SA4-(AH) Video SWG post 132S4aV250050

Online, 17 June – 7 July 2025

**Source: Philips International B.V.**

**Title: [FS\_Beyond2D] pCR on Multi-view plus Depth Evaluation**

**Spec: 3GPP TR 26.956 1.0.0**

**Agenda item: 3.7**

**Document for: Agreement**

**1. Introduction**

At the SA4#132 meeting evaluation of multi-view plus depth was added to version 0.0.9 of the PD.

**2. Reason for Change**

This pCR adds the same information to the TR.

**4. Proposal**

It is proposed to agree the following changes to 3GPP TR 26.956.

\* \* \* First Change \* \* \* \*

Modify section 2 as follows:

# 2 References

(…)

[MD-4] ISO/IEC 23090-10:2022 (Amd1), “Information Technology — Coded Representation of Immersive media — Part 10: Carriage of Visual Volumetric Video-Based Coding Data”

[MD-5] Guede et al., IBC 2023, “Efficient Delivery and Rendering on Client Devices via MPEG-I Standards for Emerging Volumetric Video Experiences”. https://www.ibc.org/technical-papers/ibc2023-tech-papers-efficient-delivery-and-rendering-on-client-devices-via-mpeg-i-standards-for-emerging-volumetric-video-experiences/10277.article

[MD-6] Dziembowski, B. Kroon, J. Jung (Eds.), Common test conditions for MPEG immersive video, ISO/IEC JTC 1/SC 29/WG 04 N 0659, April 2025, Online.

[MD-7] D. Mieloch (Ed.), Verification test report of MPEG immersive video, ISO/IEC JTC 1/SC 29/WG 04 N 0341, April 2023, Antalya.

[MD-8] B. Brand, Michel Bätz, Joachim Keinert, Camorph: a toolbox for conversion between camera parameter conversions, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, url: https://github.com/Fraunhofer-IIS/camorph, 2022.

[MD-9] ISO/IEC 14496-15:2024: Carriage of network abstraction layer (NAL) unit structured video in the ISO base media file format.

[MD-10] ISO/IEC 23090-12:2023: MPEG immersive video.

[MD-11] ITU-T H.265:2024 | ISO/IEC 23008-2:2024, Annex G: Multiview high efficiency video coding (MV-HEVC).

[MD-12] J.Y. Jeong, J. Kim, B.H. Lee, K.J. Yun, W. Cheong, S.H. Yoo, [INVR] Multiview dataset Classroom and Bartender for 3D INVR activity, ISO/IEC JTC1/SC29/WG4 MPEG VC/M69151, Sapporo, Japan, July 2024.

[MD-13] Overview and Efficiency of Decoder-Side Depth Estimation in MPEG Immersive Video, IEEE Transactions on Circuits and Systems for Video Technology, doi: 10.1109/TCSVT.2022.3162916, ode: <https://gitlab.com/mpeg-i-visual/ivde>.

[MD-14] Encoder guidelines for MPEG immersive video, ISO/IEC JTC 1/SC 29/WG 04/N 660, April 2025, url: https://www.mpeg.org/wp-content/uploads/mpeg\_meetings/150\_OnLine/w25085.zip, Online.

\* \* \* Next Change \* \* \* \*

Modify section 7.4.5 as follows:

### 7.4.5 Performance Metrics and Requirements

The tests are run for a chosen level as described in clause 7.4.6. Bitstreams are provided. Camera calibration, depth estimation, and encoding are not evaluated.

The test will have four rate points and QP values are selected for each sequence to approximately match the 5 to 50 Mbps range. When saturation occurs before 50 Mbps a lower value may be chosen in consultation. When there are multiple video components or packed regions then the other QP values need to be directly derived from the texture QP using an equation or a look-up table. (They cannot depend on the sequence.)

