Expert viewing and activities in MPEG Visual Quality Assessment

PD Dr.-Ing. Mathias Wien RWTH Aachen University

Convenor of ISO/IEC JTC 1/SC 29/AG 5 MPEG VQA

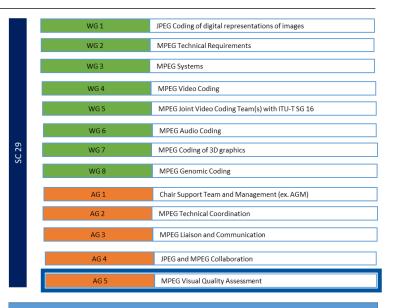


MPEG Visual Quality Assessment (VQA)

- New structure of MPEG since Oct. 2020: Foundation of ISO/IEC JTC
 1/SC 29/AG 5 MPEG VQA as one of 10 MPEG groups (WGs and AGs)
- Scope

2

- Select and design VQA methodologies and objective quality metrics for the assessment of visual coding technologies
- Guidance and support on VQA at start of and during standardization projects (e.g. CfE, CfP), and for verification testing
- Liaison with ITU and other organizations on VQA standards creation and improvement
- 4 Focus Groups: SDR, HDR, 360, Immersive Video
- 2 Ad-Hoc Groups:
 - Learning-based quality metrics
 - Guidelines for subjective visual quality evaluation



MPEG VQA Advisory Group Convenor: Mathias Wien				
SDR Video Chair: Vittorio Baroncini	HDR Video Chair: Andrew Segall	360 Degree Video Chair: Yan Ye	Immersive Video (MIV) Chair: Joel Jung	

ISO: International Standardization Organization | IEC: International Electrotechnical Commission | JTC1: Joint Technical Committee | MPEG: Moving Pictures Experts Group SC29: Sub-committee 29 "Coding of Audio, Picture, Multimedia and Hypermedia Information" | AG: Advisory Group | WG: Working Group | CfE: Call for Evidence | CfP: Call for Proposals



VQA: Visual assessment at CfP stage

R3

R2

- Call for Proposals:
 - Evaluation of responses (competitive stage of standardization process)
 - Defined test set, encoding constraints, reporting requirements (algorithm description, objective quality metrics, complexity metrics, ...)
 - Formal assessment (two or more laboratories)
 - MOS results
 - Ranking of proposals
- Example: CfP for video based dynamic mesh coding
 - CfP October 2021, evaluated in April 2022, DCR test
 - Collaborative standardization phase since then

Test Dataset Random R2 R1 15 12 Access Longdress 8 11 14 10 Longdress Basketball player 16 Basketball player Mitch 17 12 Football 12 Football 10.00 9.00 (a) 8.00 7.00 6.00 MOS 5.00 4.00 3.00 2.00 1.00 R4: D (b) R3: E R2: ---Α Ranking

Texture

Map Size

2k x 2k

R3

Precision

Table 2 Target bitrates in Mbit/s for Random Access and All Intra

Attribute

Test Dataset All

Figures from MPEG doc m59772

Table 1 Test material datasets

Faces

Precision

Vertices

MPEG Verification Tests

 MPEG traditionally conducts verification tests for its media coding standards

- AG5N39: Guidelines for Verification Testing of Visual Media **Specifications**
 - Purpose and goals of the verification tests
 - Procedural steps in a verification tests
 - Selection of test material
 - Preparation of bitstreams and rate points
 - Conduction of visual tests
 - Reporting

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 N39 Online - October 2021

Guidelines for Verification Testing of Visual Media Specifications AG 5 MPEG Visual Quality Assessment Status

Serial Number

1 Introduction

This document defines guidelines f ongoing and future verification test a visual media output. In order to ass formal subjective evaluation of the ompressed bitstream is mandatory MPEG verification tests has been sir of a new standard. Nevertheless, ti Groups induces the need for a proce test results between the relevant W understanding of previous verification into the new organizational structure The description currently relies o reconstructed. For visual media wh 6DoF), additional steps may be requ

2 Purpose and Goals of Verif

MPEG verification tests are conduct the fulfillment of the goals of the achieved compression capability of t scenarios as well as the successful standardization activity.

To this end, the verification test to performance of the relevant predeces which has been used for developing rification tests typically target one a step-by-step fashion to support the

3 Procedural steps in a Verifi

- The WG develops the work item a final state (FDIS).
- The WG consults with AG5 to p of the verification test plan is su AG5 appoints a test coordinator
- for coordination of the verificat laboratories to perform the form The verification test plan is appr - The laboratories report the resul
- The WG produces the verification

4 Selection of Test Material

For each application scenario which is supposed to be tested, a set of representative test sequences shall For each application scenario which is suppose to or tested, a set or representative two sequences than be collected. It is strongly recommended to exclude sequences which have already been used in the oe consecent. It is strongly recommended to exclude sequences which have already been used in the development plass of the standard. The number of test sequences should be balanced with the overall effort. A set of 4 to 5 test sequences for an application scenario may be considered adequate. ethor. A set of a 100 o sea soquences for an approximon scenario may or considered norquiae.

