VIRTUAL REALITY AND DASH

D. Podborski[1], E. Thomas[2], M. M. Hannuksela[3], S. Oh[4], T. Stockhammer[5], S. Pham[6]

[1]Fraunhofer Heinrich Hertz Institute, Germany
[2]TNO, the Netherlands
[3]Nokia, Finland
[4]LGE, S. Korea
[5]Qualcomm, Germany
[6]Fraunhofer FOKUS, Germany

ABSTRACT

Virtual Reality (VR) has lately gained significant attention primarily driven by the recent market availability of consumer devices, such as mobile phone-based Head Mounted Displays (HMDs). Apart from classic gaming applications, the delivery of 360° video is considered as another major use and is expected to be ubiquitous in the near future. However, the delivery and decoding of high-resolution 360° videos in desirable quality is a challenging task due to network limitations and constraints on available end device decoding and processing. In this paper, we focus on aspects of 360° video streaming and provide an overview and discussion of possible solutions as well as considerations for future VR video streaming applications. This paper mainly focuses on the status of the first standardization activities to support interoperable 360° video streaming. More specifically, MPEG's ongoing work on Omnidirectional MediA Format (OMAF) is introduced -- aiming at harmonization of VR video platforms and applications. The paper also discusses the integration in MPEG Dynamic Adaptive Streaming over HTTP (MPEG-DASH), which is considered a cornerstone of 360° video streaming services with OMAF content. In the context of the general OMAF service architecture, three distinct delivery approaches and considerations for content protection are discussed.

1 INTRODUCTION

In recent years, there has been a lot of activity around Virtual Reality (VR) as evidenced by large industry engagement. Many important players in the computer industry have demonstrated support for VR and introduced Head Mounted Displays (HMDs), such as Oculus Rift, HTC Vive, Samsung GearVR, Sony PlayStation VR, and Google Daydream. Expecting that the increasing popularity of consumer VR HMDs will lead to an increased demand for VR content, various companies have also started to develop omnidirectional cameras to allow capturing of 360° video content. There are many low-cost consumer solutions, such as Ricoh Theta, Samsung Gear 360, and LG 360 Cam, as well as more

expensive professional 360° cameras, such as Nokia OZO, GoPro Omni, and Fraunhofer OmniCam360, already available on the market. At the same time, major multimedia streaming platforms, such as YouTube and Facebook, have already launched support for 360° video streaming for VR devices and there have even been several successful live game broadcastings of professional American sports leagues [1] or live event streaming in VR [2]. Motivated by the industry interest on 360° video delivery for VR, several industry forums and standardization bodies have started work. In February 2016, MPEG launched an activity on Omnidirectional MediA Format (OMAF) [6] that aims at standardizing the storage and delivery format for 360° audio-visual content by the end of 2017 in order to avoid market fragmentation. In early 2017, the Virtual Reality Industry Forum (VR-IF) [3] was established to support high quality interoperable VR experiences. Also, the 3rd Generation Partnership Project (3GPP) [4] has an ongoing study item on VR that could possibly lead to a normative work item starting in the course of 2017. Moreover, the W3C WebVR Community Group [5] is specifying APIs for accessing VR devices on the web.

This paper provides an overview of the key concepts of OMAF in Section 2. Section 3 describes three different streaming approaches, which are under consideration for inclusion in OMAF specification and discusses their advantages and disadvantages. In Section 4, considerations on security and content protection are discussed. Finally, conclusions are provided in Section 5.

2 OVERVIEW OF OMAF

2.1 Content processing architecture of OMAF

An overview of the envisioned OMAF content processing architecture and its components are depicted in Figure 1. VR systems enable users to navigate through 360° video and create an immersive user experience. A scene (A) is captured by multiple video cameras pointing in different directions and results in a video signal (Bv). Before the encoding (Ev) is carried out, the content, which is captured by several cameras, is stitched together and the stitched image is projected onto a three-dimensional projection structure (e.g. a sphere). The image data on the projection structure is further arranged onto a two-dimensional projected picture and regions of the picture are mapped onto a packed picture, assuming the optional region-wise packing is used. After encoding, the content is encapsulated into ISO Base Media File Format (ISOBMFF) segments (F) together with additional metadata information that provides additional signalling for DASH clients. The segments are then delivered to the client over HTTP using unicast, multicast or broadcast delivery. After downloading the file segments (F’), the client decapsulates the coded bit-streams (E’v) and extracts the corresponding metadata. Finally, the video is decoded and rendered on the client device using the extracted metadata according to the viewing orientation of the user.

