Versatile Video Coding of 360-Degree Video using Frame-based FoV and Visual Attention

J. Carreira*†, Sergio M. M. de Faria*†, Luis M. N. Tavora†, Antonio Navarro*‡, Pedro A. A. Assuncao*†
*Instituto de Telecomunicações, Portugal
†Instituto Politécnico de Leiria, Portugal
‡Universidade de Aveiro, Portugal
jcarreira@co.it.pt

***Abstract*—High quality omnidirectional video requires ultra high resolution formats encoded with very high bit rates to guarantee acceptable QoE in video delivery services. Since in general the full FoV, i.e., 360°, is not required at once by users, rather than agnostic encoding of the whole 360° video, this work proposes a flexible coding approach where the full FoV is mapped into video frames and efficiently encoded using intra-FoV prediction. By avoiding inter-FoV prediction, the proposed approach enables independent decoding of one or more FoVs extracted from a single compressed stream containing the full FoV video. To achieve improved quality in those FoVs which attract more visual attention, non-uniform coding is proposed for the new Versatile Video Coding standard (VVC), using perceptually-driven quantisation for each FoV. This strategy, makes use of visual attention maps to decrease the overall bit rate without compromising the quality of the most relevant regions. The simulation results show that the proposed coding mechanism achieves consistent quality gains in the relevant FoV without significant losses in the remaining ones. In comparison with the reference VVC, the proposed method is able to achieve average quality gains up to 1.56 dB and to efficiently adapt the coding parameters to the visual attention information.**

***Index Terms*—Omnidirectional video coding, independent FoV decoding, visual attention-based coding.**

I. Introduction

In recent years, omnidirectional video has been increasingly integrating applications and devices in the consumer market, providing enhanced multimedia experiences and pushing forward the quality requirements towards greater resolutions, e.g., Ultra-High Definition (UHD), such as 4K or 8K at 60Hz or 120Hz. The huge amount of data that is necessary to represent the full Field-of-View (FoV), i.e., 360°, of any arbitrary visual scene imposes the use of highly efficient compressed formats. The forthcoming Versatile Video Coding (VVC) [1] that is under development by Joint Video Experts Team (JVET), is particularly suited to deal with such requirements of 360° video, as it aims to double the coding efficiency of the current High Efficiency Video Coding (HEVC) standard at UHD resolutions.

In general, multimedia users of omnidirectional video (e.g., applications using Head Mounted Display (HMD) are able to change their viewing direction, interactively at time, to focus on certain specific regions of a spherical scene. This means that only a limited region of a whole spherical scene is required for display at the end user device at any specific time instant. Although the observer might want to observe the whole spherical scene, only a narrow FoV is required at each time. In the case of HMDs, this is imposed by the allowed FoV, which typically ranges between 90° and 210° [2]. Therefore, since only a reduced amount of visual information is required at each time, there is no need for the whole 360° video content to be delivered and decoded because only a reduced part (i.e., a viewport) is sent to the display at the receiver-side. In this case, a great deal of transmission bandwidth and decoding resources can be saved by using some flexible coding and delivery mechanisms for reduced FoV but still allowing dynamic adaptation to users needs [3].

Different approaches have been proposed to deal with similar problems as described above. For instance tile-based approaches map the whole 360° video scene into tiles which are independently processed [4]–[6], allowing receivers to select any tile according to the viewing area or relevant viewport. In [7] tiles are combined with different levels of quality and spatial resolution. The impacts of the delay introduced by using adaptive tile-based 360° video streaming was also evaluated, revealing that the end-to-end delay is a critical factor in the overall quality of experience (QoE). Although these approaches provide flexible encoding solutions, the bit streams only encode the relevant part of the 360° which poses limitations in multi-user application scenarios because not all of them may be receiving the same viewport. In [8] a scalable coding approach is proposed based on a combination of a down-sampled layer in Equirectangular Projection (ERP) format and a higher-resolution layer using Cube-Map Projection (CMP) format. While the former provides a low-quality full FoV representation, the latter provides a high-quality representation of a narrow region. The results indicate that using this approach a significant bit rate reduction can be achieved but only considering an isolated viewport corresponding to a single cube face.