It is recommended to collect hashes (e.g. md5sums) of the uncompressed source sequences to enable A to recommended to context masters (e.g., macrounne) or use uncompressed source sequences to enable manifoliums identification. This is especially helpful if multiple variants of a test sequence are

5 Preparation of Bitstreams and Rate Points

The verification test typically compares the performance of the new standard to the performance of the The verification is a typically compares are performance of the new seminary to the performance of the relevant predecessor (the anchor). Comparable configurations must be used for both, the anchor and the the same provided the similar is comparative voltages and no many or uses not south, the auctor and me may scheme. The configurations shall be aligned as closely as possible. All relevant tools available in the anchor shall be used. tne ancnor snats ne used.

It is recommended to define 4-5 rate points, covering the MOS range. Two strategies are possible:

- Matching rates between anchor and new scheme. This allows for direct assessment of the quality improvement of the new scheme relative to the anchor;
- quanty improvement or use new scinene reserve to me ancier;

 Matching quality of rate points. This supports a more reliable computation of rate savings over

 Accordance of the Disservation of the Disservation on the stancing quanty or rate points. This supports a more retunite computation of rate savings over the covered rate range using the Bjoutegaard Delta Rate method [3]. Depending on the performance relation between the two schemes, it may be possible to achieve both, rate and

Bitstreams should be generated and crosschecked by independent parties to assert the correctness of the Distreams shown be generated and crossenected by independent parties to assert the correctness of the encoding process and the compliance to the defined encoder configurations of both, anchor and new

sometime.

The typical procedure for determining rate points includes the definition of candidates based on The typical procedure for determining rate points includes the definition of caminates based on objective metrics with a suggested preselection w.r.t. suitability for subjective evaluation, followed by objective metrics wing a juggested preselection w.e.t. standardly for subjective evaluation, non-river of dry-run experiments which may include experts viewing sensions or viewing sensions with native subjects. Formal subjective assessments with native subjects are recommended.

The formal subjective evaluation is coordinated by a test manager who is appointed by AG5 in The format suspective evaluation is coordinated by a test manager who is appointed by AG5 in agreement which the WG. The test manager is responsible for all required logistic, technical, and design activities in the context of the formal subjective evaluation for the verification test.

The coordination activity of the test manager comprises

- ooramation activity or the eest manager companies.

 The selection, the direction, the coordination and the instructions to the test laboratory(ies);
- Lies serviction, the unrecommune coordination and the instructions to the test aboratory (see: If and when required, any action necessary to assert qualification of new laboratories as capable to perform state-of-the-art formal subjective evaluations meeting the requirements of the
- ventication ees.

 The selection of the evaluation conditions and the assertion of their suitability according to the
- The design, the supervision of the conduction (or conduction) of the formal subjective test. The collection and the statistical analysis of the data resulting from the test, to be submitted in Co-editing of the verification test report.

7 Reporting

The report of the verification test shall include a description of:

- The verification test activity, including scope, results, and test procedures;
- The test procedure

- The test logistic (e.g. equipment, subjects involved, voting scheme and other details); the test organic (e.g. equipment, suspecia surverse, voting sentents and other security including MOS and CI, plots of the MOS results grouped by test
- Summary of the BD rates values that allow to provide proof of the bit rate savings. It is recommended to highlight any specific features of the new scheme which differentiate it from the

AG5N39: https://www.mpeg.org/wp-content/uploads/mpeg_meetings/136_OnLine/w20975.zip



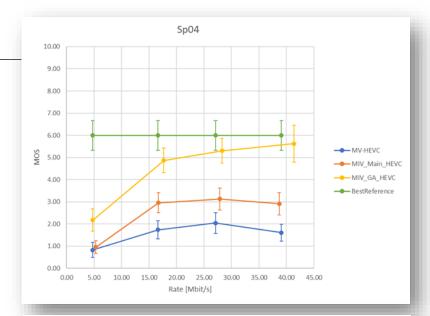
VQA: Verification Testing Activities

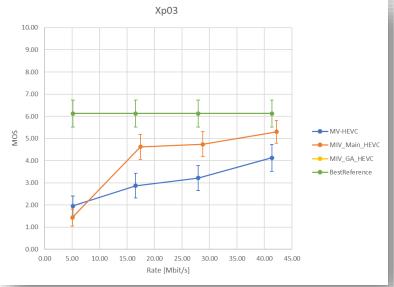
- Completed verification test activities (mostly DCR)
 - MPEG-5 Part 1: Essential Video Coding (EVC):
 SDR and HDR. MPEG-5 Part 2: Low-Complexity
 Enhancement Video Coding (LCEVC): AVC,
 EVC, HEVC, VVC + enhancement layer
 - MPEG-I Part 3 / ITU-T H.266
 Versatile Video Coding (VVC): SDR, HDR, 360° video
 - MPEG-I Part 4: V-PCC
 - MPEG-I Part 12: MPEG Immersive Video (MIV) (ACR test, example)

