An Overview of the OMAF Standard for 360° Video

Miska M. Hannuksela[*], Ye-Kui Wang[+] and Ari Hourunranta[*]

[*]*Nokia Technologies*
Hatanpään valtatie 30
33100 Tampere, Finland
{miska.hannuksela,ari.hourunranta}@nokia.com

[+]*Huawei Technologies*
10180 Telesis Ct
San Diego, CA 92121, USA
yekui.wang@huawei.com

***Abstract**:* Omnidirectional MediA Format (OMAF) is arguably the first virtual reality (VR) system standard, recently developed by the Moving Picture Experts Group (MPEG). OMAF defines a media format that enables omnidirectional media applications, focusing on 360° video, images, and audio, as well as the associated timed text, supporting three degrees of freedom (3DOF). This paper gives an overview of the first edition of the OMAF standard.

1. Introduction

Virtual reality (VR) has been researched and trialed for many years [1][2]. Thanks to the growth of computing capability in devices and network bandwidth as well as advances in the technology for head-mounted displays, VR's wide deployment became possible only recently. Facebook's two-billion-dollar acquisition of Oculus purchase in 2014 seemed to be a start and a catalyst to the fast proliferation of VR research and development, device production, and services throughout the globe. Almost suddenly, VR became a buzzword everywhere in the world, many companies started to have VR as an important strategic direction, and all kinds of VR cameras and devices started to be available in the market.

Unavoidably, numerous different, non-interoperable VR solutions have been designed and used. This called for standardization, for which the number one target is always to enable devices and services by different manufactures and providers to interoperate.

The Moving Picture Experts Group (MPEG) started to look at the development of a VR standard in October 2015. This effort led to the arguably first VR system standard, called Omnidirectional MediA Format (OMAF) [3]. The first edition of OMAF was finalized in October 2017. Unless otherwise stated, by OMAF we mean OMAF first edition.

OMAF defines a media format that enables omnidirectional media applications, focusing on 360° video, images, audio, and timed text. In this paper, we focus on video and won't discuss the other media types beyond this point. OMAF supports only a simple version of VR that enables three degrees of freedom (3DOF). Only rotations around any coordinate axes are supported, whereas purely translational movement of the user would not result in different omnidirectional media being rendered. Although OMAF specifies a format that enables rendering of omnidirectional media, it does not specify the rendering process.

OMAF content authoring includes pre-processing, encoding, and file and segment encapsulation. OMAF supports both projected and fisheye omnidirectional video. For the former, the pre-processing necessarily includes stitching and projection, and may also include region-wise packing. An OMAF player performs media file and segment decapsulation, media decoding, and media rendering. For fisheye video the image stitching is a part of the OMAF player. Representation formats of omnidirectional video are discussed in detail in Section 2.

2375-0359/19/$31.00 ©2019 IEEE
DOI 10.1109/DCC.2019.00050

The key underlying technologies for file/segment encapsulation and delivery for OMAF are ISO Base Media File Format (ISOBMFF, ISO/IEC 14496-12) and Dynamic Adaptive Streaming over HTTP (DASH, ISO/IEC 23009-1). OMAF specifies file format and DASH extensions in a backward compatible manner to enable efficient signaling and delivery of omnidirectional media. ISOBMFF and DASH basics as well as their OMAF extensions are overviewed in Section 3. Note that OMAF also specifies signaling and delivery of omnidirectional media over MPEG Media Transport (MMT, ISO/IEC 23008-1).

Specifics of the use of media codecs are defined in OMAF media profiles, which are specified as requirements and constraints on media coding as well as on signaling and encapsulation of the media data in an ISOBMFF file. Moreover, OMAF specifies the use of DASH with media profiles. The OMAF video profiles are overviewed in Section 4.

In Section 5 we discuss some implementation aspects of OMAF, for both content authoring and player. After that, in Section 6, we draw a conclusion and take a look at future VR standardization work in MPEG.

