 **ISO/IEC JTC 1/SC 29/WG 2** **N0157**

**ISO/IEC JTC 1/SC 29/WG 2  
MPEG Technical requirements   
Convenorship: SFS (Finland)**

**Document type:** Output Document

**Title:** Use cases and requirements for MIV – edition-2 (final)

**Status:** Approved

**Date of document:** 2022-01-21

**Source:** ISO/IEC JTC 1/SC 29/WG 2

**Expected action:** None

**Action due date:** None

**No. of pages:** 7 (with cover page)

**Email of Convenor:** igor.curcio@nokia.com

**Committee URL:** <https://sd.iso.org/documents/ui/#!/browse/iso/iso-iec-jtc-1/iso-iec-jtc-1-sc-29/iso-iec-jtc-1-sc-29-wg-2>

**INTERNATIONAL ORGANIZATION FOR STANDARDIZATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC 1/SC 29/WG 2**

**MPEG TECHNICAL REQUIREMENTS**

**ISO/IEC JTC 1/SC 29/WG 2** **N0157**

**Online – January 2022**

|  |  |
| --- | --- |
| **Title** | **Use cases and requirements for MIV – edition-2 (final).** |
| **Source** | **WG 2, MPEG Technical requirements** |
| **Status** | **Approved** |
| **Serial Number** | **21128** |

# Abstract

This document provides the finalized requirements of MPEG Immersive Video (MIV) – edition 2.

# Introduction

The MPEG Immersive Video (MIV) [1] is part of the MPEG-I ISO/IEC 23090 family of standards. This standard is designed to efficiently compress representation from real or synthetically generated dynamic, volumetric scenes ingested as 2D video sources. Unlike a 3DoF scene where a viewer can navigate in the scene by head rotation, MIV edition-1 also provides the ability for limited translation (about the scale of head size movement) of viewing position in 3D space.

# Definition of terms

|  |  |
| --- | --- |
| **source** | a term used to describe the video material or some of its *attributes* before encoding. |
| **6DoF** | A scene in which a user is able to move around freely with 6 degrees of freedom. |
| **viewpoint** | triple of x, y, z characterizing the position in the Cartesian coordinatesof a user who is consuming the immersive visual content |
| **viewing space** | domain constraints for an intended *viewport* rendering; the domain is defined in the 3D global space and related to the *viewing orientation*; it defines a scale between 0 and 1 for every point in space for a given direction of the viewport, to be used by the application |
| **viewer** | The person or entity whose position and orientation are used for rendering. |

# List of use-cases

1. **6DoF omni-directional** **immersive content: Immersive content where a viewer can walk around using untethered XR head-mounted devices.**

This use case reflects the full 6DoF immersive content experience enabled through real-life captured and synthesized content. Full 6DoF immersive content allows the viewer to navigate his/her location within the content space, with only limitations imposed by a content creator with natural viewpoint change of the corresponding rendered video. The captured visual data could involve technologies such as sparse and dense light fields. The amount of data to be encoded would be so large, that efficient scene understanding, segmentation and smart coding would become a necessity. The result of using these new technologies should provide a high quality, photo-realistic immersive experience.

1. **Coding of camera captured natural scenes with non-Lambertian characteristics.**

When natural scenes with non-Lambertian surfaces are captured, e.g. specular surfaces, transparent objects, etc, the appearance of the scene varies depending on the viewpoint within the viewing volume, from which the scene is consumed. The coding of the scene with non-Lambertian surfaces will require first determining those regions of the scene that express such characteristics and coding of additional meta-data information that will help a renderer in the client device to represent the scene in a photorealistic manner, regardless of the rendering technology used.

1. **Bullet time**

A viewer watches a video stream. Their device has relatively limited decoding and rendering capabilities such as those of a mid-level mobile device. Preferably, the video feed is an immersive video (6DoF) video, but it may also be 2D video.

As an intermediate solution between 2D and 6DoF video, the video stream may also be a multi-view (multi-track) video stream consisting of separate (rectified) views but lacking geometry information. In that case the viewer can switch between (rectified) source views in a logical order following increasing or decreasing angle, thus achieving a look-around effect, albeit not smoothly.

