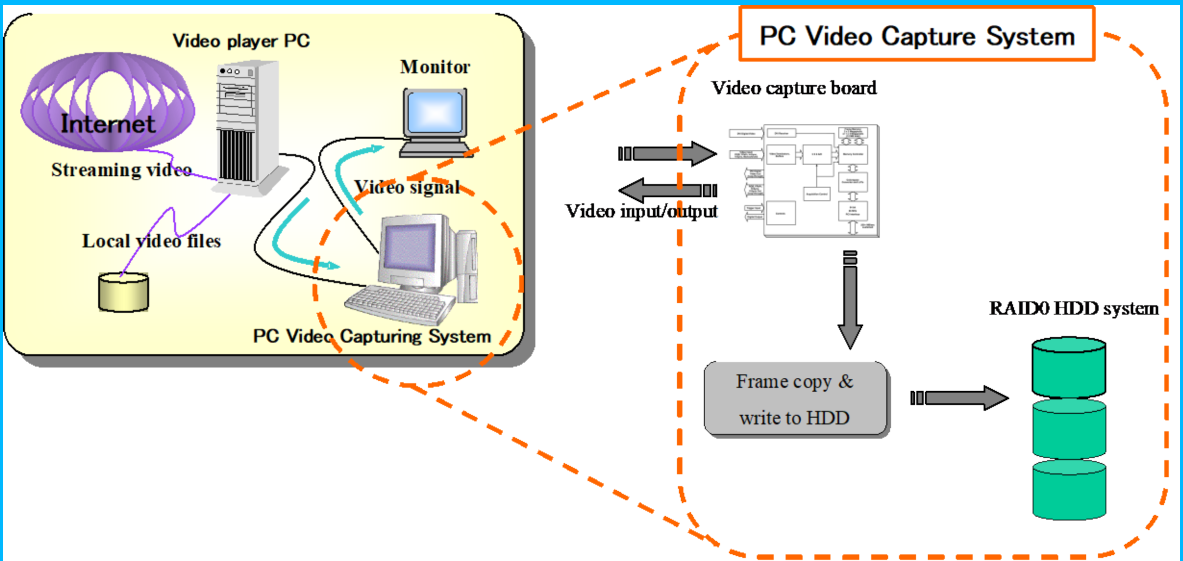
Recording at variable frame rate reduces the amount of redundancy in video frames during pausing without skipping events. There are two possibilities to play back the PVS on the display:
- Using a proprietary Player (SQAviPlayer), which reads AVI file at variable frame rate and time stamps from LOG file.
- Using a standard player e.g. QuickTime connected to an output of SQVRtoCR

SQVRtoCR converts variable to constant frame rate PVSs.



NTT method PIFREC 1.0 (Lossless PC Video & Voice Recorder)

The PC capture system uses a capture board to receive the signals passed from a PC to its monitor, without adding any processing load to the PC, and stores them while retaining high video quality. So, video service providers can evaluate and monitor video quality, an operation which is particularly necessary if the video service is charged for, without imposing a processing load on the receiving terminal, a penalty which has conventionally been unavoidable.



Product composition:

PC video & voice recording software	Frame detection, and storing of video and voice data
PC video capture board	High-resolution video capture
Video capturing PC set (PC, hard disc and other peripherals. Monitor is not included)	Video play-back

Specification:

Input format	Analog signal/digital signal (DVI)
Output format	AVI format Video: uncompressed video, reference video Voice: uncompressed audio
Maximum recording time	1 hour (in the case of VGA and 30fps)
Recording performance	VGA, 30fps* and full color (24 bits)

*Frame rate: the number of frames displayed on the monitor each second. 30fps for example means that the display is refreshed 30 times each second. The higher the value, the smoother the video looks. The frame rate of television (NTSC) is 30fps.

7. Objective Quality Models

7.1. Model Type

VQEG MM has agreed that Full Reference, Reduced Reference and No Reference models may be submitted for evaluation. The side-channels allowable for the RR models are:

• PDA/Mobile (QCIF):	(1kbit/s, 10kbit/s)
• PC1 (CIF):	(10kbit/s, 64kbit/s)
• PC2 (VGA):	(10kbit/s, 64kbit/s, 128kbit/s)

Note that for each side-channel condition the limits defined here represent the maximum allowable side-channel data rate. For example, where the side-channel is limited to 10 kbit/s, then valid side-channels are those that use a data rate of <=10 kbit/s and any data rates above 10 kbit/s are invalid.

Proponents may submit one model of each type for all image size conditions. Thus, any single proponent may submit up to a total of 13 different models (one FR model for QCIF, one FR model for CIF, one FR model for VGA; one NR model for QCIF, one NR model for CIF, one NR model for VGA; two RR models for QCIF, two RR models for CIF, three RR models for VGA). Note that where multiple models are submitted, additional model submission fees may apply.

7.2. Model Input and Output Data Format

Video will be full frame, full frame rate. The progressive video format will be used in the multimedia test. Models will use the same 8s SRC and PVS seen by viewers (see section 4.1.1 & 6.1.1).

7.2.1 Full reference Models

The FR model will be given an ASCII file listing pairs of video sequence files to be processed. Each line of this file has the following format:

```
        <source-file>    <processed-file>
```

where <source-file> is the name of a source video sequence file and <processed-file> is the name of a processed video sequence file, whose format is specified in section 6.1.7 of this document. File names may include a path. For example, an input file should adhere to the following naming convention:

```
/video/vXX_YYY.avi    /video/vXX_YYY.avi          (Unix)
or
\video\vXX_YYY.avi    \video\vXX_YYY.avi          (Windows)

/video/cXX_YYY.avi    /video/ cXX_YYY.avi         (Unix)
or
\video\cXX_YYY.avi    \video\cXX_YYY.avi          (Windows)

/video/qXX_YYY.avi    /video/qXX_YYY.avi          (Unix)
or
\video\qXX_YYY.avi    \video\qXX_YYY.avi          (Windows)
```

where v represents VGA, c represents CIF, q represents QCIF, XX indicates the test number and YYY represents the video sequence index. The leading characters (v,c,q) and all extensions ("avi" and "dat") should be in lower cases.

The output file is an ASCII file created by the model program, listing the name of each processed sequence and the resulting Video Quality Rating (VQR) of the model. The contents of the output file should be flushed after each sequence is processed, to allow the testing laboratories the option of halting a processing run at any time. Each line of the ASCII output file has the following format:

< source-file > <processed-file> VQR

Where < source-file > is the name of the source file run through this model, without any path information; and <processed-file> is the name of the processed sequence run through this model, without any path information. VQR is the Video Quality Ratings produced by the objective model. For the input file example, this file contains the following:

v01_ref.avi   v01_001.avi 0.150

v01_ref.avi   v01_002.avi 1.304

v01_ref.avi   v01_003.avi 0.102

v01_ref.avi   v01_004.avi 2.989

Each proponent is also allowed to output a file containing Model Output Values (MOVs) that the proponents consider to be important. The format of this file will be

v01_001.avi 0.150 MOV_1 MOV_2,… MOV_N

v01_002.avi 1.304 MOV_1 MOV_2,… MOV_N

v01_003.avi 0.102 MOV_1 MOV_2,… MOV_N

v01_004.avi 2.989 MOV_1 MOV_2,… MOV_N

7.2.2 Reduced-reference Models

In an effort to limit the amount of variations and in agreement with all proponents attending the VQEG meeting consensus was achieved to allow only downstream video quality models.

7.2.2.1 Downstream Model Original Video Processing:

The software (model) for the original video side will be given the original test sequence in the final file format and produce a reference data file. The amount of reference information in this data file will be evaluated in order to estimate the bit rate of the reference data and consequently assign the class of the method (Section 7.1). The input file format of the full-reference model will be used for the RR model for the original video side. Deterministic RR models for the original video side may ignore the processed video file name which is the second argument. For example, given an input file:

/video/qXX_YYY.avi   /video/qXX_ZZZ.avi   (Unix example, for Windows OS the path conforms to Windows format)

Then, the model should produce reference data files whose file names are made in the following way:

/video/qXX_YYY_BBB.dat   (deterministic models) or

/video/qXX_YYY_ZZZ_BBB.dat (deterministic and non-deterministic models)

where BBB indicates side-channel bandwidth in kbps. For example, for a VGA RR model with the 10kbps side channel, the output file names should be as follows:

vXX_001_10.dat
vXX_002_10.dat

or

vXX_001_023_10.dat
vXX_002_100_10.dat.

