Adaptive Streaming of VR/360-degree Immersive Media Services with high QoE

Christian Timmerer[†,‡], Mario Graf[‡], and Christopher Mueller[‡]
[‡]Bitmovin Inc., [†]Alpen-Adria-Universität
San Francisco, CA, USA and Klagenfurt, Austria
[‡]{*firstname.lastname*}@bitmovin.com, [†]{*firstname.lastname*}@itec.aau.at

Abstract – *Real-time entertainment services deployed over the open, unmanaged Internet – streaming audio and video – account now for more than 70% of the evening traffic in North American fixed access networks and it is assumed that this figure will reach 80 percent by 2020. More and more such bandwidth hungry applications and services are pushing onto the market including immersive media services such as virtual reality and, specifically 360-degree videos. The dynamic adaptive streaming of VR/360-degree immersive media services with high Quality of Experience becomes an important issue and is subject to various research and development efforts. In this paper, we describe a adaptive streaming approach built on top of existing, standardized formats with minor extensions and perform comprehensive evaluations which can be used as guidelines for the further development of such services. Finally, we also provide a brief overview about ongoing standardization activities in this domain.*

Introduction and Background

Universal media access [1] as proposed in the late 90s, early 2000 is now reality. It is very easy to generate, distribute, share, and consume any media content, anywhere, anytime, and with/on any device. A major technical breakthrough and enabler was certainly the adaptive streaming over HTTP resulting in the standardization of MPEG-DASH [2], which is now successfully deployed in HTML5 environments thanks to corresponding media source extensions (MSE). Note that besides MPEG-DASH also Apples' HTTP Live Streaming (HLS) is an important format in the adaptive media streaming landscape but the recent MPEG format referred to as Common Media Application Format (CMAF) aims at harmonizing both approaches – but that's another story.

One of the next big things in adaptive media streaming is most likely related to virtual reality (VR) applications and, specifically, omnidirectional (360-degree) media streaming, which is currently built on top of the existing adaptive streaming ecosystems. These kinds of applications and services are commonly referred to as immersive media services. In this paper, we describe basic principles of adaptive streaming of immersive media services, encoding options available, and evaluations with respect to bitrate overhead, bandwidth requirements, and quality aspects. Before going into details, we will briefly highlight some background in this exciting area.

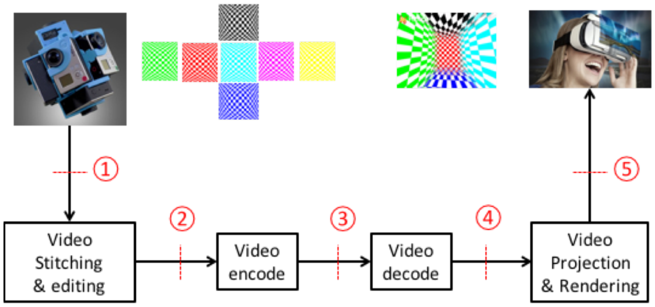

Figure 1. Basic System Architecture and Interfaces in an Immersive Media Ecosystem.

The basic system architecture including major interfaces of an immersive media ecosystem is shown in Figure 1. It starts with multiple videos being captured including various metadata ①, stitched together, and further edited before entering the encoding process ②. The encoding – typically a single video – considers projection and interactivity metadata and utilizes an appropriate storage and/or delivery format ③ before it will be decoded on the target device. After decoding – again typically a single video –, various projection and interactivity metadata ④ will guide the rendering process which interacts with the corresponding input/output technology ⑤.

In theory, various projection formats have been proposed (e.g., equirectangular, cube maps, pyramid maps, frustum maps, equal-area projection) while currently, in practice, mainly equirectangular projection is used. Equirectangular projection adopts a constant spacing of latitudes and longitudes which allows for simple, efficient processing but introduces horizontal stretching in the projected panorama near the poles. It is supported in most of the available content generation tools (i.e., camera hardware + stitching software) which explains its popularity.

