 ISO/IEC JTC 1/SC 29/AG 5 N00118

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

**MPEG Visual Quality Assessment**

**Convenorship: DE**

**Document type:** Output Document

**Title:** **Visual quality comparison of ECM/VTM encoding**

**Status:** Approved

**Date of document:** 2024-07-8

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

**Expected action:** Information

**Action due date:** None

**No. of pages:** 9 (without 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 N00118**

**Rennes, FR – April 2024**

|  |  |
| --- | --- |
| **Title** | **Visual quality comparison of ECM/VTM encoding** |
| **Source** | AG 5 MPEG Visual Quality Assessment |
| **Status** | Approved |
| **Serial Number** | 23963 |
| **Authors** | M. Wien, J.-R. Ohm, V. Baroncini |

# Abstract

The results of the second subjective assessments comparing the compression performance of the ECM-12.0 to the VTM-11 with ECM-comparable settings are reported. The rate points for the tests were selected from a set of ECM simulations. The corresponding VTM bitstreams were generated using a one-time QP switch (QP+=1) within the sequence to match the VTM bitrate to the ECM. The switching points were chosen such that the ECM rate should never be higher than the VTM rate and that the distance should not exceed about 2% of the VTM rate. Three UHD sequences were evaluated using the Random Access configuration and three HD sequences were evaluated using the Low Delay B configuration. In addition to the configurations following the common testing conditions, two of the UHD sequences and one HD sequence were additionally evaluated using a configuration variant without inter template-matching tools (denoted as ECM-nTM subsequently). Three tests were conducted on the provided test set: An expert viewing test with 24 viewers at the Rennes meeting site, a laboratory test with 10 naïve test subjects by VABTech, and a laboratory test with 5 expert viewers by GBTech. Both, the laboratory test conducted with naïve viewers and the on-site test with experts demonstrate a clear visual benefit of the ECM when compared to the VTM for a significant number of cases across the three tests. Indicative BD rate measurements from the on-site tests in Rennes show that both, the rate savings computed by PSNR and by MOS are in a comparable range. Indicative BD rate gains are in the range of about 25-29% for the UHD test sequences when considering both, the ECM and ECM-nTM configurations. For the HD test sequences, the indicative BD rate gains are around 20% for Beatriz and GTAVs900. On the DOTA sequence, no gain is observed. For the test sequence BeachMountain2, the subjective results of the on-site tests in Rennes reveal potential issues with the performance of the template matching tools in the ECM. These results seem to suggest that further study of such tools may be useful. Similar indications may be deducted from the laboratory tests.

# Introduction

This document reports the results of the second subjective assessment of the compression performance of the ECM (Enhanced Compression Model, an exploration software on technology targeting higher video compression capability) compared to the VTM (VVC Test Model reference software) as an anchor. This is the second test of this kind, with the first one being reported in JVET-AB2029 [1]. ECM-12.0, which was the latest version available for the task and VTM-11 with ECM-comparable settings (VTM-11ecm-12) were used for this purpose. Besides the ECM-12.0 software using the set of coding tools according to the common testing conditions, a second version was used where template matching tools were switched off. It is noted that the software version for this configuration includes some bug fixes on top of ECM-12.0 (git hash 6da105bb). In the following the three compared schemes are referred to as VTM, ECM, and ECM-nTM, respectively. The software packages can be found in [2][3].

Tests are presented for the test configurations UHD-RA and HD-LDB, using the random-access (RA) and low-delay (LDB) configuration files available with the corresponding VTM and ECM software. The rate points for the tests were selected from a set of ECM simulations [4]. The corresponding VTM bitstreams were generated using a one-time QP switch (QP+=1) within the sequence to match the VTM bitrate to the ECM. The switching points were chosen such that the ECM rate should never be higher than the VTM rate and that the distance should not exceed about 2% of the VTM rate. The same rate matching operation was applied to match the ECM-nTM version to the ECM.

Three tests were conducted on the provided test set: An expert viewing test with 24 viewers at the Rennes meeting site, a laboratory test with 10 naïve test subjects by VABTech, and a laboratory test with 5 expert viewers by GBTech, both in Rome.

