The MPEG Immersive Video Coding Standard

Vinod Kumar Malamal Vadakital

Ofinno • ofinno.com • April 2022

A Brief Overview of the MPEG Immersive Video Coding Standard

Introduction

The MPEG immersive video standard (MIV) [1] is one among a suite of MPEG-I standards which is focused on the coded representation of immersive media. MIV started as an exploration activity after the standardization of the Omnidirectional Media Format (OMAF) version 1 [3]. The mandate was to investigate coded immersive video representations that allow for more than a simple three degrees of freedom (3DoF) visual interaction, allowing viewers some amount of head translation motion; in other words, allowing for six degrees of freedom (6DoF) visual interaction. This meant handling disocclusions and inter-view redundancies. All through the process of exploration of new technologies, as well as the actual standardization process, one thing that was clear and agreed upon was the use of traditional two-dimensional (2D) video codecs for encoding the video data. After reviewing contributions and evaluating evidence of technology, a Call-for-Proposal (CfP) was issued in January 2019. Subjective and objective quality evaluation of responses was done in March 2019, and a common test model that combined concepts from multiple responses was established. The standard reached the Final Draft International Standard (FDIS) stage of the standardization process in July 2021.

3DoF Versus 6DoF

Degrees of Freedom (DoF), in the context of immersive video coding, refers to the number of dimensions available to a viewer to interact with a three-dimensional (3D) scene. In 3DoF, a viewer can interact with the scene using rotational motion about the three cardinal axes of a 3D scene as illustrated in Figure 1. These rotations are called yaw, pitch, and roll in literature. The origin of rotation itself is fixed to some constant position in 3D space. 6DoF, in addition to the rotational motion supported by 3DoF, can also support translational motion along the three cardinal axes. These additional translation motions, illustrated in Figure 2, are sometimes called surging, strafing, and elevating. The reference axis for each of these six motions can vary based on the conventions used by applications. By providing additional dimensionality to visually interact with the scene, 6DoF provides viewers with an improved perception of immersion. A 3DoF video also lacks the proper representation of motion parallax and disparity cues resulting in sickness when viewing for some users. However, due to the sheer amount of data to be processed, it is also more challenging to code than a 3DoF video.

Inputs to an MIV Codec

The input to an MIV encoder is a set of videos captured from multiple cameras, where each camera can have arbitrary pose. The set of videos captured from one camera is called a view. Furthermore, each view associated with a camera can have multiple components. The components are broadly classified

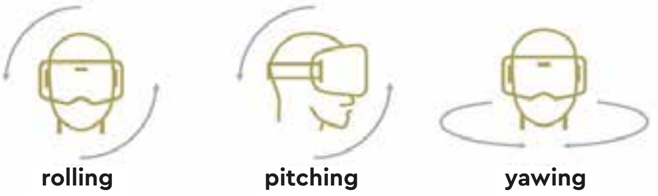

FIG. 1: Visual interactions supported by 3DoF videos

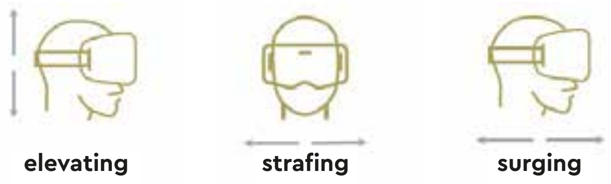

FIG. 2: Additional visual interactions supported by 6DoF videos

as geometry and attributes in the MIV. Geometry defines, for each pixel in a view, the position in 3D space where a ray passing through the pixel intersects an object in 3D space. The spatial position of the pixel in one frame of a view, along with a camera transformation matrix provide two of the three coordinates in 3D space; the third coordinate is provided by a depth map. Furthermore, for each projected point, attributes such as texture, normal, and transparency can also be provided as component videos. Figure 3 illustrates the relation between a view and its component videos. MIV, in edition 1, supports equirectangular, orthographic, and perspective projections.

Brief Overview of the Codec Design

The inputs to an MIV encoder are videos of some n>1 views, where each view comprises at least a texture attribute component. Additionally, a geometry component (represented as depth-maps), and other attributes such as transparency, normal and material information that is useful for reconstructing a viewport at the client, can also be provided as inputs. The encoder analyses and processes the input views to generate some k≥1 atlas video streams per component. An atlas is a composition of patches where each patch is generated by projecting parts of the 3D scene onto some know projection planes. The generation of patches are optimized to minimize inter-view redundancy and the number of pixels to be decoded, while maximizing the quality of the content presented to viewer for a given bitrate.