Although not considered in this paper, OMAF allows for circular images to be captured by one or multiple fisheye cameras. In this case, stitching, projection, and region-wise packing are omitted in favour of directly rendering the fisheye video on the client side. However, the delivery schemes described in this paper, especially the ones that adapt to the current user viewport do not relate to this alternative approach. It is relevant to mention that a full immersive experience for VR is built on 3 cornerstones: maximum video quality, full 3D spatial audio and system interaction providing minimum latency. OMAF addresses all three aspects with the availability and inclusion of MPEG-H audio for full 3D sound

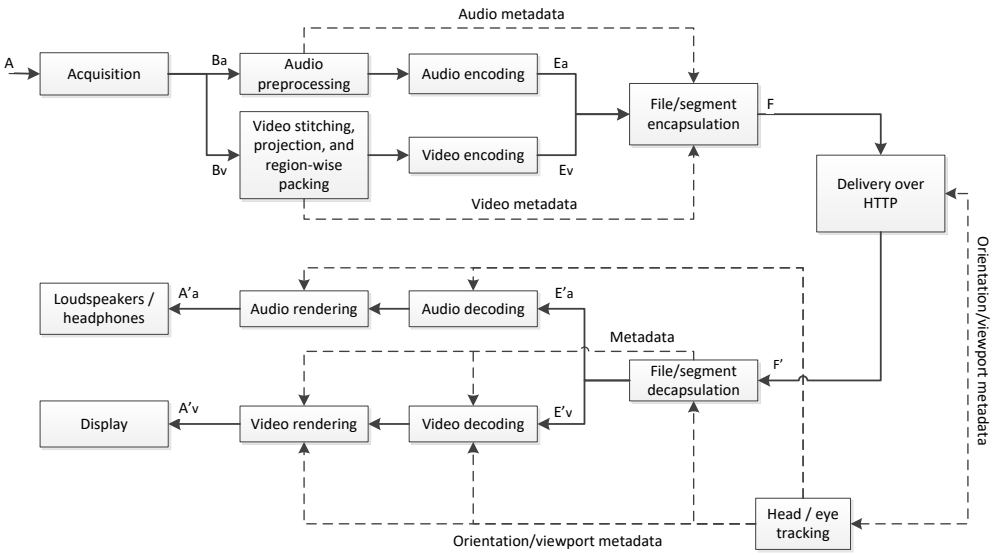

Figure 1. VR streaming service architecture [6].

capabilities. Full immersion is provided by sufficiently good rendering engines that enable motion-to-photon latencies in the range of 20ms. Provided that, after video decoding, a full 360° video and a full 3D audio scene is available, modern HMD architectures can provide full immersion and movement in the audio-visual scene. In the remainder of the paper we focus on the most challenging media tasks, namely video delivery, decoding and rendering.

2.2 Projection

For video processing, assuming the availability of full 360° video, projection enables mapping of a 360° image to a 2D representation. This is carried out by first defining a three-dimensional projection structure, such as a sphere or a cube, which the 360° video is projected onto, and second, forming a two-dimensional plane (projected picture) onto which the defined structure is mapped. The most commonly used projections are the Equirectangular Projection (ERP) and the Cube Map Projection (CMP). In addition, several projections have been proposed recently. Viewport-agnostic projections represent the full 360° video in equal quality, independent of the viewport, whereas in viewport-dependent projections a certain area of the 360° video is represented with a higher fidelity than other areas (see for instance 2.3.3). Additional projection maps in both categories are currently being investigated for inclusion in the OMAF specification: most of these are described in [15].

ERP is the most common and basic projection method used for 360° video. Figure 2 illustrates the mapping from a sphere to a 2D plane. The sphere is sampled into several horizontal circles (circles of latitudes) and each of those is mapped to a horizontal line onto the rectangular 2D picture plane. The lines (on the plane) towards the upper and lower picture boundary are significantly stretched with respect to their respective circles on the sphere.

The main benefits of ERP are its simplicity, wide deployment and the availability of original footage in this format. However, it suffers from several issues. First, video content is affected by geometric distortion that hurts compression efficiency as motion compensation of traditional video codecs is based on linear motion models. Second, picture memory requirements (both in video coding and graphics processing components) are inflated due to the inherent oversampling towards the sphere poles.

