AVHD-AS Project: Subjective Test Procedure

Version 1.0

Feb. 29, 2016

|  |  |
| --- | --- |
| Contacts: |  |
| Margaret Pinson, NTIA | mpinson@its.bldrdoc.gov |
| Chris Schmidmer, OPTICOM GmbH | cs@opticom.de |
| Shahid Mahmood Satti, OPTICOM GmbH | ss@opticom.de |

Editorial History

|  |  |  |
| --- | --- | --- |
| Version | Date | Nature of the modification |
| 1.0 | Feb. 29, 2016 | Initial Draft, created by Shahid Mahmood Satti |
|  |  |  |
|  |  |  |

1. Introduction

This document describes the subjective test procedure for the VQEG AVHD project. The procedure will be followed by ILGs\proponents for creation of the validation databases.

1. Subjective Testing Related Details

### Viewing Distance

The instructions given to the subjects will request subjects to maintain a specified viewing distance from the display device. The proposed viewing distance range for the subjective testing is:

* Monitor/TV case: 3-5H (Living room like environment)
* Mobile case: 4-7H

where H = Picture Height (picture is defined as the size of the video window, not the physical display.)

Note: The subjective testing setup for mobiles will be described in a Section XX.

### Viewing Conditions

Preferably, each test subject will have his/her own video display. The test room will conform to ITU-P.913 requirements. The viewing environment should be reasonably quiet. Lighting conditions should be reproducible and must not change significantly when performing the test for different subjects.

It is recommended that subjects be seated facing the center of the video display at the specified viewing distance. That means that subject's eyes are positioned opposite to the video display's center (i.e. if possible, centered both vertically and horizontally). If two or three viewers are run simultaneously using a single display, then the subject’s eyes, if possible, are centered vertically, and viewers should be centered evenly in front of the monitor/TV.

## Display Specification and Set-up for TV case

Only high-end consumer TV or professional grade monitors should be used. LCD and OLED displays may be used, provided that the display meets the other specifications and is color calibrated for video [Color calibration should be discussed and specified]. If a curved TV shall be used, it must be agreed upon by ILGs/proponents. Testing display must be a full HD (1920x1080) display with aspect ratio 16:9.

Given that the subjective tests will use different HD display technologies, it is necessary to ensure that each test laboratory selects appropriate display specification and common set-up techniques are employed. Due to the fact that most consumer grade displays employ some kind of display processing that will be difficult to account for in the models, all subjective facilities doing testing shall disable any additional signal processing.

**Input requirements for display port**

* HDMI, DVI or DP

Monitors must have a diagonal between 27” to 32”. TVs must have a diagonal between 40” and 55”.

Labs must post to the reflector what display they plan to use; VQEG members have 2 weeks to object.

## Subjective Test Method: ACR(without Hidden Reference)

The 5-point ACR scale will be used. Reference conditions such as original SRC will not normally be used in tests, unless as anchor conditions.

In the ACR test method, each PVS is presented once for subjective assessment. The test presentation order is randomized according to standard procedures. Subjective ratings are reported on the five-level quality scale:

5 Excellent

4 Good

3 Fair

2 Poor

1 Bad

The quality labels must be translated to equivalent labels of the working language, e.g., the same language in which they receive the instructions. Viewers will not have the option of re-playing a PVS.

## Length of Sessions

The time of actively viewing videos and voting will be limited to 40 minutes per session. Total session time, including instructions, warm-up, and payment, will be limited to 1.5 hours. [till here it was edited in the VQEG meeting]

## Subjects and Subjective Test Control

Each test will require at least 24 valid subjects.

Video sequences can 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).
* familiarity with the language sufficient to comprehend instruction and to provide valid responses using the semantic judgment terms expressed in that language.

## 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 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.

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.

## 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, i.e., 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]

## 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.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | 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 |

1. Source Video Sequences

# Selection of Source Sequences (SRC)

Proponents cannot have any knowledge of the source sequences selected by the ILG.

The following video formats are of interest to this testing:

* 1080p 23-60 fps progressive (letterboxed content is allowed)

SRC reuse in subjective test. Proposal needed. One way is to go with the RSRC method presented by Lucjan/Margaret at San Diego VQEG Feb-March 2016.

## Purchased Source Sequences

Datasets that will not be made public may use source video that must be purchased (i.e., source video sequences that proponents must purchase prior to receiving that subjective dataset). Because the appropriateness of purchased source may depend upon the price of those sequences, the total cost must be openly discussed before the ILG chooses to use purchased source sequences (e.g., VQEG reflector, audio conference); and the seller must be identified. (Reminder: the scenes to be purchased must be kept secret until model & subjective dataset submission). A majority of proponents must be able to purchase these source video sequence (i.e., for model validation).

## Requirements for Camera and SRC Quality

The source video can only be used in the testing if an expert in the field considers the quality to be good or excellent on an ACR-scale. The source video should have no visible coding artifacts. 1080i footage may be de-interlaced and then used as SRC in a 1080p experiment. Down-sampling from UHD to 1080p is allowed.

At least ½ of the SRC in each experiment must have been shot originally at that experiment’s target resolution.

The ILG will view the scene pools from all proponents and confirm that all source video sequence have sufficient quality. The ILG will also ensure that there is a sufficient range of source material and that individual SRCs are not over-used.

A review of SRC quality must be conducted prior to using the SRCs for the tests.

After this phase, all scenes will be considered final. No scene may be discarded or replaced after this point for any technical reason.

## Content

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

1) movies, movie trailers

2) sports

3) music video

4) advertisement

5) animation

6) broadcasting news (business and current events)

7) home video

8) general TV material (e.g., documentary, sitcom, serial television shows)

## Scene Cuts

Scene cuts shall occur at a frequency that is typical for each content category.

## Scene Duration

Final source sequences will be upto 5 minutes. The proposal is to divide into three pools of 1, 3 and 5 minutes.

## SRC Repetition

SRC repetition within a test is not allowed. Different parts of a longer video can be used as an SCR. Overlap of SRCs between different tests is allowed. However, it must be limited to below 25%.