Each atlas conceptually comprises two parts – (a) the atlas video data, and (b) the atlas metadata. The atlas metadata annotate aspects of the atlas video data which is required to decode a coded atlas video data and render a viewport of the 3D scene. The atlas video data is like any 2D video data, except that they are a composition of patches projected from the 3D scene. They are encoded as a video bitstream with any 2-D video encoder, while the metadata is encoded using the MIV standard. Figure 4 illustrates a typical MIV encoder.

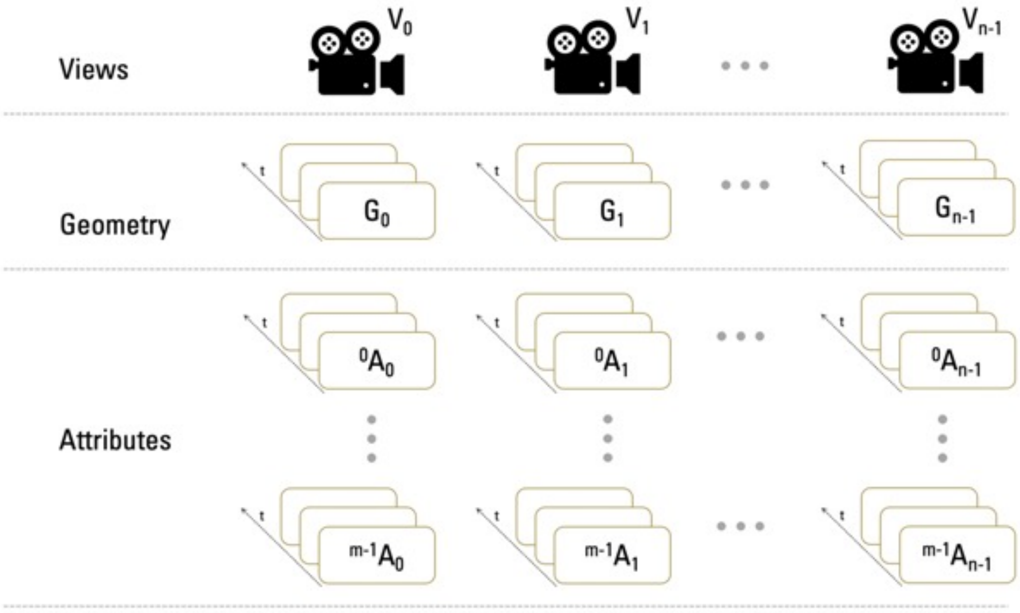

FIG. 3: An MIV scene is a capture of a 3D scene captured from multiple viewpoints shown in this illustration as the set of views {V_0 .. V_{n-1}}. Each view, in turn, comprises videos that can either be a representation of geometry, $G_{[0 .. n-1]}$, or attributes $^{[0 .. m-1]}A_{[0 .. n-1]}$. Geometry is represented as depth maps, and texture is an example of an attribute.

Processing of Views to Generate Atlas Video Data

Broadly speaking, the primary goal of an MIV encoder is to reduce interview redundancies – that is, to identify points that are captured by multiple views and encode them only once. A way to achieve this is to first identify one or more views, called basic views in MIV, that are likely to have the least number of redundant points between them. The other views are called non-basic views. The basic views are used as anchors for identifying redundant points from non-basic views. A similarity matching process for each pixel in a non-basic view is performed by unprojecting (into 3D space) and reprojecting them back onto the basic views. This process results in a binary mask, called the pruning mask, for each non-basic view. The pruning mask informs an encoder of regions in a non-basic view that are safe to prune, because it is available in one of the basic views.

With the pruning mask available, a process of patch generation is performed where regions that cannot be pruned are cropped, and then formatted into coherent regions called patches. All basic views are considered as patches in this context. The patches are collected and composed into one or more sets of videos called atlases. The number of atlases that the patches are composed into is decided based on the application/use-case design.

Most current generation video codecs use predictive coding mechanisms, both spatially as well as temporally. To assist a video encoder to perform its task optimally, it is important that the temporal consistency of patches across time is maintained. The reference implementation of MIV handles this by performing the pruning and packing process over a group of consecutive frames.

MIV Bitstream Format

MIV inherits the V3C bitstream format specified in ISO/IEC 23090–5 [2]. A V3C bitstream is a sequence of coded V3C sequences (CVS). Each CVS is in-turn a sequence of coded V3C sub-sequences (CVSS). A high-level illustration of the composition of syntactic structures in an MIV bitstream is shown in Figure 5.