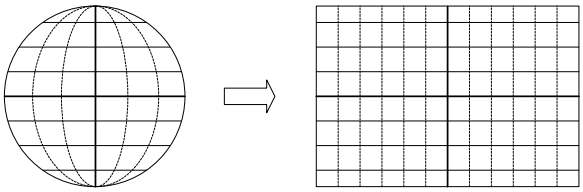

Figure 2. Equirectangular Projection

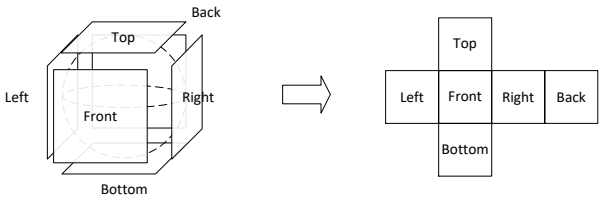

Figure 3. CubeMap Projection

Another widely used projection is CMP, for which the 360° video is projected onto the six faces of a cube and the faces are subsequently arranged onto a rectangular 2D plane. An example is shown in Figure 3. The compression performance of ERP and CMP was compared in [14], suggesting that on average CMP is slightly more efficient than ERP with relatively large sequence-wise differences. CMP comes along with good support in rendering frameworks such as OpenGL. In addition, the rectangular nature of CMP also naturally lends itself to tile-based approaches as described later in this paper. A considerable number of more sophisticated projections for 360° video is discussed in the various standardization activities, however, none of these projections has yet gained significant traction or adoption [15].

2.3 Region-Wise Packing

The region-wise packing process may be carried out to projected pictures prior to encoding. The selected projection format and the region-wise packing metadata are stored jointly with the encoded video signal (Ev) in a media file (F) so that the inverse process can be applied at the receiver side. For each region, the metadata defines a rectangle in a projected picture, the respective rectangle in the packed picture, and an optional transformation of rotation by 90, 180, or 270 degrees and/or mirroring. As the sizes of the respective rectangles can differ in the projected and packed pictures, the mechanism infers region-wise resampling. At the time of writing this paper, OMAF only specifies rectangular packing.

Among others, region-wise packing provides signalling for the following usage scenarios:

1) Additional compression for viewport-independent projections is achieved by increasing the sample density of different regions to achieve more uniformity across

the sphere. For example, the top and bottom parts of ERP are oversampled and region-wise packing can be applied to down-sample them horizontally as illustrated in Figure 4. This method provides about 4.3% bit-rate reduction on average [7].

2) Arranging the faces of plane-based projection formats, such as CMP, in an adaptive manner. Different arrangements of cube faces in packed pictures have been studied in [8] and [9].

3) Generating viewport-dependent bit-streams that use viewport-independent projection formats. For example, regions of ERP or faces of CMP can have different sampling densities as illustrated in Figure 5 and the underlying projection structure can have different orientations.

4) Indicating regions of the packed pictures represented by an extractor track. This is needed when an extractor track collects tiles from bit-streams of different resolutions (see Section 3.3).

Figure 4. Example of packing of regions of an equirectangular panorama.

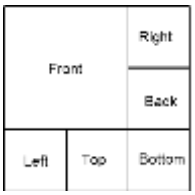

Figure 5. Example of unequal sampling densities of cube faces.

2.4 Recommended Viewport and Initial Viewpoint

The key aspect of immersion in Virtual Reality applications is the freedom given to the user to move within a 360° scene. In 360° video content, the viewer is able to look in the desired direction. However, this total freedom of viewing direction interferes with the art of storytelling because of the director's inability to control the user's gaze. In addition, 360° content may not offer interesting action all around the viewer at any point in time. Therefore, there is a risk that the viewer may miss an important part of the scene that contributes to the story, especially: when there is a scene change; when the user tunes in to live content; or when the viewer fast-forwards within a movie. For these reasons, the OMAF specification provides metadata signalling for two main use cases that are: initial viewpoint and recommended viewport. The initial viewpoint gives, for any moment in the content, the point in the content that the application should render first as the centre of the current viewport instead of conventionally the centre of the projected video frame. This ensures that the viewer is initially presented with the interesting part of the scene according to the director. The recommended viewport provides a viewpoint and a Field of View (FoV) recommended by the director. This way, the director is able to frame more precisely the interesting part of the entire 360° scene. One usage of the recommended viewport is for non-VR-capable devices such as traditional TV sets. With such metadata, a media player can present on a TV screen, the director's cut view of the 360° content, hence broadening the range of devices capable of consuming 360° content. For HMD consumption, the director’s cut may be used to provide cues (visual or audio) to users to guide them to the action.

