Special Paper
방송공학회논문지 제28권 제7호, 2023년 12월 (JBE Vol. 28, No. 7, December 2023)
https://doi.org/10.5909/JBE.2023.28.7.904
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

An Overview of MPEG-I Scene Description, a Standard for MPEG Immersive Media Integration

Youngkwon Lim[a)‡]

Abstract

MPEG has developed a new standard on scene description for immersive media, ISO/IEC 23090-14 MPEG-I Scene Description. The standard has defined technologies for two interfaces, media access interface and presentation engine interface by adding several nodes to the glTF chosen as a base technology. The standard has strong advantages in integrating MPEG developed immersive media assets and dynamic streaming of immersive media contents composed of scene description and media assets considering presentation time of them. However, the standard is behind it being supported by major authoring tools and game engines compared to other competing standards in the market such as USD and ITMF.

Keywords : Immersive media, scene description, MPEG

I. Introduction

MPEG is widely known as a standard development organization for audio visual media data compression. However, MPEG has also long history of developing scene description standards. ISO/IEC 14496-11 BiFS (Binary Format for Scene) was the first scene description standard MPEG has developed[1]. As MPEG-4 standard has considered media assets as an interactive objects there has been a need for a technology compositing the objects as a scene. The standard describes how the various objects are positioned in space and time, and also defines dynamic behavior and user interaction to create a scene using various MPEG-4 media object assets. The scene description is heavily based on the VRML-97 specification and MPEG-4 defined a set of 2-D scene description nodes to enable the implementation of low-cost 2-D only systems as the media assets are still 2-D audiovisual media when it was developed. MPEG-4 has also added binarization of the scene description and scene update commands enabling dynamic update of scene without completely redrawing a scene. MPEG-4 BiFS was the first scene description standards allowing insertion, update, or removal of nodes and fields in a time synchronous manner.

a) Samsung Research America
‡ Corresponding Author : Youngkwon Lim
E-mail: yklwhite@gmail.com
Tel: +1-214-906-3479
ORCID: https://orcid.org/0000-0003-1472-7430
· Manuscript October 10, 2023; Revised November 11, 2023; Accepted November 11, 2023.

The second scene description standard MPEG has developed is ISO/IEC 14496-20 Lightweight Application Scene Representation (LASeR) and Simple Aggregation Format (SAF)[2]. Considering that mobile implementation of BiFS is not sufficiently efficient, MPEG has developed additional scene description more suitable for resource limited devices. Scalable Vector Graphics Tiny 1.1 (SVGT 1.1) has been chosen as a base standard. As SVGT 1.1 does not have any audio-visual media support and it does not have any means to update the scene dynamically. LASeR filled such gap and ISO/IEC 14496-20 has also provided a packaging format so that the scene description and media assets can be stored efficiently in a same file and streamed according to presentation time of dynamic scene updates and media assets.

As MPEG has started a project targeting immersive media applications, MPEG Immersive (MPEG-I), new set of requirements for scene description technology for immersive media has been identified. Scene description technology for MPEG-I should efficiently support 6 degree-of-freedom media and application but two former scene description did not considered such use cases. This paper will introduce ISO/IEC 23090-14 MPEG-I Scene Description standard[3] by analyzing its requirements, base scene description technology, MPEG extensions to it in section II. In the section III, comparison with other widely used scene description standards will be provided to better understand the advantages and disadvantages of MPEG-I scene description.

II. Overview of MPEG-I scene description standard

1. Requirements of the standard

MPEG has defined list of requirements for the MPEG-I scene description standard[4] and reproduced in provided in Table 1. As MPEG defines standards for the interface between functional modules, requirements for two interface are defined in this case, presentation engine interface and media access interface. The former is the interface processing scene description data for rendering of immersive media and the latter is the interface supporting integration of MPEG media to immersive scene description. As streaming of immersive scene description is one of the most important and unique use cases supported by MPEG-I scene description, the requirements for presentation engine covers random-access, dynamic updates, and partial delivery and processing. Regarding integration of MPEG media into MPEG-I scene description, supporting of wide variety of MPEG media for synchronized presentation of immersive media is an one of the most important and unique use case. Therefore, the requirements for the media access interface covers flexible support of wide range of features of MPEG media and various types of delivery methods for synchronized presentation of them.