As of today, no special techniques for coding in the spherical domain of the video exists and, thus, the video is projected to the rectangular domain. Therefore, state of the art video codecs (AVC, HEVC, VP8, VP9) and delivery formats (DASH/HLS) can be used to deploy a basic adaptive streaming service of immersive media content. However, this is very inefficient as the typical Field of View (FoV) of many VR devices is limited and a lot of content is delivered, decoded, and rendered for nothing (e.g., what is happening outside of the users' FoV). Viewport adaptive streaming has been introduced to overcome this limitation but requires multiple versions of the same content for each view. That is,

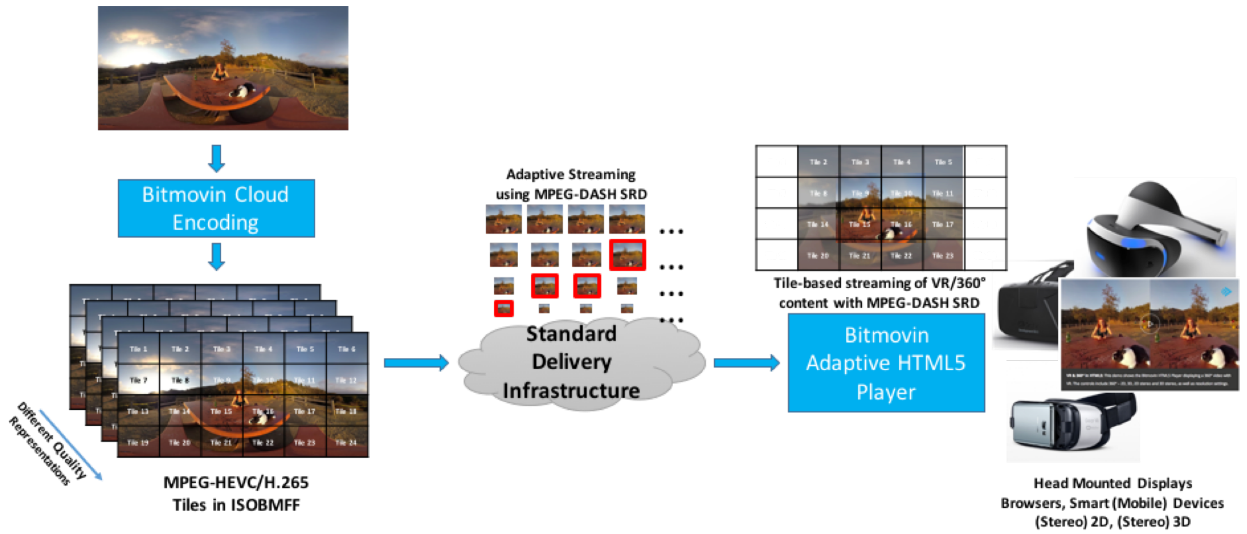
Bitmovin Cloud Encoding
Different Quality Representations
MPEG-HEVC/H.265
Tiles in ISOBMFF
Adaptive Streaming using MPEG-DASH SRD
Standard Delivery Infrastructure
Tile-based streaming of VR/360° content with MPEG-DASH SRD
Bitmovin Adaptive HTML5 Player
Head Mounted Displays
Browsers, Smart (Mobile) Devices
(Stereo) 2D, (Stereo) 3D

FIGURE 2. EXAMPLE WORKFLOW OF ADAPTIVE STREAMING OF IMMERSIVE MEDIA SERVICES.

it adopts a similar strategy as in adaptive media streaming (DASH/HLS) but the number of versions of the same content significantly increases which impacts (cloud) storage and (content delivery) network costs. A novel approach in this domain utilizes the concept of video tiles – as part of the HEVC standard [3] – and the Spatial Relationship Descriptor (SRD) of the MPEG-DASH standard [4] to enable efficient adaptive streaming of immersive media services.

ADAPTIVE STREAMING OF IMMERSIVE MEDIA SERVICES

In this section, we describe the basic concepts of adaptive streaming of immersive media services utilizing HEVC tiles and MPEG-DASH SRD. An example workflow is shown in Figure 2 comprising an encoding service which produces tile-based adaptive video content and an adaptive player utilizing DASH/HLS for the actual streaming.

The equirectangular video is encoded in (*a*) multiple rectangular tiles using MPEG-HEVC/H.265 – each tile is independent from other tiles – and (*b*) different quality representations (e.g., bitrate, resolution) which are stored in fragmented ISO base media file format (ISOBMFF) containers (i.e., fMP4) according to the new CMAF standard (compatible with both DASH and HLS). A manifest (e.g., MPD with SRD + extensions) describing the temporal and spatial relationships of the available media content is provided to the adaptive player which requests tiles and quality representations based on the given context. The main context factors include the actual client device (e.g., desktop browser, mobile apps, head-mounted displays), network capabilities and conditions (e.g., available bandwidth), and the users' field of view. The adaptation logic inside the adaptive streaming player requests media segments pertaining to both tiles and quality representations while maximizing the Quality of Experience (QoE) under the given context conditions. In addition to the adaptation logic, the player should multiplex (or rewrite) the various HEVC tiles into a single HEVC bitstream compatible with the clients' hardware HEVC decoder. Note that this step is required as most devices currently support only one HEVC hardware-enabled decoding instance to be used simultaneously which might become obsolete in the (near) future.

