VQEG 3DTV Group

Test Plan for Evaluation and Specification of Viewing Conditions and Environmental Setup for 3D Video Quality Assessment

Draft Version 1.0, 2012

Editors' note:	A blue highlight will occur before proposals that require explanation (e.g., text to be deleted). Tracked changes are used to identify proposals that have not been agreed upon. Changes that have been agreed upon are not marked. The wording of agreements made occurred during audio calls may need to be adjusted.

Contact: Marcus Barkowsky Tel: +33 2 40 68 30 22 Email: Marcus.Barkowsky@univ-nantes.fr
Second author
Third author

Editorial History

Version	Date	Nature of the modification
1.0	November 14, 2012	Initial Draft, edited by Marcus Barkowsky

Table of Contents

List of Acronyms

ACR-HRR	Absolute Category Rating with Hidden Reference Removal
ANOVA	ANalysis Of VAriance
CCIR	Comite Consultatif International des Radiocommunications
CODEC	Coder-Decoder
FR	Full Reference
GOP	Group of Pictures
HD	High Definition (television)
HRC	Hypothetical Reference Circuit
ITU	International Telecommunications Union
ITU-R	ITU Radiocommunications Standardization Sector
ITU-T	ITU Telecommunications Standardization Sector
MM	Multimedia
MOS	Mean Opinion Score
MPEG	Motion Pictures Expert Group
NR	No (or Zero) Reference
NTSC	National Television Standard Committee (60-Hz TV, used mainly in US and Canada)
PAL	Phase Alternating Line (50-Hz TV, used in Europe and elsewhere)
PVS	Processed Video Sequence
RR	Reduced Reference
SMPTE	Society of Motion Picture and Television Engineers
SRC	Source Reference Channel or Circuit
VQEG	Video Quality Experts Group

List of Definitions

<definitions on visual discomfort/fatigue etc. to be copied from GroTruQoE3D1>

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 subset 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.

Rewinding is defined as an event where the HRC playback jumps backwards in time. Rewinding can occur immediately after a pause. Given the reference sequence (A B C D E F G H I), two example processed sequence containing rewinding are (A B C D B C D E F) and (A B C C C C A B C). Rewinding can occur as a response to transmission error; for example, a video player encounters a transmission error, pauses while it conceals the error internally, and then resumes by playing video prior to the frame displayed when the transmission distortion was encountered. Rewinding is different from variable frame skipping because the subjects see the same content again and the motion is much more jumpy.

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.

Transmission errors are defined as any error resulting from sending the video data over a transmission channel. Examples of transmission errors are corrupted data (bit errors) and lost packets / lost frames. Such errors may be generated in live network conditions or through simulation.

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.

1. Introduction

This document defines evaluation tests aiming towards the standardization of the viewing environment for 3DTV visualization conducted by the Video Quality Experts Group (VQEG). It describes the subjective datasets that were used in this evaluation, the different kinds of setups that were used in the participating labs, the immediate results that were obtained from the observers and the statistical analysis that was performed on the acquired data. The text is based on discussions and decisions from meetings of the VQEG HDTV working group (HDTV) at the periodic face-to-face meetings as well as on conference calls and in email discussion.

The goal of the 3DViewing project is to analyze the impact of different viewing environments on the reliability of the obtained subjective evaluation, i.e. the obtained Mean Opinion Scores. The various influence factors related to the viewing environment such as the room illumination, the observer distance, the voting interface, the display technology, the instructions given to the observer, the training of the observers, etc. will be compared using a single subjective dataset in various scenarios in the participating labs.

Data analysis will be performed taking into consideration the differences between the labs and their setups as well as observer population differences in general and individual observer behavior if appropriate. Feedback from the observers will be considered.

The outcome of the project will be a list of restrictions for performing subjective experiments in 3DTV similar to the guidelines in ITU-T BT.500 that allow for reproducible subjective experiments across different labs. The validity of the recommendations is limited by the scope of the subjective experiment that was used in this project.

2. Overview: Expectations, Division of Labor and Ownership

2.1. Participation

The participation to the project is open for all subjective assessment labs.