Additional set of requirements has been also defined for selection of base scene description technology as several scene description technologies have been already defined and widely used by contents industry. The base scene description technology must support core features of scene description such as integration of basic audio-visual media and their synchronization including a way to integrate the media through external referencing. In addition, the base scene description must supports various customization through various nodes such as camera, texture, geometry, etc. and private extension.

2. Base scene description technology

Based on the requirements, glTF has been chosen as the base scene description technology for MPEG-I scene description standard. glTF has been developed by Khronos Group as a technology to integrate media assets using various technologies standardized by them such as OpenGL and so on. The technology has been also recently published as

Table 1. List of MPEG-I scene description requirements

Category	List of requirements
Presentation engine Interface	• It shall be possible to update the whole scene-graph, a sub-graph, or a node in the scene description • It shall be possible to correctly render a 6DoF Presentation after a random access in time • It shall be possible to perform timed scene description updates • It shall be possible to associate a scene description update with the corresponding scene description • It shall be possible to use a scene description as the entry point to a 6DoF presentation.
Media access interface	• It shall be possible to access timed and non-timed, 2D and 3D media (meshes, point clouds, audio elements, …), stored locally or over the network • It shall be possible to pre-fetch media that the presentation engine expects to be used in the presentation • It shall be possible to retrieve media depending on the desired level of detail • It shall be possible to retrieve and access referenced media partially in time and space • It shall be possible to describe position, orientation, and visual/acoustic characteristics when rendering referenced media • It shall be possible to synchronize media objects/resources and media components of a single object • Audio elements shall be rendered consistently with their corresponding visual elements, if such visual elements exist. • The specification shall enable synchronization of audio and video of users and the scene.
Base scene description selection	• The scene description shall support audio and video formats as well as other media formats standardised by MPEG (including OMAF). • The scene description shall enable the support of other visual or audio media formats. • The scene description shall support definitions to indicate how sub-graphs and objects are related in terms of their temporal, spatial and logical relationships • The scene description shall support composition of digital representations of natural and synthetic objects. • The scene description shall support synchronisation between objects and attributes in the scene. • The scene description shall support spatial and temporal random access. • The scene description should support information to enable a renderer to perform path tracing. • The scene description shall support sub-graph representation that allows modular rendering e.g. leafs in the scene description tree can also be packaged and referenced individually from a parent scene description and container. • The scene description shall support references (e.g. URLs) to external media resources in place of embedded file references • The scene description shall support a mechanism to safely customize behavior for nodes like camera, texture, geometry, audio, and object placement nodes through sandboxed, validated domain specific shaders or scripts for these nodes without affecting the functionality or forcing changes to the root node graph or other node types; i.e. provide a mechanism to safely extend the scene description.

ISO/IEC international standard, ISO/IEC 12113[5]. The standard can be implemented for free as IP rights are not supported to be asserted against other participants implementing it[6].

3. Architecture

As shown in Figure 1, conceptually a client processing MPEG-I scene description is composed of two major functional module, presentation engine and media access function. The former process scene description document and render immersive media contents according to it. When there are media required for rendering immersive media content, media access function retrieve such media data according to scene description. Such retrieval is controlled by presentation engine through media access function API. When media data retrieved by media access function appropriate buffer is created to deliver media data to presentation engine. Media data delivered through buffer is formatted to be directly rendered by presentation engine and synchronously delivered regardless of potential jitter from retrieval and processing by media access function.