360° video causes challenges for decoding and delivery due to its large resolution and bitrate. In order to mitigate these challenges, OMAF supports viewport-dependent streaming, where displayed area of the 360° video, i.e. the viewport, has higher picture quality than the remaining areas of the sphere. We review features that facilitate viewport-dependent operation through all Sections of this paper.

2. Representation formats of omnidirectional video

2.1. The coordinate system

OMAF uses a right-handed coordinate system, where the user looks from the sphere center outward towards the inside surface of a unit sphere on which the video is mapped. OMAF specifies global coordinate axes that are shared for all media types intended to be rendered together and used for determining the initial viewing orientation. Each video may use its own local coordinate axes. Rotation metadata, which consists of yaw, pitch, and roll rotation angles around coordinate axes, specifies the relation between global and local coordinate axes. The use of unaligned global and local coordinate axes can be advantageous e.g. for correcting the horizon to be exactly horizontal in the projected omnidirectional video or for improving perceived picture quality by avoiding seams between projection surfaces to cross objects of interest.

2.2. Omnidirectional projection formats

Omnidirectional projection is the most fundamental aspect of projected omnidirectional video. Projection is a necessary geometric operational process used at the content production side to generate 2D pictures from the stitched sphere signal, and an inverse operation of the projection process needs to be used in the rendering process by the OMAF player. OMAF specifies the support of two types of projection, equirectangular projection (ERP) and cubemap projection (CMP). Other projection formats were also considered but not agreed to be included in OMAF.

As illustrated in Figure 1, the process for ERP is close to how a 2D world map is typically generated, but with the left-hand side being the east instead of the west, as the viewing perspective is opposite. In ERP, the user looks from the sphere center outward towards the inside surface of the sphere, while for a world map, the user looks from outside the sphere towards the outside surface of the sphere.

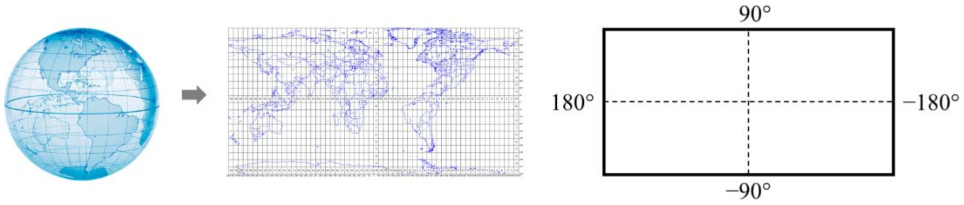

Figure 1. An illustration of the equirectangular projection.

As illustrated in Figure 2, in the CMP supported in OMAF, the sphere signal is rectilinearly projected into six square faces, which are laid out to form rectangle with 3:2 ratio of width vs. height, with some of the faces rotated to maximize continuity across face edges.

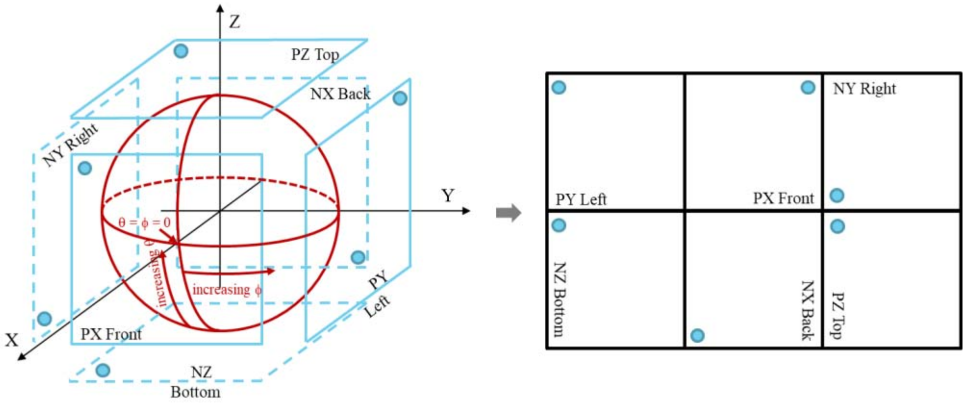

Figure 2. An illustration of the cubemap projection in OMAF [3].

