ITUPublications

International Telecommunication Union

Standardization Sector

ITU Focus Group Technical Specification

(06/2024)

ITU Focus Group on metaverse
(FG-MV)

FGMV-39

Use cases and requirements for virtual and real fusion coding in metaverse applications

Working Group 2: Applications & Services

PREPUBLISHED

Version

Technical Specification ITU FGMV-39

Use cases and requirements for virtual and real fusion coding in metaverse applications

Summary

Metaverse is an emerging research and application field with the combination of multiple technologies including digital twin, Internet of Things (IoT), digital assets, multimodal data fusion and artificial intelligence generated content (AIGC). Users need immersive experience such as playback of camera-captured 3D scenes with 6DoF of viewer position and orientation. The current video coding standard is optimized by 2D videos and the coding efficiency may not be enough. Therefore, the metaverse applications need an efficient virtual and real coding technology to support low-delay and immersive experience for users. The virtual and real coding technology can support affordable coded bit rate and high coding efficiency for immersive videos, omnidirectional videos, as well as the source content with high quality depth information. The interaction between digital human and users, online meetings, gaming, sports viewing can be the use cases benefiting from this coding technology.

This Technical Specification provides the related requirements, reference model of application system and use cases of the virtual and real fusion coding in metaverse applications.

Keywords

Virtual and real fusion; video coding; metaverse

Note

This is an informative ITU-T publication. Mandatory provisions such as those found in ITU-T Recommendations are outside the scope of this publication. This publication should only be referenced bibliographically in ITU-T Recommendations.

Change log

This document contains Version 1.0 of the ITU Technical Specification on "Use cases and requirements for virtual and real fusion coding in metaverse applications" approved at the 7th meeting of the ITU Focus Group on metaverse (FG-MV) held on 12-13 June 2024.

Acknowledgments

This Technical Specification was researched and written by Zekun Wang (China Telecom) as a contribution to the ITU Focus Group on metaverse (ITU FG-MV). The development of this document was coordinated by Yuntao Wang and Yuan Zhang, as FG-MV Working Group 2 Co-Chairs, and by Julien Maisonneuve as Chair of the Task Group on industrial metaverse, and by Zekun Wang (China Telecom) and Marcelo Moreno (Fraunhofer IIS, Germany), as Co-Chairs of the Task Group on media coding.

Additional information and materials relating to this report can be found at: https://www.itu.int/go/fgmv. If you would like to provide any additional information, please contact Cristina Bueti at tsbfgmv@itu.int.

Editor & TG Co-Chair:	Zekun Wang China Telecom China	Tel: +86-18964957670 E-mail: wangzk2@chinatelecom.cn
WG2 Co-Chair:	Yuntao Wang China Academy of Information and Communications Technology	Tel: +86 18611547086 E-mail: wangyuntao@caict.ac.cn

	(CAICT) China	
WG2 Co-Chair:	Yuan Zhang China Telecom China	E-mail: zhangy666@chinatelecom.cn
TG Co-Chair:	Marcelo Moreno Fraunhofer IIS Germany	E-mail: marcelo.moreno@iis-extern.fraunhofer.de

Table of contents

Page

Technical Specification ITU FGMV-39

Use cases and requirements for virtual and real fusion coding in metaverse applications

Introduction

This Technical Specification covers the use cases and requirements for virtual and real fusion coding in metaverse applications. The data format of virtual and real fusion coding includes immersive videos and omnidirectional videos. Virtual and real fusion coding will benefit the metaverse applications and different use cases. This technical specification will identify the common requirements and framework of application system for virtual and real fusion coding-related metaverse applications.

1 Scope

The scope of this technical specification includes:

- Requirements for virtual and real fusion video coding

- Reference model of application system for virtual and real fusion video coding

- Use cases of the virtual and real fusion video coding in metaverse application

2 References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of these Technical Specifications. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of these Technical Specifications are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within these Technical Specifications does not give it, as a stand-alone document, the status of a Recommendation.

