VERSATILE VIDEO CODING OF 360° VIDEO USING ADAPTIVE RESOLUTION CHANGE

J. Carreira[1,2], Sergio M. M. de Faria[1,2], Luis M. N. Tavora[2], Antonio Navarro[1,3], Pedro A. Assuncao[1,2]

[1]Instituto de Telecomunicações, Portugal
[2] Polytechnic of Leiria / ESTG, [3]Universidade de Aveiro, Portugal
{*jcarreira, sergio.faria*}*@co.it.pt, luis.tavora@ipleiria.pt, antonio.navarro@av.it.pt, amado@co.it.pt*

ABSTRACT

Encoding 360° video with ultra high spatial resolution requires high bitrates to guarantee acceptable QoE in video delivery services. However, since in general the full Field-of-View (FoV), i.e., 360°, is not required at once, a great deal of bandwidth can be saved if only a limited FoV is delivered, according to content relevancy or/and user demand. This work addresses this problem using the concept of Adaptive Resolution Change (ARC) defined in the forthcoming Versatile Video Coding (VVC) standard, by dynamically mapping the full FoV into multiple video frames with different spatial resolutions. Those FoVs attracting more visual attention are encoded with higher resolution while the others are encoded with lower resolution, thus without compromising the visual quality and resolution of the most relevant regions. The simulation results show that the proposed adaptive coding scheme is able to deliver high quality video for the most relevant FoV at any time instant, achieving a maximum bitrate reduction of 37.2%.

Index Terms— Omnidirectional video coding, adaptive resolution change, visual attention-based FoV resolution.

1. INTRODUCTION

Omnidirectional video requires ultra high spatial resolution encoded at high bitrates to guarantee acceptable QoE in video delivery services. However, in general, multimedia users of omnidirectional video, e.g., applications using Head Mounted Display (HMD), only focus on a limited region of a whole spherical scene at a given time, thus the whole 360° Field-of-View (FoV) does not need to be present all the time. Therefore, adaptive video coding schemes can be used to save a great deal of transmission bandwidth and decoding resources, by providing flexible selection and coding of specific FoVs. The common encoding framework for 360° video uses a planar projection of 360° ×180° video content from a 3D sphere to a 2D plane, which is defined by the so-called projection mapping function [1]. In the case of the Cube-Map Projection (CMP), each partition is a cube face that corresponds 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, easing the introduction of streaming adaptation schemes.

To cope with the huge amount of data necessary to represent the full 360° video, new highly efficient compressed formats are under development, such as, the forthcoming Versatile Video Coding (VVC) that is under development by JVET group [2]. The VVC standard aims at providing not only twice the coding efficiency of its predecessors, but also high-level syntax to support stream adaptation to different content and delivery constraints. One of such adaptive concepts is the Adaptive Resolution Change (ARC) [3], which enables dynamic variations of the spatial resolution without a refresh point. The high-level signalling defined for ARC [4] can be used to efficiently support adaptive streaming of 360° video with minimal overhead.

Different approaches have been proposed to deal with the problem of adaptive 360° video streaming. For instance, tile-based approaches map the whole 360° video scene into tiles that are independently encoded [5–7], allowing receivers to select any tile according to the viewing area or relevant viewport. 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). 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 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 bitrate 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 bitrate 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 base layer stream. However, these approaches take advantage of the scalable extension of the High Efficiency Video Coding (HEVC) standard [8–10], but this is not available in the most recent VVC standard. Thus, adaptive single-layer encoding schemes compliant with the VVC standard are yet to appear.

This work proposes an adaptive coding approach for VVC encoding of 360° video taking advantage of ARC. The spherical video represented in CMP format is mapped into video frames with different spatial resolutions, which are efficiently encoded using intra-FoV prediction. To achieve higher quality only in those FoVs that attract more visual attention, an adaptive FoV resolution scheme is proposed. This strategy relies on the visual attention maps to selectively encode with lower resolution those FoVs attracting less visual attention, while keeping the higher resolution of the most relevant. Nevertheless, the proposed scheme can also be used in applications using feedback from the observer (e.g., eye tracking with gaze prediction) to select the most relevant FoVs. Moreover, the proposed scheme takes advantage of the implicit temporal scalability present in the VVC to encode and deliver different FoVs in independent layers of different spatial resolutions, enabling flexible bit stream truncation and FoVs-based decoding.

