**Source: Nokia Corporation1, Interdigital, Philips**

**Title: [FS\_Beyond2D] Scenario on volumetric video XR real-time communication**

**Agenda Item: 9.9**

**Document for: Discussion and Agreement**

# Introduction

A new study item FS\_Beyond2D ([SP-240479](https://www.3gpp.org/ftp/TSG_SA/TSG_SA/TSGS_103_Maastricht_2024-03/Docs/SP-240479.zip)) was approved at SA#103. One of the objectives of the study is:

2. Establish and document a set of beyond 2D video end-to-end reference scenarios, including real-time communication, streaming services, split rendering, and messaging and corresponding workflows (capturing, encoding, packaging, delivery, decoding, rendering, including general constraints on latency, as well as complexity) to support 3GPP network related delivery and devices leveraging the generation or display technologies. This includes identifying and defining relevant beyond 2D formats in the context of above workflows, and representation technologies to support delivery of these formats within 3GPP networks.

In this contribution, a draft scenario on volumetric video XR real-time communication is proposed for incorporation into FS\_Beyond2D TR 26.956 as basis for future work. The scenario is structured according to the template provided in [S4aV240003](https://www.3gpp.org/ftp/TSG_SA/WG4_CODEC/3GPP_SA4_AHOC_MTGs/SA4_VIDEO/Inbox/Drafts/S4aV240003-onlineEdits.docx) including the online edits agreed at the VIDEO SWG telco on 26 March.

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

## 6.x Scenario #x: Volumetric video XR real-time communication

1. **Scenario name**

Volumetric video XR real-time communication

1. **Motivation for the scenario**

*What is the market relevance of the proposed scenario within the next few years? Are there any commercially available or pre-released products or prototypes?‘*

Hybrid working habits have become commonplace since the global pandemic. While traditional 2D-video communication systems usually get the job done, transferring the meetings into an immersive environment requires video conferencing systems to evolve to support 3D communication. Volumetric video is a technology, which enables capture of an object or scene in three dimensions and its playback independent from the original capture position(s) or orientation(s).Volumetric video offers a complete rendering of all parties involved, capturing their non-verbal body language and mannerisms. This enhances the immersiveness of the communication significantly, providing a deeper and more meaningful level of interaction**.**

**Online education** is one of the sectors which can benefit greatly from improved immersion and visual depth perspective. Perhaps a guitar teacher wants to observe their pupil’s hand positioning and finger placement on the fretboard. Another example could be a yoga instructor, who needs to see the other person from any viewpoint to evaluate the yoga poses accurately. This comprehensive viewpoint could help the teacher to remotely identify and correct any nuances that may be hindering the student’s development. This is a clear progression from the limitations that come with current 2D teleconferencing tools or solutions that rely on animating participants.

In a **professional environment**, volumetric video allows a group of people to enter a shared 3D space in the metaverse, interacting with three-dimensional CAD models and other people in real time. They could perform a dynamic examination of the model from every conceivable angle, ensuring a comprehensive understanding of spatial relationships and bridging the gap between virtual planning and tangible construction. Furthermore, enabling more immersive communication for corporate meetings may be beneficial for situations like hiring interviews. The interviewers can detect subtle changes in the interviewees reactions that may help formulate more accurate hiring decisions.

With the increasing capabilities and affordability of new multisensory devices and algorithms that allow capturing and experiencing volumetric content with high fidelity, several prototypes or commercial products have emerged within the industry. Some examples are provided below:

* Nokia – <https://www.nokia.com/blog/3d-live-communication-becomes-part-of-everyday-life-with-volumetric-video/>
  + Real-time volumetric video streaming demo at the **Mobile World Congress (MWC)** 2024.
* Ericsson - <https://www.ericsson.com/en/ericsson-one/holographic-communication>
  + Details here: <https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/holographic-communication-in-5g-networks>
* Imverse - <https://imverse.com/>
* 8i - <https://8i.com/stream/>
* **Telefónica - Holographic Telepresence System with 3D Capture** demonstrated at MWC 2023
  + The collaboration involves the partners **Evercoast**, **Intel**, and **AWS**.
  + <https://www.telefonica.com/en/communication-room/press-room/telefonica-showcases-its-holographic-telepresence-with-3d-capture-at-mwc/>

NOTE 1: The examples are meant to provide motivation and demonstrate the market relevance of the scenario and not to give detailed information on the capture setup, formats or other aspects of the workflow. For the workflow description see the clause 3.

NOTE 2: The present scenario targets XR applications and volumetric video use cases that can be realized with data rates and latencies achievable over 5G networks. Holographic communication is seen as a future evolution already receiving some interest from the industry, as demonstrated in some of the examples above. However, it is out of the scope of the present scenario.