3 STREAMING APPROACHES AND DASH INTEGRATION

3.1 Overview

In a survey preceding the work on MPEG immersive media, the primary identified delivery for VR experiences was Adaptive Bit-Rate (ABR) streaming over the Internet, and DASH provides a deployed standardized framework for ABR delivery. As already mentioned in the introduction, there are different approaches for streaming a 360° video. Currently, three different streaming approaches, each targeting to send the entire 360° video to the client, are under consideration for inclusion in OMAF, namely: viewport-agnostic streaming, viewport-dependent streaming and tile-based streaming. One or more interoperability points, in the form of profiles, will be defined to support these approaches or a subset of them. The most basic approach is a viewport-independent solution and entails sending the entire 360° video in a viewport-agnostic fashion, i.e. regardless of user viewing orientation. Regular DASH clients may be used in order to deliver the video. The second solution provides a viewport-dependent solution in which the user selects an Adaptation Set based on the current viewing orientation. For this purpose, several Adaptation Sets needs to be available on the server, each for a specific user viewing orientation. Note that within each Adaptation Set, multiple Representations may be present that can be used for bit-rate adaptation. In distinct Adaptation Sets, each providing the full 360° video, for instance, one Adaptation Set may enhance the quality/resolution in the specific viewing orientation compared to the rest of the 360° video. The third solution is tile-based streaming, where the 360° video is tiled into separate spatial regions. Each region may be made available at varying quality or resolution. In the following subsections, these three approaches are described in more detail.

3.2 Viewport-Agnostic Streaming

The viewport-agnostic streaming solution is the simplest approach to stream 360° video. It is straightforward in the sense that no modifications are required in the streaming system other than at the content capturing and preparation (i.e. stitching, projection and region-wise packing) and at the rendering process of the player. With this approach, the entire 360° video is encoded as if it were a traditional video and is offered to the DASH client. The DASH client operates without feedback from the HMD orientation sensors.

In this approach, the content is projected onto a viewport-agnostic projection format: e.g. ERP or CMP as described in Section 2. Clients only need to check whether they support the projection/packing scheme used and request one of the offered alternatives based on their capabilities. After choosing the desired projection/packing scheme, rate adaptation of DASH clients occurs as in traditional video streaming, i.e. a DASH client may switch to another representation of the same projection/packing scheme corresponding to another bit-rate or resolution based on throughput characteristics. Therefore, the extensions to DASH are expected to be minimal, i.e. the projection/packing metadata of a given representation needs to be indicated to the DASH client for the content selection.

While its deployment simplicity makes viewport-agnostic streaming approaches attractive, the main issue of such an approach is that a large portion of the bandwidth and decoder resources is used for content that is not displayed at all, if only a single user accesses the

presentation. This results in a waste of bandwidth and decoder resources that could be better utilized for the part of the content that is presented to the user.

3.3 Viewport-Dependent Projection and/or Region-Wise Packing

In order to overcome the issues pointed out for the viewport-agnostic streaming solution, viewport-dependent streaming is a viable option. The key idea is to provide several Adaptation Sets at the server side, each of which emphasizes the video area associated with a given viewing orientation. Each Adaptation Set is thereby encoded, spending a higher number of bits to represent the viewport desired by the client. This can be achieved either using a viewport-dependent projection such as Truncated Square Pyramid Projection (TSP) **Error! Reference source not found.**, region-wise varying quantization step size, or applying region-wise packing with higher resolution for regions representing the desired viewport in a viewport-agnostic projection such as ERP. It was found in [10] that the multi-resolution ERP and CMP bit-streams created through region-wise packing have better streaming rate-distortion performance than pyramid-based viewport-dependent projections.

In order for this approach to work properly and provide a high quality of experience, several versions of the same content need to be made available at the server side, with each version representing a different viewing orientation. The number of versions can be quite high, e.g. many tens of streams. In addition, this number can become even bigger if devices with different FoVs need to be supported. This means that in addition to traditional switching based on throughput characteristics, the client does not only need to select an Adaptation Set that corresponds to a supported projection/packing scheme, but also need to switch Adaptation Sets based on the current viewing orientation. It also requires the DASH manifest to include region-wise quality ranking information [6] indicating the viewing orientation for which the representation has been made available.