This work was supported by Programa Operacional Regional do Centro, project ARoundVision CENTRO-01-0145-FEDER-030652 and by FCT/MCTES through national funds and when applicable co-funded EU funds under the project UIDB/EEA/50008/2020, Portugal.

The remainder of the paper is organised as follows. Section 2 presents a brief description of the ARC concept introduced in the VVC. Section 3 describes the proposed coding scheme using adaptive FoV resolution based on visual attention and performance evaluation is presented and discussed in Section 4. Finally Section 5 concludes this paper.

2. ADAPTIVE RESOLUTION CHANGE IN VVC

The requirements for future video coding implicitly defined novel techniques to be included in the VVC standard [11]. Among others, there was the requirement for adaptive streaming support by providing means of fast and seamless switching capability between representations with different properties, such as different spatial resolutions. The flexibility to modify the spatial resolution without inserting an Instantaneous Decoding Refresh (IDR) frame, not only opens the possibility to adapt the video data to dynamic channel conditions or user preference seamlessly, but also removes the beating effect caused by intra-coded pictures. This significantly increases the versatility in terms of bitrate adaptation and active speaker change in video telephony, adaptive stream switching and support for subpictures with variable resolution.

To achieve this goal, the VVC standard introduces the Reference Picture Resampling (RPR) [12] technique, where pictures in the Decoded Picture Buffer (DPB) can be stored at a different resolutions from the current frame and then resampled in order to be used for prediction. Figure 1-a shows an example where the current frame is predicted from two reference frames. While reference 0 is smaller than the current picture and is up-scaled, reference 1 is larger and requires downscale to be used for prediction.

Figure 1-b shows a simplified block diagram of the VVC encoder with support for ARC (extra functions highlighted in dark gray). The input video is selectively downscaled according to adaptation needs and then the conventional encoding procedure is applied. Then, for inter-frame prediction the reference frames stored in the DPB are scaled (up or down) as required. In order to support adaptive frame resolution, extra high-level signalling is carried out using the PPS NAL units identifying the frame spatial resolution [4, 13]. However, the existing motion compensation algorithm is similar, with changes only affecting the motion vector scaling and sub-pixel location derivations.

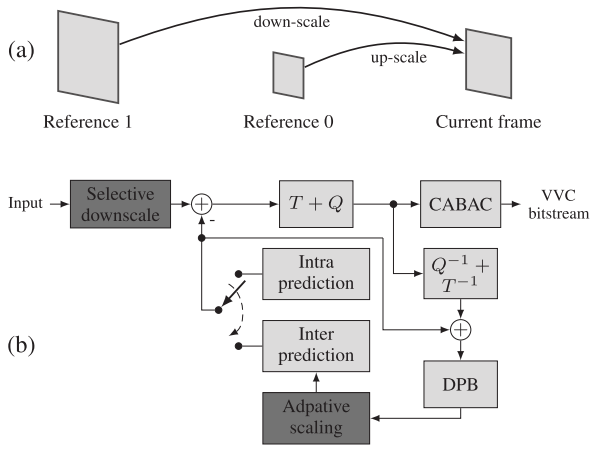

Fig. 1. Adaptive resolution change using RPR (a) and a simplified block diagram of the VVC encoder (b).

3. PROPOSED ADAPTIVE FOV RESOLUTION CODING

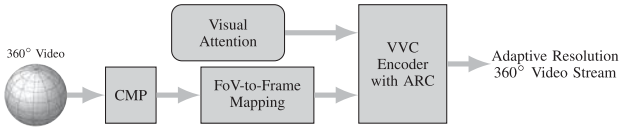

Fig. 2. Functional diagram of the proposed coding approach.