The model should save the output files in the current directory. The ILG should make sure that PVS files are not available for the software for the original video side.

7.2.2.2 Downstream Model Processed Video Processing:

The software (model) for the processed video side will be given the processed test sequence in the final file format and a reference data file that contains the reduced-reference information (see Model Original Video Processing). The input file format of the full-reference model will be used for the model for the processed video side. The format of this file will be

/video/qXX_YYY.avi /video/qXX_ZZZ.avi

where v indicates VGA resolution, XX indicates the test number, YYY represents the source video sequence index and ZZZ represents the processed video sequence index. Then, the model should make reference data file names as follows:

/video/qXX_YYY_BBB.dat   (deterministic models) OR
/video/qXX_YYY_ZZZ_BBB.dat (deterministic and non-deterministic models)

where BBB indicates side-channel bandwidth in kbps. Finally, the model should use the processed video file and the reference data file, and produce a VQR for the processed video sequence. The ILG should make sure that SRC files are not available for the software for the processed video side.

The output file format of the RR model will be identical with that of the FR model (Section 7.2.1).

7.2.2.3 Input Parameters for RR models.

Some RR models, the identical software may generate and process reference data files at various side-channel bandwidths. In this case, the software needs information on side-channel bandwidth. In order to provide the information, the software (model) for the original video side will be given two arguments as follows:

CompanyName_vRRsrc.exe vXX.txt BBB

where vXX.txt is the input file name, XX indicates the test number and BBB indicates side-channel bandwidth in kbps.

The software (model) for the processed video side will be given two arguments as follows:

CompanyName_vRRpvs.exe vXX.txt BBB

7.2.3 No-reference Models

The NR model will be given an ASCII file listing only processed video sequence files. Each line of this file has the following format:

<processed-file>

where <processed-file> is the name of a processed video sequence file, whose format is specified in section 6.1.7 of this document. File names may include a path. For example, an input file should adhere to the following naming convention:

/video/vXX_001.avi
/video/vXX_002.avi

The output file format of the NR model will take the form

<processed-file> VQR

where <processed-file> is the name of the processed sequence run through this model, without any path information. VQR is the Video Quality Ratings produced by the objective model.

NR models will be required to predict the perceptual quality of both the source and processed video files used in subjective quality tests.

7.3. Submission of Executable Model

For each video format (QCIF, CIF, and VGA), a set of 2 source and processed video sequence pairs will be used as test vectors. They will be available for downloading on the VQEG web site http://www.vqeg.org/. Each proponent will send an executable of the model and the test vector outputs to the ILG by the date specified in action item "Proponents submit their models (executable and, only if desired, encrypted source code)" of Section 5.3. The executable version of the model must run correctly on one of the two following computing environments:

- SUN SPARC workstation running the Solaris 2.3 UNIX operating system (SUN OS 5.5). [Ed. Note: The used of SUN workstation should be agreed]
- WINDOWS 2000 workstation and Windows XP.

The use of other platforms will have to be agreed upon with the independent laboratories prior to the submission of the model.

The ILG will verify that the software produces the same results as the proponent with a maximum error of plus or minus 0.0001% of the proponents reported value. See Annex X for requirements on non-deterministic models. A maximum of 5 randomly selected files will be used for verification. If greater errors are found, the independent and proponent laboratories will work together to correct them. If the errors cannot be corrected, then the ILG will review the results and recommend further action.

7.4. Registration

Measurements will only be performed on the portions of PVSs that are not anomalously severely distorted (e.g. in the case of transmission errors or codec errors due to malfunction).

Models must include calibration and registration if required to handle the following technical criteria (Note: Deviation and shifts are defined as between a source sequence and its associated PVSs. Measurements of gain and offset will be made on the first and last seconds of the sequences. If the first and last seconds are anomalously severely distorted, then another 2 second portion of the sequence will be used.):

- maximum allowable deviation in *offset* is ±20
- maximum allowable deviation in *gain* is ±0.1
- maximum allowable *Horizontal Shift* is +/- 1 pixel
- maximum allowable *Vertical Shift* is +/- 1 pixel
- maximum allowable *Horizontal Cropping* is 12 pixels for VGA, 6 pixels for CIF, and 3 pixels for QCIF (for each side).
- maximum allowable *Vertical Cropping* is 12 pixels for VGA, 6 pixels for CIF, and 3 pixels for QCIF (for each side).
- no *Spatial Rotation or Vertical or Horizontal Re-scaling* is allowed
-
- no Spatial *Picture Jitter* is allowed. Spatial picture jitter is defined as a temporally varying horizontal and/or vertical shift.

For a description of offset and gain in the context of this testplan see Annex IX. This Annex also includes the method for calculating offset and gain in PVSs.

No Reference Models should not need calibration

Reduced Reference Models must include temporal registration if the model needs it. Temporal misalignment of no more than +/-0.25s is allowed. Please note that in subjective tests, the start frame of both the reference and its associated HRCs are matched as closely as possible. Spatial offsets are expected to be very rare. It is expected that no post-impairments are introduced to the outputs of the encoder before transmission. Spatial registration will be assumed to be within (1) pixel. Gain, offset, and spatial registration will be corrected, if necessary, to satisfy the calibration requirements specified in this test plan.

The organizations responsible for creating the PVSs shall check that they fall within the specified calibration and registration limits. The PVSs will be double-checked by one other organization. After testing has been completed any PVS found to be outside the calibration limits shall be removed from the data analyzes. ILG will decide if a suspect PVS is outside the limits.

8. Objective Quality Model Evaluation Criteria

This chapter describes the evaluation metrics and procedure used to assess the performance of an objective video quality model as an estimator of video picture quality in a variety of applications.

8.1. Evaluation Procedure

The performance of an objective quality model is characterized by three prediction attributes: accuracy, monotonicity and consistency.

The statistical metrics root mean square (rms) error, Pearson correlation, and outlier ratio together characterize the accuracy, monotonicity and consistency of a model's performance. The calculation of each statistical metric is performed along with its 95% confidence intervals. To test for statistically significant differences among the performance of various models, the F-test will be used.

The statistical metrics are calculated using the objective model outputs and the results from viewer subjective rating of the test video clips. The objective model provides a single number (figure of merit) for every tested video clip. The same tested video clips get also a single subjective figure of merit. The subjective figure of merit for a video clip represents the average value of the scores provided by all subjects viewing the video clip.

Objective models cannot be expected to account for (potential) differences in the subjective scores for different viewers or labs. Such differences, if any, will be measured, but will not be used to evaluate a model's performance. "Perfect" performance of a model will be defined so as to exclude the residual variance due to within-viewer, between-viewer, and between-lab effects

The evaluation analysis is based on DMOS scores for the FR and RR models, and on MOS scores for the NR model. Discussion below regarding the DMOS scores should be applied identically to MOS scores. For simplicity, only DMOS scores are mentioned for the rest of the chapter.

The objective quality model evaluation will be performed in three steps. The first step is a monotonic rescaling of the objective data to better match the subjective data. The second calculates the performance metrics for the model and their confidence intervals. The third tests for differences between the performances of different models using the F-test.

8.2. PSNR

PSNR will be calculated to provide a performance benchmark. Proponents are encouraged to calculate PSNR Ideally, PSNR should be calculated with a spatial registration accuracy 0.1 pixel. If this is not possible, then a maximum registration tolerance of 0.5 pixel spatial accuracy is required. The Pearson correlation evaluation metric defined in this section will be applied to determine the predictive performance of PSNR and this will be reported in the final report.

