Omnidirectional Video Streaming Using Visual Attention-Driven Dynamic Tiling for VR

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

*V-SENSE, School of Computer Science and Statistics, Trinity College Dublin, Ireland.

†Grupo de Tratamiento de Imágenes, Information Processing and Telecommunications Center
and ETSI Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain.

***Abstract*—This paper proposes a new adaptive omnidirectional video (ODV) streaming system that uses *visual attention* (VA) maps. The proposed method benefits from a novel approach to VA-based bitrate allocation algorithm and *dynamic tiling*, providing enhanced virtual reality (VR) video experiences. The main contribution of this paper is the use of VA maps: (i) to distribute a given bitrate budget among a set of tiles of a given ODV and, (ii) to decide an optimal tiling structure (*i.e.,* tile scheme) per chunk. For this, a novel objective metric is proposed: the visual attention spherical weighted (VASW) PSNR. This metric operates in the spherical domain and by means of a VA probabilistic model aims at capturing the quality of the actual areas observed by the users when navigating through the ODV content. We evaluate the proposed system performance with varying bandwidth conditions and the tracked head orientations from disjoint user experiments. Results show that the proposed system significantly outperforms the existing tiled-based streaming method.**

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

I. Introduction

Recent significant industrial investment in *virtual reality* (VR) technologies has led to immersive VR video streaming experiences using *omnidirectional video* (ODV), also known as 360° video. However, the delivery of perceptually acceptable quality level is a challenging task for ODV because of the high amount of data involved, the limitations of the present Internet, and the processing and decoding constraints on the available client devices.

Existing HMDs have a viewable field of view (FoV) and use only a fraction of the given ODV at a given time, namely *viewport*. Therefore, transmission of ultra-high resolution of ODV (*e.g.,* $\geq$ 8K×4K) is needed to obtain a decent VR video quality level. In this context, viewport-dependent solutions [1]–[4] can enhance the level of QoE for ODV streaming. However, in scenarios with delay-prone communication pipelines [5] and rapid head orientation activities, such solutions are merely inefficient to comply with the motion-to-photon latency requirement, thus penalizing the quality of experience (QoE).

Streaming solutions relying on tile-based MPEG-Dynamic adaptive streaming over HTTP (DASH) solutions [1] have been proposed to deal with the ODV delivery problem. Tiles are self-decodable spatial regions that allow the client to select which portions have to be extracted from a given bitstream.

Recent research works focus on adaptive ODV streaming using fixed-sized tiles. Various practical adaptation strategies, for instance, were discussed in [1] using the tile-based encoding. In addition, an optimal spatial-temporal smoothness approach was proposed in [6] for tile-based adaptive streaming. Similarly, a viewport-adaptive video delivery system was developed in [7] that uses tiles and different DASH representations that differ by their bitrate and different scene regions. An optimal DASH representation for each tile was requested in a viewport aware manner in [4]. Also, several version of DASH representations were generated for different viewport positions in [8], where the opposite areas of the defined viewport were set to black to reduce the encoding bitrate. However, none of the described works consider dynamic tiling and viewport-based VA maps, which are the major contributions of our work.

This paper proposes a novel adaptive streaming system to obtain a decent VR video quality through delay-prone networks, *e.g.,* Internet. The objective is to provide high video quality by finding the most appropriate tiling structure and target encoding bitrate level for each tile of a given ODV by consulting VA maps.

Our main contribution is a novel adaptive streaming system design which determines the most appropriate tiling scheme on a chunk basis and the required encoding bitrate for each tile of a given content using a VA probabilistic model. We conducted subjective viewing sessions using an HMD to estimate this model which is first used in a novel bitrate allocation algorithm for the tiles. In a second step, we propose a VA-driven objective quality metric, VASW-PSNR, to determine the optimal tiling scheme for each DASH representation on a chunk basis. To verify our method, we recorded head orientations from participants in disjointed viewing sessions using an HMD and compared our proposed method with reference solutions, which are based on naive tiled-based adaptive streaming approaches.