OMAF supports both mono and stereo omnidirectional video content. Stereoscopic content is realized by frame packing, with three frame packing arrangement types supported: side-by-side, top-bottom, and temporal interleaving.

2.3. Region-wise packing (RWP)

RWP is an optional step after projection (at the content production side). It enables manipulations (resize, reposition, rotation by 90°, 180°, or 270°, and vertical/horizontal mirroring) of any rectangular region of the packed picture before encoding.

The RWP metadata indicates the interrelations between regions in the projected picture (e.g. an ERP picture) and the respective regions in the packed picture (i.e. the picture in the coded video bitstream) through the position and size of the regions in both projected and packed pictures as well as indications of the applied rotation and mirroring, if any. When RWP has been applied, the decoded pictures are packed pictures characterized by RWP metadata. Players can map the regions of decoded pictures onto projected pictures and consequently onto the sphere by processing the RWP metadata.

RWP can be used for many purposes, including: 1) indicating the exact coverage of content that does not cover the entire sphere (see Figure 3a); 2) generating viewport-specific video or extractor tracks with region-wise mixed-resolution packing or overlapping regions (see

Section 4.2); 3) compensating the pole area oversampling of ERP e.g. as illustrated in Figure 3b, and 4) arranging the cube faces of CMP in an adaptive manner.

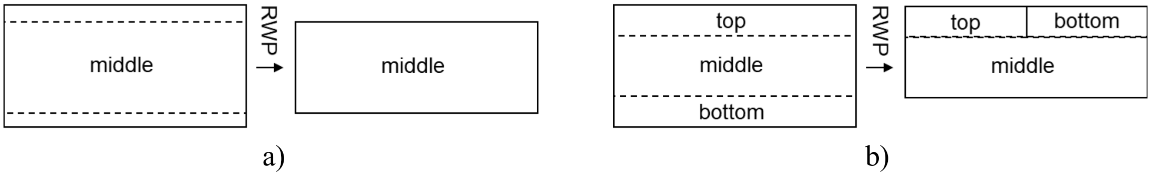

Figure 3. Examples of RWP for ERP; a) for indicating coverage of 135° in elevation; b) for compensating polar area oversampling.

2.4. Fisheye video

As mentioned in Section 1, OMAF also specifies the support of omnidirectional media with fisheye video. Fisheye video does not use projection or region-wise packing. Rather, for each picture, the circular images captured by fisheye cameras are directly placed into a rectangle. Parameters indicating the characteristics of the fisheye video are signaled and used for correct rendering. Compared to projected omnidirectional video, the main advantage of fisheye video is its support of low cost, user generated omnidirectional media content by mobile terminals.

3. Format extensions for omnidirectional video

The ISOBMFF is a popular media container format for audio, video, and timed text. ISOBMFF compliant files are often casually referred to as MP4 files. A basic building block in ISOBMFF is called a box, which is a data structure consisting of a four-character-code (4CC) type, the byte count of the box, and a payload, whose format is determined by the box type and which may contain other boxes. An ISOBMFF file consists of a sequence of file-level boxes. Each stream of timed media or metadata is logically stored in a track. A sample entry of a track describes the coding and encapsulation format used in the samples and includes a 4CC sample entry type and contained boxes that provide further information of the format or content of the track. A restricted video sample entry type (`'resv'`) is used for video tracks that require post-processing operations after decoding to be displayed properly. The type of post-processing is specified by one or more scheme types associated with the restricted video track. For example, ISOBMFF specifies a restricted video scheme type for frame-packed stereoscopic video.

OMAF requires the use of restricted video tracks for omnidirectional video. The restricted video scheme types specified in OMAF are summarized in Table 1.