8.3. Data Processing

Prior to any data analysis, the ILG will perform an inspection of the subjective test data. Any source sequences presented in the test with a MOS rating of <4 will be identified and the file will be examined. If, in the opinion of the ILG the poor MOS vaues for these source sequences are due to inferior quality then they shall be removed and not included in the subsequent data analysis. This data inspection will be completed prior to proponents submitting their objective data to the ILG.

8.3.1. Calculating DMOS Values

The data analysis will be performed using the difference mean opinion score (DMOS). DMOS values are calculated on a per subject per PVS basis. The appropriate hidden reference (SRC) is used to calculate the DMOS value for each PVS. DMOS values will be calculated using the following formula:

DMOS = MOS (PVS) – MOS (SRC) + 5

In using this formula, a DMOS of 5 indicates 'Excellent' quality and a DMOS of 1 indicates 'Bad' quality. Any DMOS values greater than 5 (i.e. where the processed sequence is rated better quality than its associated hidden reference sequence) will be considered valid and included in the data analysis.

8.3.2. Mapping to the Subjective Scale

Subjective rating data often are compressed at the ends of the rating scales. It is not reasonable for objective models of video quality to mimic this weakness of subjective data. Therefore, in previous video quality projects VQEG has applied a non-linear mapping step before computing any of the performance metrics. A non-linear mapping function that has been found to perform well empirically is the cubic polynomial given in (1)

$$DMOSp = ax^3 + bx^2 + cx + d$$

Where DMOSp is the predicted DMOS, and the VQR is the model's computed value for a clip-HRC combination. The weightings *a*, *b* and *c* and the constant *d* are obtained by fitting the function to the data [DMOS, VCR]. The mapping function will maximize correlation between DMOSp and DMOS :

$$DMOSp = k(a'x^3 + b'x^2 + c'x) + d$$

with constant k = 1, d = 0

This function must be constrained to be monotonic within the range of possible values for our purposes.

Then root mean squared error is minimized over k and d.

a = k*a'

b = k*b'

c = k*c'

This non-linear mapping procedure will be applied to each model's outputs before the evaluation metrics are computed.

Proponents, in addition to the ILG, may compute the coefficients of the mapping functions for their models and submit the coefficients to ILGs. The proponent who submits the coefficients should also submit his mapping tool (executable) to ILGs so that ILGs can use the mapping tool for other models. It is desirable that the proponent also submit the coefficients of the mapping functions for all the other proponents' models. If a proponent chooses not to exercise this option to compute the coefficients of the mapping functions, the ILG will compute the coefficients of the mapping functions. The ILG will use the coefficients of the fitting function that produce the best correlation coefficient provided that it is a monotonic fit.

Any and all mapping algorithms used for the official data analysis must be referenced.

8.3.3. Averaging Process

Primary analysis of model performance will be calculated per processed video sequence. Secondary analysis of model performance may be calculated and reported on (1) averaged data, by averaging all SRC associated with each HRC ($\mathrm{DMOS}_{\mathrm{H}}$), and on (2) averaged data, by averaging all HRC associated with each SRC ($\mathrm{DMOS}_{\mathrm{S}}$).

8.3.4. Aggregation Procedure

The evaluation of the objective metrics is performed in two steps. In the first step, the objective metrics are evaluated per experiment. In this case, the evaluation/statistical metrics are calculated for all tested objective metrics. A comparison analysis is then performed based on significance tests. In the second step, an aggregation of the performance results is considered. The aggregation will be performed by taking the average values for all three evaluation metrics for all experiments (see section 8.3).

The ILG has the option open to use some of the secret tests to replicate ILG experiments (i.e., run viewers through another lab's experiment). Models' performance evaluation will follow the procedures laid out above. The data collected will be not be considered an additional experiment for the purposes of model comparison--i.e. no double weight for any single experiment. The first experiment run will provide the data to be used for model performance evaluation. The replication data will be used for other analyses. This data will be shared with the proponents.

8.4. Evaluation Metrics

Once the mapping has been applied to objective data, the three evaluation metrics: root mean square error, Pearson correlation coefficient and outlier ratio are determined. The calculation of each evaluation metric is performed along with its 95% confidence interval.

8.4.1. Pearson Correlation Coefficient

The Pearson correlation coefficient R (see Equation 2) measures the linear relationship between a model's performance and the subjective data. Its great virtue is that it is on a standard, comprehensible scale of -1 to 1 and it has been used frequently in similar testing.

X

$$R=\frac{\sum_{i=1}^{N}(Xi-\overline{X})*(Yi-\overline{Y})}{\sqrt{\sum(Xi-\overline{X})^{2}}*\sqrt{\sum(Yi-\overline{Y})^{2}}} \qquad (2)$$

Xi denotes the subjective score DMOS and Yi the objective DMOSp one. N represents the total number of video samples considered in the analysis.

It is known [1] that the statistic z (3) is approximately normally distributed and its standard deviation is defined by (4). Equation (3) is called Fisher-z transformation.

$$z=0.5\cdot\ln\left(\frac{1+R}{1-R}\right) \qquad (3)$$

$$\sigma_z=\sqrt{\frac{1}{N-3}} \qquad (4)$$

The 95% confidence interval (CI) for the correlation coefficient is determined using the Gaussian distribution, which characterizes the variable z and it is given by (5)

CI = $z \pm 2 * \sigma_z$ (5)

NOTE. If the mean is based on less than thirty samples (ie., N < 30), then the Gaussian distribution must be replaced the appropriate Student's t distribution, depending on the specific number of samples in the mean [1].

.

8.4.2. Root Mean Square Error

The accuracy of the objective metric is evaluated using the root mean square error (rmse) evaluation metric.

The difference between measured and predicted DMOS is defined as the absolute prediction error Perror (6)

$$Perror(i) = DMOS(i) - DMOS_p(i) \tag{6}$$

where the index *i* denotes the video sample.

The root-mean-square error of the absolute prediction error Perror is calculated with the formula (7)

$$rmse = \sqrt{\left(\frac{1}{N-d}\sum_N Perror[i]^2\right)} \tag{7}$$

.

Where N denotes the number of samples and d the number of degrees of freedom of the mapping function (1).

The root mean square error is approximately characterized by a χ^2 (n) [1 [Ed. Note: a page number or equation should be given here], where n represents the degrees of freedom and it is defined by (8)

$$n = N - d \tag{8}$$

where N represents the total number of samples.

Using the χ^2 (n) distribution, the 95% confidence interval for the rmse is given by (9) [1]

$$\frac{rmse * \sqrt{N-d}}{\chi^2_{0.95}(N-d)} < rmse < \frac{rmse * \sqrt{N-d}}{\chi^2_{0.05}(N-d)} \tag{9}$$

[Ed. Note: Header is missing here—see an earlier version]

The consistency attribute of the objective metric is evaluated by the outlier ratio OR which represents number of “outlier-points” to total points N.

$$OR = \frac{TotaNoOutliers}{N} \tag{10}$$

where an outlier is a point for which

$|Perror(i)| > 1.96 * \sigma(DMOS(i))$/sqrt(NumberOfObservers)

(11)

where σ(DMOS(i)) represents the standard deviation of the individual scores associated with the video clip i. The individual scores are approximately normally distributed and therefore twice the σ value represents the 95% confidence interval. Thus, 2 * σ(DMOS(i))value represents a good threshold for defining an outlier point.

The outlier ratio represents the proportion of outliers in N number of samples. Thus, the binomial distribution could be used to characterize the outlier ratio. The outlier ratio is represented by a distribution of proportions [1] characterized by the mean (12) and standard deviation (13)

$$p = \frac{TotalNoOutliers}{N} \qquad (12)$$

$$\sigma_p = \sqrt{\frac{p*(1-p)}{N}} \qquad (13)$$

For N>30, the binomial distribution, which characterizes the proportion p, can be approximated with the Gaussian distribution . Therefore, the 95% confidence interval (CI) of the outlier ratio is given by (14)

CI = $\pm 2 * \sigma_p$ (14)

NOTE. If the mean is based on less than thirty samples (ie., N < 30), then the Gaussian distribution must be replaced the appropriate Student's t distribution, depending on the specific number of samples in the mean [1].

