Visual Attention-Aware Omnidirectional Video Streaming Using Optimal Tiles for Virtual Reality

Cagri Ozcinar*, Julián Cabrera†, and Aljosa Smolic*

***Abstract*—Owing to its interactive look around nature and very large resolution requirement, providing immersive *omnidirectional video* (ODV) streaming experiences in *virtual reality* (VR) applications demands cost-effective solutions to meet both the content delivery network and device constraints. In this paper, we introduce an adaptive omnidirectional video (ODV) streaming pipeline that optimizes DASH representations of ODV content considering their visual attention (VA) maps. The main contribution of this paper is the use of VA maps: (*i*) to compute a novel objective quality metric that captures the fact that not all of the ODV is actually watched by users: the visual attention spherical weighted (VASW)-based objective quality measurement, (*ii*) to define optimal tile representations of the ODV frames, namely tiling schemes, which are composed of variable-sized and non-overlapping tiles, and (*iii*) to efficiently distribute a given bitrate budget among the set of tiles within a tiling scheme for an ODV. We evaluate the proposed system performance with varying bandwidth conditions and the tracked head orientations from user experiments. Results show that the proposed system significantly outperforms the existing non-tiled and tile-based streaming solutions.**

***Index Terms*—omnidirectional video, visual attention, tiles, adaptive streaming, virtual reality.**

I. Introduction

Tremendous activity can be observed in the video industry these days in terms of offering immersive *virtual reality* (VR) video experiences using *omnidirectional video* (ODV), also known as 360° video. The recent trial of live ODV streaming at the 2018 World Cup tournament [1] by BBC is a decisive proof of relevance. This emerging video representation is captured typically by multiple cameras which cover 360° of a scene and rendered through head-mounted displays (HMDs) which allow viewers to look around a scene from a central point of view in VR, resulting in a more immersive and interactive visual experience than that available through traditional 2D video.

Delivery of ODV at a perceptually acceptable quality level, however, is a challenging task due to limitations in processing and delivery pipelines, as well as constraints imposed by the available end-user devices. ODV is typically produced and stored in a planar representation, *e.g.,* an equirectangular projection (ERP), in order for it to remain compatible with existing video delivery pipelines, and then it is projected back onto a spherical surface when rendered. However, existing HMDs have a viewable field of view (FOV) and use only a fraction of the given content at a given time, namely the *viewport*. Transmission of ultra-high resolution ERP, *e.g.,* $\geq$ 8K, is, therefore, needed to obtain a decent VR video quality level. In this context, transmission of a current region-of-interest [2]–[4] or viewport-dependent streaming solutions [5]–[10] can be utilized to enhance the perceived VR video quality. However, in video transmission scenarios with delay-prone communication pipelines [11] and rapid head orientation activities [12], such solutions are merely inefficient in terms of complying with the motion-to-photon latency requirement; thus penalizing the quality of experience (QoE).

Given its look-around viewing nature and very large resolution requirement, transmission of ODV demands cost-effective compression and adaptive streaming solutions to meet network constraints. MPEG-dynamic adaptive streaming over HTTP (DASH) [13], for example, can be used to provide seamless video streaming experiences through networks. In DASH, each given video has a set of DASH representations which contain its different bitrate levels. Each DASH representation, which has its bitrate level, consists of multiple self-decodable time segments, namely chunks, which can be requested individually and decoded by DASH players. In this context, in order to reduce both the bitrate consumption of the end-user and the visual distortion of the viewport, as well as to improve the bitstream decoding performance using parallel decoding features, ODV frames in each chunk can also be divided into self-decodable spatial regions [14], namely, *tiles*. To this end, the spatial representation description (SRD) feature in DASH [13], [15] is used to provide the necessary signaling to transmit each tile of a given content and reconstruct the full 360° of the scene for VR.

Although tile-based encoding brings benefits to ODV streaming pipeline, the selection of the tiling scheme impacts its compression performance. The tile-based encoding technique introduces several opportunities for ODV content [3], [4], [14]–[16]. For example, cost-effective video coding [6], [17], [18], partial decoding [15], [19]–[21], parallel decoding [14], [21], and utilizing *visual attention* (VA) maps, which describe how the users consume a given video at a given time [12], [22], in video streaming can be made possible using tiles. However, the selection of the tiling scheme, which represents the spatial partitioning structure [21] containing a set of non-overlapping tiles, impacts the compression efficiency. More clearly, usage of larger resolution tiles (*i.e.,* smaller number of tiles) can increase the coding gain for some content at some bitrates by exploiting a large number of redundant pixels, but they provide less flexibility in terms of exploiting

C. Ozcinar and A. Smolic are with V-SENSE, School of Computer Science and Statistics, Trinity College Dublin, Ireland. Emails: ozcinarc@scss.tcd.ie and smolica@scss.tcd.ie

J. Cabrera is with Grupo de Tratamiento de Imágenes, Information Processing and Telecommunications Center and ETSI Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain. Email: julian.cabrera@gti.ssr.upm.es

Manuscript received DD, MM, 2018; revised DD, MM, 2018.

the redundant pixels outside of the viewport region [7]. In contrast, using small tiles may decrease the coding efficiency by exploiting fewer spatial redundancies. It is, therefore, necessary to find an optimal tiling scheme based on viewport trajectories (*e.g.,* VA) for a given ODV; thus, a smart delivery strategy can save network bandwidth and improve the overall QoE performance for VR video applications.

This paper introduces an adaptive ODV streaming system design which determines an optimal tiling scheme of a given content for each chunk, and the required encoding bitrate level for each tile of the optimal tiling scheme using a VA-based probabilistic model. First, we conducted formal subjective viewing sessions using an HMD to estimate this model by collecting viewport trajectories of ODVs. This model is used to allocate a given bitrate within the tiles. Second, we propose a VA-driven spherical weighted objective quality metric, namely VASW-PSNR, to determine the optimal tiling scheme for each chunk of a DASH representation. Our work extends our short conference paper [23] by providing a new formulation for the developed VA-driven quality metric, using variable-sized tiles, and a comprehensive performance assessment with a broad range of 8K resolution ODVs with different complexities. Our proposed design does not require any modification of the existing DASH players and is entirely transparent to them. As such, we expect that our work will provide helpful input for the streaming industry in terms of considering variable-sized tiles per chunk and non-uniform bitrate distribution based on VA for transmitting the ODV content. To verify our proposed solution, we recorded eight participants' viewport trajectories in disjointed viewing sessions using an HMD and compared our proposed method with reference adaptive streaming solutions, which are based on naive tiling and non-tiling schemes. Experimental results show that the proposed optimization framework for adaptive ODV streaming demonstrates quality enhancements compared with the reference streaming solutions.

The remainder of this paper is organized as follows. A brief overview of the related work is detailed in Section II. Following that, the proposed streaming system model is described in Section III. Experimental results are provided in Section IV. Finally, the conclusion of the paper is provided in Section V.

II. Related Work

Several related studies are available for compression and streaming for ODV on one side, objective quality assessment techniques for ODV on the other side. In the following, we outline only the ones that are most related to our work.

A. Studies related to compression

ODV compression algorithms typically work by considering a tile-based representation of a given content that offers several opportunities for VR applications. Cost-effective video coding [17], [24], [25] and partial decoding [15], [20], for example, were made possible using tiles. Ozcinar *et al.* [17] estimated an optimal set of encoding parameters (*e.g.,* set of quantization and resolution pairs) for tile-based ODV streaming. Results showed that the selection of an optimal set of encoding parameters achieved significant bitrate savings, as compared with the use of reference solutions. Similarly, Xiao *et al.* [24] estimated optimal tiling schemes, wherein the estimation problem was formulated as a trade off between the costs of storing a set of tiles and the costs of serving sets of tiles that cover possible views of the chunk. Differently from our work, the usage of visual attention-aware bitrate allocation, determination of a set of optimal tiles were not part of their considerations to improve the QoE. An optimal spatio-temporal smoothness approach was also developed in [26] for tile-based streaming. With this system, the clients can request the best set of tiles according to the location of the viewport. The optimal bitrate for each tile is determined in a way to minimize spatial and temporal quality variations. However, the formulated optimization problem is based on the traditional mean square error which does not take the spherical distortion of the ODV representation into account. By contrast, our proposed system relies on the server side and our formulation considers the expected viewport quality distortion by integrating the geometrical and compression distortions.

Transmission of very high resolution ODV content not only consumes excessive network bandwidth but also requires high computational power at client locations. For this purpose, regional down-sampling was studied [25] to increase the compression gain and reduce the computational complexity of the codec. A regional down-sampling algorithm for the temporal domain was also proposed, where the intra- and inter-frames were encoded in full resolution and a regionally down-sampled format, respectively. A recent study [27] proposed a new packing arrangement for the transmission of ODV in order to provide higher-resolution video experiences. The packing arrangement guaranteed that the content within the viewer's FOV originated from 6K resolution content, while the remaining parts were covered with the content at a lower resolution.

Distortion of spherical geometry can be taken into account in the video encoding process to achieve ODV compression gain. To this end, Liu *et al.* [28] proposed a rate-control mechanism for the high efficient video coding (HEVC) standard [29] based on the spherical PSNR (S-PSNR) metric [30], which minimized the spherical distortion at a given target bitrate. Experimental results showed compression gain compared with the standard HEVC rate-control mechanism in terms of S-PSNR at identical bitrates. Similarly, a new rate-distortion optimization was introduced [31] that improves ODV coding efficiency by considering the spherical distortion within the rate-distortion optimization process of HEVC.

B. Studies related to streaming

Viewport-aware techniques in adaptive streaming systems are ideal solutions for improving the quality of the VR video experience. Early studies [3], [4], [16], [32] used a previous progressive streaming technology that can be supported by region-of-interest-coding solutions to enhance the perceived ODV quality. Although these approaches were far from achieving high-quality performance, because of the limitations of the underlying technologies, each study represents pioneering work in this area. A more recent study [7] utilized DASH-based streaming technology to transmit tiled ODVs. A viewport-aware bitrate distribution was introduced wherein an optimal bitrate level was selected for each fixed-sized tile, based on a spherical distance criterion. Similarly, two differently encoded versions of the same content were delivered to reduce the transmission rate of a given ODV using DASH [6]. The tiles, which were overlapped with the current viewport, were delivered in high resolution while the rest of the tiles were transmitted in low resolution. Furthermore, several versions of DASH representations were generated for different viewport positions [8], where the opposite areas of the defined viewport were set to uniform black to increase the compression gain. Hosseini *et al.,* [9] assigned a high-quality level to the tiles within the current viewport, which the user was most likely to use, while a low-quality level was presented outside of a user's immediate viewport tiles. Similar to the previous work, Corbillon *et al.,* [10] proposed a viewport-adaptive video delivery system, wherein DASH representations that differ by bitrate and scene region were used. Furthermore, a recent practical study [2] examined several DASH strategies and evaluated bitrate overhead and quality requirements. However, none of those mentioned above studies considered flexible-sized tiling strategy, per chunk bitrate optimization, and VA, which are the major contributions of this work. The importance of using VA [33] and benefits of chunk-based bitrate optimization at the server side [34] are also emphasized by professional streaming service providers.

Prefetching of ODV tiles is an efficient solution for saving network bandwidth, but requiring very low transmission latency for VR applications [35]–[38]. For instance, the total expected distortion of the prefetched tiles was minimized by using a developed probabilistic model [35]. Additionally, Petrangeli *et al.* developed an algorithm [36] to predict the future viewport position and to minimize quality transitions during viewport changes. End-to-end ODV streaming latency was also minimized by using a server push mechanism in DASH. Furthermore, fetching only the pixels visible to the current viewport was investigated in preliminary studies [37], [38]. Although the proposed strategies were ideal for saving of network bandwidth, extremely low end-to-end latency was required [2], [32], [39] to predict accurate future viewport positions for each client and content. Such low latency requirement is not the always the case in existing networks and VR devices.