# Test setup

## Logistics

### On-site

At the meeting site, a setup with four OLED displays was employed. This included a PC with a Decklink video board for HDMI connection and SSD drives capable of stable playout of raw YUV 10 bit data at the required frame rate.

| **Test Site** | **On-site** |
| --- | --- |
| **Display, size, connection  (resolution setting)** | 4 × Samsung TQ65S90CAT, 65”, HDMI (3840×2160) connected via an HDMI splitter to the video PC |
| **Viewing distance** | 3 viewers at 1.5H |
| **Viewing angle** | ±75°, 90° (at screen center) |
| **Total number of viewers** | 24 (video coding experts) |

Participating experts confirmed visual acuity and normal colour vision (3 females, 21 males).

### GBTech

| **Test Site** | **GBTech** |
| --- | --- |
| **Display, size, connection  (resolution setting)** | LG OLED, 65” (3840×2160) |
| **Viewing distance** | 1.5H |
| **Viewing angle** | 90° |
| **Total number of viewers** | 5 (expert viewers) |

The viewers were checked for acuity and color blindness (all male), age below 35.

### VABTech

| **Test Site** | **VABTech** |
| --- | --- |
| **Display, size, connection  (resolution setting)** | LG OLED, 65” (3840×2160) |
| **Viewing distance** | 1.5H |
| **Viewing angle** | ±75°, 90° (at screen center) |
| **Total number of viewers** | 10 (naïve viewers) |

The viewers were checked for acuity and color blindness (2 females, 8 males), age below 23.

## Test sequences and quantizer settings

For evaluation a set of sequences available on the JVET ftp site were selected. All bitstreams were produced and crosschecked by JVET volunteers.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **Sequence name** | **fps** | **Frames** | **md5 hash** |
| UHD-RA | BeachMountain2 | 30 | 30 | d5f013f3b24934b3d0bd66bdedad6e80 |
| UHD-RA | CrossRoad2s500 | 50 | 500 | 42a24083fc6e46d77f8f0f8005d195a7 |
| UHD-RA | DrivingPOV3 | 60 | 600 | e81b65724c4235128b2749ccb3b0fb4a |
| HD-LDB | Beatriz | 50 | 500 | fe74cd5046fa033b4f743f42b29e69cd |
| HD-LDB | DOTA2s360 | 60 | 550 | 999ae67b022892a7b70ab4dcc70f1043 |
| HD-LDB | GTAVs090 | 60 | 600 | 43643d686a91a51758c63346fab7eca3 |

The UHD test sequences were selected to reflect a variety of global and local motion, illumination, and detail. These test sequences were encoded using the random access configuration. The HD test sequences represent conversational and gaming type applications. Correspondingly, these sequences were encoded using the low-delay B configuration according to the JVET Common Testing Conditions.

The sequences under evaluation are of 9.5-10sec length. The original files for the tested rate points are available on the JVET ftp site under ahg/JVET-AH\_ECM-test/. For all sequences, four QP values were selected to cover the impairment range from slight to strong impact. The QPs were determined for the ECM.

It was agreed to perform rate matching for the VTM and ECM-nTM rate points in this comparison. For this purpose, it was suggested that the ECM rate should not be higher than the VTM rate and that the distance should not exceed about 2% of the VTM rate. The same rule was applied for the ECM-nTM bitstreams. It was agreed to keep QP values in the case where the overshoot of the ECM rate was less than 1% of the VTM rate.

In order to accommodate time constraints for the on-site tests in Rennes, the ECM-nTM test points were evaluated only for half of the sequences. The assignment of test points is indicated below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Category** | **Sequences** | **VTM** | **ECM** | **ECM-nTM** |
| UHD-RA | BeachMountain2 | x | x | x |
| UHD-RA | CrossRoad2s500 | x | x | - |
| UHD-RA | DrivingPOV3 | x | x | x |
| HD-LDB | Beatriz | x | x | x |
| HD-LDB | DOTA2 | x | x | - |
| HD-LDB | GTAVs090 | x | x | - |

The selected ECM QPs and the matching VTM QPs are provided in the accompanying spreadsheet.