The QMIV tool, available at <https://gitlab.com/mpeg-i-visual/qmiv>, is available to compute full-reference objective metrics:

- PNSR (PSNR)

- Weighted sphere PSNR (WS-PSNR)

- Immersive video PSNR (IV-PSNR)

- Immersive video SSIM (IV-SSIM)

All source views that were used for encoding are provided. Each source view is reconstructed by decoding and rendering (view synthesis). The QMIV tool is then run on all source views and the score is averaged over all views.

Depending on bit rate, quality of depth maps and rendering, either the video codec or view synthesis is the limiting factor.

There is experience in testing of multi-view plus depth video in MPEG context. The test conditions as described are a simplification and evolution of the common test conditions for MIV defined in [MD-6].

The main challenge with testing of multi-view plus depth video is that codecs are asymmetric. The input is a number of source views (with depth maps), and the output of the decoder + renderer can be any viewport within a spatial region around those source views. In the mentioned CTC two tests are used:

- Objective evaluation at source view positions

- Subjective evaluation of pose trace videos (dynamic viewports)

This has resulted in a lack of correlation between objective and subjective results, but despite that it is the best-known approach. Alternatives that have been tried and dismissed (for now):

- Objective evaluation at dynamic viewports: It includes view synthesis in the reference condition and this skews the results towards a specific renderer. It prevents an A/B comparison of different renderers.

- Subjective evaluation at source view positions: This is not how the end-user will interact with the content, and it does not evaluate artifacts due to viewport dynamics.

For this test, because the aim is to prove feasibility of a scenario, objective evaluation may be sufficient, especially when supplemented with (informal) real-time demonstration of the same bitstreams that were used for objective evaluation.

\* \* \* Next Change \* \* \* \*

Replace section 7.4.7 as follows:

### 7.4.7 Test Sequences

The evaluation has been performed on the sequences listed in Table 7.4.7-1. Only the first 65 frames were used of each test sequence.

Table 7.4.7-1: Test sequences for the evaluation of the scenario

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sequence | Provider | Frames | Resolution | Bit depth | Color format |
| Breakfast  (Figure 7.4.7.1-1) | InterDigital | 97 @ 30 Hz  3.2 s | 1920 x 1080  5 x 3 views | texture: 10b  depth: 16b | texture: 4:2:0 BT.709  depth: 4:2:0 full range linear |
| Bartender  (Figure 7.4.7.2-1) | ETRI | 300 @ 30 Hz  10.0 s | 1920 x 1080  21 views | texture: 10b  depth: 16b | texture: 4:2:0 BT.709  depth: 4:2:0 full range linear |
| DanceMoves  (Figure 7.4.7.3-1) | Philips | 449 @ 15 Hz  29.9 s | 1920 x 1080  6 views | texture: 10b  depth: 16b | texture: 4:2:0 BT.709  depth: 4:2:0 full range linear |

#### 7.4.7.1 Test sequence Breakfast

The Breakfast sequence (Figure 7.4.7.1-1) is part of the MIV CTC [MD-6] but has not been used for the development of ISO/IEC 23090-12:2023. It consists of 97 frames @ 30 fps, and was captured using a 5 x 3 planar rig of BlackMagic Micro Studio cameras. The total size of the rig is about 1 m wide and 0.5 m tall. The cameras have a field of view of about 66° by 40°. The scene has been shot in the dining room of Chateau de la Ballue, 35560 Bazouges-la-Pérouse, France. While InterDigital originally provided depth maps, later these have been replaced by depth maps that were produced by ETRI using their internal tools, comprising block matching based on plane-sweeping, cost aggregation, semi-global matching, and depth refinement.

For the purpose of this study, the license has been updated. The test sequence and derived work are hosted by InterDigital on their Aspera server.

A group of people sitting at a table

AI-generated content may be incorrect. A group of people sitting at a table

AI-generated content may be incorrect.