The proposed design does not require any modification of the existing DASH players, being entirely transparent to them. As such, we expect that our work will provide a beneficial input for the streaming industry considering the transmission of dynamic tiled ODV for each chunk and varying bitrate distribution for each tile with the help of VA maps. The remainder of this paper is organized as follows: Sec. II describes the proposed system. Then, we present evaluation results in Sec. III, and we conclude the paper in Sec. IV.

II. Proposed system

We consider an end-to-end adaptive streaming system for VR applications to deliver very high resolution of ODV over

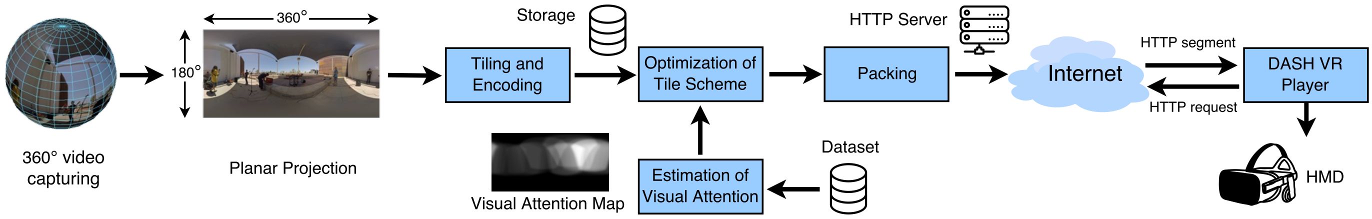


Fig. 1: Schematic diagram of the proposed adaptive ODV streaming system for VR.

the Internet as depicted in Fig. 1. The server side of the proposed system contains tiling, encoding, optimization of tiling scheme, and packing for adaptive streaming.

Each captured ODV is first mapped onto a 2D plane using the commonly used equirectangular projection (ERP) [9], for backward-compatibility purpose with the existing video coding standards. A given ERP content is then divided into a predetermined number of *tiles* which in turn are encoded at various bitrate levels. For adaptive streaming purposes, each encoded tile is segmented into chunks and stored in the HTTP server. To this end, a set of several predetermined tiling schemes is available to prepare DASH representations at different bitrates.

In addition, a VA map is computed for each ODV. This map represents the probability that the user watches a pixel within the ERP representation. For a given tiling scheme, a target bitrate is calculated for each tile by our proposed VA map based bitrate distribution method. Here, the objective is to reinforce the quality of those parts of the ODV that are more likely to be seen, maximizing the expected quality of the actual content watched by the user.

Finally, the proposed system selects the optimal tiling scheme according to the VAWS-PSNR metric. The selection of optimal tiling schemes is carried out on a chunk basis. Hence, full 360° DASH representations are delivered to the clients upon their HTTP requests.

A. Estimation of visual attention maps

To estimate VA maps, we utilized the developed test-bed in [10], [11] which gathers user's viewport tracking data for the given videos. Given a collection of tracked head orientations, a viewport-based VA map is estimated for each chunk of an ODV. In this work, each VA map serves as a bidimensional histogram for the pixel locations of the ODV content, and its values represent the number of times that clients have paid attention to the analogous pixels in the ODV. The higher the VA map value is, the more times the pixel at that position in the ODV has been watched. For this, from each recorded head position (a pair of *yaw* and *pitch* values), the viewport area is estimated with a mask, where the pixels within the viewport are one and pixels outside are zero.

The VA map for a given chunk k, V_A^k, averages all the contributions of the set of clients considered, $\mathcal{C}$. Finally, the probability of each pixel location $v_a^k(i,j)$ is estimated as:

$$v_a^k(i,j) = V_A^k(i,j)/\mathcal{N}, \tag{1}$$

where $\mathcal{N}$ is the number of times that the head orientation is tracked during the chunk

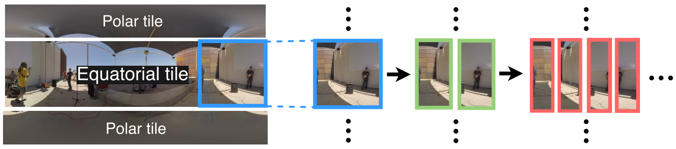


Fig. 2: The used tiling scheme with its different structures.