## Test method and test design

The Degradation Category Rating (DCR) method was applied for the subjective evaluation [5]. The test sequences were evaluated using the 11-grade scale shown in Figure 1 below.

Ein Bild, das Text, Screenshot, Schrift, Zahl enthält.

Automatisch generierte Beschreibung

Figure 1: Meaning of the 11 grades numerical scale

Each basic test cell (BTC) is structured as followed:

**“Original” (1sec) – [uncompressed sequence] (10sec) – “A” (1sec) – [PVS] (10sec) – “Vote <N>” (5sec)**

PVS denotes the processed video sequence under evaluation.

### On-site test

Two test sessions were designed with 38 scores for the UHD sequences and 34 scores for the HD sequences. All test sessions were designed inserting a stabilization phase as suggested in ITU-R BT.500-14. The session duration was chosen to be no longer than 20 minutes to avoid a fatigue impact. The test sessions included trapping BTCs where the original uncompressed sequence was shown for evaluation. To accommodate the different frame rates of the test sequences at both resolutions, the playout scripts were designed to include sufficient synchronization time for the displays.

The participating experts were trained with one training session for UHD resolution and one session for HD resolution (6 votes each). All test sequences under evaluation were occurring at least once in the training sessions, and a selection of rate point representing the expected impairment range was presented for both resolutions. Before the presentation of the training sessions, the experts were instructed on the meaning of the impairment scale. Any occurring requests and questions on the test procedure or the scale were answered. For the training, the following method was used:

- Explain procedure and scoring scale.   
- Run test session (6 BTCs here, 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.  
 - Answering of remaining questions, no further communication of scores.

The UHD and HD test sessions were both presented with a viewing distance of 1.5H of the UHD display for the center seat. The HD sequences were displayed without scaling in the center of the UHD area with a mid-gray padding around them. Thereby, an effective 3H viewing distance for the HD content was achieved.

### Laboratory test with expert viewers (GBTech)

For the whole experiment four test sessions were designed each made of 24 BTCs. Sessions were all 10” long to maintain low the fatigue level.

During the test run with five expert viewers, only the two UHD test sessions were run.

Experts didn’t run any formal training other than a quick informal view to the most relevant cases, run before the forma test, but to overcome the lack of formal training the two UHD sessions were run twice averaging the scores.

No special viewing trick was applied (slow motion, frame freeze, repeat).

### Laboratory test with naïve viewers

This laboratory viewing test was carried out with ten naïve viewers that run the four HD and UHD test sessions. In this experiment each session was made of 27 BTC (i.e. inserting three dummy BTC at the beginning of each session, as recommended by ITU-R BT.500-14.

All viewers were screened for acuity and color blindness.

Viewers were trained with one training session made of UHD clips followed by HD clips, for a total of 16 BTCs during which a comprehensive range of quality was presented at both resolutions.

The kind of test (impairment scale) was explained to the naïve viewers as well as the timing of the voting process.

Questions were answered before and after the training session.

# Results and analysis

## Data processing

### On-site test

As a first step, the complete set of votes experts was screened for the trapping sequences. In both, UHD and HD test sessions, a single viewer below score of 8 for the original. The results of corresponding session with that trapping sequence were not regarded for the affected viewer. The data then was further screened applying z-scores with a threshold of 2.5. As a final processing step, isolated outliers -considered to be obvious errors- were removed.

### Laboratory test with expert viewers

The results were double checked for inconsistency applying a “score slope check” (SSC), to be in line with compression rates; in other terms high compression rate quality cannot be scored better than low compression rate quality.

This new data post screening method was applied also to the other test conducted using naïve viewers.

### Laboratory test with naïve viewers

The post screening of data included Pearson analysis and the above described SSC analysis; for each viewers only one case of SSC fault was admitted (more than one fault leads to subject exclusion); no viewers was excluded after this analysis.

## Subjective results

The measured MOS values of the reconstructed video on the 11-grade scale are plotted over the bit rate of the corresponding bitstream. The ±95% confidence intervals (CIs) for the MOS values are indicated.

The MOS results as well as the bitrates and the PSNR values are provided in the accompanying spreadsheet.