The metadata for projected omnidirectional video are stored as boxes in the sample entry. The metadata can include projection format, frame packing type, RWP metadata, rotation metadata, and content coverage (i.e. the sphere regions covered by the content).

The default viewing orientation to start displaying the omnidirectional video is along the X-axis of the global coordinate axes. Content authors can override the default behavior by using an initial viewing orientation timed metadata track.

Table 1. Restricted video scheme types specified in OMAF

Scheme	Description
podv	Open-ended scheme type for any projected omnidirectional video without limitations on projection format or region-wise packing. Both monoscopic and stereoscopic content are allowed, and the content coverage can be less than 360°.
erpv	Closed scheme type constraining 'podv' to equirectangular projection only and a maximum of one packed region per view for region-wise packing.
ercm	Closed scheme indicating constraining 'podv' to either the equirectangular or cubemap projection.
fodv	Open-ended scheme type for any fisheye omnidirectional video.

OMAF specifies region-wise quality ranking (RWQR) metadata as a basic mechanism to enable viewport-dependent content selection. Quality ranking metadata can be provided for sphere regions and for rectangular regions on decoded 2D pictures. Quality ranking values are given for indicated regions and describe the relative quality order of the regions: when region A has a non-zero quality ranking value less than that of region B, region A has a higher quality than region B. The quality ranking values need not represent any absolute quality scales, and value 0 indicates an undefined quality ranking value.

DASH specifies a Media Presentation Description (MPD) format for describing the content available for streaming and Segment formats for the streamed content. In the Segment format for ISOBMFF, each Segment consists of one or more self-contained movie fragments. The MPD is specified as an Extensible Markup Language (XML) schema and contains one or more Adaptation Sets, each containing one or more Representations. A Representation corresponds to an ISOBMFF track, and an Adaptation Set contains Representations of the same content between which the player can select e.g. based on the available bitrate. Each Segment can be split to Subsegments, which are indexed within the Segment itself. This design enables clients to fetch the Segment indexing first and then conclude byte ranges for Subsegments.

OMAF specifies MPD extensions assisting in selection of omnidirectional content. Table 2 summarizes the descriptors that can be used for characterizing omnidirectional video content.

Table 2. MPD descriptors for characterizing omnidirectional video

Descriptor name	Description
FramePacking element	Stereoscopic frame packing format
PF descriptor	Indicates projected omnidirectional video and the used projection format(s)
RWPK descriptor	Indicates whether region-wise packing has been applied
CC descriptor	Content coverage
SRQR descriptor	Region-wise quality ranking information for sphere regions
2DQR descriptor	Region-wise quality ranking information for rectangular regions on decoded pictures
FOMV descriptor	Indicates fisheye omnidirectional video. Additionally, common two-lens setups (monoscopic 360°, stereoscopic 180°) can be indicated.

4. OMAF Video Profiles

4.1. Overview of OMAF Video Profiles

OMAF supports viewport-agnostic content authoring and streaming, where existing encoding and player implementations for 2D video can be used with straightforward modifications. The viewport-agnostic approach is simple and enables re-using implementations built for 2D video, but has the following disadvantages:

1. The bitrate required for high-quality viewport-independent 360° video content is often higher than the network throughput in many access networks.
2. The decoding capacity of many today's devices is limited to approximately 4K real-time decoding (e.g., 4096×2048 ERP resolution), while high-quality 360° video content can have higher resolution and many head-mounted displays are also suitable for viewports extracted from ERP content of higher than 4K resolution.

OMAF provides tools and media profiles for viewport-dependent streaming to overcome the above-described disadvantages. In viewport-dependent streaming the viewport or a superset of the viewport covering slightly more than the viewport is covered by higher quality video than the remaining areas of the 360° video.

Table 3 summarizes the OMAF video profiles. The required video codec level approximately corresponds to the decoding capacity given in the table. The closed scheme types listed in the table specify which projection formats are allowed and which constraints, if any, apply to region-wise packing. The scheme types also imply which file format metadata must or can be present. The OMAF video profiles also require equivalent metadata to be present as supplemental enhancement information in the video bitstreams.