All the above mentioned methods can reduce the overall bit rate in 360° video streaming by encoding with high quality only small size regions, i.e., small FoVs, while the remaining visual content is either not encoded or encoded with lower quality in the main stream. Also, since they are all based

This work is funded by FCT/MEC through national funds, under project ARoundVision SAICT-45-2017-POCI-01-0145-FEDER-030652, PTDC/EEICOM/30652/2017, and when applicable co-funded by FEDER PT2020 partnership agreement under the project UID/EEA/50008/2019.

978-1-7281-5606-4/19/$31.00 ©2019 IEEE
DOI 10.1109/ISM46123.2019.00020

in adaptive streaming, which requires network feedback and stream adaptation, they are not the most suitable solutions for low-delay or real-time streaming applications. Moreover, these approaches take advantage of the scalable extension of the HEVC standard [8]–[10], but this is not available in the most recent VVC standard. Thus, flexible single-layer encoding schemes, compliant with the VVC standard are yet to appear.

This work advances one step further in this direction, by proposing a flexible coding approach for VVC encoding of 360° video. The spherical scene (i.e., full FoV), represented in CMP format, is mapped into video frames which are efficiently encoded using intra-FoV prediction. By avoiding inter-FoV prediction, the proposed approach enables independent decoding of one or more FoVs extracted from a single stream containing the full FoV video. To achieve improved quality in those FoVs which attract more visual attention, asymmetric compression is proposed using perceptually-driven quantisation for each FoV, based on visual attention maps. As the proposed scheme does not require stream adaptation and user feedback is not necessary, this can be used in a wider range of applications than existing approaches. Moreover, this scheme takes advantage of the implicit temporal scalability present in the VVC to encode and deliver different FoVs in different independent layers (i.e., substreams), which enable flexible bit stream truncation and FoV decoding.

The remainder of the paper is organised as follows. Next section presents a brief description of the reference VVC framework for 360° video encoding. Section III describes the proposed frame-based FoV coding scheme using visual attention maps. Section IV presents and discusses the experimental results and Section V concludes this paper.

II. Reference encoding framework for 360° video

The common encoding framework the 360° video, uses a planar projection of 360°×180° video content from a 3D sphere to a 2D plane, which is usually defined by the so-called projection mapping function [11]. The projection mapping also includes a down-sampling step in order to remove the bias towards the original representation format. Then, after such projection mapping, the resulting planar 2D video data is encoded. The decoder output is mapped back to the spherical domain, from which viewports are extracted for display. This framework is presented in Figure 1, where it is shown that the 3D sphere can be projected using different mapping functions, producing different planar representations, e.g, ERP, CMP, etc. The JVET group recently started to explore the use of other projection mapping functions based on polyhedrons, such as, Octahedrom Projection (OHP) [12] and Icosahedrom Projection (ISP) [13], which extend the number of faces beyond the six of CMP. The polyhedron-based projections are obtained by arranging small polyhedrons corresponding to a small FoV in order to form a rectangular shape suitable for coding. Therefore, these projections naturally divide the omnidirectional video into partitions corresponding to small FoVs. In the case of the CMP each partition is a cube face, corresponding to a 90°×90° FoV in the entire omnidirectional scene. In order to form a projection cube around the sphere , each FoV is 90° apart from each other. The CMP provides a planar format suitable for partitioning the omnidirectional scene into multiple FoVs without further processing.

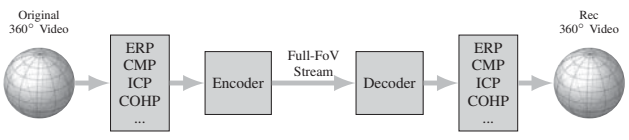

Figure 1. Reference omnidirectional video coding framework.