8.5. Statistical Significance of the Results

8.5.1. Significance of the Difference between the Correlation Coefficients

The test is based on the assumption that the normal distribution is a good fit for the video quality scores' populations. The statistical significance test for the difference between the correlation coefficients uses the H_0 hypothesis that assumes that there is no significant difference between correlation coefficients. The H_1 hypothesis considers that the difference is significant, although not specifying better or worse.

The test uses the Fisher-z transformation (3) [1]. The normally distributed statistic (15) is determined for each comparison and evaluated against the 95% t-Student value for the two–tail test, which is the tabulated value t(0.05) =1.96.

$$Z_N = \frac{z1 - z2 - \mu_{(z1-z2)}}{\sigma_{(z1-z2)}} \qquad (15)$$

where $\mu_{(z1-z2)} = 0$ (16)

and Ed. Note: eqn missing: copy from ver 13 or hdtv testplan (17)

σ_{z1} and σ_{z2} represent the standard deviation of the Fisher-z statistic for each of the compared correlation coefficients. The mean Ed. Note: investigate this error: see ver 1.13 ask D.Hands **Error! Reference source not found.**(16) is set to zero due to the H_0 hypothesis and the standard deviation of the difference metric z1-z2 is defined by (17). The standard deviation of the Fisher-z statistic is given by **Error! Reference source not found.**(18):

$$\sigma_z = \sqrt{1/(N-3)} \quad (18)$$

where N represents the total number of samples used for the calculation of each of the two correlation coefficients.

8.5.2. Significance of the Difference between the Root Mean Square Errors

Considering the same assumption that the two populations are normally distributed, the comparison procedure is similarly to the one used for the correlation coefficients. The H_0 hypothesis considers that there is no difference between rmse values. The alternative H_1 hypothesis is assuming that the lower prediction error value is statistically significantly lower. The statistic defined by (19) has a F-distribution with n1 and n2 degrees of freedom [1].

$$\zeta = \frac{rmse_{\max}}{rmse_{\min}} \quad (19)$$

rmse,max is the highest rmse and rmse,min is the lowest rmse involved in the comparison. The ζ statistic is evaluated against the tabulated value F(0.05, n1, n2) that ensures 95% significance level. The n1 and n2 degrees of freedom are given by N1-1, respectively and N2-1, with N1 and N2 representing the total number of samples for the compared average prediction errors.

8.5.3. Significance of the Difference between the Outlier Ratios

As mentioned in paragraph 8.3.3, the outlier ratio could be described by a binomial distribution of parameters (p, 1-p), where p is defined by (12). In this case P is equivalent with the probability of success of the binomial distribution.

The distribution of differences of proportions from two binomially distributed populations with parameters (p1, 1-p1) and (p2, 1-p2) (where p1 and p2 correspond to the two compared outlier ratios) is approximated by a normal distribution for N1, N2 >30, with the mean:

$$\mu_{(p1-p2)} = \mu(p1) - \mu(p2) = p1 - p2 = 0 \quad (20)$$

and standard deviation:

$$\sigma_{p1-p2} = \sqrt{\frac{\sigma(p1)^2}{N1} + \frac{\sigma(p2)^2}{N2}} \quad (21)$$

The null hypothesis in this case considers that there is no difference between the population parameters p1 and p2, respectively p1=p2. Therefore, the mean (20) is zero and the standard distribution (21) becomes equation (22)

$$\sigma_{p1-p2} = \sqrt{p*(1-p)*(\frac{1}{N1}+\frac{1}{N2})} \quad (22)($$

where N1 and N2 represent the total number of samples of the compared outlier ratios p1 versus p2. The variable p is defined by 23

$$p = \frac{N1 * p1 + N2 * p2}{N1 + N2} \qquad (23)$$

References

[1] M. Spiegel, “Theory and problems of statistics”, McGraw Hill, 1998.

9. Recommendation

The VQEG will recommend methods of objective video quality assessment based on the primary evaluation metrics defined in Section 8. The Study Groups involved (ITU-T SG 12, ITU-T SG 9, and ITU-R SG 6) will make the final decision(s) on ITU Recommendations.

10. Bibliography

10. Bibliography

- VQEG Phase I final report.
- VQEG Phase I Objective Test Plan.
- VQEG Phase I Subjective Test Plan.
- VQEG FR-TV Phase II Test Plan.
- Vector quantization and signal compression, by A. Gersho and R. M. Gray. Kluwer Academic Publisher, SECS159, 0-7923-9181-0.
- Recommendation ITU-R BT.500-10.
- document 10-11Q/TEMP/28-R1.
- RR/NR-TV Test Plan

ANNEX I
INSTRUCTIONS TO THE SUBJECTS

Notes: The items in parentheses are generic sections for a Subject Instructions Template. They would be removed from the final text. Also, the instructions are written so they would be read by the experimenter to the participant(s).

(*greeting*) Thanks for coming in today to participate in our study. The study's about the quality of video images; it's being sponsored and conducted by companies that are building the next generation of video transmission and display systems. These companies are interested in what looks good to you, the potential user of next-generation devices.

(*vision tests*) Before we get started, we'd like to check your vision in two tests, one for acuity and one for color vision. (*These tests will probably differ for the different labs, so one common set of instructions is not possible.*)

(*overview of task: watch, then rate*) What we're going to ask you to do is to watch a number of short video sequences to judge each of them for "quality" -- we'll say more in a minute about what we mean by "quality." These videos have been processed by different systems, so they may or may not look different to you. We'll ask you to rate the quality of each one after you've seen it.

(*physical setup*) When we get started with the study, we'd like you to sit here (point) and the videos will be displayed on the screen there. You can move around some to stay comfortable, but we'd like you to keep your head reasonably close to this position indicated by this mark (point to mark on table, floor, wall, etc.). This is because the videos might look a little different from different positions, and we'd like everyone to judge the videos from about the same position. I (the experimenter) will be over there (point).

(*room & lighting explanation, if necessary*) The room we show the videos in, and the lighting, may seem unusual. They're built to satisfy international standards for testing video systems.

(*presentation timing and order; number of trials, blocks*) Each video will be (*insert number*) seconds (minutes) long. You will then have a short time to make your judgment of the video's quality and indicate your rating. At first, the time for making your rating may seem too short, but soon you will get used to the pace and it will seem more comfortable. (*insert number*) video sequences will be presented for your rating, then we'll have a break. Then there will be another similar session. All our judges make it through these sessions just fine.

(*what you do: judging -- what to look for*) Your task is to judge the quality of each image -- not the content of the image, but how well the system displays that content for you. The images come in three different sizes; how you judge image quality for the different sizes is up to you. There is no right answer in this task; just rely on your own taste and judgment.

(*what you do: rating scale; how to respond, assuming presentation on a PC*) After judging the quality of an image, please rate the quality of the image. Here is the rating scale we'd like you to use (*also have a printed version, either hardcopy or electronic*):

5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

Please indicate your rating by pushing the appropriate numeric key on the keyboard (button on the screen). If you missed the scene and have to see it again, press the XXX key. If you push the wrong key and need to change your answer, press the YYY key to erase the rating; then enter your new rating. [Note, this assumes that a program exists to put a graphical user interface (GUI) on the computer screen between video presentations. It should feed back the most recent rating that the subject had input, should have a "next video" button and an "erase rating" button. It should also show how far along in the sequence of videos the session is at present. The program that randomly chooses videos for presentation, records the data, and contains the GUI, should be written in a language that is compatible with the most commonly used computers.]

(*practice trials: these should include the different size formats and should cover the range of likely quality*)
Now we will present a few practice videos so you can get a feel for the setup and how to make your ratings. Also, you'll get a sense of what the videos are going to be like, and what the pace of the experiment is like; it may seem a little fast at first, but you get used to it.

(*questions*) Do you have any questions before we begin?