MPEG-I group addressed the needs of the storage and transmission of ODV by developing the media representation for ODV, called omnidirectional media format (OMAF) [40]. This format mainly enables the consumption of ODV by considering coding, encapsulation, and presentation for adaptive streaming. At this stage, OMAF specified two projection formats, namely ERP and cubemap projection. The system implementation of the OMAF targets the distribution of video and audio signals from the capturing side to rendering and presentation on the client side [41].

Further studies investigated additional aspects of VR video streaming, such as an efficient ODV content preparation strategy [42] and a cost-optimal downloading method [43] for ODV streaming. In order to consume less bandwidth while maintaining the user's experience, Dambra *et al.* [42] improved the streaming of ODV by editing the film in such a way as to limit the client's motion. The content editing strategy and streaming module were integrated within the MPEG DASH-SRD player. Rossi *et al.* proposed a navigation-aware transmission strategy in [43]. Although we share the similar problem formulation in the sense of modeling the distortion of ODV, both algorithms have different perspectives. The optimization logic in [43] is at the client side, providing user-centric optimization. In their work, the client has to solve an optimization problem to find out the optimal bitrate for each tile. It is more a theoretical work which contains several approximations like infinite playback buffers or exact channel estimation. Differently, our bitrate optimization is at the server side of the streaming pipeline, providing a global solution for all the clients in the network and able to work in a real scenario.

C. Studies related to quality assessment

Most quality assessments take its spherical characteristics into account to evaluate the quality of a given VR content. Yu *et al.* considered the problem of evaluating the coding efficiency of omnidirectional content and subsequently developed the S-PSNR metric [30]. The planar projected content was mapped onto the spherical representation in order to compute the observed distortion of a given VR content. However, pixel mapping from the planar to the spherical representation requires pixel interpolation to obtain values in specific positions. Craster parabolic projection (CPP-PSNR) was also introduced [44], which computed the distortion between the reference and impaired content mapped by the projection. More recently, Sun *et al.* introduced weighted spherical - mean square error (WS-MSE) and WS-PSNR [45]. As the sphere-to-plane projection techniques distort the spherical representation, the distortion between a given reference and impaired content is calculated using the weights according to the pixel position on the spherical surface. Each weight is computed from the projected 2D plane (*e.g.,* ERP) by considering the impact of stretching distortion for each pixel. However, none of the described metrics elaborately measured the perceived quality of a given ODV, nor considered the look-around nature of the ODV viewing experience. For this, we extended the concept of the WS-based objective quality metric, and introduced a new VA-weighted objective quality metric for ODV.

Considering that users can only see one part of the scene at a time in HMDs, visual attention and perception have become fundamental research topics for measuring the quality of omnidirectional content [12], [22], [46]–[49]. The perceived quality impact of a given omnidirectional image content in peripheral vision, for instance, was studied in one investigation [50], in which various compression levels and spatial resolutions were evaluated. The quality at the central region of the viewport was fixed, and the quality of near and far peripheral regions were degraded until the subject noticed the distortion. However, this work only considered the region inside of the viewport by changing the quality levels. More recently, Xu *et al.,* [47] addressed quality assessment required in ODV compression and ultimately developed two objective quality assessment methods. One of the proposed methods weighted the distortion of pixels with regard to distances from the center. The other proposed method predicted viewing directions. However, distortion was measured on planar projected frames, and the non-uniform sampling density on the sphere was not considered in the study. Besides, the computational saliency prediction algorithm used, which was proposed for traditional 2D video, might not provide an accurate prediction performance for ODV [12].

III. Proposed System Model

We consider an adaptive streaming pipeline to deliver very high-resolution of ODVs to VR clients over the Internet, as illustrated in Fig. 1. The proposed system enables each client to navigate through a delivered ODV and creates an immersive VR video experience using a given HMD. Our proposed design consists of three nodes: source, media platform, and delivery. The source node is responsible for capturing and post-processing (*e.g.,* sphere-to-plane mapping). The media platform estimates for each DASH representation, an optimal pair of tiling scheme and a set of tiling bitrates for each c chunk according to the target bitrate of the DASH representation and the VA map.

To achieve this objective, the media platform handles the tiling, estimation of VA, bitrate allocation and encoding, optimization of the tiling scheme, packing (*i.e.,* encapsulation) and transferring the prepared video data to the content delivery network (CDN). CDN is a cloud-based video streaming system which delivers ODVs to the edge servers in such a way as to connect effectively with the end users. Finally, in the delivery node, each end user requests and receives an appropriate DASH representation containing a set of tiles.

The source node of the proposed system captures each ODV v and maps it onto a 2D planar plane (*i.e.,* rectilinear) using the projection techniques (*e.g.,* ERP) to keep it compatible with the existing video pipelines. While the proposed system is compatible with other projection techniques (*e.g.,* cubemap), in the following we consider the ERP as the input ODV format, which is the most widely deployed ODV representation [51]. ERP contains full panoramic 360° horizontal and 180° vertical views of the captured scene.

The media platform divides a given content into a predefined set of tiles, $\mathcal{T}$, of different sizes. Then, each tile, t, is encoded at different predefined bitrates $\{r_i^t\}$ and segmented into chunks of a predefined duration. From $\mathcal{T}$, several different tiling schemes can be built for each chunk. A tiling scheme for chunk k, s_k, consists of a subset of non-overlapping tiles that reconstruct the full 360° scene.

The target bitrate r of a DASH representation is then allocated within each tile $t \in s_k$ based on the proposed VA-based bitrate distribution method, and considering the set of predefined encoding bitrates available for each tile $\{r_i^t\}$. Here, the objective is to reinforce the quality of those parts of the ODV which are more likely to be seen, maximizing the expected quality of the actual content watched by the user. For this objective, a VA map is computed using the recorded viewport trajectories for each chunk, and represents the probability that a set of users watches a tile within the ERP representation.

An optimal tiling scheme s_k^* is then determined for each chunk according to the proposed VASW-based objective quality metric, which considers the VA map and spherical representation of a given content. The selection of the optimal tiling scheme is carried out on a chunk basis, and the selected tiles are packed and stored on servers building up each optimized DASH representation. Hence, an optimal set of tiles, which is composed of non-overlaping and variable-sized (or fixed-sized) tiles which have optimal bitrate levels for a given content, is delivered to the clients when they make their HTTP requests for any DASH representation.

A. Optimization of tiling scheme

Let $\mathcal{S}_k$ be the set of all possible tiling schemes that can be built upon the set of tiles $\mathcal{T}$ on chunk k stored in the server.

To determine the optimal tiling scheme $s_k^* \in \mathcal{S}_k$ for allocating the target bitrate r within a set of tiles, an exhaustive search is applied on $\mathcal{S}_k$. This exhaustive search aims to find the tiling scheme that incurs in the lowest distortion for the considered chunk according to the VASW-based objective quality metric. For that purpose, the exhaustive search algorithm performs the following steps:

1) For each tiling scheme $s_k^i \in \mathcal{S}_k$:
   a) Perform the bitrate allocation among the tiles that conform s_k^i according to the proposed tile bitrate allocation algorithm based on VA map data.
   b) Compute the distortion of the chunk, $D_{chunk}(s_k^i)$, according to the proposed VASW-based objective quality metric.
2) Select the tiling scheme that incurs in the minimum distortion, s_k^*.

Hence, once the bitrate allocation algorithm has determined the target bitrate each tile in chunk s_k^i, its distortion can be calculated as in Eq. (1):

$$D_{chunk}(s_k^i) = \sum_{f=1}^{F} \overline{D}_{V_p}(f), \tag{1}$$

where F is the number of frames in chunk s_k^i and $\overline{D}_{V_p}(f)$ represents the VASW-based quality measurement of frame f within the chunk.

Finally, s_k^* is calculated as in Eq. (2):

$$s_k^* = \underset{s_k^i \in \mathcal{S}_k}{\arg\min}\, D_{chunk}(s_k^i) \tag{2}$$

To achieve our objective in Eq. (2), the proposed optimization formulation requires inputs from main components: *i*) estimation of VA, *ii*) VASW-based objective quality measurement, and *iii*) tiling and bitrate allocation. These algorithms are described in the following sections:

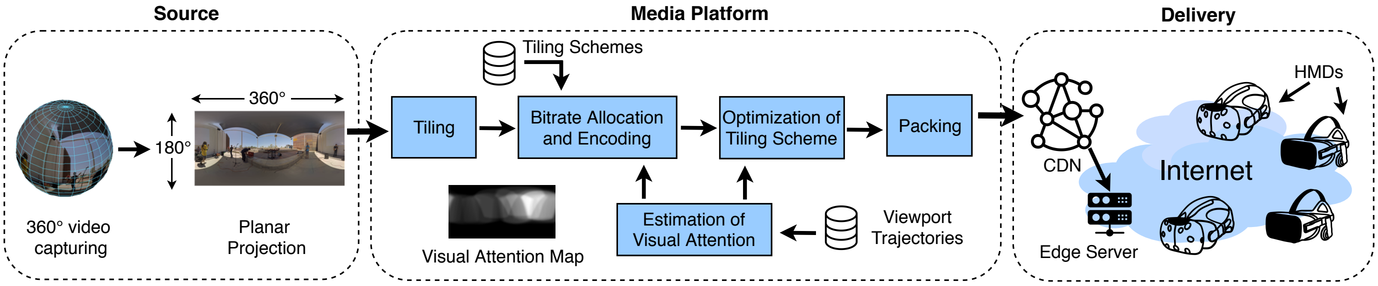


Fig. 1: Schematic diagram of the proposed adaptive ODV streaming pipeline for VR.

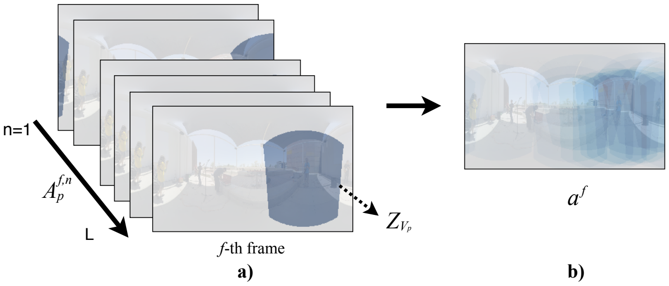


Fig. 2: Estimation of viewport-based visual attention (VA) map using the collected viewport trajectories: *a*) viewport mask for the f-th frame of the n-the client, *b*) VA probability for the f-th frame.

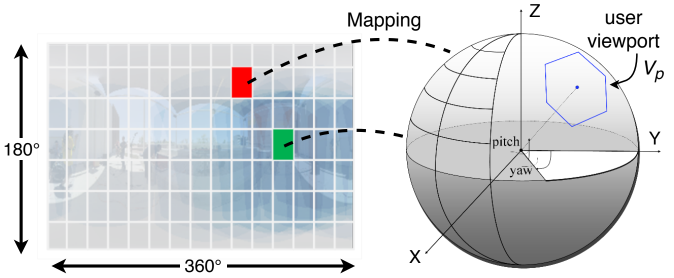


Fig. 3: Illustration of a mapping from a planar ERP to the spherical surface [45].

B. Estimation of visual attention

To estimate VA, we collected a set of viewport trajectories from participants following the previous work done by [12], in which a limited number of tracked viewport trajectories were collected using only selected ODVs with limited number of participants. Our extended dataset contains viewport trajectories from 25 participants watching 17 ODVs. This dataset is available at our project page[1].

Given a set of tracked viewport trajectories, a viewport-based VA map is estimated for each ODV frame. In this work, each VA map serves as a bi-dimensional histogram for the pixel locations of its corresponding ODV frame, and its values represent the number of times that clients have paid attention to the analogous pixels in the frame. For a given pixel position, the higher the VA map value is, the more times the pixel at that position in the frame has been watched. Fig. 2 illustrates the estimation of viewport-based VA using a set of viewport trajectories from participants.

The computations for the VA maps require the estimation of the users' viewport position. By using each recorded viewport trajectory (*azimuth*, *elevation*, and *roll*), which is available for each frame, the viewport's corresponding pixel area is estimated on a planar surface. In order to accomplish this task, a mask is used wherein the pixels within the viewport are assigned the value one, and the pixels outside of the viewport are assigned the value zero.