III. Frame-based FoV coding using visual attention

In this section the proposed flexible coding approach for 360° video for the new versatile video encoding standard VVC is described. Figure 2 illustrates the functional diagram of the proposed coding scheme. The 360° video is converted into a planar representation by using the polyhedron-based Cube-Map Projection (CMP), which results in a 2D mapping with six different FoVs. Subsequently, FoV to frame mapping is used to convert FoVs into frames and the resulting temporal frame sequences form the input of the VVC encoder. This is a frame-based FoV representation of 360° video with granularity of 90°×90°. Moreover, from the 360° video a visual attention map is determined which is used to optimise the encoding process. As described next, the frame-based FoV representation enables asymmetric coding of FoVs, with the help of the attention map.

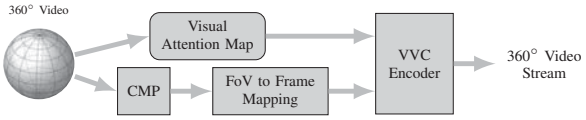

Figure 2. Functional diagram of the proposed coding approach.

Figure 3 (1) shows the CMP representation, with the six cube faces, corresponding to six FoVs identified by their initials. Conventional coding schemes arrange all cube faces into a rectangular matrix (i.e., one frame with 2×3 cube faces) to encode the whole 360° video. This is illustrated in Figure 3 (2). However, since the six FoVs do not have a great deal of spatial correlation among them, the coding gain at FoV level mostly arises from temporal correlations. This is also a rigid coding configuration because it does not allow a functionality to truncate the coded stream for extraction and decoding of the only one or few FoVs. If each cube face is represented as a frame, then a straightforward encoding structure can be the one shown in Figure 3 (3). However, due to low correlation between different FoV this is not very efficient.

The proposed coding configuration is shown in Figure 3 (4), where each cube face is also encoded as a frame sequence, but temporal predictions are constrained to use only intra-FoV predictions, as illustrated in Figure 3 (4). Such prediction structure

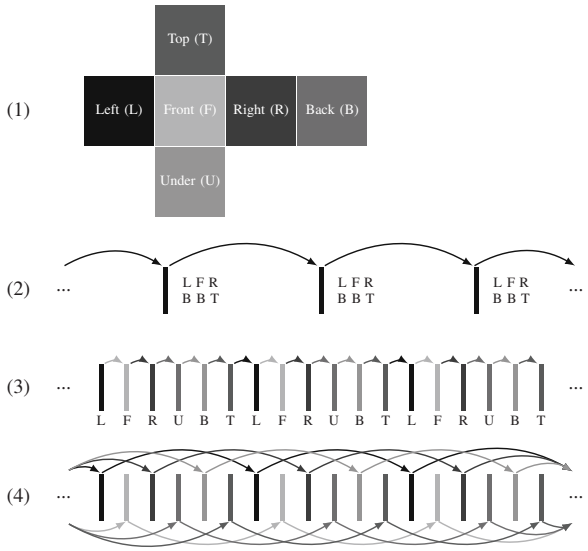

Figure 3. Coding dependencies using different approaches.

does not allow neither spatial nor temporal prediction across different FoVs. Thus, by avoiding inter-FoV predictions, the proposed approach enables independent extraction, deliver and decoding of one or more 90° FoVs by simple truncation of the single compressed stream containing the full FoV video. In order to accomplish the desired prediction structure, the reference frames used are explicitly signalled in the bit stream, which is already supported by the VVC standard [1].

In the aforementioned approach, the multiple FoVs are encoded and multiplexed into a single coded stream, but taking advantage of the implicit temporal scalability present in the VVC. By encoding different FoVs in different temporal layers (i.e., substreams), flexible bit stream truncation and FoV decoding is enabled, by only looking for the layer identifier at the Network Abstraction Layer (NAL) unit headers. Finally, this coding prediction scheme is complaint with the VVC standard.