B. Optimization of tiling scheme

The proposed system works by dividing a given ODV into tiles using the tiling scheme, paying attention to the typically lower importance and low-motion characteristics of the poles and the dominant viewing adjacency of the equatorial region. As the poles occupy the largest regions of redundant pixels, in those areas, a larger tile resolution size is used to compress them using a lower bitrate. Also, as the equator is associated with the most dominant viewing adjacency, it is further divided horizontally into several tiles in a dynamic way to achieve higher coding gain. Fig. 2 illustrates the tiling scheme considered in this paper.

Each given ODV consists of a set of tiles, $\mathcal{T}$, of various sizes and each *t-th* tile, $t \in \mathcal{T}$, has m different target encoding bitrates $\{R_t^1 \dots R_t^m\}$. Thus, to build the set of DASH representations, an optimal pair of tiling scheme and bitrate allocation for its tiles has to be determined.

A tiling scheme, T_s, can then be defined as a selection of a subset of tiles in $\mathcal{T}$ which cover without overlapping the whole 360° area. We consider that there exists a predefined set of n_t possible tiling schemes, $\mathcal{T_S}$, that is evaluated in the optimization process. Therefore, let R_r be the target bitrate of the *r-th* DASH representation. For each k chunk, the optimal tiling scheme $T_{s,k}^*$ can then be calculated as follows:

$$T_{s,k}^* = \max_{T_s \in \mathcal{T_S}} Q(T_s, V_A^k, B(R_r, V_A^k)), \tag{2}$$

where V_A^k represents the VA map for chunk k, $Q(\cdot)$ represents the quality of T_s taking into account V_A^k and following a bit allocation function, $B(\cdot)$, which in turn takes into account V_A^k and is subject to a target bitrate R_r.

1) Visual attention-based bitrate allocation: For any tiling scheme, T_s, to be considered for optimization, our proposed VA-based bitrate allocation algorithm distributes a given target bitrate R_r among the set of T tiles for each chunk by utilizing VA map. For this, we first estimate a weight, φ_t, for each tile as follows:

$$\varphi_t = \frac{\sum_{i=x_t}^{M_t} \sum_{j=y_t}^{N_t} v_a^k(i,j)}{M_t N_t}, \tag{3}$$

where M_t and N_t are the width and height pixel number of the t-th tile, x_t and y_t are horizontal and vertical positions of the top-left corner of the t-th tile, and $v_a^k(i,j)$ is the pixel VA probability. Then, the bitrate assigned to the t-th tile is:

$$b_t = \widetilde{\varphi}_t R, \quad \text{where} \quad \widetilde{\varphi}_t = \frac{\varphi_t}{\sum_{t=1}^{T} \varphi_t} \quad \text{with} \quad \sum_{t=1}^{T} b_t \leq R \tag{4}$$

Finally, to accommodate the DASH representations, from the set of encoded versions of each tile, the ones with the bitrate closest to the computed values are selected, on the condition that the total bitrate is not higher than the target bitrate, R.

2) Visual Attention quality metric: The observation space of ODV for VR applications can be defined as a sphere. However, in this work, we have considered the use of the ERP representation for the ODV content for backward-compatibility with the existing video coding and transmission standards. As stated in [12], this projection involves a non-linear transformation; thus the relationship between pixels in the 2-D plane and the spherical surface is not linear. Therefore, results from objective distortion metrics for traditional video such as MSE may differ significantly in the ERP plane compared to those in the sphere domain. To account for this effect, we take as reference metric the WS-MSE formulation proposed in [12], where correction weights are proposed for each pixel in the ERP representation.

Similarly, we argue that any quality metric for ODV should also consider the fact that only part of the content (the current viewport) is presented to the user at any given time. Thus, not all the pixels of the ODV should contribute equally to the computation of the quality metric. In this sense, we propose to use the VA probability model to capture this effect and extend the WS-MSE metric to the the visual attention spherical weighted metrics: VASW-MSE and VASW-PSNR.