Table 3. Overview of OMAF video profiles.

OMAF video profile	**Codec**	**Bit depth**	**Decoding capacity**	**Scheme types**
HEVC-based viewport-independent OMAF video profile	HEVC Main 10 Profile	≤10 bits	4K @ 60 Hz	`erpv`
HEVC-based viewport-dependent OMAF video profile	HEVC Main 10 Profile	≤10 bits	4K @ 60 Hz	`erpv`, `ercm`
AVC-based viewport-dependent OMAF video profile	AVC Progressive High Profile	8 bit	4K @ 30 Hz	`erpv`, `ercm`

4.2. Viewport-dependent streaming enabled by OMAF video profiles

This section provides more details on how HEVC- and AVC-based viewport-dependent OMAF video profiles support viewport-dependent streaming.

Methods to achieve viewport-dependent streaming can be categorized into the following approaches:

1. Viewport-specific 360° streams: Several versions of 360° content are encoded for different viewport orientations and different bitrates or qualities. The player chooses the 360° stream that covers the current viewport at higher quality and suits the network

throughput. Note that the 360° stream also covers areas other than the current viewport, albeit at lower quality.

2. Tile-based viewport-dependent 360° streaming: Projected pictures are partitioned into tiles that are coded as motion-constrained tile sets (MCTSs). Several versions of the content are encoded at different bitrates using the same MCTS partitioning. Each MCTS sequence is stored as an HEVC-compliant sub-picture track (with sample entry type `'hvc1'`) and made available for streaming as a Representation. The player selects on MCTS basis which bitrate or quality is received. A subdivision of the tile-based viewport-dependent streaming schemes is provided in Table 4. Further details of how OMAF supports this category of viewport-dependent streaming are provided below.

Table 4. Tile-based viewport-dependent 360° streaming approaches.

Approach	Description	OMAF support?	E.g.
Region-wise mixed quality (RWMQ)	Several versions coded with the same tile grid and different bitrate / picture quality. Players choose high-quality MCTSs for the viewport.	yes	[4], OMAF Annex D
Viewport + 360° video	A complete low-resolution/low-quality omnidirectional picture and high-resolution MCTSs covering the viewport are received.	yes	OMAF Annex D
Region-wise mixed resolution (RWMR)	Tiles are encoded at multiple resolutions. Players select a combination of high resolution tiles covering the viewport and low-resolution tiles for the remaining areas.	yes	[5], OMAF Annex D

HEVC enables partitioning of a picture into tiles along a grid of tile columns and rows. Encoders can restrict the encoding in a manner that a motion-constrained tile set (MCTS) references data only within the same MCTS in the current and reference picture(s). AVC does not include the concept of tiles, but the operation like MCTSs can be achieved by arranging regions vertically as slices and restricting the encoding similarly to encoding of MCTSs. Examples of viewport-dependent schemes for AVC are available in Annex D of OMAF. For simplicity, the terms tile and MCTS are used in this paper but should be understood to apply to AVC too in a limited manner.

Viewport-dependent profiles enable players to use either a single decoder instance or one decoder instance per sub-picture track, depending on the capability of the device and operating system where the player runs. The use of single decoder instance is enabled through extractor tracks as explained in the next paragraph. To facilitate multiple decoder instances, viewport-dependent OMAF profiles use sub-picture tracks rather than HEVC tile tracks (sample entry type `'hvt1'`). While the bitrate of respective tile and sub-picture tracks is similar, sub-picture tracks have the advantage that multiple decoder instances can be used without the need of manipulating the tracks or changing the decoder operation.

