**3GPP TSG-SA WG4 Meeting #130S4-241841r1**

**USA, Orlando, 18 – 22 November 2024**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *CR-Form-v12.2* | | | | | | | | |
| **CHANGE REQUEST** | | | | | | | | |
|  | | | | | | | | |
|  |  | **CR** |  | **rev** |  | **Current version:** |  |  |
|  | | | | | | | | |
| *For* [***HE******LP***](http://www.3gpp.org/3G_Specs/CRs.htm#_blank)*on using this form: comprehensive instructions can be found at* [*http://www.3gpp.org/Change-Requests*](http://www.3gpp.org/Change-Requests)*.* | | | | | | | | |
|  | | | | | | | | |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Proposed change affects:*** | UICC apps |  | ME |  | Radio Access Network |  | Core Network |  |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | | | |
| ***Title:*** | pCR on Scenario: Streaming of Beyond 2D Produced Content | | | | | | | | | |
|  |  | | | | | | | | | |
| ***Source to WG:*** | Nokia, Philips, Interdigital, Deutsche Telekom, Fraunhofer HHI, Sony, China Mobile, Huawei | | | | | | | | | |
| ***Source to TSG:*** |  | | | | | | | | | |
|  |  | | | | | | | | | |
| ***Work item code:*** |  | | | | |  | ***Date:*** | | | 2024-11-12 |
|  |  | | | |  | |  | | |  |
| ***Category:*** | B |  | | | | | ***Release:*** | | | Rel-19 |
|  | *Use one of the following categories:* ***F*** *(correction)* ***A*** *(mirror corresponding to a change in an earlier release)* ***B*** *(addition of feature),* ***C*** *(functional modification of feature)* ***D*** *(editorial modification)*  Detailed explanations of the above categories can be found in 3GPP [TR 21.900](http://www.3gpp.org/ftp/Specs/html-info/21900.htm). | | | | | | | | *Use one of the following releases: Rel-8 (Release 8) Rel-9 (Release 9) Rel-10 (Release 10) Rel-11 (Release 11) … Rel-16 (Release 16) Rel-17 (Release 17) Rel-18 (Release 18) Rel-19 (Release 19)* | |
|  |  | | | | | | | | | |
| ***Reason for change:*** | | An evaluation scenario is proposed that handles the streaming of produced Beyond 2D (B2D) content that provides experiences beyond what is achievable with traditional 2D video. The scenario allows to evaluate the streaming of high-quality, professionally captured and produced B2D video content. | | | | | | | | |
|  | |  | | | | | | | | |
| ***Summary of change:*** | | New clause with scenario description on streaming of B2D produced content | | | | | | | | |
|  | |  | | | | | | | | |
| ***Consequences if not approved:*** | | Streaming of B2D produced content cannot be evaluated. | | | | | | | | |
|  | |  | | | | | | | | |
| ***Clauses affected:*** | |  | | | | | | | | |
|  | |  | | | | | | | | |
|  | | **Y** | **N** |  | | | |  | | |
| ***Other specs*** | |  |  | Other core specifications | | | | TS/TR ... CR ... | | |
| ***affected:*** | |  |  | Test specifications | | | | TS/TR ... CR ... | | |
| ***(show related CRs)*** | |  |  | O&M Specifications | | | | TS/TR ... CR ... | | |
|  | |  | | | | | | | | |
| ***Other comments:*** | | Initial text on this scenario is included in the PD (v0.0.4) clause 2.2. | | | | | | | | |
|  | |  | | | | | | | | |
| ***This CR's revision history:*** | |  | | | | | | | | |

**== CHANGE 1 (all new) ===**

## 7.x Scenario x: Streaming of Beyond 2D Produced Content

### 7.x.1 Scenario name

Streaming of Beyond 2D Produced Content

### 7.x.2 Motivation for the scenario

The proposed scenario handles the streaming of produced Beyond 2D (B2D) content that provides experiences beyond what is achievable with 2D content. The scenario allows to evaluate the streaming of high-quality, professionally captured and produced B2D video content.