With this approach, bandwidth requirements are reduced compared to viewport-agnostic streaming and decoder resources are more efficiently used for content that is actually shown to the user. However, the main drawback of this approach is that more storage is required at the server side and the approach is less cache-efficient, since many different versions are made available at the server and more encodings need to be performed at the content generation side. This can be costly, especially for live streaming services. In addition, when this solution is utilized, fast switching needs to be enabled. Therefore, a low end-to-end (E2E) latency is required and frequent Random Access Points (RAP) need to be made available for the client to switch to the proper Adaptation Set for the current viewing orientation.

3.4 Tile-Based Streaming

Another solution that can be used for viewport-based adaptation is tile-based streaming. The benefit of tile-based streaming, in comparison with the viewport-dependent streaming described, is that the number of versions of the content made available at the server side only depends on the tiling granularity and the number of different representations (varying in qualities or resolutions) of each of the tiles. This number is typically much lower than the number considered for the previous case, albeit only for the same resolution. Besides, the storage capacity required by each of the tiles is much lower than that of each of the

versions made available as described for the previous solution. With this approach, the content is spatially segmented into several tiles. In order to achieve higher bit-rate savings, each of the tiles is typically made available at different resolutions. Clients download the tiles that correspond to the user's viewing orientation at a higher resolution than the other tiles covering the space outside the current viewport. DASH already supports signalling of spatially subdivided content via the Spatial Relationship Description (SRD) [11]. SRD expresses the position and size of tiles relative to a 2D plane, which in this case correspond to the entire 360° video projected onto a rectangle shape as explained in Section 2.2.

If clients have multiple decoders, the tiles can be independent video bit-streams that the clients decode simultaneously. However, care needs to be taken for playback synchronization, especially if the number of tiles increases and such clients are not typically available. Another solution is to use a single decoder that solves the synchronisation issue. For this purpose, tiles need to be aggregated into a single HEVC-compliant bit-stream. This operation can be performed using ISOBMFF extractors for HEVC [12], which enable the creation of tracks that convert individual motion-constrained tile sets into a single conformant video bit-stream so that a regular single decoder can be used for decoding a video subsection or a combination thereof. Essentially, this entails extracting video slice payload data and prepending proper video slice header data [12].

As already mentioned, the number of different versions that need to be encoded for this solution is smaller than for the previous solution. A further benefit of this solution in comparison with the previous one is that the number of streams does not depend on the target FoV of the devices. Larger FoV devices can be supported by just retrieving more high resolution tiles, while for the previous case a version for each target FoV size needs to be made available.

As for the previous solution, a low end-to-end (E2E) latency is required and frequent Random Access Points (RAP) need to be made available for the client to switch to the proper representation of the tiles according to the current viewing orientation of the client. It is also important to mention that only by the combination of tiles with region-wise packing, a single HEVC decoder can be used for a beneficial performance.

3.5 Comparison of the streaming Approaches

The following table summarizes the three presented approaches:

	Viewport-Agnostic	Viewport-dependent	Tile-based streaming
Server Storage Requirement	Low	High	Medium
Cache Storage Requirements	Low	High	Medium
Content Processing Complexity[1]	Low	High	Medium
Bandwidth-Quality Performance	Low	Good	Good
Required stream change latency[2]	N/A	Low	Low
File Format Extensions	Minimal (projection/mapping)	OMAF Projection region-wise packing	Projection Region-wise packing Tiling file format Extractors
DASH Extensions	Projection/Mapping scheme signalling	- Projection/Mapping scheme signalling - Viewport orientation of the representation	- Projection/Mapping scheme signalling - Spatial Subdivision - Aggregation of Tiles

[1]The content processing complexity indicates the amount of processing work necessary on the content creation side with contributing factors such as number and resolution of encodings per video and renderings of viewport-dependent video.

[2]The low latency requirement is meant to ensure that a higher quality is shown in the viewport. Note that the viewport-agnostic approach has no latency requirement, since no quality switches are required for changing viewing orientation.

4 SECURITY CONSIDERATIONS

Web apps and web browsers are becoming popular client platforms, because they can be easily deployed across different devices. With W3C HTML5 Media Source Extensions (MSE) and Encrypted Media Extensions (EME), MPEG DASH and Common Encryption (CENC) [13] enable delivery of protected media content to web browsers. In general, the same concepts of protecting media content can be applied to 360° video. When it comes to projection (ERP or CMP) in the web browser, most solutions today rely on WebGL (an OpenGL implementation targeted at web browsers) and the Canvas API. W3C WebVR specification uses these APIs to provide support for VR devices such as HMDs. However, Canvas causes problems, as secure implementations prevent frames being extracted from a protected media stream. This is often referred to as a 'trusted media path'. As a result, protected media cannot currently be projected in web browser clients. Not only web apps are affected by this problem. DRM implementations on mobile (iOS, Android) or CE devices are increasingly adhering to the model of a trusted media path, where media frames cannot be extracted or manipulated after decoding. Currently, for web browser clients, the W3C WebVR Community Group is specifying web APIs for VR devices, but relies on Canvas (VRCanvasLayer) as well.