Figure 7.4.7.1-1: Breakfast sequence (view 7, frame 0)

#### 7.4.7.2 Test sequence Bartender

The Bartender sequence (Figure 7.4.7.2-1) was originally provided by ETRI for the 3D Implicit Neural Video Representation (3D-INVR) activity in SC 29/WG 4 [MD-12], and it is currently being used as a reference sequence in the MPEG Gaussian splat coding (GSC) exploration in WG 4/WG 7. The sequence has not been used for the development of ISO/IEC 23090-12 MIV. It consists of 300 frames @ 30 fps, and was captured using a 7 x 3 planar rig of BlackMagic Micro Studio 4K cameras. The total size of the rig is about 3 m wide and 0.5 m tall. The calibration of the camera system was done using Reality Capture, mainly relying on bundle adjustment.

Permission was obtained form ETRI to use this sequence for this study. The license requires citation of [MD-12]. Because the cited document is not publicly available, it was agreed with ETRI to distribute the cited document with the test sequence.

As the sequence was provided without depth maps, Philips has generated depth maps using Immersive Video Depth Estimator (IVDE) 8.0 [MD-13] by Pozań University of Technology. We have used the default parameters of the software, except that we disabled temporal enhancement. Each frame was estimated in parallel and the resulting YUV files were concatenated.

NOTE: Visual inspection of the depth maps shows that the quality of the depth maps is below that of Breakfast, and this has an impact on the evaluation.

ETRI and Philips have agreed that Philips will provide the source material, original MPEG contribution, copyright license, and derived work on a Philips-managed OneDrive server, and participants of this study may request access based on their e-mail address, by sending an e-mail to the contact person: Bart Kroon <<bart.kroon@philips.com>>.

Figure 7.4.7.2-1: Bartender sequence (view 10, frame 0)

#### 7.4.7.3 Test sequence DanceMoves

The DanceMoves sequence was captured by Philips on location in Veghel with the help of a professional production company, a self-built capture rig, and volunteer actors. The sequence has not been used for the development of any standard. It consists of 449 frames @ 15 Hz, and was captured using a linear rig of six Azure Kinect cameras. The total size of the rig is about 60 cm wide. Due to limitations with the capture system (a single laptop with USB 3 interface) the frame rate was limited to 15 fps.

As the depth maps of the Azure Kinect cameras were cropped compared to the texture frames, Philips has generated new depth maps using Immersive Video Depth Estimator (IVDE) 8.0 [MD-13] by Pozań University of Technology. We have used the default parameters of the software, except that we disabled temporal enhancement. Each frame was estimated in parallel and the resulting YUV files were concatenated.

NOTE: Visual inspection of the depth maps shows that the quality of the depth maps is below that of Breakfast, and this has an impact on the evaluation.

Philips will provide the source material, copyright license and derived work on a Philips-managed OneDrive server, and participants of this study may request access based on their e-mail address, by sending an e-mail to the contact person: Bart Kroon <<bart.kroon@philips.com>>.

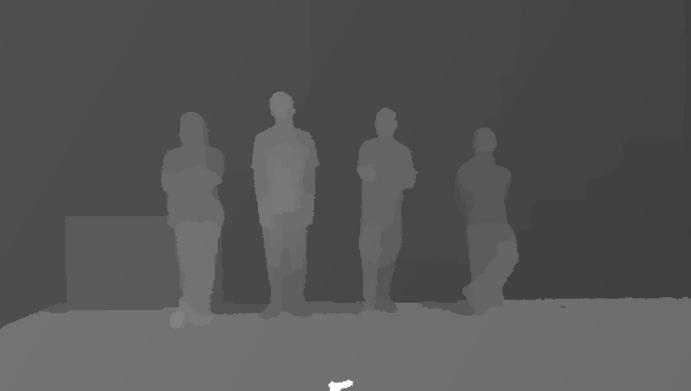
 

Figure 7.4.7.3-1: DanceMoves sequence (view 3, frame 0)

\* \* \* Next Change \* \* \* \*

Remove section 7.4.8 completely.