The VA map for a given f-th frame, V_A^f, averages the contributions of the L users considered for that frame as defined in Eq. (3):

$$V_A^f = \sum_{n=1}^{L} A_p^{f,n}, \tag{3}$$

where $A_p^{f,n}$ is the viewport mask for the f-th frame of the n-th client which is obtained at (i,j) by using Eq. (4):

$$A_p^{f,n}(i,j) = \begin{cases} 1 & (i,j) \in Z_{V_p} \\ 0 & (i,j) \notin Z_{V_p} \end{cases} \tag{4}$$

where Z_{V_p} represents the area of the viewport on the planar surface.

Finally, the probability of VA for each pixel location, $a^f(i,j)$, is estimated as (Eq. (5)):

$$a^f(i,j) = \frac{V_A^f(i,j)}{L} = \frac{\sum_{n=1}^{L} A_p^{f,n}(i,j)}{L}. \tag{5}$$

C. VA-driven spherical weighted quality metric

Its spherical representation and interactive exploration (look-around) nature are two key elements for the quality measurement of the ODV content. The observation space of the ODV can be modelled as a sphere surrounding the end users, and each user at any given instant, is able to observe a part of the sphere, corresponding to the actual user's viewport.

Geometrical distortion is involved because of its non-linear geometrical transformation from the spherical to the planar surface [45]. As the spherical ODV content is first projected onto a 2D planar surface, involving a non-linear geometrical

[1] https://v-sense.scss.tcd.ie/research/va-aware-odv-streaming/
https://github.com/cozcinar/optimalTiles

transformation between the pixels of both representations [45] results in geometrical distortion. As an example of this geometrical distortion, the content on the poles of the sphere is severely stretched to be able to fill up the upper and bottom areas of the ERP representation. Fig. 3 illustrates the non-linear geometrical transformation from ERP to spherical surface with an example of a given texture frame and its VA map.

In this context, when computing a traditional pixel-based distortion measurement, such as MSE, the result may differ significantly in the projected plane (*e.g.,* ERP) compared to that obtained in the sphere domain. To account for this effect, we take as reference metric the WS-MSE formulation proposed by Sun *et al.* [45], where correction weights are introduced for each pixel of a given planar projection, defined as $w(i, j)$ for ERP. These weights are in range from zero to one, so ERP pixels in the stretched areas get a lower value than pixels in the equatorial zone. For a given ERP frame, each pixel weight is estimated using the stretching ratio as in Eq. (6):

$$w(i,j) = cos\frac{(j+0.5-N/2)\pi}{N} \qquad \forall i \in M, \quad \forall j \in N, \tag{6}$$

where $M \times N$ is the resolution of a given ODV frame. WS-MSE is then formulated as shown in Eq. (7) [45]:

$$\text{WS-MSE} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} e(i,j)^2 w(i,j)}{\sum_{i=1}^{M}\sum_{j=1}^{N} w(i,j)}, \tag{7}$$

where $e(i, j)$ is the pixel error of an ERP frame of $M \times N$ pixels, and $w(i, j)$ is the weight associated to pixel (i, j) for mapping the ERP pixel to the sphere.

Additionally, in this study, we argue that any quality metric for ODV should also consider the fact that the user cannot observe all the content but only part of it when navigating through it. At any given time, the user can just observe the area corresponding to the location of the viewport on the sphere, although the viewport position changes along time according to the user's head movement. Thus, each viewing of the ODV content consists of a subset of the total pixels that define the ODV, and that subset of pixels is the one that should be considered when measuring the user's quality (*i.e.,* fidelity) on a particular viewing of the ODV. In this sense, we propose to use the VA probability model to capture this effect and extend the WS-MSE metric.

Let $D_{V_p}(f, n)$ be the distortion of the actual part of frame f that has been observed by the user n of the set of users. Considering the WS-MSE as the distortion metric, and taking into account the viewport mask for that frame $A_p^{f,n}$, $D_{V_p}(f, n)$ can be computed as in Eq. (8):

$$D_{V_p}(f,n) = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} e(i,j)^2 w(i,j) A_p^{f,n}(i,j)}{\sum_{i=1}^{M}\sum_{j=1}^{N} w(i,j) A_p^{f,n}(i,j)}, \tag{8}$$

In Eq. (8), the numerator computes the projection-compensated squared error of the pixels that belong to the viewport area, while the denominator represents the area of the viewport in the sphere domain, S_{V_p}. Although in the ERP planar representation, the actual area of the viewport depends on its location due to the geometrical distortion of the projection, in the sphere domain, the area of each viewport does not change with the viewport location. Thereby, Eq. (8) can be re-written as in Eq. (9):

$$D_{V_p}(f,n) = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} e(i,j)^2 w(i,j) A_p^{f,n}(i,j)}{S_{V_p}}, \tag{9}$$

By using Eq. (5), the average $D_{V_p}(f, n)$ value over the total number of users (L), $\overline{D}_{V_p}(f)$, can be calculated as in Eq. (10):

$$\begin{aligned} \overline{D}_{V_p}(f) &= \frac{1}{L}\sum_{n=1}^{L} D_{V_p}(f,n) \\ &= \frac{1}{LS_{V_p}}\sum_{n=1}^{L}\left(\sum_{i=1}^{M}\sum_{j=1}^{N} e(i,j)^2 w(i,j) A_p^{f,n}(i,j)\right). \end{aligned} \tag{10}$$

From Eqs. (5) and (10), we can express $\overline{D}_{V_p}(f)$ using the proposed visual attention probability map in Eq. (11):

$$\begin{aligned} \overline{D}_{V_p}(f) &= \frac{1}{S_{V_p}}\left(\sum_{i=1}^{M}\sum_{j=1}^{N} e(i,j)^2 w(i,j) \frac{1}{L}\sum_{n=1}^{L} A_p^{f,n}(i,j)\right) \\ &= \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} e(i,j)^2 w(i,j) a^f(i,j)}{S_{V_p}} \end{aligned} \tag{11}$$

We refer to $\overline{D}_{V_p}(f)$ as the the visual attention based spherical weighted distortion metric, namely, VASW-MSE which is used in Eq. (2) to determine optimal tiling schemes. Finally, VASW-PSNR is defined from VASW-MSE as:

$$Q_{V_p}(f) = 10log\left(\frac{\chi^2}{\overline{D}_{V_p}}(f)\right), \tag{12}$$

where χ is the maximum possible intensity level of a given ODV frame. For example, this value is 255 for eight bit-depth ODV video content, *i.e.,* $2^{\text{bit-depth}} - 1$. The proposed quality measurement represents the ratio between the maximum possible power of a signal and the noise power based on weighted error with VA in the spherical observation space.

D. Tiling and bitrate allocation

The proposed system works by dividing a given ODV into tiles following a generic tiling architecture that consists of two large tiles for the poles of width equal to that of the ERP, and a set of tiles with different sizes for the equatorial region [7]. Fig. 4 illustrates the used tiling architecture and its possible tile sizes considered in this study.

The used tiling scheme is motivated firstly, by the lower importance due to the stretching that suffers the ERP in the poles together with the usual low-motion characteristics of the scene in them. Secondly, for the dominant viewing adjacency of the equatorial region, we propose the use of a set of tiles to achieve higher coding gain. This set of tiles consists of tiles of different sizes and overlap can occur among some of them.

To generate various number of tile sizes to be used in optimization, each ODV frame is divided into g number of tiles. Two tiles are used for the poles, while for the equatorial area several tiles of different sizes are used. We consider first a tile covering the whole equatorial area, and then we recursively

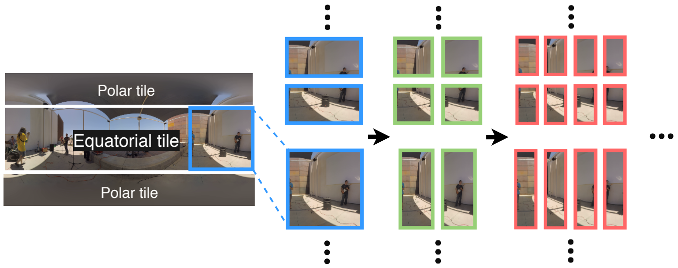


Fig. 4: The used tiling structure with its different schemes.

generate new tiles dividing by half the width and the height of the previous tiles until a predefined number of divisions is reached in each dimension: g_{ver} for the height and g_{hor} for the width. The higher these predefined constants (g_{hor} and g_{ver}) are, the higher the number of schemes used in the tiling scheme optimization is. In other words, it is a given trade-off between granularity of the tiling schemes and computational complexity of the search engine of optimization in Eq. (2).

For a given chunk k and for the given target bitrate $r \in \mathcal{R}$, the objectives are to select the most suitable subset of tiles, called as the optimal tiling scheme s_k^*, among the available ones that cover without overlapping the ODV scene, and to define the bitrate allocation scheme, taking into consideration the characteristics of the VA.

In order for a tiling scheme $s_k \in \mathcal{S}_k$ to be considered for optimization, our proposed VA-based bitrate allocation algorithm first distributes the given target bitrate $r \in R$ within each tile $t \in s_k$ chunk by utilizing a chunk-based VA map. Here, the tiling scheme s_k can be defined as a selection of a subset of tiles in a set of available tiles which cover without overlapping the whole 360° area.

To allocate a given target bitrate r within each tile $t \in s$, we first estimate a weight, φ_t, for each t-th tile as in Eq. (13):

$$\varphi_t = \frac{\sum_{i=x_t}^{M_t} \sum_{j=y_t}^{N_t} w(i,j) P_a^k(i,j)}{\sum_{i=x_t}^{M_t} \sum_{j=y_t}^{N_t} w(i,j)} \qquad i \in M, j \in N, \tag{13}$$

where M_t and N_t are the width and height of the resolution size of the t-th tile, x_t and y_t are the horizontal and vertical pixel positions of the top-left corner of the t-th tile, respectively. $w(i,j)$ represents the weight associated to pixel (i,j) for mapping the projected pixel to the sphere, as defined in Eq. (6). $P_a^k(i,j)$ is the pixel VA probability at the (i,j) pixel location for the c-th chunk, which is estimated as in Eq. (14):

$$P_a^c(i,j) = \sum_{f=1}^{F} a^f(i,j), \tag{14}$$

where F is the total number of frames in the given c-th chunk.

Finally, to accommodate a given DASH representation that has target bitrate $r \in \mathcal{R}$, a bitrate for each tile, r_t, from the set of available encodings, $\{r_i^t\}$, can be selected under the condition that the total bitrate is not higher than the given target bitrate r, as defined in Eq. (15):

$$\begin{aligned} r_t \approx \tilde{\varphi}_t r \quad \forall t \in s_k, \quad &\text{where} \quad \tilde{\varphi}_t = \frac{\varphi_t}{\sum_{t=1}^{T} \varphi_t}, \\ &\text{subject to} \quad \sum_{t=1}^{T} r_t \le r, \\ &\text{with} \quad R_{min}^s \le r_t \le R_{max}^s, \end{aligned} \tag{15}$$

where T is the total number of tiles for a given $s_k \in \mathcal{S}_k$, and R_{min}^s and R_{max}^s are constraints for minimum and maximum bitrate allocation for each tile of s-th tiling scheme.

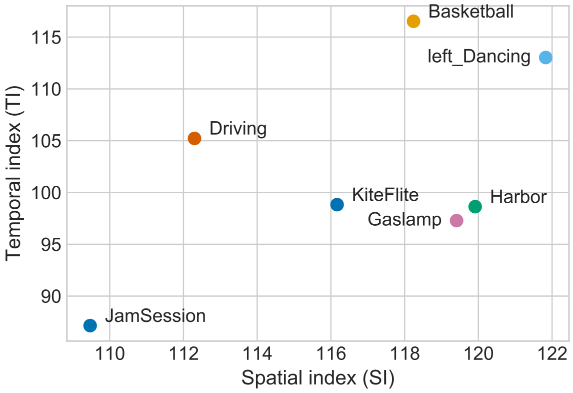


Fig. 5: Video statistics: average SI and TI of ODV sequences used in the experiment.