2.2. Release of Video Sequences, Subjective Data and the Official Data Analysis

The video sequences are made publicly available by VQEG for download by any partner who wants to perform a subjective assessment in the context of this work or any associated study.

VQEG will make available each individual viewer's scores (i.e., including rejected viewers) and the associated data as provided by the participating labs. This viewer data will not include any indication of the viewer's identity, and should indicate the following data: (1) whether the viewer was rejected, (2) country of origin, which indicates frame rate that the viewer typically views, (3) gender (male or female), (4) age of viewer (rounded to the nearest decade would be fine), (5) experience of the viewer for 3D, (6) type of video that the viewer typically views (e.g., standard definition television, HDTV, IPTV, Video Conferencing, mobile TV, iPod, cell phone).

A standard questionnaire may be established that lists the questions asked of all viewers. This questionnaire may include other questions, and must take no longer than 5 minutes to complete. If possible, the questionnaire should be automated and (after translation) be used by all viewers.

2.3. Using the Data in Publications

Publications based on the collected data are highly appreciated. Appropriate citations and eventual co-authorship should be respected and welcomed.

3. Video sequence data set

The video sequence data set that shall be used in this project contains 3DTV sequences of high quality in Full-HD resolution per view. Adaptations to the display environment are allowed if necessary, i.e. for polarized line interlaced displays. Corresponding signal processing shall be documented by the lab conducting the experiment.

The degradations were chosen in such a way that they exhibit mostly degradations on the scale of “Video Quality”. In particular, degradations on “visual comfort” and “depth quality” and “depth quantity” were avoided whenever possible. Nevertheless, degradations on those scales may be perceived by the subjects as it is difficult to isolate the scales. For example, a strong coding degradation also decouples the degradations perceived in the left and right view and therefore introduce binocular rivalry leading to “visual discomfort”.

Detailed information about the video sequence data set can be found in the following publication:

Matthieu Urvoy, Jesús Gutiérrez, Marcus Barkowsky, Romain Cousseau, Yao Koudota, Vincent Ricordel, Patrick Le Callet and Narciso García, "NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences", Fourth International Workshop on Quality of Multimedia Experience, Yarra Valley, July 2012.

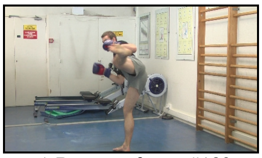
Available online at: http://hal.archives-ouvertes.fr/hal-00717865

3.1. Reference Source Sequences (SRC)

The following sequences have been selected: (more details to be added later)

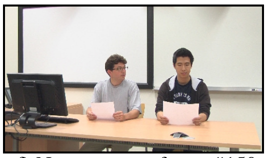

a) Barrier gate - frame #245

b) Basket - frame #285

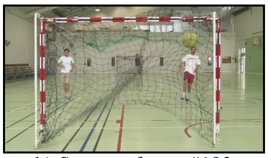

c) Boxers - frame #189


d) Hall - frame #200


e) Lab - frame #390


f) News report - frame #150


g) Phone call - frame #181


h) Soccer - frame #193


i) Tree branches - frame #200


j) Umbrella - frame #230

3.2. Hypothetical Reference Circuits (HRC)

The following degradations were selected: (more details to be added later)

HRC	Impairments & Degradations	
	Type	Parameters
0	None – Reference sequence	
1	Video coding (H.264)	QP 32
2	Video coding (H.264)	QP 38
3	Video coding (H.264)	QP 44
4	Still image coding (JPEG2k)	2 Mb/s
5	Still image coding (JPEG2k)	8 Mb/s
6	Still image coding (JPEG2k)	16 Mb/s
7	Still image coding (JPEG2k)	32 Mb/s
8	Reduction of resolution	↓4 downsampling
9	Image sharpening	Edge enhancement
10	Downsampling & sharpening	HRC 8 + HRC 9

***Table 1.** List of HRCs of NAMA3DS1-CoSpaD dataset*

3.3. Availability of the SRC sequences

The sequences are available for free download at the following location:

http://www.irccyn.ec-nantes.fr/spip.php?article1052&lang=en

3.4. Subset selections

Several subsets have been identified for subjective experiments that require a smaller set of PVS for evaluation. These are listed in this section.