\* \* \* Next Change \* \* \* \*

Add a new section 9.x as follows:

## 9.x Scenario 3: Streaming of Multi-view plus depth Produced Content

### 9.x.1 Evaluation Overview

### 9.x.2 Reference Sequences

The reference sequences that were introduced in section 7.4.7 were selected.

### 9.x.3 Performance Metrics

The performance metrics that were introduced in section 7.4.5 were selected.

### 9.x.4 Candidate Solutions

#### 9.x.4.1 Solution 1: HEVC Main10 MIV Main

##### 9.x.4.1.1 Introduction

The evaluation framework is available in the mvd/ folder of <https://github.com/5G-MAG/rt-beyond2d-evaluation-framework>, [tag v0.1.0](https://github.com/5G-MAG/rt-beyond2d-evaluation-framework/tree/v0.1.0), under the 5G-MAG public license.

[Ed.(BK): To be published as version 0.1.0.]

##### 9.x.4.1.2 Reference Software

The software that has been used for the evaluation of the scenario is listed in Table 9.x.4.1.2-1. All software has been built from source using Python 3.12, LLVM 18.1.8 with help of the install.py script of TMIV, as follows:

*# environment with python, clang and clang++ on the path*

git clone https://gitlab.com/mpeg-i-visual/tmiv.git

cd tmiv

python -m venv venv

. venv/bin/activate

python -m pip install --upgrade pip

pip install -r requirements.txt

scripts/install.py clang-release

Table 9.x.4.1.2-1: Software used for the evaluation of the scenario

|  |  |  |
| --- | --- | --- |
| Software | URL | Version |
| Test model for MPEG immersive video (TMIV) | <https://gitlab.com/mpeg-i-visual/tmiv> | 24.0 |
| HEVC test model (HM) | <https://vcgit.hhi.fraunhofer.de/jvet/HM> | 18.0 |
| Quality metrics for immersive video (QMIV) | <https://gitlab.com/mpeg-i-visual/qmiv> | 2.0 |

HM 18.0 and Kvazaar 2.3.1 have been compared in MPEG context for the coding of MIV video sub-bitstreams [MD-14]. HM 18.0 was selected for this study because it has a better rate-distortion characteristic in general. However, because HM lacks support for delta QP maps, packed video support was disabled in TMIV.

##### 9.x.4.1.3 Parameter Settings

For this study, content was encoded using TMIV and HM. Encoding of MIV bitstreams using TMIV and HM involves three steps:

1. Run the TMIV encoder to output a raw YUV video file for each video sub-bitstream, and a partial MIV bitstream with patch parameters and video parameters. The main work of the TMIV encoder is to prune pixels, patch patches, and generate atlas frames.

2. Run HM TAppEncoder to encode each YUV file.

3. Run the TMIV multiplexer to combine the partial MIV bitstream and the coded video sub-bitstream into a full MIV bitstream (a V3C sample stream).

All sequences have been encoded using the configurations in Table 9.x.4.1.3-1. The purpose of having multiple configurations is to illustrate the impact of pixel rate on rate-distortion characteristics. Because this is a new representation there is no anchor.

- The *full views* (FV) condition codes the texture and geometry video component of each view as a separate HEVC Main 10 video sub-bitstream. This condition gives the highest quality, but also the largest pixel rate and bit rate. It serves as an upperbound of what can be achieved with the current test sequences and software if pixel rate is not a concern. The MIV level depends on the input.

- The *MPEG* *MIV main* (A) condition is part of the MIV CTC anchor, defined in ISO/IEC JTC 1/SC 29/WG 04 N 0659. It results in two atlases, each with a texture and geometry component, thus resulting in four video sub-bitstreams. It causes TMIV to select a number of source views based on an available pixel budget. The resulting bitstreams have MIV level 3.5. Some source views are selected to be basic views and they are fully coded. Some other views are selected as semi-basic views and they are placed in full in the atlas, but then some patches can be placed on top. Finally there are additional views from which only patches are taken (Figure 9.x.4.1.5.1-1).