At specific moments in the stream, the viewer has the option to request the (generation and) download of a bullet time frame. The bullet time frame will generally have a larger viewing space and a higher visual quality, and it enables the viewer to appreciate a relevant moment in more detail.

In a variant on this use case, the viewer may pause the video stream at any frame, after which the viewer is presented with the option to request a bullet time frame of that frame, or the nearest available frame, from the server.

The viewer has the option to progress in time for both the normal video and for bullet frames. For the bullet time frames a next frame or previous frame is downloaded on request of the viewer.

Having viewed the bullet time feed, the viewer chooses to return to the regular feed:

* In the case of a live event, it is preferable that the audio continues while the bullet time frame is viewed, because the viewer does not want to miss out on any live action.
* For pre-recorded events, the video may be paused while the bullet time frame is consumed, and video will restart from that frame.
* Switching back to the regular feed the global characteristics of the viewport is preserved as much as possible. For instance, in case the user had translated the virtual camera into the scene, a zoomed-in virtual camera likely best approximates the bullet time viewport.

1. **Object-based MIV content composition and compression**

This use case reflects a metaverse scenario of casting the full 6DoF immersive content experience enabled through real-life captured and synthesized objects from the sources that may or may not be homogeneous (e.g., different locations, different representations (RGB-D, mesh, point cloud, etc.), different projections, FOVs, etc). Just like game broadcasting, the hyper-real metaverse casting will require high-quality 6DoF content experience that may require substantial compression efficiency, include real-time encoding and decoding capability with high throughput, and provide a scalable encoding functionality depending on the viewer’s location, orientation, gesture, preferences, and so on. The 6DoF objects in this use case may have different viewing characteristics (e.g., lifetime or duration, frame rates, projection, camera parameters, etc.).

1. **Coding of natural scenes captured with moving cameras and/or adjustable focus**

It is common for ordinary 2D video capture to utilize moving cameras, pan/tilt/zoom, and auto-focus, which correspond to camera extrinsic and/or intrinsic parameters that change over time. In this use case, immersive video scenes with texture and depth are captured with dynamic extrinsic and/or intrinsic parameters.

1. **Geometry representation using colour**

Commercially available depth sensors can capture depth from natural scenes. Some available depth sensors provide depth capture data in a format that represents depth using color, using RGB 4:4:4 or YUV 4:2:0. In the case of YUV 4:2:0, the U and V component values are used for the depth representation.

1. **Unaligned texture and geometry**

When capturing natural content with video cameras and time-of-flight depth sensors, a video camera and depth sensor cannot occupy the exact same physical location. Immersive video source content captured by temporally synchronized video cameras and depth sensors may have different camera parameters for texture and geometry.

# Requirements

## **General**

1. The edition-2 of the specification shall be an extension/modification of the MIV edition 1.
2. The edition-2 of the specification shall include technology that efficiently uses existing tools of standardized video codecs.
3. The edition-2 of the specification shall be compatible with relevant MPEG-I system standards.

## **Binocular disparity and Parallax**

1. The edition-2 of the specification shall enable client implementations to support head-motion parallax and binocular disparity in rendering of images and video, where the user perceives natural viewing conditions consistent with head motion.
   1. The edition-2 of the specification shall enable client implementations to support head-motion parallax in rendering regardless of the viewing orientation.
      1. The edition-2 of the specification shall support viewing orientations for head-motion parallax to azimuth and elevation ranges between 0 degrees and 360 degrees and 0 degrees and 180 degrees, respectively.
   2. The edition-2 of the specification shall enable defining a viewing-space that is larger than than head-scale 6DoF motion.

Note: For example, such Viewing Space may enable immersive experience for the following viewing situations:

* A user sitting in a natural way, on a wheeled chair which can freely move.
* A user standing in a natural way and taking a couple of steps in any orientation in 3D space.
  + 1. The edition-2 of the specification shall support changing of the viewing-space over time.
  1. The support for head-motion parallax in rendering shall be enabled regardless of the distance between the viewer and the objects that are in the viewport provided that no objects are within the viewing-space.
  2. The edition-2 of the specification shall enable content authoring in a manner that head-motion parallax and binocular disparity are supported in rendering without annoying artefacts while the user is in the viewing-space.