A. Visual attention based quantisation

To deal with variable importance of the omnidirectional visual content, non-uniform VVC encoding is proposed to achieve improved quality in those FoVs which attract more visual attention. Visual attention maps are used as input parameters and can be estimated either from the input 360° video signal or based on user feedback. Regardless of their origin, the use of visual attention maps in 360° video is of utmost importance because the whole visual scene is not meant to be observed all at once, thus homogeneous encoding of all FoVs is not a very efficient approach.

Figure 4 shows an example of one attention map (AM) in the CMP format, where the dark regions indicate the regions which have low attention and white regions indicate the regions which have high attention. As can be seen in Figure 4, the visual attention is not uniformly distributed across the entire spherical domain, but rather more concentrated on two FoVs. This characteristic is used for non-uniform bit rate allocation by adjusting the quantisation parameter of each FoV based on a perceptual weight derived from the visual attention map, which allows to provide higher quality to the FoVss with higher saliency values.

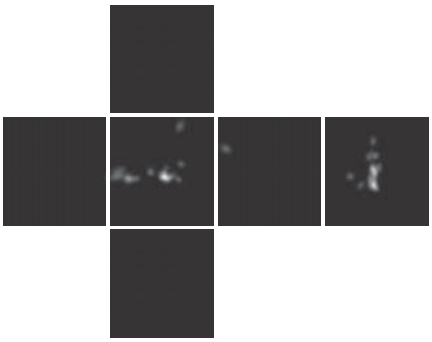

Figure 4. Example of a visual attention map in the CMP format.

The proposed perceptually-driven quantisation method is based on the energy of the visual attention signal in each Fov. Firstly the average energy ratio (E_R) of the visual attention of each FoV is calculated as follows:

$$E_R(f) = \frac{E(f)}{\sum_{i=1}^{6} E(i)}, \tag{1}$$

$$E(f) = \sum_{\mathbf{p}} \mid AM(f, \mathbf{p}) \mid^2, p \in \Omega, \tag{2}$$

where f is the FoV index and $\mathbf{p}$ is a pixel-based visual attention value belonging to the visual attention map of the current FoV (Ω). The value of E_R indicates how visual attention is distributed across the omnidirectional scene. Subsequently, the bit rate is allocated to different FoVs according to the energy of the corresponding visual attention map, which is used to define higher Quantisation Parameter (QP) values for those frames (i.e., FoV) that have lower visual attention, as follows,

$$\Delta QP(f) = \alpha \left(1 - \frac{E_R(f)}{max\{E_R(f)\}}\right). \tag{3}$$

The underlying principle of this approach is to maintain the quantisation unchanged for the FoV with the highest saliency energy value, i.e., $E_R(f) = max\{E_R(f)\}$, and increase the quantisation for the remaining ones. The normalisation by the maximum value in (3) removes the impact of saliency differences between different video contents. Finally, the α parameter in (3) is used to control the maximum variation introduced in the QP. For those cases where a saliency map is not available, a constant QP is used for all FoVs.

Summarising, the proposed non-uniform FoV coding leads to an overall bit rate saving by increasing the quantisation parameter of FoVs with lower importance in terms of visual attention, while maintaining high quality the in most relevant FoV corresponding to the highest users' attention.

Table I
TEST SEQUENCES USED IN THE EXPERIMENTS.