- The *Synthesize center view* (SCV) condition was designed for this study because the pixel rate of the MIV CTC may be too high for mobile devices. The atlas has a single synthesized center view plus patches of the source views. The aim of this condition is to provide a MIV level 2.5 result by lowering the pixel rate compared to the A condition (Figure 9.x.4.1.5.1-1-2).

Table 9.x.4.1.3-1: Encoder conditions

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Condition | Profile | Level | Abbreviation | Directory name |
| Full views | HEVC Main 10 | - | FV | config/full\_views |
| MIV main anchor | HEVC Main 10 MIV 2 (FDIS 23090-12:—) | 3.5 | A | config/miv\_main\_anchor |
| Synthesize center view | HEVC Main 10 MIV Extended (23090-12:2023) | 2.5 | SCV | config/synthesize\_center\_view |

Encoding was performed by running the encode.py script of TMIV with appropriate parameters. For all sequences the first 65 frames were encoded. It executes the TMIV Encoder, HM, and the TMIV Multiplexer with appropriate parameters. For example:

*TMIV\_DIR*/bin/encode.py -i *INPUT\_DIR* -o out -s D02 -n 65 \

-r RP0 -f 0 -v HM -j 4 -t *TMIV\_DIR* \

--config-dir share/config \

-c config/synthesize\_center\_view/SCV\_1\_TMIV\_encode.json \

-m config/synthesize\_center\_view/SCV\_3\_TMIV\_mux.json \

-C share/config/hm/encoder\_randomaccess\_main10.cfg

The only substantial difference between the encoder conditions is the TMIV encoder configuration because the TMIV multiplexer configuration is trivial and the HM configuration is kept to the same random-access configuration for all conditions.

The rate point RP0 is a result without coding of the video sub-bitstreams that can be used to determine how much quality is lost by the pixel pruning prior to video coding. Rates RP1 .. RP4 correspond to the following QP values in Table 9.x.4.1.3-2. The same QP values are used for all sequences and all encoder conditions. The geometry QP is derived from the texture QP as done for the MIV CTC [MD-6]. Hence, virtually no QP tuning has been performed at all.

Table 9.x.4.1.3-2: QP values for all sequences and encoder conditions

|  |  |  |
| --- | --- | --- |
| Rate point | Texture | Geometry |
| RP1 | 20 | 2 |
| RP2 | 30 | 10 |
| RP3 | 40 | 18 |
| RP4 | 50 | 26 |

##### 9.x.4.1.4 Distribution

##### 9.x.4.1.5 Evaluation Results

###### 9.x.4.1.5.1 Example atlas frames

The full views (FV) condition encodes each component of each view separately, e.g. resulting in 30 separate 1920 x 1080 videos for the Breakfast sequence. Figure 9.x.4.1.5.1-1 and Figure 9.x.4.1.5.1-2 provide examples of atlas frames for the MIV main anchor (A) and synthesize center view (SCV) conditions. A comparison of pixel rates is provided in Tabe 9.x.4.1.5.2-1. Note that the size of each atlas depends on the sequence and on the encoding condition. This is because TMIV calculates the atlas frame size based on a number of inputs.

A group of people sitting at a table

AI-generated content may be incorrect. A collage of a room with many objects

AI-generated content may be incorrect. A group of people sitting at a table

AI-generated content may be incorrect. A black and white image of a person in a room

AI-generated content may be incorrect.

Figure 9.x.4.1.5.1-1: Video components of condition A with left to right: texture for atlas 0 and 1, geometry for atlas 0 and 1

A group of people sitting at a table

AI-generated content may be incorrect. A group of people sitting at a table

AI-generated content may be incorrect.