This section describes the proposed coding approach using adaptive FoV resolution for 360° video and the new ARC functionality of the VVC standard. 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 CMP, which results in a 2D mapping with six different FoVs. Subsequently, FoV-to-frame mapping converts FoVs into frames, resulting in temporal frame sequences for the input of the VVC encoder. This is a frame-based FoV representation of 360° video with granularity of 90° ×90°. This allows for simple extraction of a particular 90° FoV thus, by combining multiple frame-based FoVs, any arbitrary viewport can be extracted. Moreover, visual attention provides a measure of the instantaneous FoV relevance, for reducing the global encoding rate by adapting the spatial resolution of each FoV. The six cube faces of the CMP representation, corresponding to six FoVs are identified by their initials in Figure 3-a. Conventional coding schemes arrange all cube faces with equal spatial resolution into a rectangular matrix (i.e., one frame with 2×3 cube faces) to encode the whole 360° video, as illustrated in Figure 3-d. This is a rigid coding configuration because it does not allow flexibility to extract and decode only one or few FoVs. Alternatively, if each cube face is represented as a frame, as shown in Figure 3-b, then each FoV is simply encoded as a frame sequence with uniform spatial resolution.

Moreover, in the proposed method, the spatial resolution of each FoV can be adaptively selected as shown in Figure 3-c, enabled by the ARC and RPR functions of the VVC standard. In this example, the front FoV is coded at full resolution while the remaining ones are downscaled. The coding configuration for this case, where each cube face is encoded as a frame sequence with different spatial resolutions, is represented in Figure 3-e. As can be observed, temporal predictions are constrained to use only intra-FoV predictions, not allowing 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 extracting part of a single compressed stream

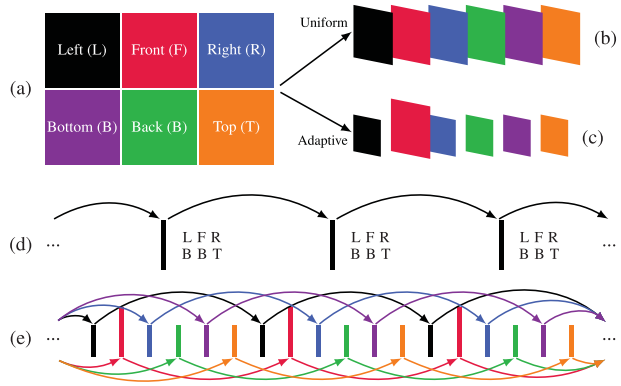

Fig. 3. Uniform and adaptive FoV resolution examples.

containing the full FoV video. In order to accomplish the necessary prediction structure, the reference frames are explicitly signalled in the stream, which is already supported by the VVC standard [14]. By using the above prediction scheme in combination with ARC, more efficient multiplexing of the different FoVs is achieved in comparison with a multi-encoder approach, both in terms of redundant signalling and decoder complexity.

3.1. Visual attention based FoV resolution

To deal with variable importance of the omnidirectional visual content, adaptive resolution encoding is proposed to provide higher resolution in those FoVs that attract more visual attention. For this purpose, visual attention maps are used as input parameters, which can be estimated either from the input 360° video signal or generated from user feedback. Regardless of their origin, the use of visual attention in 360° video is of utmost importance because the whole visual scene is not meant to be observed all at once, thus uniform encoding of all FoVs is not the most efficient approach.

An example of one attention map in the CMP format is shown in Figure 4, where the dark regions indicate regions with low attention and white regions represent regions with high attention. As can be seen in the figure, 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 rate allocation by adjusting the spatial resolution of each frame-based FoV (f) at time instant t based its the perceptual relevance, derived from the visual attention map M, as follows:

$$P_{f,t} = \frac{E_{f,t}}{\sum_{i=1}^{6} E_{i,t}}, \tag{1}$$

$$E_{f,t} = \sum_{j=t-2}^{t+2} \sum_{\mathbf{p}\in\Omega_f} \mid M_j(\mathbf{p}) \mid^2, \tag{2}$$

where $\mathbf{p}$ is the spatial position of attention values in the attention map of FoV f (Ω_f). In order to ease the impact of short-term transitions in the visual attention, in (2) the energy of a given FoV is determined using a sliding window of 5 frames ($-2 \leq t \leq 2$), approximately 100 milliseconds. The value of $P_{f,t}$ indicates how visual attention is distributed across the omnidirectional scene. In the proposed method the value of $P_{f,t}$ is used to rank the FoVs based on their visual relevance. Then, the FoV with the highest relevance is encoded at full resolution while the remaining ones are downscaled, in order to reduce the overall bitrate. Since the ARC definition supports several different spatial resolutions different downscale factors might be used. In this work only two levels have been used, corresponding to downscaling factors of $\alpha = 1.5$ and $\alpha = 2.0$. By using different spatial resolutions, the proposed method not only provides high resolution in the most visually important FoV but also enables switching without compromising the visual quality.