1. **Description of the scenario**

*This provides a description of beyond 2D video end-to-end workflows, which includes identifying and defining beyond 2D formats being used in the context and representation technologies to delivery these formats. The following aspects may be considered for each workflow:*

1. *Capturing and processing*

This scenario considers volumetric video that is captured and processed using one or more cameras. Zero or more of those cameras may be range-sensing cameras, and one or more of these cameras have colour sensors. In the case of two or more cameras that are not rigidly connected, camera extrinsics are online calibrated.

Captured content is real-time converted to a representation format that is suitable for encoding. The parameters of the source cameras may or may not be part of the representation format. At a minimum inter-view consistent depth information is estimated, but more processes like reprojection, pruning, refinement, meshing and texturing steps may be needed depending on the representation format and application-specific constraints.

1. *Encoding*

The representation format(s) are real-time encoded by using conventional 2D video codec(s) and metadata stream(s).

1. *Packaging and delivery*

The multiple streams may be multiplexed or provided as separate tracks.

1. *Decoding*

The decoder(s) will make use of hardware video decoder capabilities for all pixel data, and a small amount of metadata is decoded by a CPU. No dedicated hardware is needed for real-time decoding.

1. *\*Post-processing*
2. *Rendering*

Rendering is typically performed on a GPU without dedicated hardware.

When a viewing space is used, then:

* What is rendered is one or two viewports with perspective projection and with 6 degrees of freedom (3-D position and 3-D orientation).
* The pose of the viewport is within a viewing space that can be signalled or implicitly determined from a decoded frame. A viewing space can limit both position, orientation or both in combination. For instance, it is generally not intended for a viewport to intersect with scene elements.
* When a viewport would be rendered that is outside of the viewing space, then the renderer has to perform a mitigation to avoid a viewing experience that is not intended by the content provider.

1. *General constraints on latency, bandwidth, reliability and complexity*

The volumetric frames are organized using a low-delay reference frame structure.

All encoder, decoder and renderer processes are real-time and may have a latency in the order of at most a few frames.

1. **Supporting companies and 3GPP members**
2. *This documents the 3GPP members that support this scenario in terms of providing the information, test material, test requirements and the characterization for the tests. For each of the identified necessities, a tick box is created in the template.*
3. *Preferably several 3GPP members are included in the support, and in addition a video service provider may be included (not necessarily a 3GPP member).*
4. *Cross-verification is preferably done by the supporters of the scenario*

Nokia, Interdigital, Philips

1. **Source format properties**

*This defines a clear range of the considered and relevant source formats, including the signal properties, but also the characteristics of the content. As an example, the texture and depth format properties of the source may be used which include:*

1. *Spatial resolutions*

**TBD**

1. *Chroma Format*

**TBD**

1. *Chroma Subsampling*

**TBD**

1. *Aspect ratios*

**TBD**

1. *Frame rates*

**TBD**

1. *Colour space formats*

YCbCr 4:2:0

1. *Transfer Characteristics*

Video carrying texture information will have a BT.709 transfer function. All other video will have a linear transfer.

1. *Bit depth*

Colour will be 8 or 10 bit. Effective geometry bit depth may be higher or implicitly represented (not as video samples).

1. *Viewpoints*

The viewpoints are within a viewing space that can be signalled or implicitly determined from a decoded frame.

1. *Other signal properties*

**TBD**

1. **Encoding and decoding constraints and settings**

*Typical encoding constraints and settings such as*

1. *Relevant Codec and Codec Profile/Levels according to TS26.119*
2. *Random access frequency*
3. *Error resiliency requirements*
4. *Bitrates and quality requirements*
5. *Bitrate parameters (CBR, VBR, CAE, HRD parameters)*
6. *ABR encoding requirements (switching frequency, etc.)*
7. *Latency requirements and specific encoding settings*
8. *Encoding context: real-time encoding, on device encoding, cloud-based encoding, offline encoding, etc.*
9. *Required decoding capabilities*
10. *Synchronization requirements*

**TBD**

1. **Performance Metrics and Requirements**
2. *A clear definition on how the performance needs to be evaluated including metrics, etc addressing the main KPIs of the scenario.*
3. *Objective measures such as PSNR, VMAF, etc, may be used.*
4. *Justification on whether objective metrics are sufficient and representative of the subjective performance.*

**TBD**

1. **Interoperability Considerations for the application**
2. *Streaming with DASH/HLS/CMAF/QUIC*
3. *RTP based delivery*

**TBD**

1. **Test Sequences**

*A set of selected test sequences that are provided by the proponents in order to do the evaluation. They should cover a set of source format properties*

**TBD**

1. **Detailed test conditions**

*Provides a proposal for detailed test conditions, for example based on a reference software together with the sequences and configuration parameters.*

**TBD**

1. **External Performance data**

*References to external performance data that can be added, for example other SDOs, public documents and so on.*

**TBD**

1. **Additional Information**