3.4.1. Split into two subsets

This subset selection has been done by Acreo. The goal was to divide the sequences into two equally distributed subsets in terms of perceived quality. A common set of one SRC and one HRC has been introduced.

The following repartitioning has been used:

3.4.2. Reduction to x PVS

4. Subjective Rating Tests

Subjective tests will be performed using as many different subjective assessment setups and environments as possible.

4.1. Subjective assessment variables

The subjective assessment environment specified by the ITU-T BT.500 has been criticized as being either not applicable or, at least, not sufficiently detailed for evaluating the quality of 3D video content.

In order to understand the impact of the changes that need to be implemented in a recommendation for 3D, a variety of variables need to be considered. The following is an incomplete list of possible variables:

- Viewing distance
- Display technology (Polarized, Active Shutter, Autostereoscopic, …)
- Background lighting (15% maximum brightness of screen, dark background, ambient lighting)
- Voting interface (on screen, on separate screen, in 3D, …)
- Voting methodology (ACR, DSCQS, SAMVIQ, …)
- Stimulus in between presentations (2D gray, 3D pattern, …)
- Session duration
- Number of observers in parallel for a single screen
- …

4.2. Interaction with participants

For subjective video quality assessments in 2D, each observer has an established ground truth due to their experience in daily life. For 3D, this does not necessarily apply, therefore differences may be evaluated such as:

- Training with explanations of typical 3D artifacts
- Extension of the test introduction
- Adding additional stimuli to the training
- …

The observer screening needs to be extended by at least:

- Depth acuity (Randot Stereo test or equivalent)
- Far vision (Snellen Chart or equivalent)
- Near vision (Snellen Chart or equivalent)
- Color vision (Ishihara Plates or equivalent)

Further ophthalmologic tests may be conducted in addition such as:

- Vergence facility (prism test)
- Phoriae
- Extent of Panum's Area
- Break-off point in near
- …

A questionnaire regarding the experience of the subjects with 3D viewing experience shall be asked to the subjects. The questions are listed in Annex A. (to be provided by IRCCyN)

Further information about the health state of the subjects prior and post-test may be acquired in an additional questionnaire. An example is provided in Annex B. (to be provided by IRCCyN)

4.3. Documentation

A detailed documentation of the experiment shall be provided by each lab. It shall include as much information of the viewing setup and environment as possible. Photos of the environment may help understanding the setup.

4.4. Subjective Test Method: ACR-HR

Most of the VQEG 3DTV subjective tests will be performed using the Absolute Category Rating Hidden Reference (ACR-HR) method. Other subjective assessment methods are welcomed as additional information.

The selected test methodology is the Absolute Rating method – Hidden Reference (ACR-HR) and is derived from the standard Absolute Category Rating – Hidden Reference (ACR-HR) method [ITU-T Recommendation P.910, 1999.] The 5-point ACR scale will be used.

Hidden Reference has been added to the method more recently to address a disadvantage of ACR for use in studies in which objective models must predict the subjective data: If the original video material (SRC) is of poor quality, or if the content is simply unappealing to viewers, such a PVS could be rated low by humans and yet not appear to be degraded to an objective video quality model, especially a full-reference model. In the HR addition to ACR, the original version of each SRC is presented for rating somewhere in the test, without identifying it as the original. Viewers rate the original as they rate any other PVS. The rating score for any PVS is computed as the difference in rating between the processed version and the original of the given SRC. Effects due to esthetic quality of the scene or to original filming quality are "differenced" out of the final PVS subjective ratings.