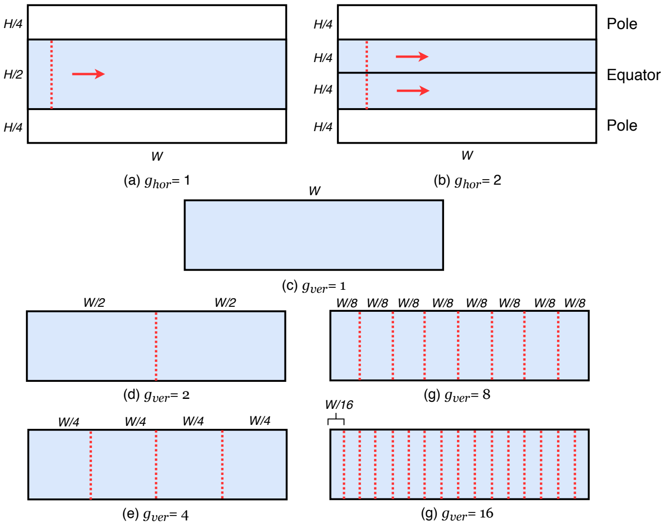


Fig. 6: The used tiling architectures with example height (g_{ver}) and width (g_{hor}) divisions to generate number of tiling schemes with variable-sized tiles.

IV. Results

A. Setup

1) Source video: We used the following six *uncompressed* ODVs from the joint video exploration team (JVET) of ITU-T VCEG and ISO/IEC MPEG: $\mathcal{V}$ = {*Basketball*, *left_Dancing*, *Harbor*, *KiteFlite*, *Gaslamp*, *Driving*, *JamSession*} [52]–[54]. We selected only test data with 8K to provide high quality for a given viewport. Each ODV is in ERP and YUV420p format, 30 fps, and of 10 *sec.* length. Also, a set of ODV was chosen to represent a broad range of content complexities. To choose variety of content types, we calculate spatial and temporal indices, SI and TI, of each ODV based on the ITU recommendation P.910 [55]. SI indicates the amount of spatial details of each video frame, where a higher value represents a more spatially complex image. The estimated TI value presents the amount of motion difference between pixel values at successive frames. More motion in adjacent frames result in higher value of TI. Fig. 5 shows the average value of SI and TI for each ODV sequence used in the experiment.

2) Collection of viewport trajectories: First, we utilized our developed test-bed [12] to collect the viewport trajectories for $\mathcal{V}$ from the L participants. The test-bed was implemented using two APIs, namely, `three.js` [56] and `WebVR` [57]. The former enabled us to create and display GPU-accelerated 3D graphics in a web browser. The latter enabled the creation of fully immersive VR experiences in a web browser, allowing us to display each ODV without the use of any other specific software. The participants viewed each ODV on the Oculus Rift consumer version as HMD and Firefox Nightly as a web browser.

Subjective tests were performed as *task-free* viewing sessions, *i.e.,* each participant was asked to look naturally at each presented 360° video while seated in a freely rotatable chair. Each session, which lasted approximately 15 *min.*, was split into a training and a test session. During the training session, the *Trolley* video sequence [53] was played to ensure a sense of familiarity with the viewing setup. Then, during the test session, each ODV randomly displayed while the individual viewport trajectories were recorded for each participant. We have only considered first views of the content by the participants. Between two successive ODVs, we inserted a five *sec.* short break period with a gray screen. Also, before playing each ODV, we reset the HMD sensor to return the initial position.

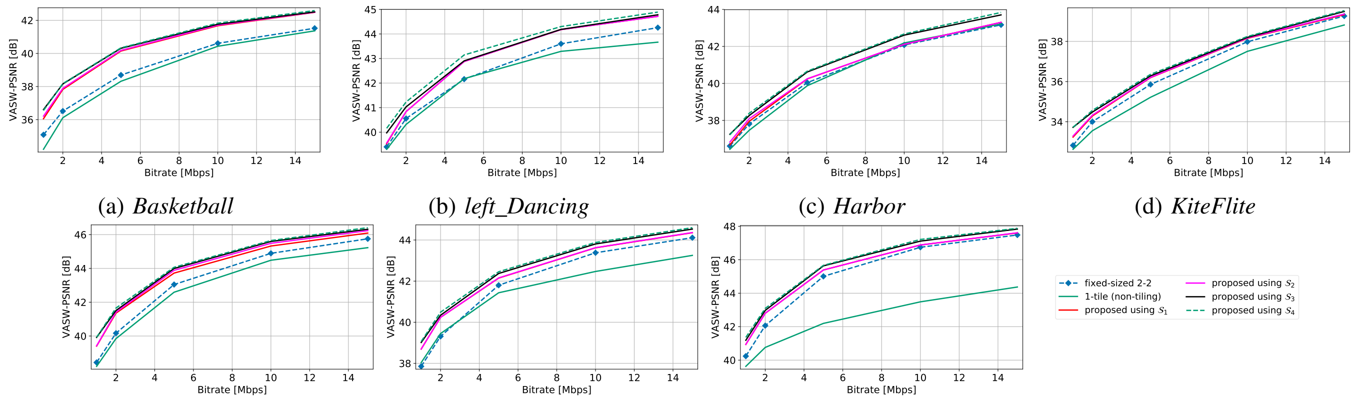
Subjective experiments were conducted with 25 participants (18 males and seven females). The participants were aged between 22 to 46 with an average of 28.2 years. Five of the participants were researchers on the VR project, and the others were naïve viewers; 60% of the participants had a medium familiarity with visual attention studies; 15% and 25% of the participants had no and high familiarity with visual attention studies respectively. Furthermore, eight participants wore glasses, and all of the participants were screened and reported normal or corrected-to-normal visual acuity. Participants were split into two groups for, (*i*) modelling of visual attention data and (*ii*) validation of the proposed approach, consisting of 17 and eight participants, respectively.

3) Tiling: Each given ODV was split into tiles based on the described tiling structure in Section III-D; two fixed-sized tiles at the poles and the varying number of tiles at the equatorial region. Due to the fact that the most professional HMDs have an approximate 90° of field-of-view, and most ODV contents contain the most salient objects in the region of the 90° longitude span of the equatorial segment, we consider the equatorial region as a 90° longitude segment, which is $H/2$. Hence, we assigned the remaining 2× 45° longitude segment for the polar regions, meaning H/4 for each pole. The equator was further split into horizontal and vertical divisions. For this, we used $\{1, 2\}$ for the height division, g_{hor}, and $\{1, 2, 4, 8, 16\}$ for the width division, g_{ver}, to generate number of equatorial tiles. In total, ten fixed-sized tiling schemes, $g_{hor} \cdot g_{ver}$, were generated, and by combining non-overlapping tiles, numerous variable-sized tiling schemes were further formed to be used as $\mathcal{S}$ in our optimization problem. Fig. 6 shows the used tiling architecture to generate a number of tiles.

4) Encoding: We used the HEVC standard [58] to encode each tile of a given ODV. For this, we used the *libx265* in the FFmpeg software (*ver.* N-85291) [59] to encode each tile. We considered the video on demand TN2224 recommendations for Apple devices [60] and encoded each tile using two-pass with 200 percent constrained variable bitrate configurations to ensure smooth perceptual video quality frame by frame for a wide range of devices. We also defined buffer size during encoding which limit the output bitrate to two times of maximum bitrate to handle large bitrate spikes. Our proposed method in this paper is flexible; it could be also used with different encoding settings, *e.g.,* unconstrained or constrained encoding settings. This software was chosen over the HEVC test model [61] reference software because of its faster encoding performance and easy control to choose the target bitrates. We selected 22 different target bitrate levels in between 1 and 50 *Mbps* to encode the ODV content, and the target bitrate for each tile is then distributed proportionally to its tile size. Hence, $R^s_{min} = \frac{1}{g_{hor} \cdot g_{ver}+2}$ *Mbps* and $R^s_{max} = \frac{50}{g_{hor} \cdot g_{ver}+2}$ *Mbps* for the s-th tiling scheme. Each bitstream was divided into 2 *sec.* chunks to perform adaptive streaming. In that end, it is also important to mention that our proposed method is video codec agnostic; it can be easily utilized with different video codecs, *e.g.,* H.264/advanced video coding, VP9, and AV1.

5) Benchmarks: We examined our proposed method with different reference streaming solutions which use non-tiling and fixed-sized tiling schemes (tiling schemes where all its equatorial tiles are of the same size), namely *1-tile (non-tiling)* and *fixed-sized* g_{hor}-g_{ver}. For the fixed-sized tiling schemes, the total encoding bitrate level for each tile is equally distributed by dividing the target bitrate level (r) to a given number of tiles (g), *i.e.,* $b_t = \frac{r}{g_{hor} \cdot g_{ver}+2}$. For the 1-tile (non-tilling) method, each ODV is uniformly encoded according to a given target bitrate level. These two reference solutions stay with their fixed-sized tiling schemes through the streaming session. Such tiling strategies are regularly used for adaptive ODV streaming algorithms in academia and industry [2], [9].

We evaluated the *proposed method* by utilizing four different sets of tiling schemes, $\{\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3, \mathcal{S}_4\}$. Each solution was based on the formulated optimization in Section III-A that can offer per chunk bitrate allocation and tiling scheme adaptation. The sets of $\mathcal{S}_1$ and $\mathcal{S}_2$ contain only fixed-sized tilling schemes, and the sets of $\mathcal{S}_3$ and $\mathcal{S}_4$ contain both variable-sized (tiling schemes where the equatorial tiles can be of different size) and fixed-sized tiling schemes ($\mathcal{S}_3 \supset \{\mathcal{S}_1, \mathcal{S}_2\}$



(a) *Basketball* (b) *left_Dancing* (c) *Harbor* (d) *KiteFlite*

(e) *Gaslamp* (f) *Driving* (g) *JamSession*

Fig. 7: red
Performance comparison using the rate-distortion curves computed with the average VASW-PSNR metric for the best-fixed sized scheme, non-tiling scheme, and proposed selection scheme using different sets.