(*subject consent form, if applicable; following is an example*)
The Multimedia Quality Experiment is being conducted at the (*name of your lab*) lab. The purpose, procedure, and risks of participating in the Multimedia Quality Experiment have been explained to me. I voluntarily agree to participate in this experiment. I understand that I may ask questions, and that I have the right to withdraw from the experiment at any time. I also understand that (*name of lab*) lab may exclude me from the experiment at any time. I understand that any data I contribute to this experiment will not be identified with me personally, but will only be reported as a statistical average.

Signature of participant Signature of experimenter
Name of participant Date Name of experimenter

ANNEX II
EXAMPLE EXCEL SPREADSHEET

lab	test	type	subject #	month	day	Year	session	resolution	rate	age	gender	order	scene	hrc	acr score
ntia	mm1	compression	1000	10	3	2005	1	vga	30	47	m	1	susie	hrc1	4
ntia	mm1	compression	1000	10	3	2005	1	vga	30	47	m	1	susie	hrc2	2
ntia	mm1	compression	1000	10	3	2005	1	vga	30	47	m	1	susie	hrc3	1
ntia	mm1	compression	1000	10	3	2005	1	vga	30	47	m	1	susie	reference	5
ntt	mm2	robust	2003	10	18	2005	2	cif	25	38	f	2	calmob	pktloss1	1
ntt	mm2	robust	2003	10	18	2005	2	cif	25	38	f	2	calmob	pktloss2	2
ntt	mm2	robust	2003	10	18	2005	2	cif	25	38	f	2	calmob	biterror1	1
ntt	mm2	robust	2003	10	18	2005	2	cif	25	38	f	2	calmob	biterror2	3
ntt	mm2	robust	2003	10	18	2005	2	cif	25	38	f	2	calmob	reference	4
yonsei	mm3	livenetwork	3018	10	21	2005	1	qcif	30	27	m	1	football	ip1	4
yonsei	mm3	livenetwork	3018	10	21	2005	1	qcif	30	27	m	1	football	ip2	3
yonsei	mm3	livenetwork	3018	10	21	2005	1	qcif	30	27	m	1	football	reference	5

ANNEX III
BACKGROUND AND GUIDELINES ON TRANSMISSION ERRORS

Introduction

Transmission errors should be created to emulate a real video service to ensure that the proponents’ models are trained and tested with realistic video material. There are three major types of transmissions used for video services today:

Packet switched radio network

This kind of transmission is typical for video service in so called 3G mobile networks. Examples of services are video streaming service, such as streaming news and sports video clips to a mobile phone, mobile TV and video shared in parallel with a normal speech call. The transmission errors are characterized by packet delays, which can be in the range of 10 ms to several seconds, and packet losses that could be massive (ranging from no losses to 50%). The packet delay might case packet to be dropped by the video client because they are received too late, or causing the buffer to run empty in the client. If the buffer runs empty it causes frame freezing (not currently included in the test plan). Packet losses will cause image artifacts in the video and possibly video frame jitter.

Transport errors should be created by running a video streaming service over a real-time link simulator, where packets can be delayed in with a delay pattern as in a typical mobile radio network. The link simulator should also be able to drop packets. Typically packets are dropped when a buffer somewhere in the network is full, and new packets arriving at the buffer are dropped. This situation can occur when the link to the mobile has a lower bandwidth than required by the video stream.

Packet losses are normally bursty, causing the video quality to vary a lot. A short video streaming sequence might even be played with best possible quality, even if the bandwidth is limited. Therefore, video streaming sequences should be longer than 8 to 10 seconds. An 8 to 10 seconds video clip can be cut out from the longer video sequence, from the part where the transmission errors have caused the desired video quality degradation. Note also that the packet size is related to video quality degradation for a certain packet loss ratio.

Wireline Internet

Typical service is video streaming to a PC with fixed Internet connection. Network congestion causes packet losses in the network switches. Random and periodic packet loss *can* occur due to faulty equipment. However, bursty packet losses are the most common loss type. Packet loss ratio is in the range from 0% to 50%. Packets are delayed with delay ranging from 2 ms to several seconds.

Transmission errors should be created with a bursty packet loss model, as expected for Internet bottlenecks.

Circuit switched radio network

A typical service is video telephony. The transmission errors are characterized by bursty loss of data. Chunks of data (packets are not used in circuit switched transmission) are lost. Block (radio blocks) error rates are typically ranging from 0.2% to 5% when averaged over a couple of seconds. Momentarily the error rate can be 100%.

Transport errors should be created by applying error masks on a bit stream. Errors in the mask should have a bursty pattern to mimic a radio interface, such as a WCDMA 64 kbps circuit switched radio bearer. Note that the size of the blocks over the simulated transport link is correlated to video quality. Within limits the larger block size the better quality for a certain block error rate. Block size can for example be 160 or 300 bytes.

Summary of transmission error simulators

Transport link type	Model	Typical error rates
Packet switched radio network	Link simulator delaying and dropping packages. Delay based on bit/block errors over a radio link. Drop based on overflow in a network buffer due to low bandwidth. The packet delay should be introduced as in a real radio network. Typical target networks are GSM, WCDMA or CDMA radio networks.	Packet delay in the range from 10 ms to 5 s. Bursty packet loss in the range 0% to 50% (for an average over one or a few seconds)

Wireline Internet	Link simulator dropping packets, as expected when the buffer in an Internet switch overruns. As described in literature packet losses can be modeled with a Markov chain with two states representing no loss/loss. See for example [2] below for example of link model.	Packet delay in the range 2 ms to 5 seconds (high value when for example a satellite link Is used). Bursty packet losses in range from 0% to 50%
Circuit switched radio network	Link simulator dropping chunks of data. Alternative is to apply an error mask to a bit stream. The error mask should have been made by simulating a radio link. The bit stream should be a H.223 bit stream, which is used for video telephony. See reference [1] below.	Typical block error rates (over a radio link) are ranging from 0.2% to 5% (average over a couple of seconds)

Table 1 Summary of transmission error simulators

Note: A video service might use multiple transport links. Thus, it is possible to use a combination of simulators to get realistic transport errors. A combination of wireline and wireless IP link simulators can be used to simulate a service, such as video streaming over Internet and a radio link

Logging parameters

Table 2 below describes the parameters to be logged when introducing transmission errors with a simulator. All parameters are required, except those explicitly described as “optional”.

Logging Category	**Logging details**
Simulator description	• Type of simulator (packet simulator, circuit switched simulator) • Simulated network (GSM/WCDMA/CDMA/Wireline Internet) • Version of simulator • Hardware/system it was run on • General description of how transport errors are introduced
Input parameters to simulator (depends on type of simulator. Only examples given here)	• Bandwidth limit • System buffer size • Block or bit error rates • Latency
Output parameters from simulator	**Packet simulator (wireline and wireless)** • Average packet loss ratio in percent • Length of window to calculate packet loss ratio • Number of total packets • Average packet delay in ms • Sequence number of lost packets (optional) • Distribution of packet delay (optional) • Packet size distribution (optional) **Circuit switched simulator** • Average block and/or bit error rate (BLER/BER) • Block size over transport link • Maximum block error rate
Decoder	• General description of decoder (name, vendor) • Version of decoder • Post filter used (if known) • Error concealment used (if known)

Table 2 Parameters to be logged when introducing transmission errors

.

References

[1] ITU-T Recommendation H.223, Multiplexing protocol for low bit rate multimedia communication.