In the ACR-HR test method, each test condition is presented once for subjective assessment. The test presentation order is randomized according to standard procedures (e.g., Latin or Graeco-Latin square or via computer). Subjective ratings are reported on the five-point scale:

5 Excellent

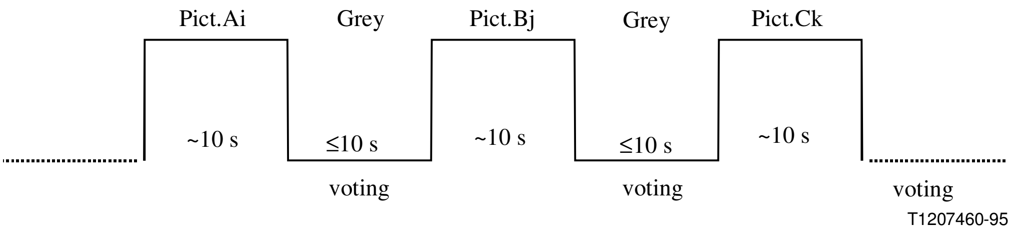
4 Good

3 Fair

2 Poor

1 Bad.

Figure borrowed from the ITU-T P.910 (1999):



Ai	Sequence A under test condition i
Bj	Sequence B under test condition j
Ck	Sequence C under test condition k

Viewers will see each scene once and will not have the option of re-playing a scene.

An example of instructions is given in Annex III.

4.5. Length of Sessions

The time of actively viewing videos and voting will be limited to 50 minutes per session. Total session time, including instructions, warm-up, and payment, will be limited to 1.5 hours.

4.6. Subjects and Subjective Test Control

Each test will require exactly 24 subjects.

The 3DTV subjective testing will be conducted using viewing tapes or the equivalent. Video sequences may be presented from a hard disk through a computer instead of video tapes, provided that (1) playback mechanism is guaranteed to play at frame rate without dropping frames, (2) playback mechanism does not impose more distortion than the proposed video tapes (e.g., compression artifacts), and (3) monitor criteria are respected.

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. At least two different randomized presentations of clips (A & B) will be created for each subjective test. If multiple sessions are conducted (e.g., A1 and A2), then subjects will view the sessions in different orders (e.g., A1-A2, A2-A1). Each lab should have approximately equal numbers of subjects at each randomized presentation and each ordering.

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. All viewers will be screened prior to participation for the following:

- normal (20/30) visual acuity with or without corrective glasses (per Snellen test or equivalent).
- normal color vision (per Ishihara test or equivalent).
- Normal stereo acuity (per Randot test or equivalent, presence of stereo vision is required but no specific acuity threshold applies, i.e. the first level of stereo acuity is sufficient)
- familiarity with the language sufficient to comprehend instruction and to provide valid responses using the semantic judgment terms expressed in that language.

4.7. Instructions for Subjects and Failure to Follow Instructions

For many labs, obtaining a reasonably representative sample of subjects is difficult. Therefore, obtaining and retaining a valid data set from each subject is important. The following procedures are highly recommended to ensure valid subjective data:

- Write out a set of instructions that the experimenter will read to each test subject. The instructions should clearly explain why the test is being run, what the subject will see, and what the subject should do. Pre-test the instructions with non-experts to make sure they are clear; revise as necessary.
- Explain that it is important for subjects to pay attention to the video on each trial.
- There are no "correct" ratings. The instructions should not suggest that there is a correct rating or provide any feedback as to the "correctness" of any response. The instructions should emphasize that the test is being conducted to learn viewers' judgments of the quality of the samples, and that it is the subject's opinion that determines the appropriate rating.
- Paying subjects helps keep them motivated.
- Subjects should be instructed to watch the entire 10-second sequence before voting. The screen should say when to vote (e.g., "vote now").

If it is suspected that a subject is not responding to the video stimuli or is responding in a manner contrary to the instructions, their data may be discarded and a replacement subject can be tested. The experimenter will report the number of subjects' datasets discarded and the criteria for doing so. Example criteria for discarding subjective data sets are:

- The same rating is used for all or most of the PVSs.
- The subject's ratings correlate poorly with the average ratings from the other subjects (see Annex II).
- 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 .

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 (see Annex I).
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.

4.8. Randomization

For each subjective test, a randomization process will be used to generate orders of presentation (playlists) of video sequences. Each subjective test must use a minimum of two randomized viewer orderings. Subjects must be evenly distributed among these randomizations. Randomization refers to a random permutation of the set of PVSs used in that test.

Note: The purpose of randomization is to average out order effects, ie, contrast effects and other influences of one specific sample being played following another specific samples. Thus, shifting does not produce a new random order , 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.