Summarising, the proposed coding method leads to an overall bitrate saving, by decreasing the spatial resolution of the FoVs with lower importance in terms of visual attention, while maintaining high resolution in the most relevant FoV that correspond to the highest users' attention.

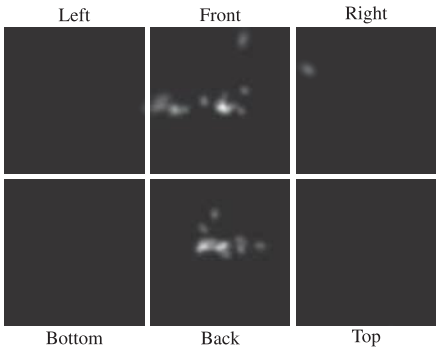

Fig. 4. Example of a visual attention map in the CMP format.

4. EXPERIMENTAL RESULTS

The performance evaluation was carried out using the CMP format as shown in Figure 3. The proposed method using adaptive FoV coding (Prop) was compared with the reference VVC encoding the full FoV (Figure 3-a) and with FoV-to-frame mapping using uniform spatial resolution (UR), i.e., the proposed method without ARC (Figure 3-b). The simulations studies were carried out using six sequences from the Salient360 database [15] shown in Table 1, which also provide ground-truth visual attention maps obtained from eye-tracking and head-mounted display information.

The omnidirectional video content has a spatial resolution of 3840×1920 in the ERP format, which was converted to CMP with resolution of 3456×2304. These sequences present different types of motion and texture complexity, as shown by the different Spatial Information (SI) and Temporal Information (TI) [16] parameters presented in Table 1. The VVC reference software, version 7.0 [17] was used with all coding modes enabled and a prediction structure using all B-frames with one reference frame (Low-delay).

Firstly, the efficiency of the proposed adaptive FoV resolution is evaluated, along the time of the sequence, by analysing the variation of the perceptual relevance ($P_{f,t}$) and the Bjøntgaard Delta PSNR (BD-PSNR) using VVC with full-FoV as reference. The perceptual relevance of each FoV (f), obtained using (1), is shown in Figure 5, for sequences Turtle and Cockpit. One can observe in the figure that Front and Back FoVs have the highest perceptual relevance and the FoV with the highest $P_{f,t}$ is not always the same along the time. The quality gains, measured by the BD-PSNR, achieved by the proposed method using $\alpha = 2$ (Prop) and uniform resolution coding (UR) are shown in Figure 6, for the same two FoVs. These results reveal that the UR method presents a near constant quality loss (i.e, negative values of BD-PSNR) in comparison with the reference VVC. This is due to the re-mapping of the omnidirectional scene into multiple frames, which increases the overhead and reduces the efficiency of inter-FoV prediction.

Table 1. Test sequences used in the experiments.

Sequence	SI	TI	Description
Abbottsford	117.8	1.21	Dorm room with one student talking with moderate motion
Cockpit	88.84	24.0	Cockpit footage with high camera vibrations
PortoRiverside	54.92	2.75	Riverside images with two standing guys and low overall motion
TeatroRegioTorino	90.18	5.07	Orchestra concert with high different people playing instruments
Touvet	91.50	5.27	High point recording of a garden with a castle
Turtle	52.17	18.3	Two women helping a turtle return to ocean