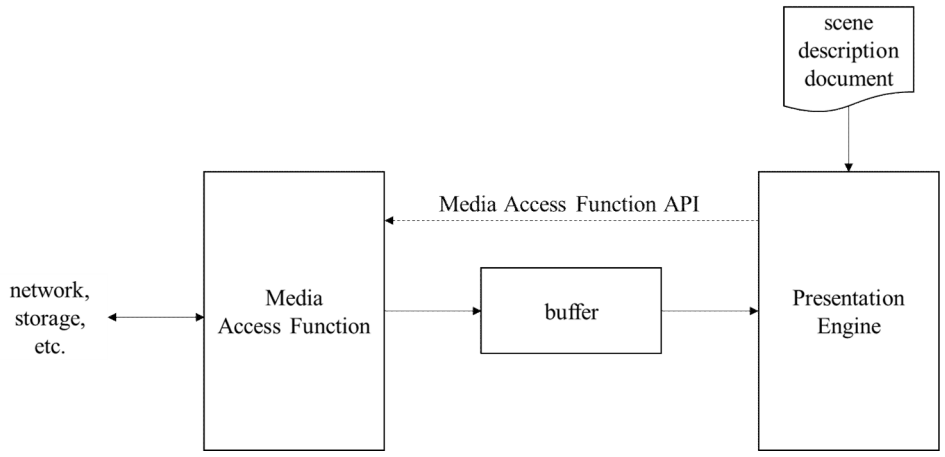

Fig. 1. Conceptual architecture of a client processing MPEG-I scene description

4. MPEG extensions to glTF

To meet the requirements described in previous section, glTF has been extended by adding MPEG extension nodes as listed in Table 2. At the top level, MPEG_media node can be used to directly integrate any MPEG media resources into scene without using any other scene objects. At the glTF scene node level, MPEG_scene_dynamic and MPEG_animation_timing are added to enable delivery of scene information through update mechanism instead of sending entire scene as a single document. In addition, delivery of any dynamic data such as audiovisual media data, scene updates or animation updates is enabled through MPEG_accessor_timed and MPEG_buffer_circular which makes synchronized continuous delivery of data through same interface instead of single delivery. MPEG_audio_spatial is introduced to add support of audio which is not naturally supported by glTF.

III. Comparison with other scene description standards Scene description standards widely used by contents industry

1. Scene description standards widely used by contents industry

Contents industry have developed various scene description technologies to create animations contents and graphics contents. One of the most widely used scene description technology is Universal Scene Description (USD)[7]. It has

Table 2. List of MPEG extension nodes

glTF node	MPEG extension node	Description
N/A	MPEG_media	Extension for referencing external media sources.
scene	MPEG_scene_dynamic	An extension to support dynamic scene updates
scene	MPEG_viewport_recommended	An extension to describe a recommended viewport.
scene	MPEG_animation_timing	An extension to control animation timelines.
node	MPEG_audio_spatial	Adds support for spatial audio.
camera	MPEG_audio_spatial	Adds support for spatial audio.
mesh	MPEG_mesh_linking	An extension to link two meshes and provide mapping information
accessor	MPEG_accessor_timed	An accessor extension to support timed media.
Texture	MPEG_texture_video	A texture extension to support video textures.
Buffer	MPEG_buffer_circular	A buffer extension to support circular buffers.

been developed by Pixar Animation Studios for integrating media assets created by various tools and exchanging intermediate data among them. Pixar Animation Studio made USD as an open source technology with W3C patent policy [7] in 2016 for free use by the industry. Recently, Alliance for OpenUSD (AOUSD) has been formed to further develop the technology as an open industry standard[8].

Another scene description technology widely used by contents industry is Immersive Technology Media Format (ITMF)[10]. It has been standardized by The Immersive Digital Experiences Alliance (IDEA) as an open standard with W3C patent policy in 2019. The core technology of it has been originally developed by the contents creation company OTOY Inc. Similar to the case of USD, it has been developed for use of creation and exchange of the media asset for their internal contents creation pipeline.

2. Comparison of key features

Using a game engine for real-time rendering of immersive contents are being actively studied and there are two game engines most widely used by the industry, Unity and Unreal[11]. USD, ITMF and glTF are well supported by the both game engines but they do not support rendering of MPEG-I scene description yet. In addition to real-time immersive content rendering, creating contents for augmented reality and rendering them on mobile devices are also important use cases of scene description. USD has been adopted by ARKit for such augmented reality use cases and glTF has been adopted by ARCore and Windows Mixed Reality Home similarly.