In the near term, it is expected that additional 'Layers' will be introduced into the web browsers or native apps to enable secure projection calculations. For this purpose, signalling necessary to enable viewport rendering needs to be made available to the media pipeline. For instance, the projection and orientation information can be signalled using the Supplemental Enhancement Information (SEI) messages, which are inserted into the bit-stream on the codec level. On the file format level, two additional boxes are specified by OMAF to signal the used projection format and region-wise packing method. Finally, for the tile-based streaming approach, the encryption can be applied only on the slice payloads, leaving the slice header data intact, which also allows for exchanging the slice headers through ISOBMFF extractors on encrypted data. This requires encryption to follow the Subsample encryption rules, with only encrypting video coding layer (VCL) data, leaving other non-VCL data unencrypted [13].

5 CONCLUSIONS

In this paper, we described interoperable 360° video streaming aspects. First, we provided an overview of the MPEG OMAF activity and explained the overall streaming architecture for VR content with details on the projection, region-wise packing and recommended viewport with initial viewpoint. Then, we discussed three different streaming approaches, which are under consideration for integration into OMAF and explained their integration in MPEG-DASH as well as described several security considerations. OMAF is well on the way to establishing itself as an international standard. Moreover, several industry forums as well as standardization bodies, are also offering many useful insights and information on the topics related to VR streaming. All the standardization and industry forum efforts together can be seen as a good prospect for the future. In the remaining work for OMAF, the different approaches are expected to be mapped to media and presentation profiles in order to provide full interoperability between content and players based on OMAF.

6 REFERENCES

[1] S. Gregory, “Watching the NBA in Virtual Reality Is Surprisingly Good”, 2016, TIME

[2] Coachella valley music and arts festival, https://www.coachella.com (08.05.2017)

[3] VR Industry Forum, http://www.vr-if.org (08.05.2017)

[4] The 3rd Generation Partnership Project, http://www.3gpp.org (08.05.2017)

[5] WebVR Community Group, https://www.w3.org/community/webvr (08.05.2017)

[6] ISO/IEC 23090-2, Information technology – Coded representation of immersive media (MPEG-I) – Part 2: Omnidirectional Media Format (OMAF) (to be published)

[7] R. G. Youvalari, A. Aminlou and M. M. Hannuksela, "Analysis of regional down-sampling methods for coding of omnidirectional video", 2016 Picture Coding Symposium (PCS), Nuremberg, Germany, 2016, pp. 1-5.

[8] M. Zhou, “AHG8: A study on compression efficiency of cube projection”, JVET-D0022 (http://phenix.it-sudparis.eu/jvet/)

[9] H.-C. Lin, J.-L. Lin, S.-K. Chang, and C.-C. Ju, "AHG8: Compact cube layout with tile partition," JVET-D0104 (http://phenix.it-sudparis.eu/jvet/)

[10] K. Kammachi-Sreedhar, A. Aminlou, M. M. Hannuksela and M. Gabbouj, "Viewport-Adaptive Encoding and Streaming of 360-Degree Video for Virtual Reality Applications", 2016 IEEE International Symposium on Multimedia (ISM), San Jose, CA.

[11] L. D'Acunto, J. van den Berg, E. Thomas, and O. Niamut, "Using MPEG DASH SRD for zoomable and navigable video", ACM MMSys 2016, New York, USA

[12] ISO/IEC 14496-15:2017, Information technology -- Coding of audio-visual objects -- Part 15: Carriage of network abstraction layer (NAL) unit structured video in the ISO base media file format

[13] ISO/IEC 23001-7 3rd Edition – Common encryption in ISO base media file format files

[14] C. Zhuang, Y. Lu, J. Li, Z. Wen, B. Cui, S. Liu, P. Lai, A. Abbas, W. Sun, “OMAF PROJ-VE: Segmented Sphere Projection (SSP) for 360 video”. M39782

[15] Y. Ye, E. Alshina, J. Boyce, “Algorithm descriptions of projection format conversion and video quality metrics in 360Lib”, JVET-F1003 (http://phenix.it-sudparis.eu/jvet/ ）