OMAF specifies the use of extractor tracks to assist in merging MCTSs into a single bitstream. Extractor tracks are specified in the ISOBMFF encapsulation format of HEVC and AVC bitstreams (ISO/IEC 14496-15). Samples in an extractor track contain instructions to reconstruct a valid HEVC or AVC bitstream by including rewritten parameter set and slice header information and referencing to byte ranges of coded video data in other tracks. Consequently, a player only needs to follow the instructions of an extractor track to obtain a decodable bitstream from tracks containing MCTSs.

In the RWMQ method described in Table 4, one extractor track per each picture size and each tile grid is sufficient. In the other two methods in Table 4, one extractor track is needed for each distinct viewing orientation.

5. Implementing OMAF

5.1. Content authoring

Preparation of viewport-agnostic 360° video content has a minor impact on the conventional video content preparation pipeline. Basically, the new properties of ISOBMFF and DASH MPD as described in Section 3 are included in the content preparation pipeline. Support for features that may not have been considered before, like a timed metadata track for initial viewing orientations, may need to be implemented.

Viewport-dependent profiles, on the other hand, require a greater number of changes to a conventional on-demand or live video authoring and stream preparation pipeline. First the viewport-dependent streaming approach needs to be selected (see Section 4.2). In the following, we concentrate on tile-based operation of HEVC-based viewport-dependent profile. The OMAF content preparation work flow includes the following steps:

1. The original 360° video is pre-processed, when needed. For example, downsampling or projection format conversion may be required.
2. The content is encoded using MCTSs. Usually multiple versions of the content are generated at different bitrates. A relatively short random access interval, e.g. in the order of 1 second, is used in encoding to enable frequent viewport switching.
3. HEVC-compliant sub-picture tracks are generated from the MCTS sequences by extracting each MCTS sequence from the bitstream and rewriting HEVC slice header and parameter set data as described in Section 4.2, approach 2.
4. One or more extractor tracks are created according to the selected viewport-dependent streaming approach (see Section 4.2) to enable composing a full 360° picture out of the sub-picture tracks. For the RWMQ scheme, a single extractor track is sufficient, whereas for the other two tiling schemes described in Section 4.2, there is a set of extractor tracks, each representing a distinct viewing direction. Further, RWP metadata is created to indicate the arrangement of the MCTSs within the decoded picture.
5. (Sub)segments are created from each track for DASH delivery, and an MPD is generated. Each extractor track forms a Representation in its own Adaptation Set. The sub-picture Representations covering the same sphere region at the same resolution but at different bitrates form an Adaptation Set. The extractor Adaptation Set(s) are associated with the corresponding sub-picture Adaptation Sets in the MPD.

The Nokia OMAF implementation [6] covers the steps 3-5 above.

5.2. Player

Processing of viewport-independent OMAF content has a minor impact on the DASH (MPD parsing) and File format parsing components of a 360° DASH video player. Further, supporting initial viewing orientation and rotation are rather simple operations for a player that already supports head mounted devices for playback.

Players are not mandated to utilize extractor tracks for viewport dependent OMAF content, but it simplifies the player operation and is hence recommended. In the following, a player operation for streaming based on extractor tracks is described, also implemented in the Nokia OMAF implementation [6]. The main steps are: 1) Selection of an extractor Representation / track; 2) Selection of sub-picture Representations / tracks, and 3) Resolving an extractor track.

When streaming over DASH, the extractor Adaptation Sets are associated with sub-picture Adaptation Sets, each possibly containing multiple Representations. SRQR descriptors of the Adaptation Sets or Representations indicate the quality ranking values for the covered sphere regions. In the RWMQ approach with a single extractor, the player selects the best quality subpicture Representations for the prevailing viewport, while in the other viewport-dependent approaches, the player needs to select one of the extractor Adaptation Sets, which then provides the optimized sub-picture Adaptation Set selection for the viewport.

To follow user's viewing orientation changes, the player must be able to switch either the used Representations (RWMQ) or Adaptation Sets (other approaches). The latency to obtain a good picture quality on the viewport after the viewing orientation changes should be kept as low as possible, while playback interruptions due to re-buffering should be avoided. Optimization of the Adaptation set and/or Representation switching logic for tile-based viewport-dependent streaming is an important research field.