Method	**Sequence**						
	Basketball	*left_Dancing*	*Harbor*	*KiteFlite*	*Gaslamp*	*Driving*	*JamSession*
fixed-sized 1-1	(2.28, 2.34, 2.52, 2.54)	(1.54, 1.61, 1.94, 2.01)	(1.5, 1.5, 1.62, 1.81)	(2.35, 2.46, 2.6, 2.69)	(1.5, 1.53, 1.66, 1.72)	(2.09, 2.1, 2.27, 2.37)	(4.73, 4.73, 4.95, 5.04)
fixed-sized 1-2	(2.21, 2.26, 2.44, 2.47)	(0.98, 1.04, 1.37, 1.45)	(0.91, 0.92, 1.03, 1.22)	(1.65, 1.76, 1.9, 1.99)	(1.03, 1.06, 1.19, 1.25)	(1.33, 1.33, 1.51, 1.61)	(1.08, 1.08, 1.29, 1.38)
fixed-sized 1-4	(1.5, 1.55, 1.73, 1.76)	(0.77, 0.83, 1.16, 1.24)	(0.38, 0.38, 0.49, 0.69)	(0.82, 0.93, 1.07, 1.16)	(0.45, 0.48, 0.61, 0.67)	(0.56, 0.56, 0.74, 0.84)	(0.41, 0.41, 0.62, 0.71)
fixed-sized 1-8	(2.11, 2.16, 2.34, 2.37)	(1.41, 1.47, 1.81, 1.88)	(0.78, 0.78, 0.9, 1.09)	(1.39, 1.5, 1.64, 1.73)	(0.78, 0.81, 0.94, 1.01)	(1.06, 1.07, 1.25, 1.35)	(1.04, 1.04, 1.26, 1.35)
fixed-sized 1-16	(2.34, 2.4, 2.58, 2.6)	(1.59, 1.66, 1.99, 2.06)	(1.01, 1.01, 1.13, 1.32)	(1.72, 1.83, 1.97, 2.06)	(0.84, 0.87, 1.0, 1.06)	(1.4, 1.41, 1.59, 1.68)	(1.48, 1.48, 1.7, 1.79)
fixed-sized 2-1	(1.92, 1.98, 2.16, 2.18)	(1.32, 1.38, 1.71, 1.78)	(1.56, 1.57, 1.68, 1.87)	(2.09, 2.2, 2.34, 2.43)	(1.37, 1.39, 1.52, 1.59)	(1.96, 1.96, 2.14, 2.24)	(3.32, 3.32, 3.53, 3.62)
fixed-sized 2-2	(**1.26**, **1.31**, **1.49**, **1.52**)	(**0.14**, **0.2**, **0.53**, **0.6**)	(**0.48**, **0.48**, **0.6**, **0.79**)	(**0.81**, **0.92**, **1.06**, **1.15**)	(**0.28**, **0.31**, **0.44**, **0.5**)	(**0.56**, **0.56**, **0.74**, **0.84**)	(**0.47**, **0.47**, **0.69**, **0.77**)
fixed-sized 2-4	(1.65, 1.7, 1.88, 1.91)	(0.97, 1.04, 1.37, 1.44)	(0.8, 0.8, 0.92, 1.11)	(1.27, 1.39, 1.52, 1.61)	(0.8, 0.83, 0.96, 1.02)	(0.88, 0.89, 1.06, 1.16)	(1.18, 1.18, 1.4, 1.49)
fixed-sized 2-8	(1.92, 1.97, 2.15, 2.18)	(1.39, 1.46, 1.79, 1.86)	(0.93, 0.93, 1.05, 1.24)	(1.5, 1.62, 1.75, 1.84)	(0.78, 0.81, 0.94, 1.0)	(1.08, 1.09, 1.27, 1.37)	(1.43, 1.43, 1.65, 1.74)
fixed-sized 2-16	(1.82, 1.88, 2.06, 2.08)	(1.32, 1.38, 1.71, 1.78)	(1.02, 1.03, 1.14, 1.33)	(1.5, 1.61, 1.75, 1.84)	(0.63, 0.65, 0.78, 0.85)	(1.19, 1.19, 1.37, 1.47)	(1.69, 1.69, 1.91, 2.0)
1-tile (non-tiling)	(1.64, 1.69, 1.87, 1.9)	(0.3, 0.36, 0.69, 0.77)	(0.65, 0.66, 0.77, 0.96)	(1.19, 1.31, 1.44, 1.53)	(0.77, 0.8, 0.93, 0.99)	(0.84, 0.84, 1.02, 1.12)	(2.68, 2.68, 2.9, 2.98)

TABLE I: BD quality (in terms of VASW-PSNR (*dB*)) saving of the proposed method using a set of schemes ($\mathcal{S}_1$, $\mathcal{S}_2$, $\mathcal{S}_3$, $\mathcal{S}_4$).

and $\mathcal{S}_4 \supset \{\mathcal{S}_1, \mathcal{S}_2\}$). Each $s \in \{\mathcal{S}_3$ or $\mathcal{S}_4\}$ was formed by gathering the non-overlapping tiles of the existing *fixed-sized* g_{hor}-g_{ver} tiles that can create the complete ODV. The used four different set of tiling schemes thorough evaluation of the *proposed method* were summarized as follows:

- *proposed method using* $\mathcal{S}_1$: Each tiling scheme was generated using the proposed tiling architecture with $g_{hor} = 1$ and $g_{ver} = 16$. The equatorial region had a set of number of *equal-sized* tiles. In total, $\mathcal{S}_1$ contains five different tiling schemes.
- *proposed method using* $\mathcal{S}_2$: Each tiling scheme was generated using the proposed tiling architecture with $g_{hor} = 2$ and $g_{ver} = 16$. The equatorial region contains a set of number of *equal-sized* tiles. Hence, in total, $\mathcal{S}_2$ contains ten different tiling schemes.
- *proposed method using* $\mathcal{S}_3$: Each tiling scheme was generated using the proposed tiling architecture with $g_{ver} = 16$. Because of high computational complexity, we started with the tiling architecture which has two identical large tiles for the equator of width equal to that of the ERP (*see* Fig. 4 (b)) and constrained the new tile searching by dividing with only width dimension g_{ver}. In this context, we generated 1513 different tilling schemes, which have *non-overlapping* and *variable-sized* tile. To have a broad range of tiling schemes, we also included the fixed-sized tiling schemes (*i.e.,* $\mathcal{S}_2$) to this scheme set; thereby, $\mathcal{S}_3$ contains 1523 different tiling schemes.
- *proposed method using* $\mathcal{S}_4$: Each tiling scheme was generated using the proposed tiling architecture with $g_{hor} = 2$ and $g_{ver} = 8$. In doing so, 5528 different tilling schemes, which contain *variable-sized* tiles, were generated. To have a broad range of tiling schemes, the fixed-sized tiling schemes (*i.e.,* $\mathcal{S}_2$) were also added to this scheme set; thereby, $\mathcal{S}_4$ contains 5538 different tiling schemes.

Experimental results were compared to each other using the Bjøntegaard Delta (BD) method [62], which describes the distance between two RD curves. In this manner, PSNR difference, namely ΔP, in dB averaged over the whole range of bitrates was identified. In addition, the observed viewport quality was measured using a set of viewport trajectories. For this purpose, we introduced the viewport-based WS-PSNR metric which considers the spherical distortion of a given ODV. Viewport-based WS-PSNR is defined for the q-th time-stamp of n-th participant as follows:

$$\text{Viewport-based WS-PSNR} = 10log\left(\frac{\chi^2}{I^q_{V_p}}\right), \tag{16}$$

where $I^{q,n}_{V_p}$ represents the MSE of a given viewport as

$$I^{q,n}_{V_p}(i,j) = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} e(i,j)^2 w(i,j) A^{q,n}_p(i,j)}{\sum_{i=1}^{M}\sum_{j=1}^{N} w(i,j) A^{q,n}_p(i,j)}, \tag{17}$$

where $A^{q,n}_p$ is the viewport mask for the q-th time-stamp of the n-th client, as defined in Eq. (4).

In addition, we evaluated the quality of each chunk with different bandwidth using the VMAF [63] quality index,

Method	**Sequence**						
	Basketball	*left_Dancing*	*Harbor*	*KiteFlite*	*Gaslamp*	*Driving*	*JamSession*
fixed-sized 1-1	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)
fixed-sized 1-2	(0.0, 5.08)	(0.0, 0.0)	(5.66, 0.0)	(7.41, 23.26)	(0.0, 7.89)	(6.52, 5.88)	(9.76, 13.64)
fixed-sized 1-4	(0.0, 0.0)	(12.0, 34.09)	(0.0, 31.82)	(11.11, 20.93)	(13.46, 52.63)	(0.0, 13.73)	(2.44, 15.91)
fixed-sized 1-8	(16.13, 20.34)	(40.0, 9.09)	(3.77, 0.0)	(0.0, 9.3)	(0.0, 10.53)	(28.26, 23.53)	(29.27, 6.82)
fixed-sized 1-16	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)
fixed-sized 2-1	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)
fixed-sized 2-2	(22.58, 11.86)	(12.0, 15.91)	(26.42, 18.18)	(14.81, 0.0)	(21.15, 0.0)	(19.57, 9.8)	(29.27, 18.18)
fixed-sized 2-4	(11.29, 11.86)	(16.0, 13.64)	(28.3, 22.73)	(3.7, 0.0)	(26.92, 13.16)	(28.26, 7.84)	(12.2, 4.55)
fixed-sized 2-8	(33.87, 50.85)	(12.0, 27.27)	(16.98, 27.27)	(48.15, 46.51)	(23.08, 15.79)	(17.39, 39.22)	(12.2, 40.91)
fixed-sized 2-16	(16.13, 0.0)	(8.0, 0.0)	(18.87, 0.0)	(14.81, 0.0)	(15.38, 0.0)	(0.0, 0.0)	(4.88, 0.0)

TABLE II: Selected optimal tiles from the fixed-sized tiling schemes (in terms of %) for the proposed method using ($\mathcal{S}_3$, $\mathcal{S}_4$).

which is widely accepted to assess visual quality of video in academia and industry. This metric accounts the temporal characteristics of video and provides perceptually accurate results for traditional 2D video and ODV [64].

For this purpose, we calculated VMAF for viewports within a FOV, called viewport-based VMAF, measured based on 2D rectilinear viewport pictures generated from reconstructed ODVs. The viewport rendered from the reconstructed ODV is compared with the viewport rendered from uncompressed ODV.

B. Experimental Results

We present the justification of the selected number of participant for generating VA maps, the measured compression gain using RD curves along with the Bjøntegaard metric [62], and the observed viewport quality using a set of viewport trajectories.

1) Impact of the used number of participants: To study the impact of the number of participants, we selected three different contents with different content complexities, and carefully evaluated our selected participants by conducting two separate experiments. Each experiment is based on Pearson's correlation coefficient (CC) [65], which is a statistical method used for measuring how correlated two variables are. In this metric, the CC range is between -1 and 1. When the correlation value is close to -1 or 1, there is almost a perfect linear relationship between the two variables.

In the first experiment, we measured the CC between the estimated visual attention map in our paper and the generation of visual attention maps using a variable number of participants. In this experiment, we randomly picked a different number of participants from our dataset. The experimental result shows that the CC score is saturated after the number of fifteen participants. We also observe that using 17 participants is sufficiently enough to generate visual attention maps.

In the second experiment, we aim to verify the fairness of our selected participants. For this, we first estimated an average visual attention map for a chunk using the selected participants in this paper. Then, we randomly selected 17 participants from our developed dataset. We repeated this random selection four times. To analyze the correlation between the random selections and our selection in this paper, we measured the CC between the initial average visual attention map and each random selection. Table V illustrates CC scores for four different selections.

This analysis shows that a high correlation exists with the results for each selection, meaning that the number of selected participants is sufficiently enough for generating representative visual attention maps.

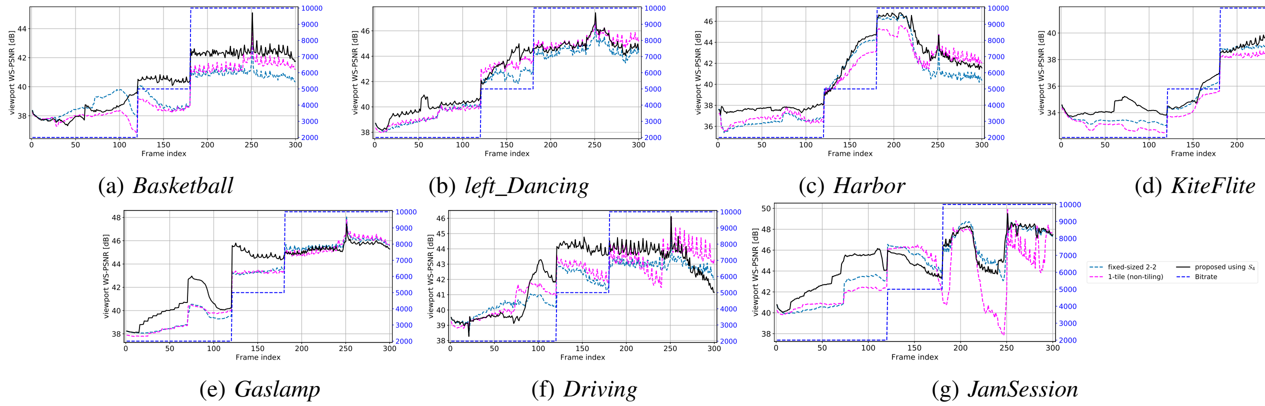
2) Assessment of coding performance: To verify and assess the expected quality (in terms of VASW-PSNR) improvements, we have compared our *proposed method* with various fixed-sized tiling and 1-tile (non-tiling) schemes using the Bjøntegaard metric in Table I.

It can be observed that for seven test sequences, quality gains ranging from 0.14 *dB* to 5.04 *dB* have been obtained with our proposed method. As evident from the results, for all of the tested contents, the proposed method provides important quality enhancements with respect to both fixed-sized tiling and 1-tile (non-tiling) schemes across a wide range of bitrates. For usage of the *1-tile (non-tiling)* scheme, sequences have lower performance than the proposed method and some of the fixed-sized tiling approaches. This is somehow expected as the underlying tiling architecture of the used tiling-based solutions reduce the bit-budget of the polar regions with increasing number of tiles [7], [9]. The polar regions contain significant amount of redundant pixels and those regions have insignificant impact on visual attention [12]. We also observed that the *fixed-sized 2-2* method is the best performing one in the category of fixed-sized tiling schemes.