A V3C bitstream can be in one of two formats: the V3C unit stream (V3C-US) format or the V3C sample stream (V3C-SS) format. The V3C-US format is simplistic in nature, lacking information that makes it unsuitable as a stand-alone format for applications. A V3C-US format comprises a sequence of syntactic elements called V3C units (V3C-U). The V3C units in a V3C-US is in decoding order with constraints imposed by the ISO/IEC 23090–5 specification. A V3C-SS is constructed from a V3C-US by encapsulating the V3C-U in the V3C-US prepended

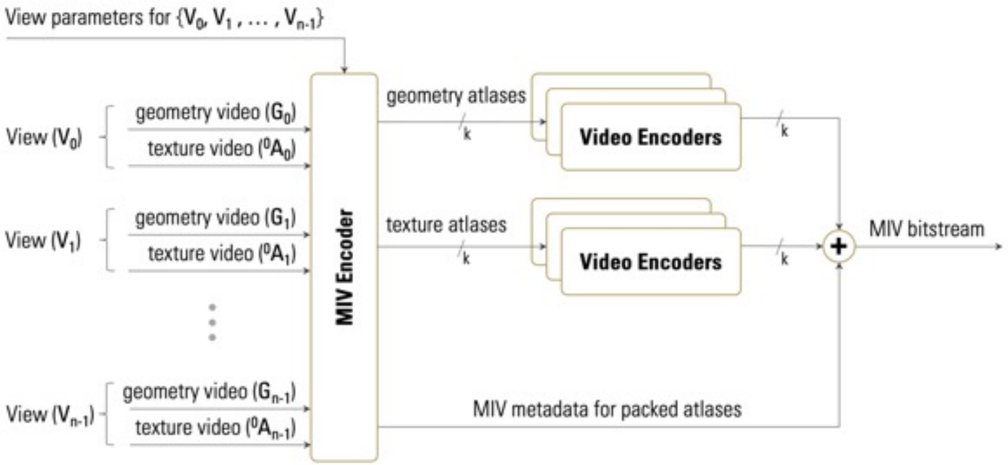

FIG. 4: A typical MIV encoder

with the size of the V3C-U. A V3C-SS always starts a sample stream header that signals the precision in bytes required to signal the size of a V3C-U in the bitstream. A V3C-US can be extracted from the V3C-SS by extracting the size information and the subsequent V3C-U.

Like traditional video coding standards produced by MPEG, information that does not vary often in a coded bitstream are collated into syntactic structures called parameter sets. ISO/IEC 23090–5 specifies the following parameter sets:

- V3C parameter set (V3C-VPS)
- Atlas sequence parameter set (ASPS)
- Atlas frame parameter set (AFPS)
- Atlas adaptation parameter set (AAPS)
- Common atlas sequence parameter set (CASPS)

A CVS of an V3C bitstream either starts with a V3C-VPS, or a decoder is provided access to a V3C-VPS through external means. In both cases, an MIV decoder requires access to a relevant V3C-VPS prior to decoding any other information of a CVS. A V3C-VPS can be used by one or more CVS of a V3C bitstream. A V3C-VPS provides information such as the number of atlases in the CVS, the presence and the nature of CVSSs in the CVS, and the profile-tier-level that the bitstream conforms. The ASPS, AFPS, and the AAPS all pertain to information relevant to a particular atlas present in the CVS. MIV edition-1 does not use the AAPS. CASPS, however, is used by MIV to signal information that are common to all atlases in the CVS – this includes the MIV view parameter and the depth quantization parameters.

Metadata relevant to patches that are packed into an atlas video frame is provided in the patch data unit (PDU). The PDU is in turn a payload of the atlas tile layer (ATL) NAL unit. The metadata for a patch includes all relevant information that is required to unproject it back into 3D space.

Information that is not required for the normative decoding process, but deemed to be useful for applications, are provided in NAL units called the