Emerging activities

- Multilayer VVC verification testing
- Visual evaluation of film gain characteristics SEI message









VQA: Expert viewing and investigations

- AG5N40: Guidelines for remote expert viewing (REV)
 - Remote testing in focus due to pandemic situation
 - Typically A-B comparisons, 4-grade force choice, 7-grade scale
 - Typically testing test model against variant vs. proposed change
- Investigations on efficient methods for (expert) viewer training
 - Goal: reliable results, good differentiation / ranking of proposals under test
- Investigations on objective metrics Goal: recommendation (and potentially development) of metrics suitable for use in decision making process of standardization context (across all types of MPEG visual media)
 - Collecting data bases, studying metrics

AG5N40: https://www.mpeg.org/wp-content/uploads/mpeg_meetings/136_OnLine/w20976.zip

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 N00040 Online - October 2021

Guidelines for remote experts viewing sessions AG 5 MPEG Visual Quality Assessment

Serial Number

J. Jung (Tencent), M. Wien (RWTH Aachen University), V. Baroncini

1 Introduction

This document provides guidelines for in MPEG. In 2D video coding, objecti compared to the uncompressed origina subjective assessment of proposals is the subjective quality while this is r deblocking or other adaptive loop filte past as part of the decision process for of immersive video, such as MIV, the following reasons: 1- both compressio is available for displayed views, maki recommended to perform remote expa

At the time of writing, two activities n process: JVET and MIV. While the these guidelines is to provide them me The general process in an MPEG acti viewing session. The content is prepar and made available to the test coordin and adapt it to the test methodology. session. The scores sent to the test coc

2 Test preparation

A contribution is reviewed. When all the group is inclined to adopt the proviewing session. In preparation of the

- 1. Test coordinator: a test coordinator
- 2. Content: the content to be vi which rate points, which conselect the content. The numbe typically 15 minutes maximu approximately similar, and no are rendered, the group agree preparation of the mp4 files.
- Cross-checker: at least one cr the correctness of the mp4 fil the naming of the files). The
- Viewers: volunteer viewers :

5.2 General process

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

- If there are not enough viewers for the test, it is postponed to a next call.
- The test coordinator allocates a few minutes to answer viewers' remaining questions, and a short The test coordinator provides the password of the zip file containing the content and the playlist.
- The viewers unzip the test set and get the VLC playlist.
- The test session is conducted: each viewer 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
- Another test can be handled, with the same process.

5.3 Test methodology

It is recommended to use the Comparison Category Rating (CCR) method [2]. It is a variant of the Absolute Category Ranking (ACR) and Degradation Category Rating (DCR) methods [3]: the goal is not to give a quality score to a sample like in the single stimulus ACR method, but rather to compare not to give a quanty score to a sample one in the single substitute (the samples, to check if there is a visual improvement that justifies the adoption of the proposed

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:

- "A" (on a mid-gray background, 1 s presentation)
- "B" (on a mid-gray background, 1 s presentation)
- "A" (on a mid-gray background, 1 s presentation)
- 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
- "Original" (on a mid-gray background, 1 s presentation)
- Original Video

	"A"	Video A								
	-	TIDEO A	-8"	Video B	"A"	Video A	"B"	Video B	"Vote ⇔"	í (a
g" Source	"A"	Video A	*9*		_					

"8" Video B "A" Video A "8" Video B "Vote cb" (b) Figure 1: Structure of a BTC (a) without and (b) with insertion of an uncompressed original sequence

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

A 7-grade scale as shown in Figure 2 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 part una correspond to an expected instant to the sect. He there may make the appared to the 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

Volumeers provide two jungitiments with one response. These samples was vested spiriture, two jungitiments with one response. The subjects are asked to rate the impairment of the second stimulus in relation to the first

Date saved: 2022-06-28



VQA: Expert viewing

- Expert viewing tests often used at JVET / MPEG meetings
 - Decision making for visually sensitive coding tools (e.g., deblocking / loop filters)
 - Dry-run testing in preparation for CfE / CfP / verification tests
 - Exploration of coding tools (e.g., for JVET EE1/Neural Networkbased Video Coding)
- Viewing performed and evaluated on-site at the meeting (or remote, see previous slide)
 - Request for suitable room at meeting
 - Time critical, tests compete with other meeting sessions
 - Use of one or more 65" OLED displays intended
 - Driven by PC with suitable player software





Test setups for expert viewing at JVET meetings in Mainz (top) and Antalya (bot.)