Sequence	SI	TI	Description
Abbottsford	26.1	1.21	Dorm room with one student talking with moderate motion
Cockpit	17.9	24.0	Cockpit footage with high camera vibrations
PlanEnergyBioLab	24.0	3.67	Laboratory with several people moving around
PortoRiverside	24.7	2.75	Riverside images with two standing guys and low overall motion
TeatroRegioTorino	19.3	5.07	Orchestra concert with high different people playing instruments
Turtle	21.2	18.3	Two women helping a turtle return to ocean

IV. EXPERIMENTAL RESULTS

The performance of the proposed frame-based FoV non-uniform coding scheme (Prop) was evaluated by comparison against using FoV to frame mapping with homogeneous coding (HC), i.e., the proposed method without QP variations based on the visual attention, and against the reference VVC encoding as illustrated in Figure 1 (Ref).

The six 360° video sequences presented in Table I, obtained from the Salient360 database [14], were used in the simulations. The omnidirectional video content has a spatial resolution of 3840×1920 in the ERP format and ground-truth visual attention maps were obtained from head-mounted display and eye-tracking. This test material has different types of motion and texture complexity, as shown by the different spatial (SI) and temporal information (TI) [15] presented in Table I. The VVC reference software, version 5.1 [16] was used with all coding modes enabled, an IDR period of 16 frames and a prediction structure using all B-frames with one reference frame (Low-delay). The Common Test Condition [17] were followed to conduct the experiments and a coding resolution of 3456×2304 was used when converting from ERP to CMP. All methods under comparison use the CMP format to ensure a fair comparison.

Firstly, the Rate-Distortion (R-D) performance of the proposed approach is compared against the reference VVC. Table II shows the Bjøntgaard delta PSNR calculated from the quality of the single FoV that corresponds to the highest visual attention as given by the energy of the saliency map (BD-Single-Fov-PSNR). The results indicate that the homogeneous encoding (HC) of all FoVs leads to poor R-D performance in comparison with the case where all faces are encoded as a single frame, i.e., without using FoV to frame mapping (reference method). Taking into account the SI in Table I and relating with the results in Table II, one can notice that the quality degradation increases with the spatial complexity. This is due to the impact of re-mapping the omnidirectional scene into multiple frames on the context used for entropy coding, which reduces its efficiency and consequently that of the inter-FoV predictive coding. On the contrary, for low spatial complexity the inter-FoV prediction is able to reduce

Table II
BJØNTGAARD DELTA PSNR OF A SINGLE FOV WITH THE HIGHEST E_R.

Sequence	BD-Single-FoV-PSNR		
	HC	Prop-$\alpha = 5$	Prop-$\alpha = 10$
Abbottsford	-0.33	0.69	1.35
Cockpit	-0.16	1.05	2.03
PlanEnergyBioLab	-0.37	0.54	1.06
PortoRiverside	-0.24	0.86	1.68
TeatroRegioTorino	-0.12	0.86	1.55
Turtle	-0.26	0.90	1.67
Average	-0.24	0.82	1.56

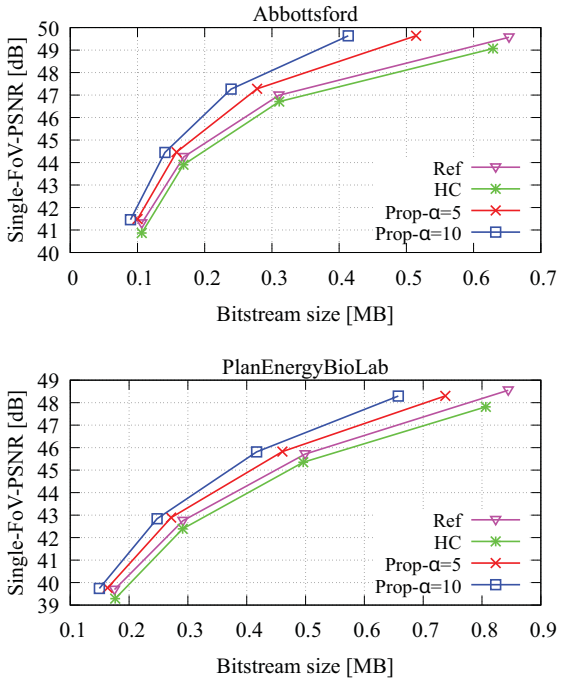

Figure 5. Rate-distortion results for the methods tested.

the impact of the increased overhead and achieve a small quality reduction for the same bit rate.