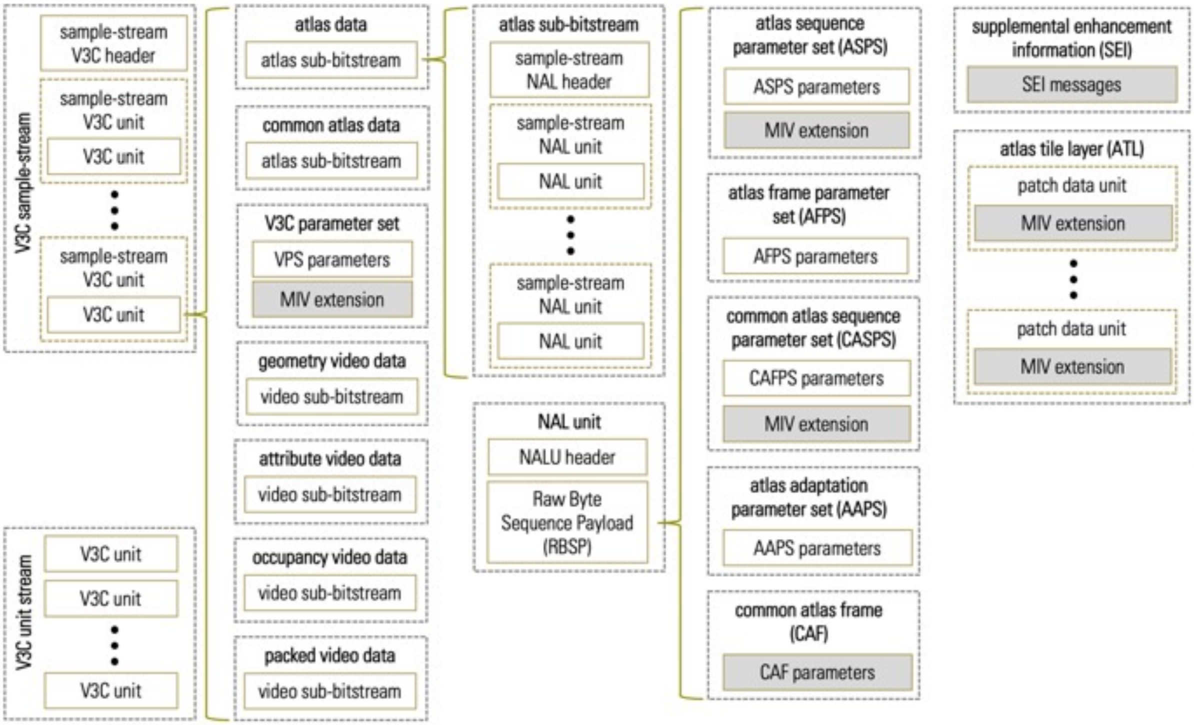

FIG. 5: MIV bitstream structure

supplemental enhancement information (SEI). ISO/IEC 23090–5 (V3C) provides several such SEI messages and ISO/IEC 23090–12 added six more of them that are relevant for MIV use-cases. A good example of an SEI message that is relevant in the MIV context is the viewing space SEI message which provides information about the 3D space from within which the coded 3D scene can be reconstructed without any significant disocclusion artefacts. Further information about the SEI messages supported by V3C and MIV can be found in [2] and [1] respectively.

Future Improvements for MIV

A working draft for the second edition of MIV was started in the 137th meeting of MPEG that took place in January of 2022. Apart from a constant effort to improve coding efficiency, the edition-2 of the MIV specification also plans to support some new use cases. These include:

- support for scenes with surfaces that exhibit non-Lambertian light transport characteristics;
- support for unaligned geometry and attribute videos obtained from capture sensors that do not share the same extrinsic and intrinsic camera parameters;
- include support for handling colourized depth, where multiple colour channels are used to represent depth; and
- extending the supported inputs for MIV to also handle point-cloud video.

Conclusions

The first edition of the MPEG immersive video (MIV) standard can be considered as a step towards enabling efficient immersive 6DoF coding. It is designed to be flexible and extensible, accommodating new use-cases easily. By decoupling the encoding of atlas video data from the metadata and focusing on specifying the metadata required to reconstruct a viewport from the coded atlases, it has positioned itself as a specification with a generic appeal.

References

[1] ISO/IEC DIS 23090–12, Information technology — Coded Representation of Immersive Media — Part 12: MPEG immersive video

[2] ISO/IEC DIS 23090–5, Information technology — Coded Representation of Immersive Media — Part 5: Visual volumetric video-based coding (V3C) and video-based point cloud compression (V-PCC)

[3] ISO/IEC DIS 23090–2, Information technology — Coded Representation of Immersive Media — Part 2: Omnidirectional media format

About the Author:

Vinod Kumar Malamal Vadakital received the B.E. degree in computer science and engineering from Bangalore University, Bengaluru, India, in 1998, and the M.S. degree in information technology and the Ph.D. degree in signal processing from the Tampere University of Technology, Tampere, Finland, in 2005 and 2012, respectively. He is currently a Principal Scientist with the Video Coding Research Team of Ofinno LLC, Tampere. His research interests lie in the area of video signal processing, computer vision, and XR technologies. Dr. Malamal Vadakital has previously been an Editor of the High Efficiency Image File Format (HEIF) version 1 specification. He is currently a Vice-chair of the MPEG Immersive Video ad-hoc group within ISO/IEC JTC 1/SC 29/WG 4, MPEG Video Coding.

About Ofinno:

Ofinno, LLC, is a research and development lab based in Northern Virginia, that specializes in inventing and patenting future technologies. Ofinno's researchers create technologies that address some of the most important issues faced by wireless device users and the carriers that serve them. Ofinno's inventions have an impressive utilization rate. Ofinno's research involves technologies such as 5G Radio and Core networks, IoT, V2X, and ultra-reliable low latency communications. Our innovators not only create the technologies, they oversee the entire process from the design to the time the technology is sold. For more information about Ofinno, please visit www.ofinno.com.