B2D content may be in the form of volumetric video, which is a frame-based immersive experience whereby each frame represents a volumetric region in 3D space in which any point is either non-occupied or having a color that may depend on the viewing direction.Volumetric video has the potential to provide an immersive and interactive experience for use cases in diverse domains such as e.g. education, entertainment, and industrial monitoring. Streaming of volumetric video has been previously considered in 3GPP for some use cases described in TR 26.928 [R1] and TR 26.998 [R2].

NOTE: Streaming of volumetric video with single asset is not a part of this scenario and handled separately in the scenario described in clause 7.y.

This scenario is based on the multi-view video representation formats that is defined in clause 4.3.4. Capturing setups and production software are available as described in the related representation format definitions. Contribution, compression and storage formats for multi-view video are available, see clause 7.x.3. It is expected that segmented media delivery will be used based on DASH and ISOBMFF. Carriage of coded media using ISOBMFF has been specified for MIV in [R4]. Hardware video decoder capabilities can be used for all pixel data. Rendering and display systems for multi-view video are described in clause 4.3.4.3.

### 7.x.3 Description of the scenario

This scenario considers on-demand streaming of B2D produced content to a UE. Capture setup, production tools and workflows for multi-view video capture systems and production tools are described in clause 4.3.4.2. Contribution, compression and storage formats are linked to the representation format. Well-defined contribution formats exist that carry the raw texture/depth images and camera parameters, e.g. as described in clause 4.3.4.2. Compression formats for multi-view video are described in clause 4.3.4.4. One codec that can be used to realize this scenario is MPEG Immersive Video (MIV) specified in ISO/IEC 23090-12. Below the workflow with MIV is described.

The multiple camera views and depth maps are encoded to create a unified representation. An example could be MIV constrained to one or more atlases and packed video data. The single video sub-bitstream per atlas would be encoded with HEVC Main10 profile. The bitstream contains all camera parameters that are necessary for 6DoF rendering. Each atlas is independently renderable.

NOTE: This study may include multi-view content for which some or all views lack depth information.

Figure 7.x.3-2 provides an example of a MIV encoder flow.



Figure 7.x.3-2: MIV encoder example

* Patch Extraction and Filtering: extraction of regions from the texture and depth map for the purpose of pixel-rate reduction and allowing object interactivity at the client.
* Background Sprite Extraction: The ground surface and far-away background can be represented by a single sprite texture with depth. This greatly reduces the required pixel space.
* Atlas Generation: The patches and sprite are packed in an atlas such that both the pixel area is optimally used and the temporal correlation is retained to guarantee an acceptable bitrate.

An example of multi-view video encoding has been described in the paper [R5].

The encoded bitstream is encapsulated to ISOBMFF according to the rules of the used codec.

For example, an MIV bitstream may be packaged in one track, or multiple tracks where the packed video data is one track, common atlas data is one track, and atlas data is another track. ISO/IEC 23090-10 [R4] specifies how to map MIV (V3C) onto ISOBMFF, file format and DASH.

When a scene is represented by multiple atlases, only one of them may be decoded based on the viewing position. This is called atlas-level sub-bitstream access. In the case of DASH, switching atlas would amount to changing tracks.

The decoder(s) will make use of hardware video decoder capabilities for all pixel data, and metadata describing information needed for rendering is decoded/parsed by a CPU.

Rendering and display systems for multi-view video are described in clause 4.3.4.3.

In the case of MIV, efficient rendering can be performed directly from the atlas after decoding to GPU memory:



* Patch Depth Binning: patches are warped to the target view and sorted on depth using an efficient histogram-based method
* Back-to-front View Synthesis: patches with the same depth from multiple source views are blended together using view-angle based weighting.
* Blend and Composite: After blending over views (per depth layer), layers are composited back to front.

The codec should support a random access reference frame structure.

For example, the MIV access units and the video sub-bitstream are organized using a random access reference frame structure. All sub-bitstreams could have the same prediction structure, but atlas data and common atlas data frames may be skipped.