The adaptation logic is certainly the core of the adaptive streaming player and consequently responsible for delivering an excellent QoE. In principle two approaches can be envisaged which are referred to as full and partial delivery both considering the users' FoV. The former – full delivery – streams the entire rectangular video content but with higher quality for tiles within the FoV and lower quality for tiles outside the FoV. The latter – partial delivery – streams only the tiles within the FoV while those outside the FoV remain on the server. Partial delivery allows for higher quality representations than full delivery but may suffer additional delay in case of user interactivity (e.g., user turns device or head to a section of the video where the corresponding tiles are outside of the current FoV).

The exact encoding and streaming configuration/policy depends on many factors which are still subject to research but in this paper, we would like to highlight some basic evaluation results. An important aspect in the evaluation of such systems is the user and quality of experience. Adopting existing metrics known from traditional video coding such as PSNR or SSIM may not work for the reasons given as follows. For example, let's consider PSNR (due to its simplicity) which cannot be adopted – in the rectangular domain – as obviously different projection formats will delivery completely different PSNR values when compared with each other. Thus, something like a PSNR in the spherical domain is needed to exclude projection format issues. Furthermore, novel streaming approaches like the one described in this paper cannot be compared in the entire spherical domain as the quality depends on the viewport. Therefore, a metric called V-PSNR [5] has been introduced which will be used for basic quality evaluation of the proposed tile-based adaptive streaming approach.

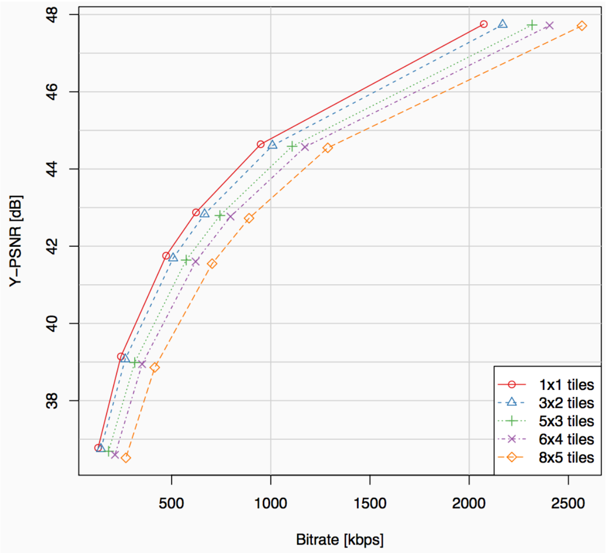

FIGURE 3. TILE OVERHEAD FOR RESOLUTION 1920X960.

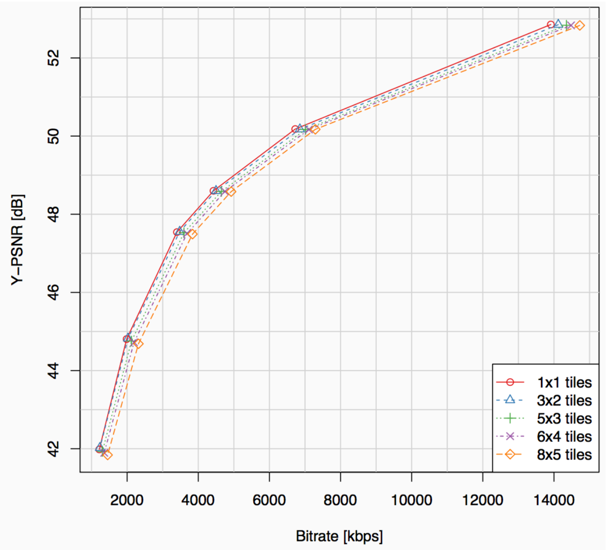

FIGURE 4. TILE OVERHEAD FOR RESOLUTION 7680X3840.

EVALUATION RESULTS

We evaluated our tile-based adaptive streaming approach using different content configurations and we performed a series of experiments to determine bitrate overhead, bandwidth requirements, and quality using V-PSNR.