The results of the proposed method (Prop) reveals an improvement of the overall R-D performance beyond the reference VVC. As shown in column 3 and 4 of Table II the proposed method achieves quality gains up to 2.03 dB for the Cockpit sequence. As for the HC method, higher quality is achieved for sequences with low spatial complexity (see SI in Table I). The proposed method is analysed for two different values of α and the results confirm that when higher QP variations occur between high and low visual attention, higher quality gains are also obtained. Overall, the proposed method consistently outperforms the reference VVC with an average quality gain up to 1.56 dB.

Figure 5 shows a graphical comparison of the rate-distortion between the different methods under evaluation. Since there are a different number of frames and different frame rates between the tested methods, the bit stream size is used rather

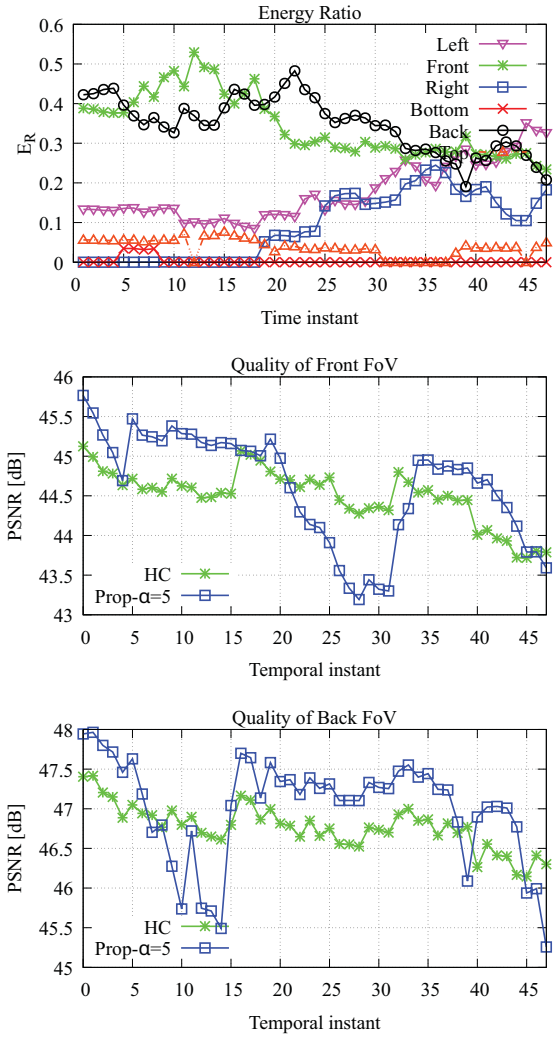

Figure 6. Temporal analysis of the visual attention and quality for the sequence Turtle.