[ITU-T H.264] ITU-T Recommendation H.264 (2003), *Advanced video coding for generic audio visual services.*

[ITU-T H.265] ITU-T Recommendation H.265(2013), *High efficiency video coding.*

[ITU-T H.266] ITU-T Recommendation H.266(2020), *Versatile video coding.*

3 Definitions

3.1 Terms defined elsewhere

These Technical Specifications use the following terms defined elsewhere:

3.1.1 Reconstructed feature [b-MPEG-VCM]

The feature tensor obtained by decoding the bitstream by the decoder.

3.1.2 Bin [b-MPEG-VCM]

The symbols that makes up a binary string including 0 and 1.

3.1.3 Bitstream [b-MPEG-VCM]

The binary data stream formed by encoding the original data.

3.1.4 Feature [b-MPEG-VCM]

A tensor data that includes channel, width and height in three dimensions.

3.1.5 Tensor [b-tensor]

In mathematics, a tensor is an algebraic object that describes a multilinear relationship between sets of algebraic objects related to a vector space. Tensors may map between different objects such as vectors, scalars, and even other tensors.

3.2 Terms defined in these Technical Specifications

None.

4 Abbreviations and acronyms

These Technical Specifications use the following abbreviations and acronyms:

3DoF	3 degrees of freedom tracking
6DoF	6 degrees of freedom tracking
PC	Personal Computer
VR	Virtual Reality
V3C	Visual volumetric video-based coding
V-PCC	Video-based point cloud compression
V-DMC	Video-based dynamic mesh coding

5 Conventions

In this Recommendation:

The keywords “is required to” indicate a requirement that must be followed strictly and from which no deviation is permitted if conformance to this document is to be claimed.

The keywords “is recommended” indicate a requirement that is recommended but is not absolutely required. Thus, this requirement needs not be present to claim conformance.

These terms are not intended to imply that the vendor's implementation must provide the option and that the feature can be optionally enabled by the network operator/service provider. Rather, it means that the vendor may optionally provide the feature and still claim conformance with the specification.

6 Introduction of the virtual and real fusion video codec

Virtual and real fusion video, also known as immersive video, is an important trend for the user’s experience. Broadly speaking, immersive video creates an immersive feeling through audio and video technology. When users put on a VR helmet, they can select viewing angles in any direction and position through interactive ways such as body sensation, line of sight, gestures, touch and buttons. The user will have a very strong feeling of being immersed in the scene. Two important technologies used are 3 degrees of freedom tracking (3DoF) and 6 degrees of freedom tracking (6DoF) to achieve the immersive experience. 3DoF and 6DoF immersive experiences provide users with interactive and changing visual and auditory dimensions, including perspective, lighting, focal length and field of view, through computational reconstruction, making users thousands of miles away feel as if they are there. Resolution and time delay are factors affecting the users’ feeling. First, the resolution is higher and higher. For example, the monocular resolution of VR helmet trend starts from 1.5K*1.5K to 4K*4K, the data volume is huge when transmitting. The second issue is time delay, the consistency of perspective switch caused by time delay also affects users’ experience. To solve these problems, an effective codec for immersive videos is necessary. The existing codec related to metaverse, use cases, requirements, reference model of application system are introduced accordingly.

7 Codec standard related to metaverse

In metaverse applications, the input data formats can be immersive video and point cloud generated in cloud server or local server. These types of data have large amount of data size compared to images or 2D videos. The process of transmitting to the users' mobile phone, personal computer (PC) or other display devices requires a high bandwidth of network. Therefore, metaverse applications are required to have efficient codec standard for data transmission in metaverse applications. For example, when wearing a Virtual Reality (VR) head display, playback of camera-captured 3-D scenes with 6 degrees of freedom tracking (6DoF) of viewer position and orientation to achieve high fidelity immersive experience has become one of the research directions of immersive video coding standard.