I. Dataset

The dataset used for the evaluation comprises video sequences of different genres and each video sequence is configured (encoding and packaging) as follows:

- **Segment length / GOP size**: it is assumed that the adaptation can be done at segment boundaries. If users change the viewing direction quickly, shorter segments allow for faster adaptation of the new viewport but lead to decreased coding efficiency. For tiled content, we used a segment length of 1s and for non-tiled content we used 1, 2, and 4 seconds (both at 25fps).
- **Tiling pattern**: smaller tiles enable better matching with the actual viewport but decrease coding efficiency whereas larger tiles increase coding efficiency but does not match the requested viewport effectively. Therefore, we have selected various tiling patterns comprising 1x1, 3x2, 5x3, 6x4, and 8x5 tiles, i.e., from one tile (no tiling at all) to 40 tiles in the most advanced configuration.
- **Resolution**: the spatial resolution comprises 1920x960, 3840x1920, and 7680x3840.
- **Projection format**: equirectangular.
- **Quantization parameter**: 22, 27, 32, 37, 42.

The used configuration above has been applied on the given video sequences and resulted in 210 different video configurations used in the subsequent evaluation. Additionally, for each sequence we recorded three viewing directions from three different people using a head-mounted display. These head motions were used to generate videos based on the proposed adaptation behavior and used to evaluate our tile-based adaptive streaming based on V-PSNR.

II. Bitrate Overhead

The goal of this evaluation was to analyze the coding overhead introduced by using HEVC tiles. Therefore, the content has been encoded according to the tiling pattern defined in the previous section at different resolutions and quantization parameters. The rate-distortion curves for the resolutions 1920x960 and 7680x3840 are shown in Figure 3 and Figure 4 respectively. It clearly shows that, the more tiles are used, the higher the bitrate at the same PSNR value. Please note that Figure 4 has a different scale and, thus, the difference looks smaller than it is. However, based on these results it is difficult to identify the most appropriate tiling pattern and, thus, a more detailed analysis with respect to the bandwidth requirements (adopting different adaptation logics) has been conducted which is presented in the next section.

III. Bandwidth Requirements

In order to determine the bandwidth requirements, we conducted several experiments for different tiling patterns and adaptation logics for both traditional streaming (i.e., monolithic tile configuration with 1s, 2s, and 4s segment lengths) and tile-based streaming (i.e., full – all tiles considered equally; only viewport – partial delivery; viewport adaptive – full delivery). In this paper, we show results for 3x2 and 6x4 tiles as these configurations seem to have the most promising tradeoff in terms of overhead.

The results are shown in Figure 5 and Figure 6 respectively. Obviously, the “tiles full” configuration has the highest bandwidth requirements as it adopts a traditional adaptation logic which does not exploit the tile feature and,

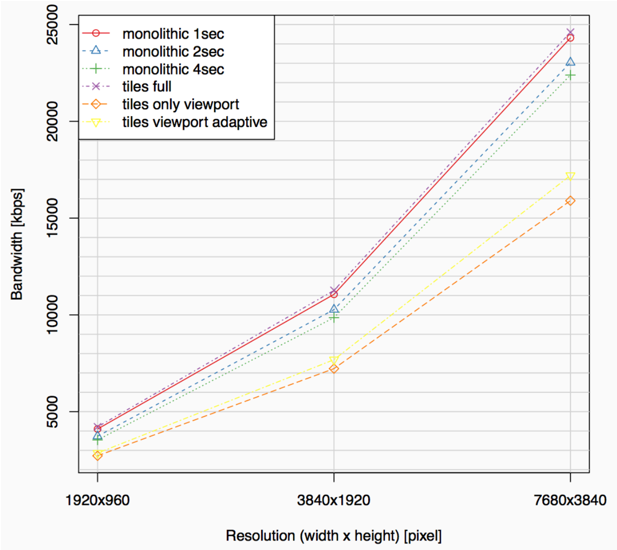

FIGURE 5. BANDWIDTH REQUIREMENTS, 3X2 TILES.