We then computed rate-distortion curves for all the schemes using our proposed VASW-PSNR measurement at the bitrates of the target DASH representations $\{1, 2, 5, 10, 15\}$ *Mbps*. Fig 7 illustrated the best performing fixed-sized tiling (*fixed-sized 2-2*), 1-tile (non-tiling), and our proposed dynamic tiling approach using different set of tiling schemes.

The results show how our approach outperforms by a significant margin the *fixed-sized 2-2* as well as the *1-tile (non-tiling)* scheme for each tested content. As was expected, our per chunk tiling optimization framework, which uses a set of tiling schemes $\{\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3, \mathcal{S}_4\}$, can reinforce those areas of the ODV scene that are more likely to be watched by selecting optimal bitrate levels and tile sizes, resulting in better compression performance for all test sequences in terms of VASW-PSNR (dB) measurement.

Table II shows the distribution of the decisions made by the proposed per chunk optimization algorithm. It can be observed that the proposed method selected various sized non-overlapping tiles from fixed tiling scheme to create the complete ODV. The selection was based on the formulated



(a) *Basketball* (b) *left_Dancing* (c) *Harbor* (d) *KiteFlite*

(e) *Gaslamp* (f) *Driving* (g) *JamSession*

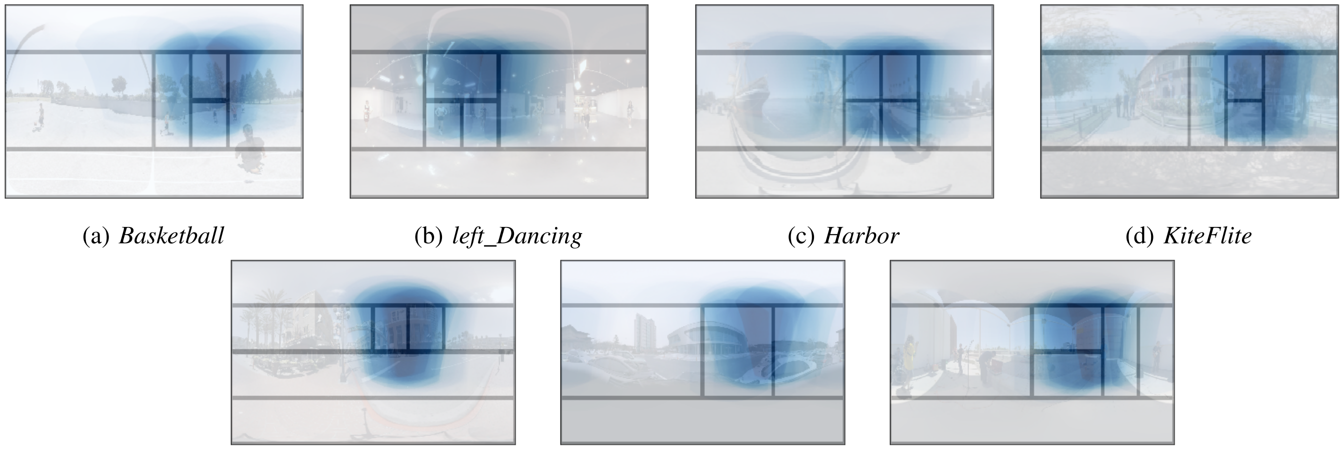
Fig. 8: Performance comparison using viewport-based WS-PSNR quality over frame between the proposed method using $\mathcal{S}_4$ and the reference methods on *varying bandwidth*).

Method		Sequence						
	Measure	*Basketball*	*left_Dancing*	*Harbor*	*KiteFlite*	*Gaslamp*	*Driving*	*JamSession*
fixed-sized 2-2	Mean	(36.71, 39.2, 40.81)	(39.32, 42.97, 44.04)	(37.12, 39.73, 42.82)	(33.37, 35.96, 38.65)	(38.66, 43.49, 45.82)	(39.4, 42.27, 43.18)	(40.26, 46.34, 46.41)
	Std	(0.82, 0.36, 0.09)	(0.33, 0.86, 0.12)	(1.14, 0.54, 1.02)	(0.14, 0.15, 0. 39)	(0.35, 0.14, 0.08)	(0.11, 0.14, 0.16)	(0.39, 0.68, 0.33)
1-tile (non-tiling)	Mean	(36.32, 38.76, 41.06)	(39.71, 42.85, 42.97)	(37.11, 39.79, 42.71)	(33.18, 35.4, 37.6)	(39.09, 43.38, 44.8)	(40.29, 41.89, 42.03)	(40.83, 45.37, 46.19)
	Std	(0.91, 0.16, 0.24)	(0.43, 1.61, 1.84)	(0.72, 0.91, 0.92)	(0.12, 0.52, 0.78)	(0.11, 0.08, 0.55)	(0.07, 1.06, 0.34)	(1.1, 0.78, 0.65)
variable-sized using $\mathcal{S}_1$	Mean	(37.28, 40.38, 42.19)	(40.02, 42.8, 44.24)	(37.83, 39.58, 43.04)	(34.06, 35.97, 37.9)	(40.43, 44.47, 45.52)	(40.8, 42.09, 43.54)	(42.77, 45.77, 46.5)
	Std	(0.56, 0.41, 0.1)	(0.63, 1.23, 0.67)	(1.32, 1.6, 1.25)	(0.69, 0.82, 1.27)	(0.29, 0.37, 0.1)	(0.39, 1.87, 0.13)	(0.93, 0.97, 0.24)
variable-sized using $\mathcal{S}_2$	Mean	(37.32, 40.38, 42.19)	(39.79, 43.01, 44.12)	(37.34, 39.58, 42.7)	(33.92, 36.17, 37.9)	(40.51, 44.69, 45.52)	(40.77, 41.94, 43.39)	(42.42, 46.04, 46.5)
	Std	(0.52, 0.41, 0.12)	(0.81, 1.34, 0.56)	(1.62, 1.6, 1.14)	(0.66, 0.93, 1.27)	(0.31, 0.37, 0.1)	(0.43, 1.73, 0.22)	(1.53, 0.97, 0.24)
variable-sized using $\mathcal{S}_3$	Mean	(37.53, 40.38, 42.19)	(40.05, 43.25, 44.12)	(**37.45**, 39.7, 42.71)	(34.22, 36.17, 37.9)	(**40.74**, 44.69, 45.43)	(**40.84**, 42.19, 43.32)	(**42.8**, 45.79, 46.5)
	Std	(0.51, 0.41, 0.09)	(0.77, 1.32, 0.56)	(1.55, 1.08, 0.64)	(0.84, 0.93, 1.27)	(0.39, 0.37, 0.14)	(0.41, 1.31, 0.05)	(1.93, 0.87, 0.24)
variable-sized using $\mathcal{S}_4$	Mean	(**37.61**, **40.59**, **42.27**)	(**40.17**, **43.38**, **44.24**)	(37.4, **40.05**, **42.99**)	(**34.3**, **36.19**, **37.9**)	(40.69, **44.85**, **45.52**)	(40.81, **42.31**, **43.62**)	(42.77, **46.04**, **46.5**)
	Std	(0.49, 0.1, 0.1)	(0.7, 1.05, 0.18)	(1.33, 1.55, 0.91)	(0.88, 1.14, 1.27)	(0.36, 0.33, 0.15)	(0.33, 1.13, 0.12)	(1.83, 0.85, 0.49)

TABLE III: Mean (standard deviation) values for viewport-based WS-PSNR of reference fixed-sized tiling (*fixed-sized 2-2*), *1-tile (non-tiling)*, and proposed method which uses variable-sized tiles over eight participants. Viewport-based WS-PSNR score for each bitrate level is represented by 3-tuple: $(2Mbps, 5Mbps, 10Mbps)$.

Method	Sequence						
	Basketball	*left_Dancing*	*Harbor*	*KiteFlite*	*Gaslamp*	*Driving*	*JamSession*
fixed-sized 2-2	(66.17, 89.80, 95.41)	(82.60, 92.77, 96.21)	(79.91, 92.02, 95.93)	(70.26, 83.46, **91.52**)	(89.50, **96.09**, **97.32**)	(84.62, 94.24, 97.22)	(89.08, 96.08, 97.78)
1-tile (non-tiling)	(53.27, 89.92, 94.98)	(89.75, 95.40, 97.02)	(82.01, 92.72, **96.31**)	(50.82, 82.28, 91.47)	(90.79, 95.68, 96.96)	(88.75, 95.16, 97.13)	(90.51, 95.46, 96.65)
proposed using $\mathcal{S}_4$	(**87.51**, **95.37**, **97.75**)	(**90.64**, **96.70**, **98.06**)	(**87.39**, **92.78**, 94.43)	(**74.56**, **87.16**, 93.04)	(**91.95**, 95.41, 95.70)	(**91.59**, **96.93**, **98.38**)	(**95.26**, **97.72**, **98.35**)

TABLE IV: Viewport-based VMAF scores for reference fixed-sized tiling (*fixed-sized 2-2*), *1-tile (non-tiling)*, and *proposed method using $\mathcal{S}_4$*. Viewport-based VMAF score for each bitrate level is represented by 3-tuple: $(2Mbps, 5Mbps, 10Mbps)$.



(a) *Basketball* (b) *left_Dancing* (c) *Harbor* (d) *KiteFlite*

(e) *Gaslamp* (f) *Driving* (g) *JamSession*

Fig. 9: Examples of optimal tiling schemes along with corresponding visual attention maps of each tested ODV sequence used in the experiments, for the first chunk at 1 *Mbps* target bitrate.

Sequence	**Selections**			
	1-selection	2-selection	3-selection	4-selection
Basketball	0.9924	0.9908	0.9955	0.9970
Gaslamp	0.9892	0.9943	0.9943	0.9951
JamSession	0.9924	0.9908	0.9955	0.9970

TABLE V: Pearsons correlation coefficient scores for different selections.

optimization algorithm which is based on VA maps of each sequence. In addition, it is also worth noting that for these seven test sequences, tiles of *fixed-sized* 1-1, 1-16, and 2-1 were not suitable due to the characteristics of their VA maps.

3) Viewport-based performance evaluation: As a further evaluation, we computed the WS-PSNR of the actual viewports observed by the users for each sequence, as defined viewport-based WS-PSNR in Eq. (16). For each frame, we computed the WS-PSNR of the viewport that the users observed using the trajectories of the eight participants left for validation.

Fig. 8 shows how our approach can optimize the DASH representations based on the VA map of the sequence. For this propose, we measure the viewport-based WS-PSNR quality using the viewport trajectory of the participant #20 in our developed dataset. Here, we also simulate the varying bandwidth which is shown in blue dashed line. As can be observed, for most of the frames of the sequences and given different bitrates, the quality of the viewports that the user is watching is much higher than that of the fixed schemes. In addition, average viewport quality shows similar significant quality enhancement over eight users. Table III reports mean (standard deviation) values for viewport-based WS-PSNR of references and the proposed method using a set of schemes ($\mathcal{S}_1$, $\mathcal{S}_2$, $\mathcal{S}_3$, $\mathcal{S}_4$) for eight participants, which were left for validation.

To provide further validation, we calculate the viewport-based VMAF scores for a set of bitrate levels $\{2, 5, 10\}$ *Mbps* using the viewport trajectory of eight validation users. Table IV reports viewport-based VMAF scores for the best performed reference fixed-sized tiling scheme (*fixed-sized 2-2*), *1-tile (non-tiling)*, and *proposed method using* $\mathcal{S}_4$. As can be seen in the table, the proposed method using $\mathcal{S}_4$ can provide consistent results with our introduced quality metrics, and for most of the given bandwidth capacities, the VMAF quality score is higher than the benchmark methods.