Note: Testing methodology for visual quality of experience of rendering with head-motion parallax and binocular disparity should be developed and/or chosen during the standardization.

* + 1. The edition-2 of the specification shall enable the support of head-motion parallax and binocular disparity in rendering in a manner that any potential degradations in the quality of experience are not abrupt but graceful.
    2. The edition-2 of the specification shall enable a client implementation to detect or be informed of the viewing-space and/or viewing orientations where rendering is potentially unacceptable.

## **Scene characteristics and understanding**

1. The edition-2 of the specification shall enable coding a dynamic volumetric scene that contains non-Lambertian surfaces.
2. The edition-2 of the specification shall include technology that identifies static and dynamic parts of the scene and efficiently organizes, code and transmits the combination.
3. The edition-2 of the specification shall enable handling of heterogenous object-specific parameters (e.g. temporal sampling, duration, atlas sizes, and non-Lambertian characteristics) at the MIV bitstream level.
4. The edition-2 of the specification shall enable coding a dynamic volumetric scene captured by cameras with time-varying extrinsic and/or intrinsic camera parameters.
5. The edition-2 of the specification shall enable coding a dynamic volumetric scene in which geometry is represented in a colour format using 3 colour components.
6. The edition-2 of the specification shall enable coding a dynamic volumetric scene represented by views for which each individual view with associated camera parameters may contain geometry but not texture, or texture but not geometry, or both texture and geometry.

## **Facilities for application**

1. The edition-2 of the specification shall support transparency of visual media.
2. The edition-2 of the specification shall support minimizing the end-to-end delay for applications that need this.
   1. The edition-2 of the specification shall enable an encoding to presentation delay that is similar to traditional linear TV.
   2. The edition-2 of the specification shall support making a trade-off between efficiency and end-to-end delay.

Note: To a significant extent, this end-to-end delay is determined by factors that are beyond the present specification

1. The edition-2 of the specification shall support keeping motion-to-photon delay within the immersive limit

Note: motion-to-photon latency is probably fully determined by the rendering in the device.

1. The edition-2 of the specification shall not preclude temporally synchronizing (live) XR streams with other content, potentially distributed over an alternative distribution path, including broadcast streams.
2. The edition-2 of the specification shall support single-frame (still image) 6DoF bitstreams.
3. The edition-2 of the specification shall enable switching between (multiple) 6DoF and 2D streams with preservation of presentation order information.
4. The edition-2 of the specification shall enable switching between (multiple) 6DoF and 2D streams with preservation of viewport information.
5. The edition-2 of the specification shall enable composition of multiple MIV bitstreams into a single MIV bitstream.
6. The edition-2 of the specification shall maintain/enhance support for independent object-level encoding and decoding in the bitstream.
7. The edition-2 of the specification shall improve the quality of output viewport renders with a significant increase in MOS scores.
8. The edition-2 of the specification shall target coded bitrates in the range of 5 to 50 Mbps and 1 giga luma samples per second.

Note: This bitrate range may be relaxed only when necessary – for e.g. in complex scenes with a large viewing space.

1. The edition-2 of the specification shall enable heterogeneous 6DoF contents (at least multi-view video and point cloud) in the same video sub-bitstream of a V3C bitstream.

# Estimated timeline

|  |  |
| --- | --- |
| meeting | Events |
| MPEG 137 | Release of final requirements , and Working Draft - 1 |
| MPEG 138 | Working Draft - 2 |
| MPEG 139 | Working Draft - 3 |
| MPEG 140 | Committee Draft |
| MPEG 141 | Draft International Standard |
| MPEG 143 | Final Draft International Standard |

# References

1. J. Boyce, B. Chupeau, L. Kondrad, Text of ISO/IEC FDIS 23090-12 MPEG Immersive Video, ISO/IEC JTC1/SC29/WG04 N00111, July 2021, Online