There are several important capabilities of scene description technologies regarding creation and exchange of immersive contents as listed in Table 3. As explained in the above sections, scene description technologies are mostly developed for integration and exchange of media assets created by other tools, integrating media assets created with other content creation tools is considered as an important capability. To describe realistic immersive content scene description technology must support representation of various type of texture of the objects considering source of the lights in the scene and reflection of each of them using mathematical equations or procedure. Representing snapshots of the animated scene at specific time as key frames and interpolate the scene between them during rendering in real-time is widely used technology for animation and it is considered as a key feature of scene description. USD, ITMF and glTF are not quite different in supporting such capabilities and well supported by major 3D contents authoring tools. As MPEG-I scene description is developed by extending glTF, it also supports major scene description capabilities through glTF, too.

Supports of volumetric media assets is one of the features scene description technologies differs hugely. MPEG-I scene description supports integration of volumetric media assets represented by Visual volumetric video-based coding (V3C) standard developed by MPEG, ISO/IEC 23090-5[12]. USD supports volumetric media assets represented by OpenVDB[13] and Draco[14]. Similarly ITMF supports volumetric media assets represented by OpenVDB only. glTF does not support any volumetric media assets.

To support immersive media service in a way similar to 2D video services over Internet or broadcast network in these days, packaging scene description document together with media assets in a file format and streaming scene description documents and media assets considering their rendering time are important features. USD does not have a general packaging format defined. However, USDZ has been developed as a packaging format of USD, an uncompressed zip file format, as part of ARKit. USD cannot be generally streamed as it does not have packaging format. As zip file is not streamable in general, streaming of USDZ is not also supported. A container format has been defined as part of ITMF specification by IDEA for packaging ITMF scene description with media assets. IDEA has

Table 3. Comparison of scene description standards

	USD	ITMF	glTF	MPEG-I SD
Developer	• Originally developed by Pixar • To be standardized by AOUSD	• Originally developed by OTOY • Standardized by IDEA	• Originally developed by Khronos • ISO/IEC 12113	• Extensions to glTF developed by MPEG (23090-14)
IPR policy	• W3C	• W3C	• not to assert IP rights against other participants implementing spec.	• unknown
Major game engine support	• Unity • Unreal	• Unity • Unreal	• Unity • Unreal	• None
App Platform	• ARKit	• None	• ARCore • Windows Mixed Reality Home	• None
Major authoring tool support	• 3D Studio MAX/ Maya/Blender	• 3D Studio MAX/ Maya/Blender	• 3D Studio MAX/ Maya/Blender	• None
Scene Description Capability	• Various geometry and Alembic • Image texture • Wide range of materials • Time sample based animation • Volumetric object support through OpenVDB and Draco	• Various geometry and Alembic • Procedural texture and images • PBR and wide range of materials • Time sample based animation • Volumetric object support through OpenVDB	• Polygon mesh only • Image texture • PBR materials • Time sample based animation • Volumetric object support through Draco	• Limited to glTF • Integration with V3C defined
Packaging Format	• No packaging format defined • USDZ for ARKit	• ITMF container specification	• A simple binarization format (glb)	• Extension to ISOBMFF is defined
Streaming Supports	• Not supported	• IDEA has demonstrated streaming of ITMF based contents to game engine and real-time rendering	• Not supported	• Dynamic update of scene description is supported • Streaming of ISOBMFF is supported

actively studied streaming of ITMF based contents. During the demonstration by several members of IDEA the possibilities of taking an ITMF file, extracting the scene description and large number of media assets, then streaming to a client implemented with Unreal and Unity from a network-based server for simple watching and interactive viewing. glTF has defined a simple binarization format for packaging but its streaming has not been explored. Carriage of MPEG-I scene description in ISO base media file format has been standardized as a part of the standard itself. As streaming of contents stored in a file with ISO base media file format considering the rendering time of the samples, streaming of MPEG-I scene description can be naturally supported.