The pixel square error can be formulated as VASW-MSE by weighting VA probability value of a given ODV and averaging the size of the viewport in the spherical surface:

$$\text{VASW-MSE} = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} e(i,j)^2 w(i,j) v_a(i,j)}{S_{V_p} = \sum_{i,j \in V_p} w(i,j)}, \tag{5}$$

where $e(i,j)$ is the pixel error of an ERP frame of $N \times M$ pixels, $w(i,j)$ is the weight associated to pixel (i,j) for mapping the ERP pixel to the spherical representation [12], $v_a(i,j)$ is the probability that pixel (i,j) be watched by the user according to the VA model, and S_{V_p} represents the spherical size of a viewport. Finally, VASW-PSNR can be calculated from VASW-MSE as follows:

$$\text{VASW-PSNR} = 10 \cdot log \frac{255^2}{\text{VASW-MSE}} \tag{6}$$

III. Experiments

A. Setup

We use the following two ODVs (8K×4K ERP) from the JVET and MPEG video coding exploration experiments: $\mathcal{V}$ = {*Train*, *Basketball*} [13]. The HEVC standard was utilized to encode each tile of a given ODV. For this, we used the FFmpeg software (*ver.* N-85291) [14] for encoding purposes with two-pass and 200 percent constrained variable bitrate configurations. Also, we used a set of *target bitrates* $\mathcal{B}$ = {1, 1.20, 1.44, 1.73, 2.07, 2.49, 2.10, 3.58, 4.30, 5.16, 6.19, 7.43, 8.92, 10.70, 12.84, 15.41, 18.49, 22.19, 26.62, 31.95, 38.34, 46.01} (in terms of *Mbps*) to encode each ODV content. To enable selectively choosing various tile sizes and bitrate level combinations per chunk, each ODV frame was divided into N tiles, encoded to be independently decodable. Each bitstream was divided into 2 *sec.* chunks. Two tiles were used for the poles, and $N-2$ tiles were used for the equator. We examined our *proposed method* with the reference naive tiled-based adaptive streaming approach that used N fixed-sized tiles, namely *fixed-sized* N *tiles*. Therefore, encoded bitrate for each tile of a given video is equally distributed for the reference methods by dividing the *target bitrate* to a given N tiles.

To estimate VA maps and viewing trajectories for viewport-based quality measurements, we organized subjective experiments under task-free condition. A total of 17 participants (13 males and 4 females) took part of the test. Participants were split into two groups for, (*i*) modelling of visual attention data and (*ii*) validation of the proposed approach, consisting of twelve and five participants, respectively. In our subjective tests, we used the Oculus Rift consumer version as an HMD and Firefox Nightly as a web browser. Each participant was seated in a rotatable chair and allowed to turn freely.

B. Performance evaluation

To verify and assess the expected coding gains, we have compared our *proposed* method, with fixed-sized tiling schemes with a number of tiles N=$\{1, 4, 6, 10, 18\}$. We computed rate-distortion curves for all the schemes using our VASW-PSNR measurement at the bitrates of the target DASH representations {2, 10, 15, 20, 25} *Mbps*. Fig 3 (a,b) shows how our approach outperforms by a significant margin the fixed tiled schemes. As was expected, our dynamic tiling scheme can reinforce those areas of the ODV scene that are more likely to be watched.

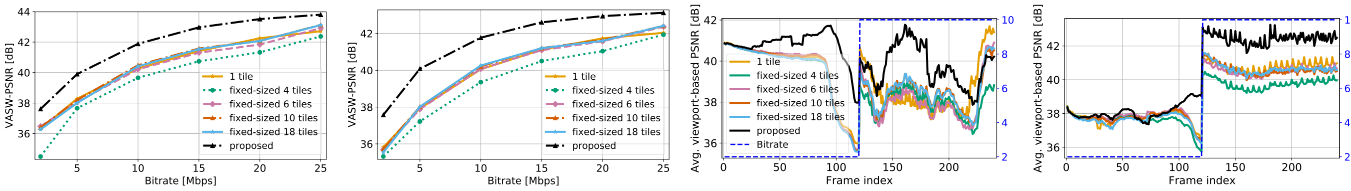
Sequence	**Fixed-size tiles**				
	N=1	N=4	N=6	N=10	N=18
Train	1.50	2.28	1.60	1.44	1.56
Basketball	1.80	2.42	1.82	1.71	1.72

TABLE I: Quality gain of the proposed method in terms of BD quality (*dB*) saving.