### On-site test

Since the confidence intervals of the expert viewing results overlap for many test points, reported BD-rate savings are considered to only be rough indications. The BD rate numbers have been computed using the RDplot tool [7].

|  |  |  |  |
| --- | --- | --- | --- |
| **Category** | **Sequence** | **ECM** | **ECM-nTM** |
| UHD-RA | BeachMountain2 | -12.65 | -29.73 |
| UHD-RA | CrossRoad2s500 | -25.26 | -/- |
| UHD-RA | DrivingPOV3 | -28.81 | -26.06 |
| HD-LDB | Beatriz | -21.54 | -16.62 |
| HD-LDB | DOTA2s360 | 1.88 | -/- |
| HD-LDB | GTAVs090 | -18.19 | -/- |

#### UHD-RA

Ein Bild, das Reihe, Diagramm, Text, Steigung enthält.

Automatisch generierte Beschreibung

#### HD-LD

Ein Bild, das Reihe, Diagramm, Steigung, Text enthält.

Automatisch generierte Beschreibung

### Laboratory test with expert viewers

For this laboratory test, since only five experts participated in the tests, no CI values are reported.

#### UHD-RA

Ein Bild, das Reihe, Diagramm, Text enthält.

Automatisch generierte Beschreibung

### Laboratory test with naïve viewers

#### UHD-RA

Ein Bild, das Reihe, Diagramm, Steigung, Text enthält.

Automatisch generierte Beschreibung

#### HD-LD

Ein Bild, das Reihe, Diagramm, Steigung, Text enthält.

Automatisch generierte Beschreibung

# Discussion and conclusions

Both, the laboratory test conducted with naïve viewers and the on-site test with experts demonstrate a clear visual benefit of the ECM when compared to the VTM for a significant number of cases across the three tests.

Due to the number of viewers, indicative BD rate figures are computed for the expert viewing results in Rennes only. The indicative BD rate measurements for the ECM generally show that both, the rate savings computed by PSNR and by MOS are in a comparable range. Indicative BD rate gains are in the range of about 25-29% for the UHD test sequences when considering both, the ECM and ECM-nTM configurations. For the HD test sequences, the indicative BD rate gains are around 20% for Beatriz and GTAVs900. On the DOTA sequence, no gain is observed.

For the test sequence BeachMountain2, the subjective results of the on-site tests in Rennes reveal potential issues with the performance of the template matching tools in the ECM. These results seem to suggest that further study of such tools may be useful. Similar indications may be deducted from the laboratory tests.

# References

1. V. Baroncini, J.-R. Ohm, M. Wien, “Visual quality comparison of ECM/VTM encoding,” Doc. JVET-AB2029, Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 5, 28th meeting, Mainz, DE, Oct. 2022.
2. JVET, VTM software package, <https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM/-/tree/VTM-11.0>.
3. JVET, ECM software package, <https://vcgit.hhi.fraunhofer.de/ecm/ECM/-/tree/ECM-12.0>.
4. M. Wien, “[AHG4] AHG meeting report on ECM expert viewing preparations,” Doc. JVET-AH0041, Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 5, 34th meeting, Rennes, FR, Apr. 2024.
5. Recommendation ITU-T P.910 (2008), *Subjective video quality assessment methods for multimedia applications*.
6. Recommendation ITU-R BT.500-14 (2019), Methodologies for the subjective assessment of the quality of television images.
7. J. Schneider, J. Sauer, M. Wien, “RDPlot - An Evaluation Tool for Video Coding Simulations”, Visual Communications and Image Processing VCIP’21, Munich, Dec. 2021.

# Annex: Expression of thanks

* ECM and VTM encoding and crosscheck:   
  Alibaba, Bytedance, Ericsson, Fraunhofer HHI, Google, Huawei, InterDigital, Kwai, MediaTek, Oppo, Qualcomm, Vivo.
* Provision of equipment at meeting site: Couvent des Jacobins, Fraunhofer HHI, RWTH, and VABTech.
* Laboratory tests: VABTech, GBTech.
* The 24 expert viewers participating in the tests.
* Christian Lehmann and Fabrice Urban for technical assistance.