[2] B. Girod, K. Stuhlmüller, M. Link and U. Horn. “Packet Loss Resilient Internet Video Streaming”. SPIE Visual Communications and Image Processing 99, January 1999, San Jose, CA

ANNEX IV

Fee and Conditions

VQEG intends to enable everybody who is interested in contributing to the work as a proponent to participate in the assessment of video quality metrics and to do so even if the proponent is not able to finance more than the regular participation fee as laid forth in this Annex (see below for details of the fees). On the other side VQEG will produce video databases which are extremely valuable to those developing video metrics. An organization cannot get access to these databases without, at a minimum, substantially participating in the VQEG work. VQEG has decided that all proponents must provide at least one database (or a comparable contribution) which fulfils the requirements laid out in Sections 4, 5 and 6 of this testplan in order to gain access to the subjective databases produced in the Multimedia tests. A comparable contribution should be agreed by the other proponents and could include such things as providing test sequences and/or running HRCs. If an organization has no facilities to create such a database by itself, it may contract a recognized subjective test facility to do so on its behalf. If an organization is lacking the financial resources to fulfil this obligation, it can ask other proponents or the ILGs to run its model on the VQEG databases. In this case the party will not be granted direct access to the video databases, but the party is still able to participate in the assessment of their models after paying the regular participation fee to the Independent Lab Group (ILG).

ANNEX V

(Owner: Kjell Brunnstrom)

Subjective Test Software Package and Required Computer Specification

Introduction

The program **acrvqwin-beta2** is a program to run subjective experiments for video quality in Windows environment, using the ACR-method It is designed to present the video on a computer screen, using the picture sizes QCIF, CIF and VGA, synchronized with the refresh of the display. This is implemented using DirectX.

System Requirements

The program uses a multithread function, so it requires a Intel based (AMD has not been tested) computer with two processors or a processor that supports multithreading. The CPU speed should be at least 2 GHz. The graphics card should support yuv to rgb conversion and have at least 256 MByte of video memory. The system should have primary memory of at least 512 MB. The program can be run on Windows 2000 and Windows XP.

Using the program

Installation and preparation

After downloading and extracting the files, the file **acrvqwin-beta2.exe** has to be placed in the same library as the five dll files: libmatlb, libmcc,libmx, libmat and libut libut (should be obtained together i.e in the same distribution, with the executable file, so versions of the dlls match the executable). The setup-file (setup.seq) can be placed anywhere as the program will let you specify the path at the start of the experiment. To be able to run the experiment program a setup-file, see also below, has to be used in order to set some of the parameters in the program and also specify the path and name of the avi-files that should be used during the experiment. To specify your avi-files you write the filenames just below the path to the directory they are in. Each filename must be written on a separate row, as the program will treat each line of text as a filename.

Running the program

The program is started in the normal way in Windows e.g double-clicking on the icon. The program will initially present a dialog box for selecting setup-file (this makes it possible to have different setup-files for various experiment stored at the same time) to use for the experiment.

After choosing the setup-file the program will enter the experiment environment, the screen will have the same grey-level as has been specified with the parameter Background. To start the first sequence you can either press the s button on the keyboard or the left button on the mouse and the video sequence will be displayed at the centre of the screen. When the sequence is finished a dialog box will appear in order to get the subjects rating of the perceived video quality. This procedure will continue until the experiment is done and the program will automatically exit. You can at any time quit the current experiment by pressing the Esc key.

The order of the sequences does not match the order of the filenames in the setup-file. The program randomises the order at the start of the program. The random numbers are drawn from a uniform distribution. The results are stored in two output-files that have the same name as the parameter SubjectId with an extension of either log or dat. The log-file shows the results in the exact order as the sequences have been shown. The dat-file sorts the data regardless of the randomisation that has been done by the program, in order to make it easier to analyse the results from a statistical point of view. The data in the dat-file are stored in exactly the same order as in the setup-file.

If the Esc key is being pressed during an experiment the program will store the necessary data in a .svd file to be able to restore the experiment later. The .svd file will be created in the folder specified by the parameter Path and have the filename specified by the parameter SubjectId. Next time the program starts it will ask you if you want to restore the aborted experiment and if selecting yes the experiment will continue from where it was aborted. If you don’t want the program to keep asking you if you want to restore an experiment, simply delete the .svd file.

For mixing 25-fps and 30-fps sequences, the parameter ScreenRefreshFreq must be set to 60 Hz. The program will then run the 30-fps sequences as normal and try to run the 25-fps sequences as close to 25 fps as possible. As 60 is evenly divided by 30, each frame will be displayed two screen refresh cycles and the timing between the frames will be constant in the 30-fps sequences. This can’t be done with the 25-fps sequences and the solution applied in this program is to alternate between displaying a frame 2 or 3 screen refresh cycles. The program uses the serie: 2, 3, 2, 3, 2. In order to distinguish between the differences in framerate of the sequences the framerate parameter in the avi-file infoheader must be valid.

Setup-file parameters

PixelDepth: The only supported mode for this version is 16 bit with the format YUV 4:2:2

ScreenHeight: The height of the screen measured in pixels.

ScreenWidth: The width of the screen measured in pixels.

ScreenRefreshFreq: The update frequency of the screen.

VideoFrameRate: The number of frames shown per second.

Background: The grey-level of the background in the experiment environment, where 0 is black and 255 is white.

NumberOfSeries: This parameter is used to set how many times each sequence will be played.

UseImageRatings: Decides whether to use an image or text to describe the different ratings for the subject.

ImagePathRatings: Specifies the path to the image if the parameter UseImageRatings is set to true.

ImagePathNoRating: Specifies the path to the image to be shown if UseImageRatings is set to true and the subject has missed to rate the sequence.

Rating5-Rating1: The text used to describe the different ratings if UseImageRatings is set to false.

NoRating: This text is shown if the subject misses to rate the sequence and UseImageRat-ings is set to false.

SubjectId: Specifies the filename for the .dat and .log file.

NoOfImagesInSeq: The number of frames that will be displayed from the avi-file. This parameter makes it possible to decide the length of the sequences from the setup-file.

Path: Specifies the path to the directory holding the avi-files to be used. The .dat and .log file will be placed here.

Example of a setup-file

```
[Sequence]
PixelDepth = 16
ScreenHeight = 1024
ScreenWidth = 1280
ScreenRefreshFreq = 60
VideoFrameRate = 30
Background = 128
NumberOfSeries =3
UseImageRatings = FALSE
ImagePathRatings = "C:\MyPath\mybitmap1.bmp"
ImagePathNoRating = "C:\MyPath\mybitmap2.bmp"
Rating5 = "Excellent"
Rating4 = "Good"
Rating3 = "Fair"
Rating2 = "Poor"
Rating1 = "Bad"
NoRating = "The sequence has not been rated"
SubjectId = "Subject"
NoOfImagesInSequence = 300
Path = "C:\MyPath"
myavi1.avi
myavi2.avi
myavi3.avi
```

Note: Experimenter is required to change the ”SubjectId” for each subject in the MM test.

Term of usage

This software is provided as is. No warranty or support is given. It may be used within the VQEG for research purposes. Copyright 2005 Acreo AB, Sweden.

List of preferred LCD Monitors for use in subjective tests:

TCO ’06:

BenQ FP241W, FP241WZ, FP241VW Q24W5

EIZO FlexScan S2110W ColorEdge CE210W S2110W

Samsung 215TW DP21**

TCO ‘03

ANNEX VI

(Owner: Quan Huynh-Thu)

Method for Post-Experiment Screening of Subjects

Method

The rejection criterion verifies the level of consistency of the raw scores of one viewer according to the corresponding average raw scores over all viewers. Decision is made using correlation coefficient. Analysis per PVS and per HRC is performed for decision.

Linear Pearson correlation coefficient per PVS for one viewer vs. all viewers:

$$r1(x,y)=\frac{\sum_{i=1}^{n}x_iy_i-\frac{\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{n}}{\sqrt{\left(\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}\right)\left(\sum_{i=1}^{n}y_i^2-\frac{\left(\sum_{i=1}^{n}y_i\right)^2}{n}\right)}}$$

Where

xi = MOS of all viewers per PVS

yi = individual score of one viewer for the corresponding PVS

n = number of PVSs

i = PVS index.