IV. Conclusion

Scene description technology plays a key role in 6 DoF immersive media service as such contents needs to be com-

posed of multiple volumetric media assets such as point clouds or mesh. Technically it is important to store a scene description together with media assets considering dynamic streaming of them considering presentation time to enable scalable media service similar to OTT service in these days. Such features were not developed much so far as scene description technology was developed mostly by contents industries for their internal contents creation and exchange needs. MPEG-I Scene Description has successfully implemented such features by extending glTF. However, it is not well adopted by the industry as it is still new to the market and its IPR policy is not aligned with the general industry practices. Special attention needs to be paid to such features in addition to further integrating new MPEG media better supporting immersive contents such as immersive audio currently under development.

References

[1] A. Eleftheriadis, "MPEG-4 systems: architecting object-based audio-visual content," 1998 IEEE Second Workshop on Multimedia Signal Processing (Cat. No.98EX175), Redondo Beach, CA, USA, 1998, pp. 535-540.
doi: https://doi.org/10.1109/MMSP.1998.739036.

[2] J. -c. Dufourd, "LASeR: The lightweight rich media representation standard [Standards in a Nutshell]," in IEEE Signal Processing Magazine, vol. 25, no. 6, pp. 164-168, November 2008.
doi: https://doi.org/10.1109/MSP.2008.929813.

[3] ISO/IEC, Information technology — Coded representation of immersive media — Part 14: Scene description, ISO/IEC 23090-14, 2023, https://www.iso.org/standard/80900.html

[4] ISO/IEC JTC 1/SC 29/WG 2, “MPEG-I Phase 2 requirements,” ISO/IEC JTC 1/SC 29/WG 2 N00230, July 2022

[5] ISO/IEC, Information technology — Runtime 3D asset delivery format - Khronos glTF™2.0, ISO/IEC 121113:2022, June 2022, https://www.iso.org/standard/83990.html

[6] Khronos Group, “Khronos Intellectual Property Framework Briefing,” May 2015, https://www.khronos.org/members/ip-framework/

[7] Pixar Animation Studios, “Universal Scene Description,” https://openusd.org/release/index.html

[8] Alliance for Open USD (AOUSD), https://aousd.org/

[9] World Wide Web Consortium, “W3C Patent Policy,” September 2020, https://www.w3.org/Consortium/Patent-Policy-20200915/

[10] Immersive Digital Experiences Alliance, “ITMF Specification Suite,” https://www.immersivealliance.org/download/download-itmf/

[11] B. Cowan and B. Kapralos, "A Survey of Frameworks and Game Engines for Serious Game Development," 2014 IEEE 14th International Conference on Advanced Learning Technologies, Athens, Greece, 2014, pp. 662-664.
doi: https://doi.org/10.1109/ICALT.2014.194.

[12] ISO/IEC, Information technology — Coded representation of immersive media — Part 5: Visual volumetric video-based coding (V3C) and video-based point cloud compression (V-PCC), ISO/IEC 23090-5, June 2021, https://www.iso.org/standard/73025.html

[13] Academy Software Foundation, OpenVDB, https://www.openvdb.org/

[14] The Draco Authors, DRACO 3D Data Compression, https://google.github.io/draco/

Introduction Authors

Youngkwon Lim

- Aug. 2012 ~ : He has been Principal Researcher of Samsung Research America, Plano, Texas, USA
- Aug. 2011 ~ Jul. 2012 : He has been a postdoctoral researcher of University of Texas at Dallas, Plano, Texas, USA
- Mar. 2002 ~ Aug. 2011 : He has received Ph. D. degree in electronics engineering from Hanyang University, Seoul, Korea
- Mar. 2000 ~ Jul. 2011 : He has been a leader of the development of domestic and international standards of Digital Multimedia Broadcasting and interactive service solution in net&tv Inc. Seoul, Korea.
- Jan. 1996 ~ Mar. 2000 : He has been researcher of ETRI, Daejon, Korea
- Mar. 1990 ~ Feb. 1996 : He has received M.S and B.S. degrees in electronics engineering from Korea Aerospace University, Seoul, Korea
- ORCID : https://orcid.org/0000-0003-1472-7430
- Research interests : immersive media systems