Moreover, this gain has been characterized in terms of the Bjøntegaard metric [15] in Table I. It can be observed that for both sequences, quality gains ranging from 1.44 *dB* to 2.42 *dB* have been obtained with our proposed method.

Sequence	**Number of tiles**				
	N=1	N=4	N=6	N=10	N=18
Train	80%	8%	0%	12%	0%
Basketball	0%	57.14%	0%	7.15%	35.71%

TABLE II: Selected tiling schemes (in terms of %) for the proposed method.



(a) *Train* (b) *Basketball* (c) *Train* (d) *Basketball*

Fig. 3: (a,b) Performance comparison using the RD curves computed with the average OmniVA-PSNR. (c,d) viewport-PSNR quality over frame between the proposed and the reference methods on *varying bandwidth (Mbps)*.

Table II shows the distribution of the decisions made by the algorithm. It can be observed that the tiling scheme choices are different according to the VA maps of each sequence. Also, it is also worth noting that for these two test sequences, $N = 6$ tiling scheme was not suitable due to the characteristics of their VA maps.

As a further assessment, we computed the PSNR of the actual viewports observed by the users at each frame of the sequence. Here, we simulate the varying bandwidth which is shown in blue dashed line. For each frame, we compute the PSNR of the viewport that the users observed using the trajectories of the five participants left for validation. Fig. 3 (c,d) shows how our approach can optimize the DASH representations based on the VA map of the sequence for a participant. 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 five users. Table III reports mean (standard deviation) values for viewport-based PSNR of references and proposed method for five participants, which were left for validation.

Sequence	**Fixed-size tiles**					**proposed**
	N=1	N=4	N=6	N=10	N=18	
Train	39.48(1.21)	39.33(1.55)	39.59(1.74)	39.63(1.84)	39.62(1.87)	**40.76**(2.03)
Basketball	38.67(2.17)	37.74(1.94)	38.23(2.15)	38.30(2.33)	38.56(2.43)	**39.85**(2.60)

TABLE III: Mean (standard deviation) values for PSNR of viewports of reference fixed-size tiling and proposed method over five participants.

IV. Conclusion

This paper introduced a new adaptive omnidirectional video (ODV) streaming system, utilizing visual attention (VA) maps. Our proposed method does not need any modification at the client side being entirely transparent to the existing DASH players. The developed system aimed at an enhanced quality of ODV streaming viewed in head-mounted displays. For this, the proposed method utilized dynamic-sized tiles per chunk and varying bitrate allocation per tile by consulting the estimated VA maps. The performance of the proposed method was verified in experimental evaluations. The results showed that our proposed method achieves significant quality gain compared to the fixed-sized tiling methods which are naive approach and used by the most existing tiling based ODV streaming solutions. As future work, we plan to extend the proposed system by considering additional tile schemes and investigating the effect of viewport quality using comprehensive user datasets and ODV sequences.

Acknowledgement

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).

References

[1] 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*, New York, NY, USA, 2017, MMSys'17, pp. 261–271, ACM.

[2] S. Heymann et al., "Representation, coding and interactive rendering of high-resolution panoramic images and video using MPEG-4," in *Panoramic Photogrammetry Workshop*, Berlin, Germany, 2005.

[3] 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)*, Sept. 2002, vol. 3, pp. III–209–III–212 vol.3.

[4] 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)*, Sep 2017.

[5] G. Cheung et al., "Multi-Stream switching for interactive virtual reality video streaming," in *2017 International Conference on Image Processing (ICIP)*, Sep 2017.

[6] 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)*, Dec 2017, pp. 1–4.

[7] X. Corbillon, A. Devlic, G. Simon, and J. Chakareski, "Viewport-adaptive navigable 360-degree video delivery," *arXiv:cs.MM 1609.08042*, vol. cs.MM, no. 1609.08042, pp. 1–7, May. 2017.

[8] C. Diaz et al., "Viability analysis of content preparation configurations to deliver 360vr video via MPEG-DASH technology," in *2018 IEEE International Conference on Consumer Electronics (ICCE)*, Jan 2018.

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

[10] 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)*, May 2017, pp. 1–6.

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

[12] 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.

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

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

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