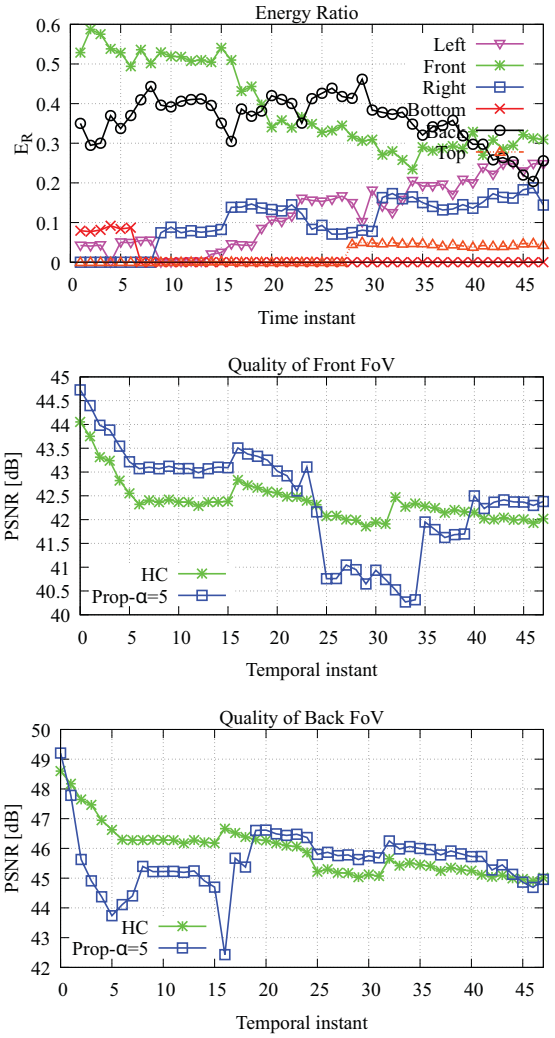

Figure 7. Temporal analysis of the visual attention and quality for the sequence PortoRiverside.

than the bit rate. This allows fair comparison between all cases because the total amount of visual data (i.e., number of pixels) is the same for all. These results show that the coding gains of the proposed method increase for higher bit stream sizes, especially for the Abbottsford sequence. Moreover, one can notice that only small differences exist between the homogeneous encoding (HC) and the reference VVC (Ref). This is because the proposed method improves the quality of the FoVs with higher visual attention, while reducing the overall bit rate. In other words, while maintaining lower overall bit rate, higher quality is achieved.

Finally, to demonstrate the effectiveness of the proposed adaptive perceptually-driven quantisation for each FoV along the time, Figures 6 and 7 show the temporal variations of the energy ratio (E_R) obtained from (1) and the FoV quality (PSNR). The figures show the PSNR obtained by the proposed encoding method and by homogeneous coding (HC), for two test sequences and two FoVs, i.e., front and back. These FoVs correspond to the ones with the highest visual attention, as shown by the value of E_R. An overall analysis of the results of both figures reveals that the HC method (green PSNR curves) maintains a quasi-homogeneous quality along the time, whereas the proposed method (Prop-$\alpha = 10$) presents PSNR variations closely related with the corresponding time evolution of E_R.

Comparing the two FoVs, one can notice opposite quality variations, as one increases (i.e., Back FoV) when the other decreases (i.e., Front FoV). This is clearly seen in Figure 6 between temporal instances 15 and 25. This occurs when the visual attention is shifted from one point of view to another, i.e., the FoV with the highest E_R changes from Front to Back FoV at time instant 20. The same behaviour is also

found in the PortoRiverside sequence (Figure 7) between the time instances 15 and 25, where the the FoV with highest E_R changes from Front to Back. The proposed method is able to account for this change and progressively increase the quality of the most important FoV. Simultaneously, in the FoVs where the visual attention decreases, a higher quantisation parameter is used, leading to quality and bit rate reduction.

V. Conclusions

In this paper a flexible coding approach for 360° video using the new versatile video encoding standard VVC was proposed. The proposed approach uses FoV to frame mapping to represent and encode omnidirectional video as multiple frames, each one corresponding to different FoVs. Non-uniform coding is further used in each FoV using perceptually-driven quantisation based on visual attention maps. It was shown that the proposed approach allows independent decoding of each FoV and enables bit stream truncation for reduced bandwidth utilisation in delivery services and applications. The results show that using visual attention maps one can decrease the overall bit rate without compromising the quality of the most relevant FoV. The consistent quality gains in the most relevant FoVs reveals that the proposed coding method is an efficient approach to enable partial delivery and decoding of 360° video streams with higher quality in the most relevant regions.

References

[1] J. Chen, Y. Ye, and S. Kim, "JVET-N1002: Algorithm description for Versatile Video Coding and Test Model 5 (VTM 5)," Joint Video Experts Team (JVET), 14th Meeting: Geneva, SW, Tech. Rep., Mar. 2019.