H.264, H.265, H.266 are used widely in the current video coding applications to save the bit rate. JPEG standards focus on the image coding aspect. Besides coding for the images and videos, MEPG also develop some standard for different format of data such as immersive video, immersive audio, point cloud and dynamic meshes. Some ITU and MEPG coding standards that have been published or are being developed can be used as references.

• H.264 Advanced video coding.

• H.265 High efficiency video coding.

• H.266 Versatile video coding.

• T.801 Information technology – JPEG 2000 image coding system – Extensions.

• T.807 Information technology – JPEG 2000 image coding system: Secure JPEG 2000.

• T.808 Information technology – JPEG 2000 image coding system: Interactivity tools, APIs and protocols,

• T.816 Information technology - JPEG 2000 image coding system: Extensions for coding of discontinuous media, T.JPEG-AI Information technology – JPEG AI learning-based image coding system.

• Audio: ISO/IEC 23090-4: MPEG-I immersive audio.

• V3C and Point Clouds: ISO/IEC 23090-5: Visual volumetric video-based coding (V3C) and video-based point cloud compression (V-PCC).

• Point Clouds: ISO/IEC 23090-9: Geometry-based point cloud compression.

• Carriage of V3C data: ISO/IEC 23090-10: Carriage of visual volumetric video-based coding data

• Immersive Video: ISO/IEC 23090-12: MPEG Immersive video.

• Dynamic Meshes: ISO/IEC 23090-29: Video-based dynamic mesh coding (V-DMC).

8 Use cases

Use case 1: Codec for digital human

From the perspective of telecommunication operators, a typical metaverse use case scenario is digital human for customer service. Different companies may have different technology to generate digital human. Two possible ways to generate virtual digital human are human-driven and computer-driven methods. For example, digital human can be generated by the local server or cloud server through 3D engine [b-3D engine] virtual space modelling. The 3D digital model is rendered to the background of real world. The facial expression, body expression and voice expression of the digital human for customer service can be generated by using the voice synthesis model, action-driven model and training data. The digital human is transmitted to the user's mobile phone or PC through efficient codec for offline interaction. The codec needs to satisfy different type of digital

human. Users can ask questions to digital customer service offline or online to get answers to common problems such as business handling.

Figure 1: One type of digital human for customer service.

Use case 2: Codec for sports viewing

Metaverse can enable users to watch the sports competition with immersive experiences. It includes multiview angle and multiple degree of freedom from the viewer's position with a mobile phone or virtual reality (VR) equipment. Users will have a near on-site immersive watching experience in the competition process and enjoy the charm of the competitive sports. Immersive video coding can provide efficient data compression, data reconstruction, data rendering with input data including multiview angle videos, depth map and camera parameters to ensure immersive experiences.