Figure 9.x.4.1.5.1-2: Video components of condition SCV with left texture and right geometry

###### 9.x.4.1.5.2 Pixel rate and MIV levels

The pixel rates per video sub-bitstreams and the aggregate pixel rate are depicted in Table 9.x.4.1.5.2-1. The MIV level is based on the luma picture size and aggregate luma sample rate level limits as provided in ISO/IEC FDIS 23090-12:— Table A.7. Note that the MIV level for the FV condition is determined mainly by the aggregate luma sample rate because the luma picture size is only 1920 x 1080 but there are many video sub-bitstreams. Note that the coding of DanceMoves for condition A is inefficient because not all space in the atlases is used.

Table 9.x.4.1.5.2-1: Pixel rates for all sequences and conditions:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Condition | Sequence | Components | Sizes | Aggregate size (# luma samples) | Aggregate luma  sample rate | MIV level |
| FV | Breakfast | 15 x texture  15 x depth | 1920 x 1080  1920 x 1080 | 62.2 M | 1.87 G/s | 3.0 |
| FV | Bartender | 21 x texture  21 x depth | 1920 x 1080  1920 x 1080 | 87.1 M | 2.61 G/s | 3.5 |
| FV | DanceMoves | 6 x texture  6 x depth | 1920 x 1080  1920 x 1080 | 24.9 M | 0.373 G/s | 2.0 |
| A | Breakfast | 2 x texture  2 x geometry | 1920 x 4608  960 x 2304 | 22.1 M | 0.664 G/s | 2.5 |
| A | Bartender | 2 x texture  2 x geometry | 1920 x 4608  960 x 2304 | 22.1 M | 0.664 G/s | 2.5 |
| A | DanceMoves | 2 x texture  2 x geometry | 1920 x 4608  960 x 2304 | 22.1 M | 0.664 G/s | 2.5 |
| SCV | Breakfast | 1 x texture  1 x geometry | 2880 x 2432  1440 x 1216 | 8.76 M | 0.263 G/s | 2.0 |
| SCV | Bartender | 1 x texture  1 x geometry | 2944 x 2368  1472 x 1184 | 8.71 M | 0.261 G/s | 2.0 |
| SCV | DanceMoves | 1 x texture  1 x geometry | 2048 x 3456  1024 x 1728 | 8.85 M | 0.133 G/s | 2.0 |

###### 9.x.4.1.5.3 Rate-distortion characteristics

A graph of a graph with different colored lines

AI-generated content may be incorrect.

Figure 9.x.4.1.5.3-1: Rate distortion curves for Breakfast for all three coding conditions

A graph with different colored lines

AI-generated content may be incorrect.

Figure 9.x.4.1.5.3-2: Rate distortion curves for Bartender for all three coding conditions

A graph of a function

AI-generated content may be incorrect.

Figure 9.x.4.1.5.3-3: Rate distortion curves for DanceMoves for all three coding conditions

###### 9.x.4.1.5.4 Pose trace videos

For each bitstream, that is for each sequence for each encoder condition and for each rate RP0 .. RP4, three pose trace videos have been rendered. A bitstream can be decoded and rendered using a command like this:

*TMIV\_DIR*/bin/TmivDecoder -j 1 -n 32 -N 128 -s D02 -r RP3 -P p01 \

-c config/synthesize\_center\_view/SCV\_4\_TMIV\_decode.json \

-p inputDirectory out -p outputDirectory out \

-p configDirectory share/config

The decoder configurations differ only in path formats: there is no out-of-band information for RP1 .. RP4.

###### 9.x.4.1.5.5 Availability of test data

The source video data (texture and depth), camera parameters, pose trace definitions, bitstreams and pose trace videos are available. For Breakfast, the information is hosted by InterDigital. For Bartender and DanceMoves, the information is hosted by Philips. Access will be provided to participants on request.

##### 9.x.4.1.6 Network Requirements

#### 9.x.4.2 Solution 2: <Name>

NOTE: While currently only MIV + HEVC has been tested, the scenario includes MV-HEVC and is open to the evaluation of other codecs.

9.x.5 Summary of Evaluation\* \* \* End of Changes \* \* \* \*