In order to provide further analysis and visualization of our approach, we show the selected tiling scheme along with the VA map of each ODV for the first chunk at 1 *Mbps* target bitrate in Fig. 9. Looking at each sub-figure, we see that the selected tiling scheme is well correlated with the VA map of each ODV sequence. This further motivates the assumption that incorporating VA into the search algorithm for an optimal tiling scheme may lead to faster and less complex computation, which will be part of our future research.

V. Conclusion

In this paper, to provide an enhanced quality of ODV streaming viewed in head-mounted displays, an adaptive ODV streaming pipeline is presented. The proposed system utilizes the characterization provided by VA maps to compute optimal DASH representations. For that, a novel objective quality measurement that captures the fact that not all the content of the ODV is actually watched by users has been proposed: the visual attention spherical weighted (VASW)-based quality measurement. Then, the use of tiling schemes to represent the ODV content is considered by means of variable-sized and non-overlapping tiles. The proposed system is able to determine optimal pairs (according to the VASW quality metric) of tiling scheme and non-uniform bitrate allocation within tiles per each chunk of every representation.

The performance of the proposed method has been verified in extensive experimental evaluations. Our solution has been compared with reference adaptive streaming solutions, which are based on naïve tiling and non-tiling schemes, and are used by most existing ODV streaming studies. The results have shown that our proposed method achieves a significant quality enhancement compared to both type of reference solutions for adaptive ODV streaming. Future work will focus on increasing the coding performance of the tiling schemes with the help of perceptual encoding techniques, modeling the peripheral vision of the viewport, and on faster and less complex search algorithms for optimal tiling schemes.

Acknowledgment

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under the Grant Number 15/RP/2776. This work has also been partially supported by the Ministerio de Economía, Industria y Competitividad (AEI/FEDER) of the Spanish Government under project TEC2016-75981 (IVME).

Appendix
Weight Distribution and Tiling for Different Projections

A. Weight distribution for cube map

For a given frame in the traditional cube map projection format, weight distribution on all faces of the cube map are the same. Therefore, each pixel weight for a cube map face is estimated using the stretching ratio as defined in Eq. (18) [45], [66].

$$w(i,j) = (1 + \frac{d^2_{cube}(i,j)}{r^2})^{-3/2} \quad \forall i \in A, \quad \forall j \in A, \tag{18}$$

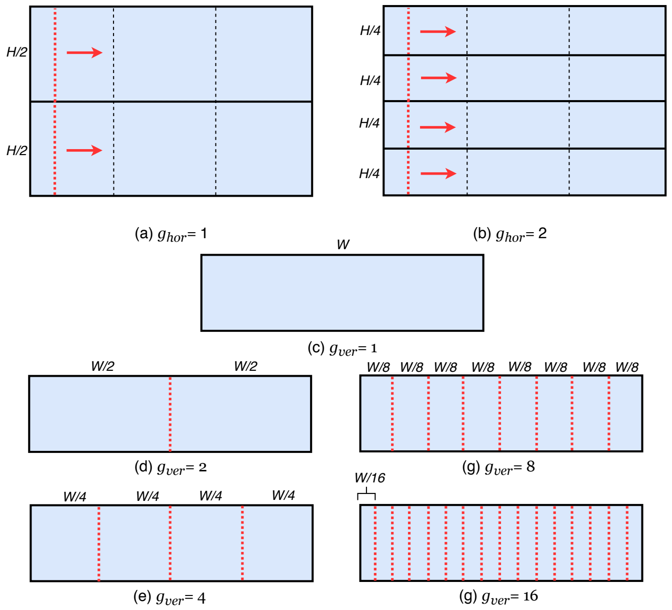
where $r = \frac{A}{2}$ is the radius, $A \times A$ is the resolution of a cube map face, and $d^2_{cube}(i,j)$ is the squared distance between (i,j) and the center of the face as defined in Eq. (19).

$$d^2_{cube}(i,j) = (1 + 0.5 - A/2)^2 + (j + 0.5 - A/2)^2. \tag{19}$$

B. Weight distribution and for equi-angular cube map

For a given frame in the equi-angular cube map projection format, similar as traditional cube map, weight distribution on all faces of the cube map are the same. Therefore, each pixel weight for a cube map face is estimated using the stretching ratio as defined in Eq. (20) [66].

$$w(i,j) = \frac{\pi^2}{\Big(1 + \big(tan(t_i)\big)^2 + \big(tan(t_j)\big)^2\Big)^{3/2} \cdot F_{ang}}, \quad \forall i \in A, \quad \forall j \in A, \tag{20}$$



(a) g_{hor}= 1
(b) g_{hor}= 2
(c) g_{ver}= 1
(d) g_{ver}= 2
(g) g_{ver}= 8
(e) g_{ver}= 4
(g) g_{ver}= 16

Fig. 10: The proposed tiling structure for cube map with its different schemes.

with

$$F_{ang} = 16 \cdot \big(cos(t_i)\big)^2 \cdot \big(cos(t_j)\big)^2. \tag{21}$$

where $A \times A$ is the resolution of a cube map face, t_i and t_j are derived as defined in Eqs. (22) and (23), respectively.

$$t_i = \frac{\pi}{4} \cdot \big(\frac{2(i+0.5)}{A} - 1\big), \tag{22}$$

and

$$t_j = \frac{\pi}{4} \cdot \big(\frac{2(j+0.5)}{A} - 1\big). \tag{23}$$

C. Tiling architecture for Cube map

This section describes a tiling structure for cube map projected ODV to be used by the proposed approach in this paper.

Each given cube map projected ODV is divided into two equal sized spatial parts to be further split for tiles with different sizes. Each part can be split into horizontal and vertical divisions. For this purpose, similarly with the ERP in this paper, we use $\{1, 2\}$ for the height division, g_{hor}, and $\{1, 2, 4, 8, 16\}$ for the width division, g_{ver}, to generate number of tiles. By combining non-overlapping tiles, numerous variable-sized tiling schemes can be generated as $\mathcal{S}$ in our optimization problem. Fig. 10 shows the proposed tiling architecture to generate a number of tiles for cube map projection.

References

[1] M. Postgate, “BBC announces live Ultra HD and VR trials for World Cup,” July 2018. [Online]. Available: https://www.bbc.co.uk/mediacentre/latestnews/2018/uhd-vr-world-cup

[2] M. Graf, C. Timmerer, and C. Mueller, “Towards bandwidth efficient adaptive streaming of omnidirectional video over http: Design, implementation, and evaluation,” in *Proceedings of the 8th ACM on Multimedia Systems Conference*, ser. MMSys’17. New York, NY, USA: ACM, 2017, pp. 261–271. [Online]. Available: http://doi.acm.org/10.1145/3083187.3084016

[3] S. Heymann, A. Smolic, K. Mueller, Y. Guo, J. Rurainsky, P. Eisert, and T. Wiegand, “Representation, coding and interactive rendering of high-resolution panoramic images and video using MPEG-4,” in *Panoramic Photogrammetry Workshop*, Berlin, Germany, Feb. 2005, pp. 24–25.

[4] C. Grunheit, A. Smolic, and T. Wiegand, “Efficient representation and interactive streaming of high-resolution panoramic views,” in *2002 International Conference on Image Processing (ICIP)*, vol. 3, Rochester, NY, USA, USA, Sep. 2002, pp. III–209–III–212 vol.3.

[5] R. Skupin, Y. Sanchez, C. Hellge, and T. Schierl, “Tile based HEVC video for head mounted displays,” in *IEEE International Symposium on Multimedia (ISM)*, San Jose, CA, USA, Dec 2016, accessed: 2017-1-16.

[6] A. Zare, A. Aminlou, M. M. Hannuksela, and M. Gabbouj, “HEVC-compliant Tile-based Streaming of Panoramic Video for Virtual Reality Applications,” in *Proceedings of the 2016 ACM on Multimedia Conference*, ser. MM ’16. New York, NY, USA: ACM, 2016, pp. 601–605. [Online]. Available: http://doi.acm.org/10.1145/2964284.2967292

[7] C. Ozcinar, A. De Abreu, and A. Smolic, “Viewport-aware adaptive 360° video streaming using tiles for virtual reality,” in *2017 International Conference on Image Processing (ICIP)*, Beijing, China, Sep 2017.

[8] C. Diaz, J. Cabrera, M. Orduna, L. Munoz, P. Perez, J. Ruiz, and N. Garcia, “Viability analysis of content preparation configurations to deliver 360vr video via mpeg-dash technology,” in *2018 IEEE International Conference on Consumer Electronics (ICCE)*, Las Vegas, NV, USA, Jan 2018.

[9] M. Hosseini and V. Swaminathan, “Adaptive 360 VR video streaming: Divide and conquer!” in *2016 IEEE International Symposium on Multimedia (ISM)*, San Jose, CA, USA, Sep 2016.

[10] X. Corbillon, A. Devlic, G. Simon, and J. Chakareski, “Viewport-adaptive navigable 360-degree video delivery,” in *2017 IEEE International Conference on Communications (ICC)*, vol. cs.MM, no. 1609.08042, May. 2017, pp. 1–7.

[11] G. Cheung, Z. Liu, M. Zhiyou, and J. Z. G. Tan, “Multi-Stream switching for interactive virtual reality video streaming,” in *2017 International Conference on Image Processing (ICIP)*, Sep 2017.

[12] C. Ozcinar and A. Smolic, “Visual attention in omnidirectional video for virtual reality applications,” in *10th International Conference on Quality of Multimedia Experience (QoMEX 2018)*, Sardinia, Italy, May 2018.

[13] ISO/IEC 23009-1, “Information technology — dynamic adaptive streaming over HTTP (DASH) — part 1: Media presentation description and segment formats,” ISO/IEC JTC1/SC29/WG11, Tech. Rep., 2014.

[14] K. Misra, A. Segall, M. Horowitz, S. Xu, A. Fuldseth, and M. Zhou, “An Overview of Tiles in HEVC,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 6, pp. 969–977, Dec 2013.

[15] O. A. Niamut, E. Thomas, L. D’Acunto, C. Concolato, F. Denoual, and S. Y. Lim, “MPEG DASH SRD: Spatial relationship description,” in *7th International Conference on Multimedia Systems*, ser. MMSys ’16. ACM, 2016, pp. 5:1–5:8.

[16] A. Smolic and P. Kauff, “Interactive 3-D video representation and coding technologies,” *Proceedings of the IEEE*, vol. 93, no. 1, pp. 98–110, Jan 2005.

[17] C. Ozcinar, A. De Abreu, S. Knorr, and A. Smolic, “Estimation of optimal encoding ladders for tiled 360° VR video in adaptive streaming systems,” in *The 19th IEEE International Symposium on Multimedia (ISM 2017)*, Taichung, Taiwan, Nov. 2017.

[18] Y. Sanchez, R. Skupin, and T. Schierl, “Compressed domain video processing for tile based panoramic streaming using HEVC,” in *2015 IEEE International Conference on Image Processing (ICIP)*, Sept 2015, pp. 2244–2248.

[19] P. Lungaro, R. Sjöberg, A. J. F. Valero, A. Mittal, and K. Tollmar, “Gaze-aware streaming solutions for the next generation of mobile +VR experiences,” *IEEE transactions on visualization and computer graphics*, vol. PP, no. 99, pp. 1–1, 2018.

[20] A. Patney, M. Salvi, J. Kim, A. Kaplanyan, C. Wyman, N. Benty, D. Luebke, and A. Lefohn, “Towards foveated rendering for gaze-tracked virtual reality,” *ACM transactions on graphics*, vol. 35, no. 6, p. 179, 11 Nov. 2016.

[21] C. Concolato, J. L. Feuvre, F. Denoual, E. Nassor, N. Ouedraogo, and J. Taquet, “Adaptive Streaming of HEVC Tiled Videos using MPEG-DASH,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, pp. 1–1, 2017.

[22] A. De Abreu, C. Ozcinar, and A. Smolic, “Look around you: Saliency maps for omnidirectional images in VR applications,” in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, Erfurt, Germany, May 2017, pp. 1–6.