VQA: Expert viewing example JVET-AD0399

- Expert viewing for the spatial scalability category of the VVC multilayer verification tests, <u>JVET-AD0399</u>
 - DCR test with 11-grade scale, for comparison of dual layer coding with a downscaled base layer at scaling ratios 1.5 and 2 to single layer coding
 - 4 test sequences (10 sec), 3 rate points (low, mid, high), 3 configurations
 - 42 BTCs (incl. 3 stabilization, 3 trap)

Test Site	On-site
Display, size, connection (resolution setting)	Samsung 65" S95B, HDMI (3840×2160), 10bit input
Viewing distance	3 viewers sitting at 1.5H, 2 views standing at 1.6H
Viewing angle	±75°, 90° (at screen center)
Total number of viewers	19 (6 female, 13 male)

Score	Impairment item				
10	Imperceptible				
9	Clichtler managetible	somewhere			
8	Slightly perceptible	everywhere			
7	Danagatible	somewhere			
6	Perceptible	everywhere			
5	Cl. 1 (11	somewhere			
4	Clearly perceptible	everywhere			
3	.	somewhere			
2	Annoying	everywhere			
1	Caranala anna aire	somewhere			
0	Severely annoying	everywhere			



VQA: Expert viewing example JVET-AD0399

Training

- Explain procedure and scoring scale.
- Run training session (6 BTCs, covering examples of the impairment range).
- Let viewers note their scores.
- Let them read out their scores. If variation is larger than 3 on the grading scale, re-run BTC
 - Viewers shall reconsider their scoring for that sequence (not enforced to change their mind)
 - No indications where to look allowed, viewers have to make it up by themselves

Data processing

- No failure on trapping sequences observed
- Single (clear) outlier value removed

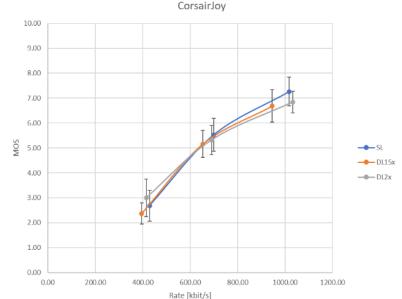


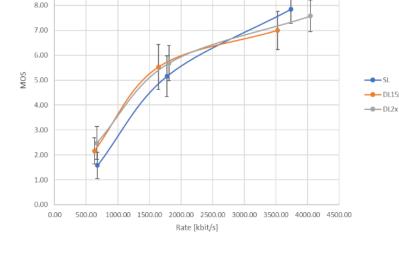
VQA: Results in JVET-AD0399

- Plots: MOS values with ±95% confidence intervals
- Results used for finalizing impairment range for tested rate points

Observation

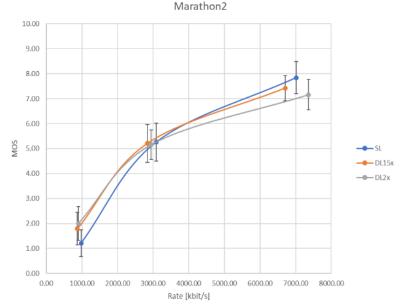
- Expert viewing session often show relatively large confidence intervals
- Applied training method seems to lead to more consistent use of grading scale by experts
- Further study intended

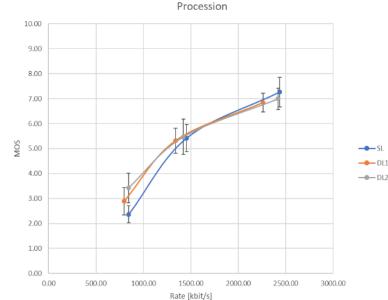




DrivingPOV3

9.00







Summary, Conclusions, Outlook

Overview of ISO/IEC JTC 1/SC 29/AG 5 MPEG VQA

- Tasks along the development cycle of visual media specifications
- Key contribution in CfE / CfP process, verification tests
- Support for Working Groups in standardization process

Key task: provide measures for decision making

- Subjective quality assessments: "live" during meetings and/or formal laboratory tests
- Provision of objective metrics suitable for the standardization process. Requirements / wish list:
 - Reproducible and understandable
 - Reliable ranking / discrimination of assessed video sequences
 - Resilient / consistent w.r.t. to unseen artifacts
- Exchange, communication and cooperation with VQEG is welcome!



Thanks for your attention!

PD Dr.-Ing. Mathias Wien RWTH Aachen University wien@lfb.rwth-aachen.de Convenor of ISO/IEC JTC 1/SC 29/AG 5 MPEG VQA

https://lists.rwth-aachen.de/postorius/lists/mpeg-vqa.lists.rwth-aachen.de/