Further, the viewport-based switching logic needs to be integrated with the adaptive bitrate (ABR) switching logic. There can be similarities in the ABR switching logic to conventional ABR, e.g. in how to take network conditions into account when selecting from which (Sub)segment to start downloading the Representation being switched to. Video streaming based on multiple sub-picture streams with short segments, however, can make bandwidth estimation even more unreliable than in conventional video streaming systems, where estimations are found challenging too [7].

A traditional DASH player can parse and decode media Segments as they arrive, possibly switching Representations based on network conditions. An OMAF player, however, must be able to download several Representations in parallel and process them in time-aligned manner, i.e., the movie fragments with the same time span can be processed only when the selected set of Representations have them available.

Resolving extractor tracks involves parsing of metadata and concatenating video data from multiple sub-picture and extractor tracks to a single HEVC bitstream. This enables decoding the merged bitstream using a conventional single video decoder subsystem.

Region-wise packing used in the RWMR and viewport+360° approaches affect the Renderer and File format parsing subsystems. The Renderer subsystem must be able to compose the output picture from several rectangular regions, possibly using conventional Graphics Programming Unit operations for upscaling, rotating, and mirroring the regions.

From run-time complexity point of view, OMAF does not bring any noticeable burden on top of a conventional 360° video player. The extractor concept introduces one or two memory copying operations for compressed video but for a 360° video player running on modern 4K-video capable devices, the impact should not be noticeable. However, some devices may have problems downloading tens of HTTP streams in parallel, each requiring bandwidth of up to several Mbps. It is therefore advisable to keep a good tradeoff between the number of required sub-picture Representations and the granularity of MCTS grids.

6. Conclusions, acknowledgment, and future work

An overview of the first edition of Omnidirectional MediA Format (OMAF), the arguably first virtual reality system standard, was provided. The overview focused on omnidirectional video, without much details on audio, image, and timed text. What was discussed in detail includes the video representation formats, the file format and DASH extensions, as well as the OMAF video profiles. In addition, the content authoring and player implementing aspects were also discussed.

The authors would like to greatly thank the numerous MPEG delegates who have contributed to the development of OMAF first edition.

At the time of writing this paper, besides maintenance of OMAF first edition, the OMAF group of MPEG is working towards the second edition of OMAF, which is expected to enable new functionalities and optimize the omnidirectional media delivery further. The expected new functionalities include capability of switching between multiple viewpoints, each corresponding to one 360° camera, support of six degrees of freedom with a limited viewing space, and overlays with various properties.

References

[1] R. S. Kalawsky, *The Science of Virtual Reality and Virtual Environments: A Technical, Scientific and Engineering Reference on Virtual Environments*, Addison-Wesley, 1993.

[2] F. Biocca and M. R. Levy (ed.), *Communication in the Age of Virtual Reality*, Lawrence Erlbaum Associates, 1995.

[3] Y.-K. Wang, M. M. Hannuksela, B. Choi, A. Murtaza, and Y. Lim (ed.), “Revised text of ISO/IEC FDIS 23090-2 Omnidirectional Media Format,” MPEG output document N17563, Apr. 2018.

[4] R. Ghaznavi-Youvalari et al., “Comparison of HEVC coding schemes for viewport-adaptive streaming of omnidirectional video,” IEEE International Workshop on Multimedia Signal Processing, Oct. 2017.

[5] A. Zare et al., “6K Effective Resolution with 4K HEVC Decoding Capability for OMAF-compliant 360° Video Streaming,” Packet Video Workshop, June 2018.

[6] https://github.com/nokiatech/omaf

[7] Y. Sani, A. Mauthe, and C. Edwards, “Adaptive Bitrate Selection: A Survey,” *IEEE Communications Surveys & Tutorials*, vol. 19, Issue 4, pp. 2985–3014, 12 July 2017.