 ISO/IEC JTC 1/SC 29/AG 5 N00031

**ISO/IEC JTC 1/SC 29/AG 5**

**MPEG Visual Quality Assessment**

**Convenorship: DE**

**Document type:** Output Document

**Title:** Guidelines for remote experts viewing sessions (draft 1)

**Status:** Approved

**Date of document:** 2021-09-03

**Source:** ISO/IEC JTC 1/SC 29/AG 5

**Expected action:** None

**Action due date:** None

**No. of pages:** 7 (with cover page)

**Email of Convenor:** wien@lfb.rwth-aachen.de

**Committee URL:** <https://isotc.iso.org/livelink/livelink/open/jtc1sc29ag5>

**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC 1/SC 29/AG 5**

**MPEG VISUAL QUALITY ASSESSMENT**

**ISO/IEC JTC 1/SC 29/AG 5 N00031**

**Online – July 2021**

|  |  |
| --- | --- |
| **Title** | Guidelines for remote experts viewing sessions (draft 1) |
| **Source** | AG 5 MPEG Visual Quality Assessment |
| **Status** | Approved |
| **Serial Number** | 20728 |
| **Authors** | J. Jung (Tencent), M. Wien (RWTH Aachen University), V. Baroncini (Vabtech) |

[Ed.Note (mw): The draft needs extension including  
- information on display tuning,  
- information on recommended viewing environment,  
- information on recommended viewing distance,  
- information on outlier detection,  
- Annex with checklists specific projects,  
- …

]

# Introduction

This contribution provides guidelines for remote expert viewing sessions for visual media coding activities in MPEG. In 2D video coding, objective metrics such as the PSNR of the reconstructed video sequence compared to the uncompressed original are well established and used for decision-taking. Nevertheless, subjective assessment of proposals is advisable for coding tools which may have significant impact on the subjective quality while this is not necessarily reflected in the objective metric. Examples are deblocking or other adaptive loop filters for which subjective assessments have been performed in the past as part of the decision process for adoption of a variant into the final specification. In the context of immersive video, such as MIV, the task of the objective metrics is particularly challenging for the following reasons: 1- both compression and synthesis artifacts are present, 2- no ground truth reference is available for displayed views, making full reference metrics inappropriate. In such conditions, it is recommended to perform remote expert viewing sessions and take the results into account for coding tool decisions.

At the time of writing, two activities make use of remote experts viewing sessions in their development process: JVET and MIV. While the description partially relates to these groups, the goal of drafting these guidelines is to provide them more broadly to activities in the area of visual media compression.

The general process in an MPEG activity is as follows: a contribution is first qualified for the remote viewing session. The content is prepared by the proponent according to the guidelines, cross-checked, and made available to the test coordinator. The test coordinator processes the content to anonymize it and adapt it to the test methodology. Then the viewing session takes place, through an online Zoom session. The scores sent to the test coordinator are processed and shared with the group.

# Test preparation

A contribution is reviewed. When all objective results according to the CTCs are available, and when the group is inclined to adopt the proposal, the group qualifies the contribution for a remote expert viewing session. In preparation of the viewing session, the following steps are applied:

1. **Test coordinator:** a test coordinator is designated by the group.

For MIV, the identity of the test coordinator is reported in the GitLab issue of the contribution. The GitLab issue is tagged with the “viewing” label.

1. **Content:** the content to be viewed is selected by the group. This includes which sequences, which rate points, which configurations. It is not the responsibility of the test coordinator to select the content. The amount of sequences to select depends on the length of the test session, typically 15 minutes maximum for a contribution. The duration of the sequences shall be approximately similar, and not smaller than 5 s or longer than 10 s. If viewports or pose traces are rendered, the group agrees on the video resolution output of the rendering process before preparation of the mp4 files.

For MIV, it is decided if the pose traces are 97 frames (with 17 frames input) or 300 frames (with 97 frames input). All this information is reported in the GitLab issue of the contribution.

1. **Cross-checker:** at least one crosschecker is appointed. The task of the crosschecker is to assert the correctness of the mp4 files generated from the sequences under consideration. The way the cross-check is done is left to the crosschecker and the proponent. This step can be made optional if the group agrees to do so (for instance to enter EEs or CEs).

For MIV, the cross-checker identity is reported in the GitLab issue of the contribution.

1. **Viewers:** volunteer viewers are selected by the group for performing the viewing task. The proponents are responsible of having sufficient volunteers for his test. If the experiment is related to an adoption decision the proponents cannot take part to the test. Per ITU-R BT.500-14 it is generally recommended to acquire at least 15 valid scores in the viewing session. Given that the test is not conducted in a lab, more volunteers would be needed, to compensate the various side effects of such remote viewing sessions. This typically allows to reject more scores based on drastic statistical criteria, and end-up with as many valid scores as in a lab experiment. Lower participant numbers may be acceptable under the constraint that the data shall be treated as indicative figures. The number of participants in this case is recommended to be at least 8.