The common atlas data with camera parameters could only change infrequently (once per second or less), for instance each time an online camera calibration is updated. While it is possible to transmit common atlas data at non-IRAP frames, this would be not desirable in this scenario.

The atlas data with patch information may be static when transmitting only full views or dynamic when an encoder selects regions of the source views for transmission based on e.g. occlusion detection or depth segmentation. In MIV, patch information is always intra-coded (unlike V3C in general).

All decoder and renderer processes are real-time. End-to-end latency may be in the range from 500 ms to multiple seconds.

Synthesis views or a view based on view control can be delivered to the client. The trade off between prediction error and bandwidth needs to be considered while selecting the number of views.

### 7.x.5 Source format properties

For this scenario, the multi-view video source format has 10 to 20 views. It is expected that most or all test data will have perspective projection (PSP), but test data with equirectangular projection (ERP) may be included.

Each view has the following components:

* Texture (color)
* Depth coded as normalized disparity

Depth information can be used in rendering e.g. by shaders for surface normal estimation. If depth is not provided, decoder-side depth estimation can be possible.

Editor’s note: Further details on depth processing is FFS.

All views have view parameters: camera ID, camera intrinsics, camera extrinsics (pose) and depth quantization parameters (optional).

Views may be undistorted, otherwise distortion parameters have to be provided.

The signal properties defined in clause 4.3.4.1 apply with no further constraints.

### 7.x.6 Encoding and decoding constraints and settings

Some constraints and settings below are given for MIV.

Codec profiles/levels:

* HEVC Main 10 MIV Main
  + MIV level 2.0 or 2.5.
* HEVC Main 10 MIV Extended
  + MIV level 2.0, 2.5 or 3.0 whereby the level 3.0 is only allowed if there is a single video sub-bitstream.

Typical random-access frequency of 32 frames can be considered. It is up to the service provider to define the exact random access frequency.

Transmission systems need to be prepared to resend data in case of data loss. If data loss still occurs or retransmitted data does not reach the receiver device in time for rendering, previous volumetric frames may be re-rendered with updated viewing poses. In case one or more of the sub-bitstreams is lost, it is up to the application to determine an optimal method for hiding the missing information.

Typically, bitrates between 5 and 50 Mbit/s may be considered.

Bitrate parameters related to video sub-bitstreams need to be configured by the streaming service provider. Transfer characteristics are signalled in the video sub-bitstreams.

No special requirements regarding ABR. Configuration is left for the service provider to determine.

Latencies between 500ms to several seconds are considered. Random access interval or segment duration are configured according to the latency requirements.

Encoding is performed by a content provider. This scenario assumes professional setting for recording and processing the content, so no real-time or encoder hardware or architecture requirements are provided.

It is expected that devices support HW accelerated video decoding.

Decoding requirements:

* HEVC Main 10
* HEVC levels are determined according to the maximum HEVC Level that is needed for a video sub-bitstream decoder to fulfill the MIV level.
  + HEVC level 5.1 for MIV level 2.0
  + HEVC level 5.2 for MIV level 2.5
  + HEVC level 6.1 for MIV level 3.0
* Video sub-bitstreams need to be independently decodable. This helps implementations on various platforms that may have only high-level APIs. For instance, geometry needs to be full range.

Samples in the sub-bitstreams should be temporally aligned.

### 7.x.7 Performance Metrics and Requirements

The tests are run for a chosen level as described in clause 7.x.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.)

[Ed.(BK): To be aligned with the agreed evaluation framework.]

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

* Weighted sphere PSNR (WS-PSNR)
* Immersive video PSNR (IV-PSNR)

All source views that were used for encoding are provided. Each source view is reconstructed by decoding and rendering (view synthesis). The IV-PSNR 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. BD-PSNR is calculated for both metrics because the metric behaves more predictably than BD-rate.

[Ed.(BK): To be aligned with the agreed evaluation framework. The discussion here on correlation of objective and subjective metrics may need to be moved to that framework after more deliberation.]