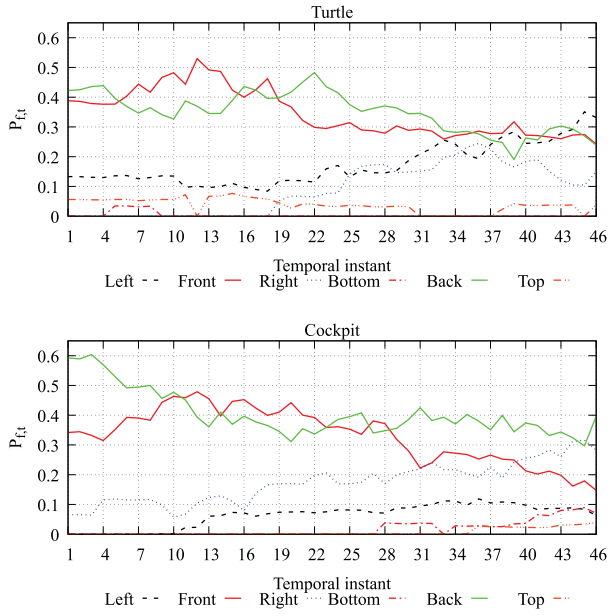

Fig. 5. Perceptual relevance of each FoV along the time.

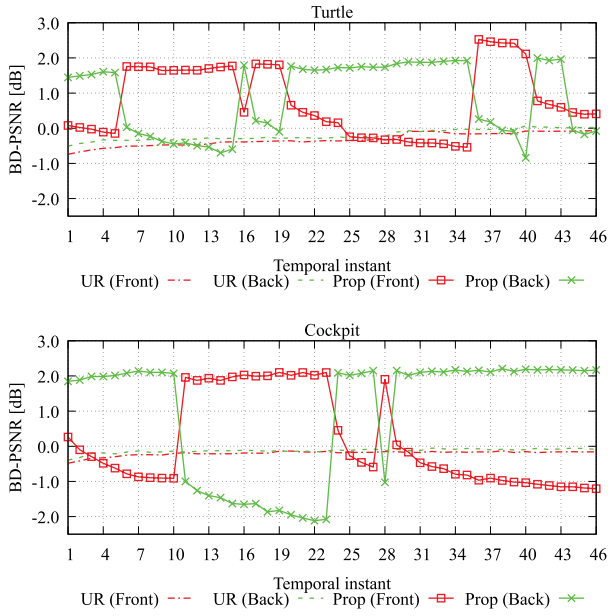

Fig. 6. BD-PSNR of methods UR and Prop for Front and Back FoVs.

Comparing the perceptual relevance (Figure 5) and the BD-PSNR (Figure 6) of the same sequence, it is noticeable that the BD-PSNR variations of the proposed method is correlated with the corresponding relevance ($P_{f,t}$). For any period of time where a particular FoV presents a greater value of $P_{f,t}$ the corresponding quality gain is greater than other FoVs. For instance, in the Turtle sequence between temporal instants 7 and 13, the Front FoV has higher $P_{f,t}$ and also higher quality gains (see top graphs of Figures 5 and 6). This is due to the adaptive FoV resolution encoding, which assigns lower resolution to the FoVs with lower $P_{f,t}$. Moreover, the results also reveal that the proposed method is able to adjust the FoV resolution whenever the visual attention switches from one FoV to another. This happens, for example, when there is an attention change in the Cockpit sequence at time instant 10 (see bottom graph of Figure 5), from the Back FoV to the Front FoV. As can be inferred from the bottom graph of Figure 6, the proposed method is able to follow the change in perceptual relevance by increasing the spatial resolution of the Front FoV (e.g., time instant 10). Simultaneously, lower quality is obtained in the FoVs where the perceptual relevance decreases, due to their lower resolution.

Table 2. Bjøntgaard Delta Rate of the FoV with the highest $P_{f,t}$.