For MIV, the identity of the viewers is recorded in the GitLab issue to facilitate the invitation to the viewing session.

# Content preparation and cross-check

The complete set of content to be viewed, as selected by the group, is prepared by the proponents.

For MIV, the content is made of pose traces, simulating a typical motion performed by a viewer. When generating the pose traces: the bitrate of the tested method should remain below the bitrate of the anchor. The proponent is allowed to use any QP for the texture (keeping a common set of QPs for all sequences).

## Mp4 files generation

For playout, the video sequences under test are converted into mp4 files by encoding them with HEVC at very high quality. For the encoding task an up-to-date version of ffmpeg must be used.

The command line call for encoding with ffmpeg raw YCBCR 4:2:0 input video is:

*ffmpeg -s:v <pix>x<lin> -c:v rawvideo -pix\_fmt <input\_pix\_fmt> -r <fps> -i <input.yuv> -c:v libx265 -crf 10 -tag:v hvc1 -pix\_fmt <output\_pix\_fmt> <output.mp4>*

where:

<pix> denotes the number of pixels per line,

<lin> denotes the number of lines per picture,

<input\_pix\_fmt> denotes the input pixel format, yuv420p10le for 10 bits, yuv420p for 8 bits,

<output\_pix\_fmt> denotes the output pixel format,

<fps> denotes the frame rate of the video.

For MIV, the input pixel format is 10 bits while the output pixel format is 8 bits.

Note: for UHD test sequences, setting “-crf 10” may be too challenging for many PCs to provide a smooth playout. In this case, the setting may be changed to “-crf 15” or even higher. It is recommended to investigate this aspect by the proponents before the test sessions and achieve consensus in the group on the applicable setting. It is recommended to use the same setting for all video sequences displayed in a test session.

## Zipping, naming and upload of the files

For provision to the test coordinator, it is recommended to pack the generated mp4 files into a single zip file with a subfolder for the anchor and a subfolder each for the proposals under test. The names of the files follow an agreed pre-defined template.

An example for MIV is given below:

**<proposal>**/**<sequence name>\_p<posetrace>\_QP<qp>.mp4**,   
*e.g.* **m56369/B\_p03\_QP4.mp4**

The zip files are uploaded by the proponent to a previously agreed location with access granted to the proponent, the crosschecker, and the test coordinator.

No deadline is defined, the proponent has the responsibility to make the data available early enough so that the next steps of the process can be smoothly handled. It is important to keep in mind that the time needed for processing the data by the test coordinator is not negligible.

The crosschecker asserts the correctness of the mp4 files.

For MIV, the crosschecker notifies this in the GitLab issue.

# Test setup

When the content has been uploaded by the proponent, and cross-checked by the cross-checker, the test coordinator is in charge of processing the content, to make it anonymous and adapted to the test methodology. This includes the following steps:

* Download of the content.
* Automatic renaming of files (a script must be used) to anonymize the sequences.
* Creation of a VLC playlist, mixing the contents in a randomized fashion, the anchor (if any), and the proposal. According to the test methodology described in section 5, the presentation of the anchor and proposed signal is randomized and not revealed to the test subjects.
* The renamed content is uploaded (the test coordinator is free to select the online drive): one single zip file includes all mp4 files. The zip file is password protected.
* Upload of a training set, reproducing the test methodology: the zip file for the training set is not password protected.
* Upload of written instructions on the viewing process (double stimulus, rating scale, etc).
* The assigned volunteers are provided with the download link by test coordinator for the contribution under consideration as well as the time slot for the test.

# Viewing session

## Requirements

Participants in the subjective evaluation must use a computer capable of smoothly playing out the provided mp4 files. Depending on the resolution of the video sequence under investigation, requirements on features of the display (*e.g.* the screen size) may be formulated by the test coordinator.

Due to general availability, the VLC media player is selected as the tool for playout. Participants are requested to have the most recent version of this software installed. The recommended configuration of VLC is as follows:

* Go to Preferences, select "Show settings: All" at the bottom right of the window
* Select Video
  + Activate Fullscreen
  + Deactivate "Show media title on video"
* Select Playlist
  + Deactivate any looping options

It is recommended to turn off other programs and any file synchronizations since they can interfere with the playout.

## General process

Once the download link is available to the participants, they can:

* Download the content
* Read the instructions carefully
* Unzip and perform the training session and ask questions to the test coordinator if needed.

The viewing session is handled via a Zoom call. When the viewing session starts:

* If there are not enough participants for the test, it is postponed to a next call.
* The test coordinator allocates a few minutes to answer participant’s remaining questions, and a short training session can be conducted, if it sounds appropriate.
* The test coordinator provides the password of the zip file containing the content and the playlist.
* The volunteers unzip the test set and get the VLC playlist.
* The test session is conducted: each volunteer notifies the test coordinator when he has finished by sending him the scores, according to a predefined template. The session is closed once all scores are available. The scores are not shared between participants.
* Another test can be handled, with the same process.