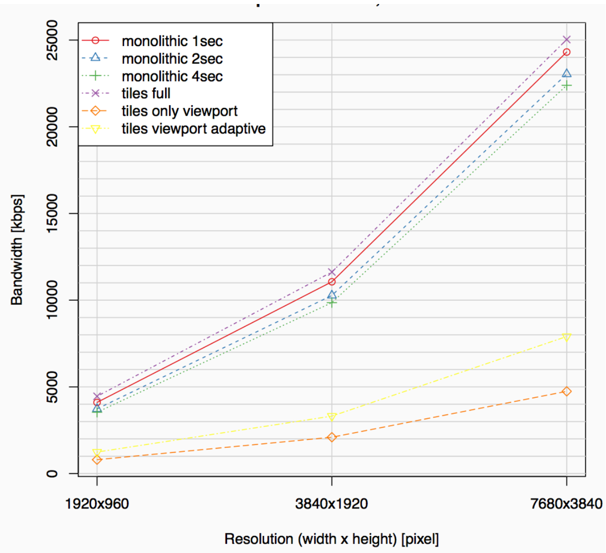

FIGURE 6. BANDWIDTH REQUIREMENTS, 6X4 TILES.

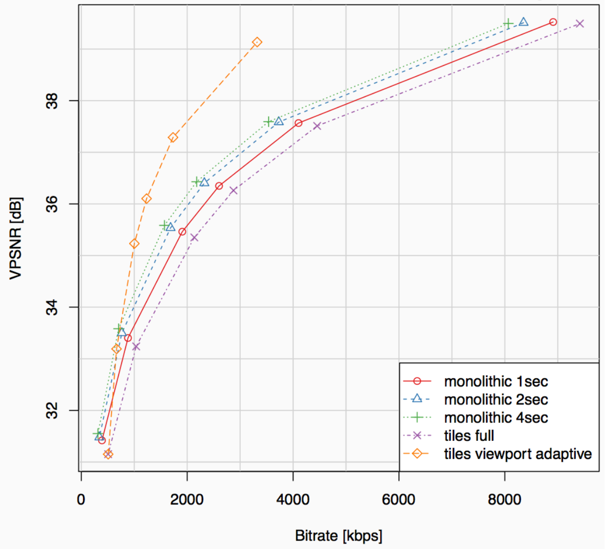

FIGURE 7. V-PSNR, 6X4 TILES, RESOLUTION: 1920X960.

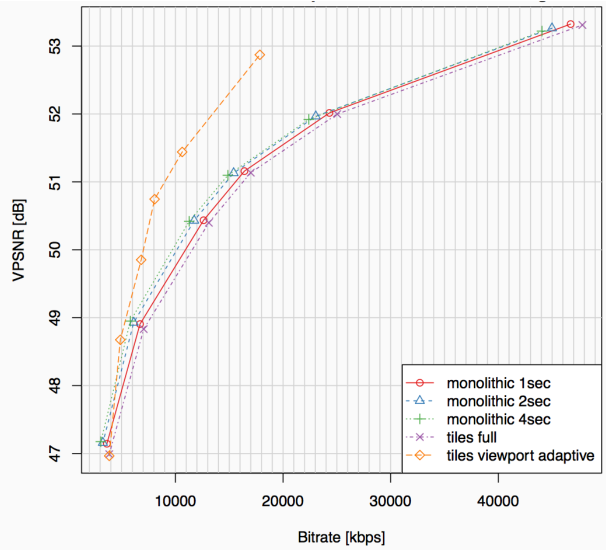

FIGURE 8. V-PSNR, 6X4 TILES, RESOLUTION: 7680X3840.

thus, requires slightly more bandwidth compared to "monolithic" using different segment lengths – due to the tile overhead. The "tiles only viewport" and "tiles viewport adaptive" have much lower bandwidth requirements than the other configurations thanks to the exploitation of the tile feature during streaming. The "tiles only viewport" has the lowest bandwidth requirement as only tiles belonging to the current viewport are streamed to the client while all other tiles remain on the server. Interestingly, the results clearly reveal that the 6x4 tile configuration has lower bandwidth requirements and, thus, is the winner configuration in terms of both tile overhead and bandwidth requirements.

IV. Quality: Viewport PSNR

The last experiment evaluates the quality of the proposed tile-based adaptive streaming approach based on V-PSNR. Therefore, we present the results of the 6x4 tile pattern at resolutions 1920x960 and 7680x3840 in Figure 7 and Figure 8 respectively. The results clearly show that the viewport adaptive tile-based streaming approach achieves superior results compared to the other approaches. The configuration "tiles only viewport" is not shown here as we assume that it will not be used in practice.

Finally, note that results for the resolution 3840x1920 are not shown in detail here as they are somewhere in between the other two resolutions and do not reveal any new findings.