There is experience in testing of immersive video in MPEG context. The test conditions as described are a simplification and evolution of the common test conditions for MIV defined in [R6].

The main challenge with testing of B2D 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.

### 7.x.8 Interoperability Considerations for the application

The B2D video bitstream needs to be carried over DASH for this scenario. It is not necessary to prove this as part of the feasibility test, if written evidence can be provided.

In the example of using MIV as a codec, there are implementations for DASH (InterDigital/Philips) and RTP + SDP (Nokia and an open source implementation by Tampere University: <https://github.com/ultravideo/uvgRTP>).

### 7.x.9 Test Sequences

Test sequences that were used during the development of a codec are discouraged because they may create a bias towards that specific codec. Sequences that were used in a verification test are permissible.

Preferably test sequences match with the intended use case both in terms of technical requirements and content semantics.

For MIV a list of available sequences is provided in [R6].

### 7.x.10 Detailed test conditions

[Ed.(BK): To be aligned with the agreed evaluation framework.]

For each candidate codec, a suitable decoder + renderer needs to be made available for testing purposes.

For MIV a reporting template or script will be provided to compute BD-PSNR based on IV-PNSR log files of all rates and sequences.

For MIV the common test conditions defined in [R6] are followed.

### 7.x.11 External Performance data

For MIV the performance data is available from the verification test report [R7].

NOTE: This performance data was based on different source view properties and the results may not translate to this study.

### 7.x.12 Additional Information

The Metaverse Standards Forum (MSF) has established a Volumetric Media Interoperability working group which aims to build a better understanding of volumetric media, to identify relevant areas of applications and compatibility requirements, and to establish common requirements for different systems. See here the WG description: <https://metaverse-standards.org/domain-groups/volumetric-media-interoperability/>

The technology is expected to be highly scalable since it uses well-established transport technologies like DASH and 2D video coding techniques.

Regarding complexity, rendering and decoding frame rates for MIV content were measured for Windows and Android platforms in [R5]. The results show that the developed platform can decode V3C content in real time on both Windows and Android. Evaluation of battery consumption (power levels) is FFS.

Streaming of B2D content has the potential to disrupt several markets including entertainment/media, education/training, retail/shopping.

Several use cases can be envisioned related to these domains. For example, in an education/training scenario, a pre-recorded video of a fitness instructor showing how to perform an exercise can help the student to better understand how the exercise is done and thus replicate in a correct way. Another example in education domain would be a mechanic giving a tutorial on how to assemble a mountain bike. The viewer can watch the movements of the mechanic from different angles and get an improved understanding of the different steps due to depth perception and different viewpoints. In the entertainment domain, users can stream a performance from their favorite band to their living room and experience greater immersion potentially together with spatial audio.

**== CHANGE 2 ===**

# 2 References

The following documents contain provisions which, through reference in this text, constitute provisions of the present document.

- References are either specific (identified by date of publication, edition number, version number, etc.) or non‑specific.

- For a specific reference, subsequent revisions do not apply.

- For a non-specific reference, the latest version applies. In the case of a reference to a 3GPP document (including a GSM document), a non-specific reference implicitly refers to the latest version of that document *in the same Release as the present document*.

…

[R1] 3GPP TR 26.928: “Extended Reality (XR) in 5G”.

[R2] 3GPP TR 26.998: “Support of 5G glass-type Augmented Reality / Mixed Reality (AR/MR) devices”.

[R4] ISO/IEC 23090-10:2022 (Amd1), “Information Technology — Coded Representation of Immersive media — Part 10: Carriage of Visual Volumetric Video-Based Coding Data”

[R5] 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>

[R6] Dziembowski, B. Kroon, J. Jung (Eds.), Common test conditions for MPEG immersive video, ISO/IEC JTC 1/SC 29/WG 04 N 0372, July 2023, Geneva.

[R7] D. Mieloch (Ed.), Verification test report of MPEG immersive video, ISO/IEC JTC 1/SC 29/WG 04 N 0341, April 2023, Antalya.