[23] C. Ozcinar, J. Cabrera, and A. Smolic, “Omnidirectional Video Streaming Using Visual Attention-driven Dynamic Tiling for VR,” in *submitted*

to IEEE International Conference on Visual Communications and Image Processing, Dec 2018, pp. 1–4.

[24] M. Xiao, C. Zhou, Y. Liu, and S. Chen, "OpTile: Toward optimal tiling in 360-degree video streaming," in *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 23 Oct. 2017, pp. 708–716.

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

[26] Y. Ban, L. Xie, Z. Xu, X. Zhang, Z. Guo, and Y. Hu, "An optimal spatial-temporal smoothness approach for tile-based 360-degree video streaming," in *2017 IEEE Visual Communications and Image Processing (VCIP)*, St. Petersburg, FL, USA, Dec 2017, pp. 1–4.

[27] A. Zare, A. Aminlou, and M. M. Hannuksela, "6K Effective Resolution with 4K HEVC Decoding Capability for OMAF-compliant 360° Video Streaming," in *Proceedings of the 23rd Packet Video Workshop*, ser. PV '18. New York, NY, USA: ACM, 2018, pp. 72–77. [Online]. Available: http://doi.acm.org/10.1145/3210424.3210425

[28] Y. Liu, M. Xu, C. Li, S. Li, and Z. Wang, "A novel rate control scheme for panoramic video coding," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, Hong Kong, China, Jul. 2017, pp. 691–696.

[29] ISO/IEC 23008-2:2013, "Information technology – high efficiency coding and media delivery in heterogeneous environments – part 2: High efficiency video coding," ISO/IEC 23008-2:2013 (HEVC), Tech. Rep., 2013.

[30] M. Yu, H. Lakshman, and B. Girod, "A framework to evaluate omnidirectional video coding schemes," in *2015 IEEE International Symposium on Mixed and Augmented Reality*, Fukuoka, Japan, Sept 2015, pp. 31–36.

[31] Y. Li, J. Xu, and Z. Chen, "Spherical domain rate-distortion optimization for 360-degree video coding," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, Jul. 2017, pp. 709–714.

[32] P. Ramanathan, M. Kalman, and B. Girod, "Rate-distortion optimized interactive light field streaming," *IEEE Transactions on Multimedia*, vol. 9, no. 4, pp. 813–825, June 2007.

[33] "Enhancing high-resolution 360 streaming with view prediction," https://code.facebook.com/posts/118926451990297/, Apr 2017.

[34] C. Chen, Y. Lin, A. Kokaram, and S. Benting, "Bitrate optimization for multi-representation encoding using playback statistics," Patent 20 180 124 146:A1, 3 May, 2018.

[35] L. Xie, Z. Xu, Y. Ban, X. Zhang, and Z. Guo, "360ProbDASH: Improving QoE of 360 Video Streaming Using Tile-based HTTP Adaptive Streaming," in *Proceedings of the 2017 ACM on Multimedia Conference*, ser. MM '17. New York, NY, USA: ACM, 2017, pp. 315–323. [Online]. Available: http://doi.acm.org/10.1145/3123266.3123291

[36] S. Petrangeli, V. Swaminathan, M. Hosseini, and F. De Turck, "An HTTP/2-Based adaptive streaming framework for 360° virtual reality videos," in *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 23 Oct. 2017, pp. 306–314.

[37] X. Liu, Q. Xiao, V. Gopalakrishnan, B. Han, F. Qian, and M. Varvello, "360° innovations for panoramic video streaming," in *Proceedings of the 16th ACM Workshop on Hot Topics in Networks*. Palo Alto, CA, USA: ACM, 30 Nov. 2017, pp. 50–56.

[38] F. Qian, L. Ji, B. Han, and V. Gopalakrishnan, "Optimizing 360 video delivery over cellular networks," in *Proceedings of the 5th Workshop on All Things Cellular: Operations, Applications and Challenges*, ser. ATC '16. New York, NY, USA: ACM, 2016, pp. 1–6. [Online]. Available: http://doi.acm.org/10.1145/2980055.2980056

[39] D. Podborski, E. Thomas, M. M. Hannuksela, S. Oh, T. Stockhammer, and S. Pham, "Virtual reality and dash," in *International Broadcasting Convention, IBC*, 2017.

[40] Y. Wang, M. M. Hannuksela, and S. Deshpande, "Wd 1 of iso/iec 23090-2 omaf 2nd edition," ISO/IEC JTC1/SC29/WG11, San Diego, USA,, Tech. Rep. N17584, Apr 2018.

[41] R. Skupin, Y. Sanchez, Y. . Wang, M. M. Hannuksela, J. Boyce, and M. Wien, "Standardization status of 360 degree video coding and delivery," in *2017 IEEE Visual Communications and Image Processing (VCIP)*, Dec 2017, pp. 1–4.

[42] S. Dambra, G. Samela, L. Sassatelli, R. Pighetti, R. Aparicio-Pardo, and A.-M. Pinna-Déry, "Film editing: new levers to improve VR streaming," in *Proceedings of the 9th ACM Multimedia Systems Conference*. ACM, 12 Jun. 2018, pp. 27–39.

[43] S. Rossi and L. Toni, "Navigation-Aware adaptive streaming strategies for omnidirectional video," in *IEEE International Workshop on Multimedia Signal Processing (MMSP 2017)*, Luton, UK, 2017.

[44] V. Zakharchenko, K. P. Choi, and J. H. Park, "Quality metric for spherical panoramic video," in *SPIE 9970, Optics and Photonics for Information Processing X, 99700C*, 2016.

[45] Y. Sun, A. Lu, and L. Yu, "Weighted-to-spherically-uniform quality evaluation for omnidirectional video," *IEEE Signal Processing Letters*, vol. 24, no. 9, pp. 1408–1412, 2017.

[46] E. Upenik, M. Rerabek, and T. Ebrahimi, "A testbed for subjective evaluation of omnidirectional visual content," in *Picture Coding Symposium (PCS)*, Nuremberg, Germany, Dec. 2016.

[47] M. Xu, C. Li, Z. Wang, and Z. Chen, "Visual quality assessment of panoramic video," *arXiv:1709.06342v1 [eess.IV]*, 19 Sep. 2017.

[48] H. Duan, G. Zhai, X. Min, Y. Zhu, Y. Fang, and X. Yang, "Perceptual quality assessment of omnidirectional images," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, Florence, Italy, May 2018, pp. 1–5.

[49] Y. Rai, J. Gutiérrez, and P. Le Callet, "A dataset of head and eye movements for 360 degree images," in *Proceedings of the 8th ACM on Multimedia Systems Conference*. Taipei, Taiwan: ACM, Jun. 2017, pp. 205–210.

[50] P. Guo, Q. Shen, M. Huang, R. Zhou, X. Cao, and Z. Ma, "Modeling peripheral vision impact on perceptual quality of immersive images," in *2017 IEEE Visual Communications and Image Processing (VCIP)*, St. Petersburg, FL, USA, Dec. 2017, pp. 1–4.

[51] B. Choi, Y.-K. Wang, M. M. Hannuksela, Y. Lim, and A. Murtaza, "Study of ISO/IEC DIS 23000-20 omnidirectional media format," ISO/IEC JTC1/SC29/WG11, Torino, Italy, Tech. Rep. N16950, Jul 2017.

[52] A. Abbas and B. Adsumilli, "Ahg8: New gopro test sequences for virtual reality video coding," JTC1/SC29/WG11, ISO/IEC, Chengdu, China, Tech. Rep. JVET-D0026, Oct 2016.

[53] E. Asbun, H. He, Y. He, and Y. Ye, "Ahg8: Interdigital test sequences for virtual reality video coding," JTC1/SC29/WG11, ISO/IEC, Chengdu, China, Tech. Rep. JVET-D0039, Oct 2016.

[54] G. Bang, G. Lafruit, and M. Tanimoto, "Description of 360 3D video application exploration experiments on divergent multiview video," JTC1/SC29/WG11, ISO/IEC, Chengdu, China, Tech. Rep. MPEG2015/M16129, Feb. 2016.

[55] "Subjective video quality assessment methods for multimedia applications," ITU-T, Tech. Rep. ITU-T P.910, 2008.

[56] "JavaScript 3D library. https://threejs.org/," https://github.com/mrdoob/three.js/, Feb 2017.

[57] "WebVR: Bringing virtual reality to the web," https://webvr.info/, Feb 2017.

[58] J.-R. Ohm and G. Sullivan, "Vision, applications and requirements for high efficiency video coding (HEVC)," ISO/IEC JTC1/SC29/WG11, Geneva, Switzerland, Tech. Rep. MPEG2011/N11891, March 2011.

[59] "x265 HEVC Encoder / H.265 Video Codec," http://x265.org/, Jan 2018.

[60] "HLS Authoring Specification for Apple Devices," https://developer.apple.com, Jan 2018.

[61] I.-K. Kim, K. McCann, K. Sugimoto, B. Bross, and W.-J. Han, "High efficiency video coding (HEVC) test model 10 (HM10) encoder description," ISO/IEC JTC1/SC29/WG11, Geneva, Switzerland, Tech. Rep. N12242, Jan. 2013.

[62] G. Bjøtegaard, "Calculation of average PSNR differences between RD-curves (vceg-m33)," VCEG Meeting (ITU-T SG16 Q.6), Austin, Texas, USA,, Tech. Rep. M16090, Apr 2001.

[63] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," *The Netflix Tech Blog*, vol. 6, 2016.

[64] M. Orduna, C. Díaz, L. Muñoz, P. Pérez, I. Benito, and N. García, "Video Multimethod Assessment Fusion (VMAF) on 360VR contents," *arXiv e-prints*, p. arXiv:1901.06279, Jan. 2019.

[65] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[66] Y. Ye and J. Boyce, "Algorithm descriptions of projection format conversion and video quality metrics in 360lib version 7," JVET of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, Ljubljana, SI, Tech. Rep. JVET-K1004, July 2018.

Cagri Ozcinar is a research fellow within the V-SENSE project at Trinity College Dublin, Ireland, since July 2016. Before he joined the V-SENSE team, he was a post-doctoral fellow in the Multimedia group at Institut Mines-Telecom Télécom ParisTech, Paris, France. He received the M.Sc. (Hons.) and the Ph.D. degrees in electronic engineering from the University of Surrey, UK, in 2011 and 2015, respectively. His current research interests include visual attention and communications for immersive media technologies. He has been serving as a reviewer for a number of IEEE transactions and conferences. He has been involved in organizing workshops, challenges, and special sessions.

Julián Cabrera received the Telecommunication Engineering and Ph.D. degrees in telecommunication from the Universidad Politécnica de Madrid (UPM), in 1996 and 2003, respectively. Since 1996, he is a member of the Image Processing Group, UPM. Since 2001, he has been a member of the faculty of the UPM, and since 2003, he has been an Associate Professor of signal theory and communications. Current research interests cover several topics related to audio-visual communications, advance video coding, 3D and Multiview scenarios, depth estimation with special focus on deep learning approaches, video subjective quality assessment for Multiview and VR360 video, and optimization of adaptive streaming techniques.

Aljosa Smolic is the SFI Research Professor of Creative Technologies at Trinity College Dublin (TCD). Before joining TCD, Prof. Smolic was with Disney Research Zurich as Senior Research Scientist and Head of the Advanced Video Technology group, and with the Fraunhofer Heinrich-Hertz-Institut (HHI), Berlin, also heading a research group as Scientific Project Manager. At Disney Research he led over 50 R&D projects in the area of visual computing that have resulted in numerous publications and patents, as well as technology transfers to a range of Disney business units. Prof. Smolic served as Associate Editor of the IEEE Transactions on Image Processing and the Signal Processing: Image Communication journal. He was Guest Editor for the Proceedings of the IEEE, IEEE Transactions on CSVT, IEEE Signal Processing Magazine, and other scientific journals. His research group at TCD, V-SENSE, is on visual computing, combining computer vision, computer graphics and media technology, to extend the dimensions of visual sensation, with specific focus on immersive technologies such as AR, VR, volumetric video, 360/omni-directional video, light-fields, and VFX/animation, with a special focus on deep learning in visual computing.