[2] W. Mason. (2015, Aug.) VR HMD Roundup: Technical Specs. [Online]. Available: https://uploadvr.com/vr-hmd-specs

[3] M. Zink, R. Sitaraman, and K. Nahrstedt, "Scalable 360° Video Stream Delivery: Challenges, Solutions, and Opportunities," *Proceedings of the IEEE*, vol. 107, no. 4, pp. 639–650, Apr. 2019.

[4] X. Zhang, X. Hu, L. Zhong, S. Shirmohammadi, and L. Zhang, "Cooperative tile-based 360-degree panoramic streaming in heterogeneous network using scalable video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[5] R. Ghaznavi-Youvalari, A. Zare, A. Aminlou, M. M. Hannuksela, and M. Gabbouj, "Shared coded picture technique for tile-based viewport-adaptive streaming of omnidirectional video," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2018.

[6] T. C. Nguyen and J. Yun, "Predictive tile selection for 360-degree VR video streaming in bandwidth-limited networks," *IEEE Communications Letters*, vol. 22, no. 9, pp. 1858–1861, Sep. 2018.

[7] A. TaghaviNasrabadi, A. Mahzari, J. D. Beshay, and R. Prakash, "Adaptive 360-degree video streaming using layered video coding," in *IEEE Virtual Reality (VR)*. IEEE, Mar. 2017, pp. 347–348.

[8] D. Liu, P. An, R. Ma, W. Zhan, and L. Ai, "Scalable omnidirectional video coding for real-time virtual reality applications," *IEEE Access*, vol. 6, pp. 56 323–56 332, Oct. 2018.

[9] T. Biatek, J. Travers, P. Cabarat, and W. Hamidouche, "Backward compatible layered video coding for 360° video broadcast," in *Picture Coding Symposium (PCS)*, Jun. 2018, pp. 318–322.

[10] R. Ghaznavi-Youvalari, A. Zare, A. Aminlou, M. M. Hannuksela, and M. Gabbouj, "Shared coded picture technique for tile-based viewport-adaptive streaming of omnidirectional video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 10, pp. 3106–3120, Oct. 2019.

[11] X. Xiu, Y. He, Y. Ye, and B. Vishwanath, "An evaluation framework for 360-degree video compression," in *2017 IEEE Visual Communications and Image Processing (VCIP)*, Dec. 2017, pp. 1–4.

[12] H.-C. Lin, C.-Y. Li, J.-L. Lin, S.-K. Chang, and C.-C. Ju, "AHG8: An efficient compact layout for octahedron format, document JVET-D0142," Joint Video Experts Team (JVET), Chengdu, CN, Tech. Rep., Oct. 2016.

[13] M. Zhou, "AHG8: Icosahedral projection for 360-degree video content, document JVET-D0028," Joint Video Experts Team (JVET), Chengdu, CN, Tech. Rep., Oct. 2016.

[14] E. J. David, J. Gutiérrez, A. Coutrot, M. P. Da Silva, and P. L. Callet, "A dataset of head and eye movements for 360° videos," in *Proceedings of the 9th ACM Multimedia Systems Conference*, ser. MMSys '18. New York, NY, USA: ACM, Jun. 2018, pp. 432–437.

[15] ITU-T, "Recommendation P.910, Subjective video quality assessment methods for multimedia applications," Apr. 2008.

[16] JCT-VC, "VVC 5.1 reference software," May 2019. [Online]. Available: https://jvet.hhi.fraunhofer.de/

[17] P. Hanhart, J. Boyce, and K. Choi, "Algorithm descriptions of projection format conversion and video quality metrics in 360Lib (Version 9), document JVET-M1004," Joint Video Experts Team (JVET), Macau, CN, Tech. Rep., Jan. 2019.