Sequence	**BD-Rate-Single-FoV**		
	UR	Prop $\alpha = 1.5$	Prop $\alpha = 2$
Abbottsford	36.7	-6.12	-3.33
Cockpit	32.6	-25.3	-35.1
PortoRiverside	17.5	-26.2	-33.0
TeatroRegioTorino	-0.84	-28.5	-32.3
Touvet	19.6	-30.3	-37.2
Turtle	22.6	-29.5	-35.0
Average	**21.4**	**-24.3**	**-29.3**

The bitrate savings of the proposed method were also evaluated using the Bjøntgaard delta rate measured from the bitrate required to encode the whole 360° video and the quality of most relevant FoV (BD-Rate-Single-FoV). The full-FoV encoding was used as reference, as above. These results are shown in Table 2 for downscaling ratios of $\alpha = 1.5$ and $\alpha = 2$. These results confirm that uniform coding of all FoVs (UR) leads to higher bitrate than full-FoV encoding, for most sequences. By establishing a relation between the results of the UR case and the SI values in Table 1, it is possible to observe that sequences with higher spatial complexity generate higher bitrates. On contrary, the proposed method (Prop) is able to achieve bitrate saving for the same sequences. As shown in column 3 and 4 of Table 2, for different values of α one can confirm that higher spatial downscaling ratios ($e.g., \alpha = 2$), achieve higher bitrate savings. Overall the proposed method is able to save an average of 29.3% bitrate, when compared to the reference VVC encoding all-FoV without ARC.

5. CONCLUSIONS

In this paper an adaptive coding approach for 360° video using the new ARC concept of VVC is proposed. The proposed approach uses FoV-to-frame mapping to encode omnidirectional video as multiple frames, each one corresponding to a different FoV. Based on visual attention, a new method is proposed to dynamically adapt the image resolution, by selectively downscaling those FoVs with lower visual relevance. The experimental results show a significant reduction in the overall bitrate, 29.3% on average, and up to a maximum of 37.2%, maintaining high quality in the most relevant FoV. The consistent bitrate savings for the same quality of the most relevant FoVs reveals that the ARC mechanism defined for the VVC is an efficient approach to enable partial delivery and decoding of 360° video streams. Furthermore, the proposed approach also allows independent decoding of each FoV and enables stream truncation for reduced bandwidth utilisation in delivery services and applications.

6. REFERENCES

[1] 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, p. 1–4.

[2] J. Chen, Y. Ye, and S. Kim, "JVET-P2002: Algorithm description for Versatile Video Coding and Test Model 7 (VTM 7)," Joint Video Experts Team (JVET), 16th Meeting: Geneva, CH, Tech. Rep., Oct. 2019.

[3] Hendry, Y.-K. Wang, J. Chen, T. Davies, A. Fuldseth, Y.-C. Sun, T.-S. Chang, and J. Lou, "JVET-M0135: On adaptive resolution change (ARC) for VVC," Joint Video Experts Team (JVET), 13th Meeting: Marrakech, MA, Tech. Rep., Jan. 2019.

[4] P. Chen, T. Hellman, B. Heng, W. Wan, and M. Zhou, "JVET-O0204: AHG 8: Adaptive Resolution Change," Joint Video Experts Team (JVET), 15th Meeting: Gothenburg, SE, Tech. Rep., Jul. 2019.

[5] 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.

[6] 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*, p. 1–1, 2018.

[7] 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, p. 1858–1861, Sep. 2018.

[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, p. 56323–56332, 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, p. 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, p. 3106–3120, Oct. 2019.

[11] "Requirements for a future video coding standard v5," ISO/IEC JTC1/SC29/WG11 N17074, Tech. Rep., Jul. 2017.

[12] Y. Sanchez, R. Skupin, K. Suehring, and T. Schierl, "JVET-P0482: AHG8 On reference picture resampling," Joint Video Experts Team (JVET), 16th Meeting: Geneva, CH, Tech. Rep., Oct. 2019.

[13] J. Samuelsson, S. Deshpande, and A. Segall, "JVET-O0204: AHG8 Adaptive Resolution Change (ARC) High-Level Syntax (HLS)," Joint Video Experts Team (JVET), 15th Meeting: Gothenburg, SE, Tech. Rep., Jul. 2019.

[14] 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.

[15] 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, p. 432–437.

[16] ITU-T, "Recommendation p.910, subjective video quality assessment methods for multimedia applications," Apr. 2008.

[17] JCT-VC, "VVC 7.0 reference software," Nov. 2019. [Online]. Available: https://jvet.hhi.fraunhofer.de/