## Test methodology

It is recommended to use the Comparison Category Rating[[1]](#footnote-1) (CCR) method. It is a variant of ACR (Absolute Category Ranking) and DCR (Degradation Category Rating) methods: the goal is not to give a quality score to a sample, but rather to compare two samples, to check if there is a visual improvement that justifies the adoption of the proposed technology.

A double stimulus method is used: volunteers are presented a pair of PVS (Processed Video Sequences) for each pose trace or test sequence. The session displays the PVS in an “A” / “B” / “A” / “B” order. In contrast to DCR, with the CCR procedure the order of the processed and unprocessed samples is chosen randomly for each test cell (50%/50%): “A” and “B” are either the anchor (provided by current test model version) or the tested video. The basic test cell (BTC) is constructed as follows:

1. “A” (on a mid-gray background, 1 s presentation)
2. Video A
3. “B” (on a mid-gray background, 1 s presentation)
4. Video B
5. “A” (on a mid-gray background, 1 s presentation)
6. Video A
7. “B” (on a mid-gray background, 1 s presentation)
8. Video B
9. “Vote *N*” (on a mid-gray background, 5 s presentation)

If an uncompressed original sequence is available (*e.g.* for 2D video), it should be inserted at the beginning of the BTC:

1. “Original” (on a mid-gray background, 1 s presentation)
2. Original Video
3. …

Graphical user interface, application

Description automatically generated

Fig: Structure of a BTC (a) without and (b) with insertion of an uncompressed original sequence

## Rating scale recommended for MIV: 7-grade rating scale

This grading scale is used in the MIV activity. The pose traces form a navigation path that correspond to an expected motion of the user. The views along the navigation path are not captured, but synthesized. Hence, no reference signal is available. The proposal is compared to the anchor, that corresponds to the output of the test model according to common test conditions.

Volunteers provide two judgements with one response: "Which sample has better quality?" and "By how much?". The subjects are asked to rate the impairment of the second stimulus in relation to the first stimulus, with the following scale:

Table

Description automatically generated

Fig: Rating scale

Instructions can contain the meaning of the scale in the specific MIV context:

* Equivalent: “I am not sure which one is the best. Either they are similar, or I can find improvements and degradations in both samples, at an equivalent level”.
* Slightly better: “It seems that I have a preference for the 1st sequence, although it seems to be a minor improvement” (parts of the 2nd sequence could be worse than the 1st sequence).
* Better: “The 1st sequence is undoubtedly of higher quality, I spot at least one place where the improvement is obvious and can hardly see areas where it is worse.”
* Much better: “I immediately noticed that the 1st sequence has a higher quality than the 2nd one”.

## Rating scale recommended for JVET: 4-grade rating scale

This grading scale is used in JVET and imposes a forced-choice approach. The test subjects are asked to rate the video sequences presented in each test cell as follows:

* “A is much better than B” (+3)
* “A is better than B” (+1)
* “B is better than A” (-1)
* “B is much better than A” (-3)

The scores are mapped to numbers reported between parenthesis.

The voting sheet is presented in the annex 1. The participants fill the voting sheet and enter their results into an online form after completion of both sessions. The participants are also asked to enter the used screen size and any specific observations in a comment field of the online form.

The voting sheets are used by the participants during the session.

# MOS processing and reporting

The scores are processed (typical post-screening of scores, outlier detections, etc), by the test coordinator and reported to the group.

From these values, the MOS and the 95% confidence intervals are calculated. A resulting MOS value of zero indicates that anchor and proposal were rated to have the same visual quality.

For MIV, the results are shared on the GitLab issue.

# References

1. Recommendation ITU-T P.910 (2008), *Subjective video quality assessment methods for multimedia applications*.
2. Recommendation ITU-R BT.500-14 (2019), *Methodologies for the subjective assessment of the quality of television images*.
3. Recommendation ITU-R BT.2100-2 (2018), *Image parameter values for high dynamic range television for use in production and international programme exchange*.
4. SMPTE ST 2084, *High Dynamic Range Electro-Optical Transfer Function of Mastering Reference Displays*, 2014.
5. SMPTE ST 2036-1, *Ultra High Definition Television – Image Parameter Values for Program Production*, 2014.
6. ETSI TS 101 154, *Digital Video Broadcasting (DVB); Specification for the use of Video and Audio Coding in Broadcast and Broadband Applications*, 2019.
7. ITU-T Tech. Paper HSTP-VID-WPOM and ISO/IEC TR 23008-8 (*Eds.*: K. Andersson, F. Bossen, J.-R. Ohm, A. Segall, R. Sjöberg, J. Ström, G. J. Sullivan, A. Tourapis), Working practices using objective metrics for evaluation of video coding efficiency experiments, July 2020.
8. RDPlot, https://github.com/IENT/RDPlot, accessed 2021-06-18.

# Annex 1: voting sheets

Table

Description automatically generated Table, calendar

Description automatically generated  
**Fig: Examples for Voting sheets for the 4-grade and 7-grade scales.**

1. Specified in ITU-T P800 [↑](#footnote-ref-1)