OVERVIEW OF MPEG STANDARDIZATION ACTIVITIES AND OPEN ISSUES

The proposed tile-based adaptive streaming approach for immersive media services utilizes standardized formats such as MPEG-HEVC/H.265 and MPEG-DASH SRD. However, some aspects are not yet covered by current standards, i.e., basically most of the metadata identified in the introduction (e.g., projection formats, interactivity metadata) as well as coding tools providing the necessary coding efficiency for the identified use cases. Therefore, MPEG started a new work item related to immersive media officially referred to as ISO/IEC 23090 (unofficially called MPEG-I although this is not yet fully confirmed and might change in the future) which – at the time of writing of this paper – foresees five parts.

The first part will be a technical report describing the scope of this new standard and a set of use cases and applications from which actual requirements can be derived. Technical reports are usually publicly available for free. The second part specifies the omnidirectional media application format (OMAF) addressing the urgent need of the industry for a standard is this area. This part will include also aspects discussed in this paper and, thus, will solve the interoperability problem required for tile-based adaptive streaming. Part three will address immersive video and part four defines immersive audio. Finally, part four will contain a specification for point cloud compression. OMAF is part of a first phase of standards related to immersive media and should finally become available by the end of 2017, beginning of 2018 while the other parts are scheduled at a later stage around 2020. Further information will become available at http://mpeg.chiariglione.org/ in due time.

An important open issue within the application domain is DRM protected content, specifically in Web environments using HTML5, MSE, and EME. Traditional adaptive media services are being massively deployed using HTML5 players thanks to MSE and encrypted content is supported thanks to EME. For VR/360-degree video content WebVR is used which requires access to the fully decrypted and decoded video frames. However, once the adaptive streaming player – typically implemented in Javascript – pushes segments into the EME it does not offer any interface to the decrypted and decoded frames as this would apparently break the trusted environment for DRM protected content. Therefore, an interface between MSE/EME and WebVR is needed – implemented within the browser exposing appropriate interfaces to the application – to enable the adaptive streaming of DRM protected VR/360-degree immersive media services.

CONCLUSIONS

In this paper, we have presented an approach for the adaptive streaming of VR/360-degree immersive media services enabling high QoE. We have described the basic system architecture (including interfaces) and an example workflow for the generation, adaptive streaming, and consumption of immersive media services adopting existing standard formats and extensions thereof. We have shown that it is possible to implement an efficient streaming mechanism using HEVC tiles and DASH spatial relationship descriptor including minor extensions for describing the projection format and other interactivity metadata.

We have evaluated our proposed approach using a comprehensive dataset where video sequences have been encoded into more than 200 different configurations which have been evaluated in terms of bitrate overhead (introduced by tiles), bandwidth requirements (for the actual streaming using various adaptation strategies), and quality using viewport PSNR (V-PSNR). The results have been presented which can be used to provide encoding and streaming guidelines for such immersive media services. Finally, we provided an overview of MPEG standardization activities in this area and discussed open issues.

Future work will include further optimizations of the proposed approach, contribution to ongoing standardization activities, and a thorough subjective quality assessment to better understand the Quality of Experience of such immersive media applications and services.

REFERENCES

[1] Mohan, Rakesh, Smith, John R., Li, Chung-Sheng, "Adapting Multimedia Internet Content for Universal Access," IEEE Transactions on Multimedia, vol. 1, no. 1, 1999, pp. 104-114.

[2] Sodagar, Iraj, " The MPEG-DASH Standard for Multimedia Streaming Over the Internet," IEEE Multimedia, vol. 18, no. 4, 2011, pp. 62-67.

[3] Sullivan, Gary J, Ohm, Jens R., Han W. J., and Wiegand, Thomas "Overview of the High Efficiency Video Coding (HEVC) Standard," IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, no. 12, 2012, pp. 1649-1668.

[4] Niamut, Omar A., Thomas, Emmanuel, D'Acunto, Lucia, Concolato, Cyril, Denoual, Franck, and Lim, Seong Yong, " MPEG DASH SRD: Spatial Relationship Description,". In Proceedings of the 7th International Conference on Multimedia Systems (MMSys '16), Klagenfurt, Austria, 2016.

[5] Yu, M, Lakshman, H. and Girod, B., "A framework to evaluate omnidirectional video coding schemes," 2015 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Fukuoka, Japan, 2015, pp 31-36.