Linear Pearson correlation coefficient per HRC for one viewer vs. all viewers:

$$r2(x,y)=\frac{\sum_{i=1}^{n}x_iy_i-\frac{\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{n}}{\sqrt{\left(\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}\right)\left(\sum_{i=1}^{n}y_i^2-\frac{\left(\sum_{i=1}^{n}y_i\right)^2}{n}\right)}}$$

Where

xi = condition MOS of all viewers per HRC, i.e. condition MOS is the average value across all PVSs from the same HRC

yi = individual condition MOS of one viewer for the corresponding HRC

n = number of HRCs

i = HRC index

Rejection criteria

1. Calculate r1 and r2 for each viewer
2. Exclude a viewer if (r1<0.75 AND r2 <0.8) for that subject

Note: The reason for using analysis per HRC (r2) is that a subject can have an individual content preference that is different from other viewers, making r1 to decrease, although this subject may have voted consistently. Analysis per HRC averages out individual's content preference and check consistency across error conditions.

xi = mean score of all observers for the PVS

yi = individual score of one observer for the corresponding PVS

n = number of PVSs

i = PVS index

R(xi or yi) is the ranking order

Final rejection criteria for discarding an observer of a test

The Spearman rank and Pearson correlations are carried out to discard observer(s) according to the following conditions:ANNEX VII

Encrypted Source Code Submitted to VQEG

Proponents are entitled to submit a file with encrypted source code along with their model's object code. This submission is not required but is offered in case there is a bug in the software that can be fixed without changing the algorithm. Normally, there would be no software updates possible after the submission of the object code.

In order for this option to be exercised the proponent must encrypt the source code with a readily available encryption program (see below for a freeware example) and send the password protected file to two ILG labs (CRC and Acreo). If it is determined by the proponent that a bug is present in the software, then the proponent must discuss the situation with the ILG Co-Chairs. If the Co-Chairs agree that a bug fix should be tried, then a procedure must be agreed to in order for the proponent to make the change to the code in the presence of the ILG member. This could be done in person or perhaps by telephone.

The proponent would make the change and the ILG member would verify that it was not an algorithm change. The code would be recompiled and tested in the presence of the ILG member. The revised code should be re-encrypted with a different password.

The encrypted file can be transported electronically or physically. It needs to be sent to both ILG contacts below:

ILG contacts:

;runnstrom Acreo Stockholm, Sweden +4686327732 Kjell.Brunnstrom@acreo.se	Filippo Speranza CRC Ottawa, Canada +1 613-998-7822 filippo.speranza@crc.ca

A good freeware encryption program:

Blowfish Advanced CS 2.57

http://www.hotpixel.net/software.html (click on Blowfish Advanced CS – Installer)

This software offers several encryption algorithms. The one that allows the largest key (448 bits) is Blowfish. It is also in German and English.

Source files should be zipped and then encrypted.

Other encryption programs can be used but if they are not free then the proponent is responsible for purchasing the program for the ILG if necessary.

ANNEX VIII

Approved Conversion Tools

Transformation of source test sequences to UYVY AVI files

Transformation of the source test sequences as described in chapter 6.1 (e.g., from Rec. 601 525-line to QCIF) shall be performed using the following tools:

Tool	Web Resources
Avisynth 2.5.5	http://www.avisynth.org/
AviSynth Filters (RawSource, KernelDeint)	http://www.avisynth.org/warpenterprises/
VirtualDub 1.6.11	http://www.virtualdub.org/
ffdshow 20050303	http://ffdshow.sourceforge.net/

For loading a sequence in raw uncompressed UYVY 4:2:2 format (as the VQEG format, see FRTV Phase I video file examples at ftp://vqeg.its.bldrdoc.gov/SDTV/VQEG_PhaseI/TestSequences/) the “RawSource” plugin for AviSynth is used. The sequences are opened with the "UYVY" mode.

De-interlacing will be performed according chapter 6.1.4 using the de-interlacing AviSynth plugin “KernelDeint”. If the de-interlacing using KernelDeint results in source sequence that has serious artifacts, the Blendfield or Autodeint may be used as alternative methods for de-interlacing.

Cropping and Rescaling is done according chapter 6.1.5, using AviSynth build-in “LanczosResize” function.

The following script can be used as a template for the the transformations. See the Avisynth documentation for detailed information on what each command does.

```
# This function limits the memory space (in MB) used by the script (should be used)
SetMemoryMax(100)

# The RawSource plugin loads raw YUV 4:2:2, 4:2:0, etc. videos
RawSource("c:\calmob.yuv",720,486,"UYVY")

# AssumeFPS() forces the framerate of the video.
AssumeFPS(30)

# Deinterlacing
KernelDeint(order=0)

# Cropping
crop(36,3,-38,-3)

# Resizing
LanczosResize(176,144)
```

The scripts are saved using the extension “.avs” and then loaded into VirtualDub (File->OpenVideoFile).

Within VirtualDub, video sequences will be saved to AVI files using Video Compression option (Video ->Compressor) "ffdshow Video Codec", configured with the "Uncompressed" decoder and the UYVY color space. For the Color Depth (Video->Color Depth), the setting “4:2:2 YCbCr (UYUV)” is used as output format. The processing mode (Video->) is set to “Full processing mode”.

The tools and scripts may be downloaded at:
Tools/Approved/ directory on ftp://mmpretest@132.163.64.167/

AviSynth Scripts for the common transformations

PAL to QCIF

```
SetMemoryMax(100)
RawSource("c:\calmob.yuv",720,576,"UYVY")
AssumeFPS(25)
KernelDeint(order=1)
crop(38,0,644,576)
LanczosResize(176,144)
```

PAL to CIF

```
SetMemoryMax(100)
RawSource("c:\calmob.yuv",720,576,"UYVY")
AssumeFPS(25)
KernelDeint(order=1)
crop(8,0,702,576)
LanczosResize(352,288)
```

PAL to VGA

```
SetMemoryMax(100)
RawSource("c:\calmob.yuv",720,576,"UYVY")
AssumeFPS(25)
KernelDeint(order=1)
crop(38,0,644,576)
LanczosResize(640,480)
```

NTSC to QCIF

```
SetMemoryMax(100)
RawSource("c:\calmob.yuv",720,486,"UYVY")
AssumeFPS(30)
KernelDeint(order=0)
crop(36,3,-38,-3)
LanczosResize(176,144)
```

NTSC to CIF

```
SetMemoryMax(100)
RawSource("c:\calmob.yuv",720,486,"UYVY")
AssumeFPS(30)
KernelDeint(order=0)
crop(36,3,-38,-3)
LanczosResize(352,288)
```

NTSC to VGA

```
SetMemoryMax(100)
RawSource("c:\calmob.yuv",720,486,"UYVY")
AssumeFPS(30)
KernelDeint(order=0)
crop(8,3,704,480)
LanczosResize(640,480)
```

HD720 to QCIF

```
SetMemoryMax(100)
RawSource("c:\calmob.yuv",1280,720,"UYVY")
AssumeFPS(30)
crop(200,0,880,720)
LanczosResize(176,144)
```

HD720 to CIF

```
SetMemoryMax(100)
RawSource("c:\calmob.yuv",1280,720,"UYVY")
AssumeFPS(30)
crop(200,0,880,720)
```

```
LanczosResize(352,288)
```

HD720 to VGA

```
SetMemoryMax(100)
RawSource("c:\calmob.yuv",1280,720,"UYVY")
AssumeFPS(30)
crop(160,0,960,720)
LanczosResize(640,480)
```

HD1080 to QCIF

```
SetMemoryMax(100)
RawSource("c:\calmob.yuv",1920,1080,"UYVY")
AssumeFPS(30)
KernelDeint(order=1)
crop(300,0,1320,1080)
LanczosResize(176,144)
```

HD1080 to CIF

```
SetMemoryMax(100)
RawSource("c:\calmob.yuv",1920,1080,"UYVY")
AssumeFPS(30)
KernelDeint(order=1)
crop(300,0,1320,1080)
LanczosResize(352,288)
```

HD1080 to VGA

```
SetMemoryMax(100)
RawSource("c:\calmob.yuv",1920,1080,"UYVY")
AssumeFPS(30)
KernelDeint(order=1)
crop(240,0,1440,1080)
LanczosResize(640,480)
```

UYVY Raw to UYVY AVI

These tools convert raw uncompressed UYVY 4:2:2 files to uncompressed UYVY AVI files. The following tools are available:

AviSynth,VirtualDub, ffdshow

Use the procedure described in “Transformation of source test sequences to UYVY AVI files” without performing the cropping and resizing step within the AviSynth script.