An example script in Matlab that creates 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
```

4.9. Subjective Data File Format

Subjective data should NOT be submitted in archival form (i.e., every piece of data possible in one file). The working file should be a spreadsheet listing only the following necessary information:

- Experiment ID
- Source ID Number
- HRC ID Number
- Video File
- Each Viewer's Rating in a separate column (Viewer ID identified in header row)

All other information should be in a separate file that can later be merged for archiving (if desired). This second file should have all the other "nice to know" information indexed to the subjectIDs: date, demographics of subject, eye exam results, etc. A third file, possibly also indexed to lab or subject, should have ACCURATE information about the design of the HRCs and possible something about the SRCs.

An example table is shown below (where HRC "0" is the original video sequence).

				Viewer ID	Viewer ID	Viewer ID	Viewer ID	…	Viewer ID
Experiment	SRC Num	HRC Num	File	1	2	3	4	…	24
XYZ	1	1	xyz_src1_hrc1.avi	5	4	5	5	…	4
XYZ	2	1	xyz_src2_hrc1.avi	3	2	4	3	…	3
XYZ	1	7	xyz_src1_hrc7.avi	1	1	2	1	…	2
XYZ	3	0	xyz_src3_hrc0.avi	5	4	5	5	…	5

5. List of participating labs

The following provides an incomplete list of labs that have indicated interest in running this subjective experiment with specific settings. The section is expected to grow or shrink in response to subjective assessments actually carried out.

5.1. IRCCyN/IVC

5.1.1. Subjective experiment in standardized environment

<details need to be copied from the publication cited above>

Main difference: consumer display 42’’ Philips Shutter glasses

Illumination: Standard D65 measured through glasses

Display (type, size): Philips 46PFL9705H 46” Shutter

Viewing distance: 3H (172cm)

Viewing angle (number of observers): 90 deg, 1 obs.

Observer screening methods: Randot, Snellen, (far, near), Ishihara, no thresholds applied

Display measurement/calibration methods: Display standard settings + EyeOneDisplay, 56cd/m2 through glasses, gamma 2.2, white point 6600K

5.1.2. Subjective experiment in living-lab style environment

Philips active shutter glasses screen, uncontrolled external daylight influence, various colors present in the viewing field of the observer

Main difference: Non standard environment

Illumination: Daylight, changing room illumination,

Display (type, size): Philips 46PFL9705H 46” Shutter

Viewing distance: larger than 3H (172cm)

Viewing angle (number of observers): 90 deg, 1 obs.

Observer screening methods: Randot, Snellen, (far, near), Ishihara, no thresholds applied

Display measurement/calibration methods: Display standard settings + EyeOneDisplay, 56cd/m2 through glasses, gamma 2.2, white point 6600K (same as IRCCyN1)

5.2. Acreo AB, Sweden

Two subjective experiments with different viewing distances on a polarized screen in standardized environment
Planned (meeting Rennes): Main difference: active consumer grade vs. passive consumer grade TV 55’’ displays

5.2.1. Viewing distance x

5.2.2. Viewing distance y

5.3. Yonsei University, Korea

One subjective experiment with polarized screen in standardized environment

Planned (meeting Rennes): Main difference: 46’’ polarized screen vs. 26’’ polarized screen, 1 viewer per test, 2 viewers per test

5.4. NTT

Planned (meeting Rennes) : Main difference: using different assessment methodologies, using different displays (polarized broadcast professional, shutter consumer display, 40’’)

5.5. AGH

Planned (meeting Rennes) : Main difference: 42’’ polarized 3D, autostereoscopic 4’’ display, outdoor environment, poor quality shutter glasses (crosstalk), room illumination control (reflection on polarized glasses)

5.6. NTIA

Planned (meeting Rennes) : Main difference: Living room vs. ITU BT.500 sound isolation room, Laptop display and ~50’’ 3D display

5.7. Orange Labs

Planned (meeting Rennes) : Main difference: Viewing distance, Display technology, room illumination

5.8. Deutsche Telekom

Planned (meeting Rennes) : Main difference: home environment, standardized environment, 52’’ display with 23’’ display shutter, eventually passive display

5.9. Technicolor

Planned (meeting Rennes) : viewing distance, viewing angle, screening methods