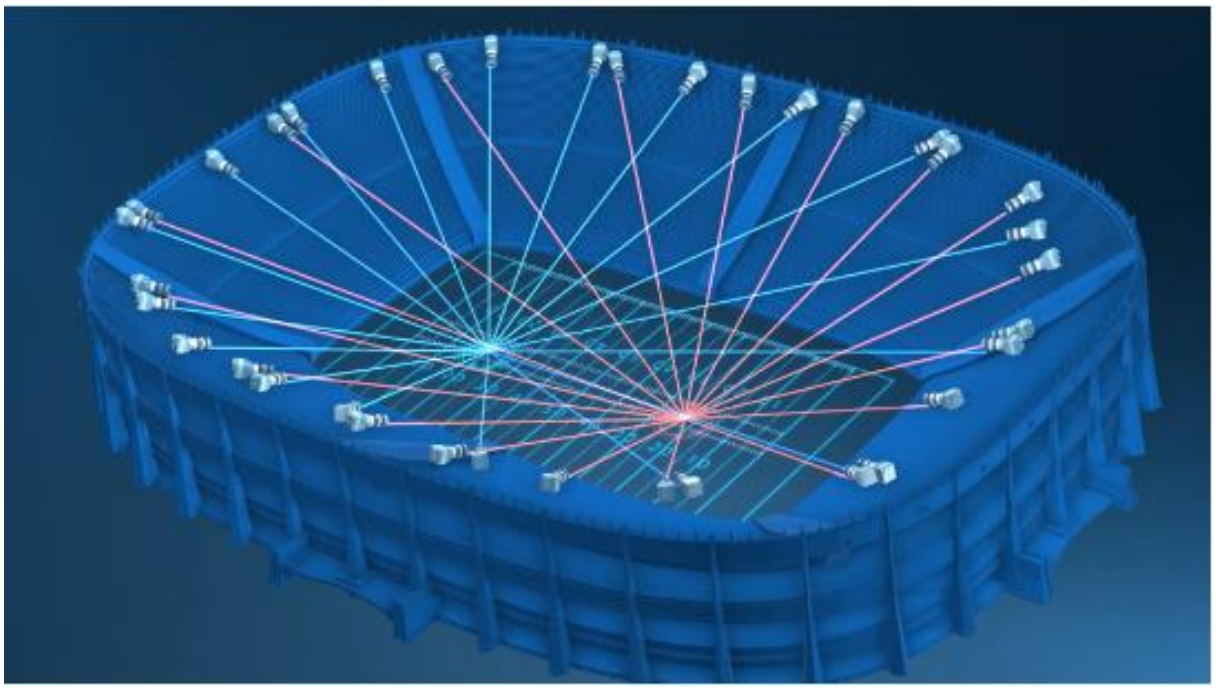

Figure 1: True view platform volumetric capture at a stadium.

Use case 3: Codec for metaverse exhibition hall

The metaverse exhibition hall provides a learning and entertainment function for teenagers and children to obtain a better understanding of the metaverse technology and applications. There are two typical forms for metaverse exhibition hall. The first form is watching immersive videos or playing games by wearing VR equipment. The second form is interacting and learning with digital human. Immersive video coding can empower such scenarios.

Figure 2: Metaverse exhibition hall.

9 Pre-processing for virtual and real fusion coding

The immersive videos are captured by camera array. Before coding, the front end needs to implement the pre-processing methods. Video pre-processing stage includes camera array calibration, image correction, depth information estimation, online registration of multiple perspective images, and packaging of multiple perspective image.

9.1 Camera array calibration

To determine the relationship between the 3D geometric position of a point on the surface of a space object and its corresponding point in the image, the camera imaging geometric modelling should be established. Under these conditions, these parameters are obtained through experiments and calculations; the process of obtaining these parameters is called camera array calibration.

9.2 Image correction

When using camera arrays to obtain 3D information of objects, the individual spatial position deviation of each camera can easily cause spatial geometric distortion of the element image, which has a significant impact on the element image array.

9.3 Depth information estimation

The basic function of depth information estimation is to input a RGB image or video steam and then return a depth value for each pixel obtained from the image, which can be synthesized into an actual image or video stream. The output image can well reflect the distance between the object and camera. The closer to the camera, the closer the colour is to warm tones. The output image is similar to an infrared thermal image. It can play a role in various fields such as autonomous driving and 3D reconstruction.

9.4 Online registration of multiple perspective images

Online registration is the process of matching multiple images captured in the same scene at different points in time from different perspectives or from different acquisition devices. Assuming there are two images captured by camera with different locations, the registration process is to find a spatial transformation that transform the moving image onto the fixed image, so that the points corresponding to the same spatial position in the two images.

9.5 Packaging of multiple perspective images

After the registration process, the multiple perspective images need to be packaged before encoding. The package process includes packet multiview image data into an encoder-recognizable encoding format and type.

10 The requirements for virtual and real fusion codec

Multiview sports viewing, multiview online gaming and multiview digital human are three metaverse use cases. To enable such kind of use cases, multiple pictures or frames are captured by fixed camera arrays. When the distances changes measured between the cameras and the objects, the multiview frames or pictures sequences are produced. There is a lot of redundancy for encoder to form a compressed bitstream.