The following tools are not approved in the MM test plan:

SwissQual

Tools/SwissQual/Conversion_Tools/UYVYIntoAVI/ directory on ftp://mmpretest@132.163.64.167/

NTIA

The “Laboratory VQM” tool available from NTIA has this capability. The UYVY AVI output file is limited to 2GB (AVI Version 1.0). This tool may be obtained at:

http://www.its.bldrdoc.gov/n3/video/vqmsoftware.htm

UYVY Raw to RGB AVI

These tools convert raw uncompressed UYVY files to uncompressed RGB AVI files. The following tools are available:

AviSynth,VirtualDub

Use the procedure described in "Transformation of source test sequences to UYVY AVI files" without performing the cropping and resizing step within the AviSynth script and save the file as described next.

Within VirtualDub, video sequences will be saved to AVI files using Video Compression option (Video ->Compressor) "Uncompressed RGB/YCbCr". For the Color Depth (Video->Color Depth), the setting "24 bit RGB (888)" is used as output format. The processing mode (Video->) is set to "Full processing mode".

The following tool is not approved in the MM test plan:

SwissQual
Tools/SwissQual/Conversion_Tools/UYVYToAVI/ directory on ftp://mmpretest@132.163.64.167/

RGB AVI to UYVY AVI

These tools convert uncompressed RGB AVI files to uncompressed UYVY AVI files using the agreed color space conversion in section 6.1.3. The following tools are available:

VirtualDub, ffdshow
The RGB Avi file will be loaded into VirtualDub (File->OpenVideoFile) and then saved according the following steps.

Within VirtualDub, video sequences will be saved to AVI files using Video Compression option (Video ->Compressor) "ffdshow Video Codec", configured with the "Uncompressed" decoder and the UYVY color space. For the Color Depth (Video->Color Depth), the setting "4:2:2 YCbCr (UYUV)" is used as output format. The processing mode (Video->) is set to "Full processing mode".

The following tool is not approved in the MM test plan:

SwissQual
Tools/SwissQual/Conversion_Tools/RGBtoYUV/ directory on ftp://mmpretest@132.163.64.167/

Processing and Editing Sequences

Two capture methods have been approved and descried in chapter 6.3.9. The tools will intercept RGB video being sent to a computer monitor and saves this video to an uncompressed AVI file.

SwissQual
Free demonstration system for capturing streaming media (the full version can be bought from SwissQual):

Tools/SwissQual/Capturing/ directory on on ftp://mmpretest@132.163.64.167/

The full system is capable of capturing QuickTime, Real Media, and Windows Media streams and includes a Media Server. The tool captures the image shown on the monitor, or more precisely from the memory (space) that Windows sends to the graphics card. A frame is captured only if new data is available. That is why there is a file created that contains the variable frame rate information. Please see:

Tools/SwissQual/VQEG_SQ_Capturing_Sequences_V2.pdf on ftp://mmpretest@132.163.64.167/

To meet the desired VQEG file format definitions, the captured video data must be converted to constant frame rate and from RGB to YUV. These conversions are already implemented in the full version of the capturing tools.

Please contact Pero Juric (pero.juric@swissqual.com) for any further information.

NTT

PIFREC 1.0 (Lossless PC Video & Voice Recorder). The PC capture system uses a capture board to receive the signals passed from a PC to its monitor, without adding any processing load to the PC, and stores them while retaining high video quality.

More Information about PIFREC can be found at: http://www.ntt-at.com/products_e/pifrec/index.html

Please contact Mr. Takawo Adachi (at-hama@ntt-at.co.jp) for any further information.

The following tool is not approved in the MM test plan:

Linux

These capture tools consist of a set of Linux system libraries that hijack the output of various players. The captured video data as well as a file with time stamps is saved to the disk. The tools developed by Marcus Barkowski from the University of Erlangen are available at this location:

Tools/playout_capure.tgz on ftp://mmpretest@132.163.64.167/

Calibration

These tools verify that the processed video sequences meet the calibration limits (e.g., temporal shifts, spatial shifts, spatial scaling, gain and level offset) specified by the test plan.

The following tool is not approved in the MM test plan:

NTIA

The “Laboratory VQM” tool available from NTIA has the ability to estimate spatial registration, temporal registration, and gain/level offset. This tool may be obtained at:

http://www.its.bldrdoc.gov/n3/video/vqmsoftware.htm

UYVY Decoder to UYVY Raw / UYVY AVI

This tool intercepts UYVY video from a video decoder (before being converted to RGB for monitor display) and saves this video to a UYVY Raw or UYVY AVI file.

The following tools are not approved in the MM test plan:

The following video decoders output some form of uncompressed YUV video and hence RGB color space conversions can be avoided:

Tool	Web Resources	Output Format
MPEG-2 Reference Codec	www.mpeg.org	Outputs YUV 4:2:2 and YUV 4:2:0

MainConcept MPEG-2 Codec	www.mainconcept.com	Since the MainConcept decoder is a DirectShow filter, its output can be an AVI file. DirectShow filter outputs a range of YUV formats in addition to RGB.
H.264 (AVC) Joint Video Team (JVT) Reference Codec		Outputs YUV 4:2:0
MainConcept H.264 codec	www.mainconcept.com	Outputs a range of YUV formats

Notes

VirtualDub is also capable of saving UYVY AVI files without using ffdshow. The following settings are used to get correct results.

The following procedure is not approved in the MM test plan:

Within VirtualDub, video sequences will be saved to AVI files using Video Compression option (Video->Compressor) "Uncompressed RGB/YCbCr". For the Color Depth (Video->Color Depth), the setting “4:2:2 YcbCr (UYUV)” is used as output format. The processing mode (Video->) is set to “Full processing mode”.

ANNEX IX

Definition and Calculating Gain and Offset in PVSs

Before computing luma (Y) gain and level offset, the original and processed video sequences should be temporally aligned. One delay for the entire video sequence may be sufficient for these purposes. Once the video sequences have been temporally aligned, perform the following steps.

Horizontally and vertically cropped pixels should be discarded from both the original and processed video sequences.

The Y planes will be spatially sub-sampled both vertically and horizontally by the following factors: 16 for VGA, 8 for CIF and 4 for QCIF. This spatial sub-sampling is computed by averaging the Y samples for each block of video (e.g., for VGA one Y sample is computed for each 16 x 16 block of video). Spatial sub-sampling should minimize the impact of distortions and small spatial shifts (e.g., 1 pixel) on the Y gain and level offset calculations.

The gain (*g*) and level offset (*l*) are computed according to the following model:

$$\underline{P} = g\underline{O} + l \qquad (1)$$

where $\underline{O}$ is a column vector containing values from the sub-sampled original Y video sequence, $\underline{P}$ is a column vector containing values from the sub-sampled processed Y video sequence, and equation (1) may either be solved simultaneously using all frames, or individually for each frame using least squares estimation. If the latter case is chosen, the individual frame results should be sorted and the median values will be used as the final estimates of gain and level offset.

Least square fitting is calculated according the following formula:

$$g = (R_{OP} - R_O R_P)/(R_{OO} - R_O R_O), \text{ and} \qquad (2)$$

$$l = R_P - g\, R_O \qquad (3)$$

where $R_{OP,}$ $\mathrm{R}_{OO,}$ R_O and R_P are:

$$R_{OP} = (1/N)\ \Sigma\ O(i)\ P(i) \qquad (4)$$

$$R_{OO} = (1/N)\ \Sigma\ [O(i)]^2 \qquad (5)$$

$$R_O = (1/N)\ \Sigma O(i) \qquad (6)$$

$$R_P = (1/N)\ \Sigma\ P(i) \qquad (7)$$