The inputs of the immersive videos or virtual and real fusion videos in such use cases can be represented by multiview angle video and depth information [b- MPEG-I]. The objects in multiview video are same with only small change of the background information. So, the representation of the multiview angle videos is important. Some techniques have been developed to generate video frames, including the preparation of source material and peer-group encoding to form geometry video data and attribute video data [b- MPEG-I]. These data can comply with the current video coding standard.

• Virtual and real fusion codec is required to support multiview angle videos captured by the camera array with fixed location.

• Virtual and real fusion codec is required to support current video coding standard, e.g., H.265-output format.

• The input format for virtual and real fusion codec are required to support immersive videos and omnidirectional videos. For some metaverse application with multiple cameras, it is recommended that the input format for virtual and real fusion codec supports multiple video and depth representation of the video data. It is recommended that the separate source view supports the representation of geometry and attribute samples.

• The compact representation of the reorganized frames of immersive videos or virtual and real fusion video is required for codec development.

• Virtual and real fusion codec are required to support flexible codec profiles for different metaverse applications.

• Virtual and real fusion codec are required to be evaluated in pre-defined datasets.

• Virtual and real fusion codec are required to be evaluated with different quantization parameters (QPs).

• Virtual and real fusion codec are recommended to support multiview video pre-processing, fast encoding and rate control.

11 Reference model of application system

The Figure 3 illustrates the high-level workflow of the immersive video coding application system. It includes feature extraction, encoder, decoder, video reconstruction and renderer blocks. The input data is multiview angle video and depth information representation of video data with each source view represented by frames of geometry and attribute samples [b-MPEG-I]. The input data may contain the camera parameters such as location and direction. The input data are transferred to feature extraction block to obtain the feature tensor. The feature tensor contains the compact visual information for encoder side to perform compression. The encoder extracts the information for scene reconstruction and removes the redundancy and packs it in a compact type into the view set in the form of patches. The encoder can multiplex the video encoder to encode the view set and re-use the subbitstream with the metadata to form a bitstream. The decoder and video reconstruction block are reverse process of encoder and feature extraction. The renderer process includes the parameter optimisation of each view of the camera. A viewport is rendered by a de-projection from each related source followed by a re-projection according to the viewport coordinates input at each frame by the application. The renderer interactively synthesizes the corresponding perspective based on the viewer's movement.

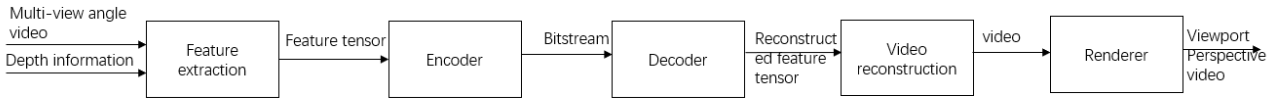

Figure 3: High level workflow of application system

11.1 Interface

The functions of the main blocks are as follows:

Feature extraction block: inputs are multiview angle video and depth information, output is feature tensor.

Encoder block: input is feature tensor, output is bitstream.

Decoder block: input is bitstream, output is reconstructed feature.

Video reconstruction block: input is reconstructed feature, output is reconstructed video.

Renderer block: input is reconstructed video, output is video from viewer's perspective.

Bibliography

[b-Immersive Video]	Boyce J M , Dore R , Dziembowski A , et al. MPEG Immersive Video Coding Standard[J]. Proceedings of the IEEE, 2021, PP(99):1-16.
[b-Metaverse]	ISO/IEC JTC 1/SC 29/WG 2 Use cases and MPEG technologies for Metaverse-related experiences 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.
[b-3D engine]	What Is a 3D Engine? https://www.easytechjunkie.com/what-is-a-3d-engine.htm
[b-MPEG VCM]	Evaluation Framework for Video Coding for Machines.
[b-tensor]	https://en.wikipedia.org/wiki/Tensor.