Rate-distortion optimized motion estimation for on-the-sphere compression of 360 videos

Alban Marie, Navid Mahmoudian Bidgoli, Thomas Maugey, Aline Roumy

RATE-DISTORTION OPTIMIZED MOTION ESTIMATION FOR ON-THE-SPHERE COMPRESSION OF 360 VIDEOS

Alban Marie, Navid Mahmoudian Bidgoli, Thomas Maugey, Aline Roumy

Inria, Univ Rennes, CNRS, IRISA, *Rennes, France*

ABSTRACT

On-the-sphere compression of omnidirectional videos is a very promising approach. First, it saves computational complexity as it avoids to project the sphere onto a 2D map, as classically done. Second, and more importantly, it allows to achieve a better rate-distortion tradeoff, since neither the visual data nor its domain of definition are distorted. In this paper, the on-the-sphere compression [1] for omnidirectional still images is extended to videos. We first propose a complete review of existing spherical motion models. Then we propose a new one called tangent-linear+t. We finally propose a rate-distortion optimized algorithm to locally choose the best motion model for efficient motion estimation/compensation. For that purpose, we additionally propose a finer search pattern, called spherical-uniform, for the motion parameters, which leads to a more accurate block prediction. The novel algorithm leads to rate-distortion gains compared to methods based on a unique motion model.

1. INTRODUCTION

A 360° video is the acquisition of a natural scene in all directions. There is a growing interest in using 360° visual data in several fields: learning [2], astrophysics [3], augmented/virtual reality [4], immersive image processing [5, 6], cultural heritage [7]. Despite its great interest, the full field of view representation comes at the price of a huge image size (typically 12K) requiring very large bandwidth. At the same time, the inherent spherical geometry makes traditional compression schemes inefficient [1]. Even worse, most of them consider the omnidirectional content as a planar 2D image (with the help of various projection mappings) which leads to severe problems such as non-regular sphere sampling, surface discontinuities and radial distortions.

In order to circumvent the drawbacks of sphere mapping, the spherical image can be directly processed on the sphere. This leads to two main questions: i) how to sample the sphere uniformly and 2) how to export the classical compression tools (blocking, transforms, etc.) onto a non-euclidean domain such as the sphere. The recent work in [1] has proposed a first solution for still image compression on the sphere, based on HEALPix sampling [8] and graph-based signal processing theory [9]. In a nutshell, the sphere is first pixelized uniformly, and using the hierarchical structure of HEALPix, the authors in [1] define "spherical blocks" on which they are able to compute inter-block predictions and graph-based transforms. They have demonstrated that compressing the 360° image directly one the sphere leads to better rate-distortion performance than the existing methods based on projection mapping.

In this paper, we propose to extend the "on-the-sphere" still image compression method [1] to videos. More precisely, we propose a novel motion estimation/compensation method for omnidirectional format. The difficulties are twofold: first, the motion needs to be characterized for an image defined on a non-euclidean grid; second, the motion of an object requires 6 *degrees-of-freedom* (DoF), not all of which can be sent to the decoder, as it would significantly impact the compression rate. Motion estimation/compensation have already been studied for omnidirectional videos. First, they were developed for projected 360° data (either with equirectangular [10, 11, 12, 13], cubemap [14, 15] or others [16, 17]). Second, they propose different dimensionality reduction methods of the motion characterization, leading to various pros and cons. In this paper, the proposed method combines the pros of both dimensionality reduction methods, thanks to a rate-distortion optimization. Second, we perform the motion compensation directly on the sphere. For doing that, we first formalize motion models for omnidirectional data and propose a high level analysis of existing 3D transformation algorithms in this general framework (Section 2). This analysis allows us to propose a novel motion model (called tangent-linear+t) to have a more complete set of possible motion models. Then, we describe the proposed rate-distortion optimized motion estimation for spherical data (Section 3). Finally, the benefits of the proposed method is evidenced in the experimental section (Section 4). To have a fair comparison between different motion compensation methods, since all of the existing methods are projection-independent and work with any type of projections, in our experiment, we chose a projection that we believe is the fairest (*i.e.*, HEALPix [8]), meaning that the distribution of pixels on the sphere surface is uniform. To neutralize the effect of different search patterns introduced in different models, we also propose a novel search pattern and use this pattern for all methods.

This work was supported by the Cominlabs excellence laboratory with funding from the French National Research Agency (ANR-10-LABX-07-01).

2. MOTION MODELS

2.1. Definitions

In most video coders, the image at time t, denoted by I_t, is divided into *blocks*. In [1], the latter are built using the hierarchical HEALPix sampling property [8]. Let $\mathcal{B}$ denotes one of these blocks and let $\mathbf{p} \in \mathcal{B}$ be one of the pixels inside it.

The *motion compensation* operation consists in estimating the pixels of a block with some well-chosen pixels of the image at time $t-k$ (with $k \geq 1$). More formally,

$$\forall \mathbf{p} \in \mathcal{B}, \quad \widetilde{I}_t(\mathbf{p}) = I_{t-k}(\tau_{\mathcal{H}}(\mathbf{p})) \tag{1}$$

where $\widetilde{I}_t$ is the motion compensated image at time t (an estimation of I_t), and $\tau_{\mathcal{H}}$ is a function whose output is a pixel coordinate on the sphere, and where $\mathcal{H}$ are motion parameters. Note that unlike the classical methods on a 2D grid, where the displacement is represented by a pixel shift $\mathbf{p}+\delta$, we denote the displacement with the pixel coordinate $\tau_{\mathcal{H}}$ directly because $\mathbf{p}+\delta$ is not defined on the non-euclidean HEALPix domain.

The *motion estimation* consists in finding, for each block, the best motion parameters $\mathcal{H}$, *i.e.*,

$$\mathcal{H} = \text{argmin}_{\mathcal{H}'} \sum_{\mathbf{p}\in\mathcal{B}} |I_t(\mathbf{p}) - I_{t-k}(\tau_{\mathcal{H}'}(\mathbf{p}))| \,. \tag{2}$$

In the equations above, the function τ depends on the *motion model* that is adopted. This model sets the relation between the coordinate shift of a pixel between two time instants and the motion of the corresponding object in the 3D world. This model rules two important aspects, namely,

- the *object shape*: the shape in the 3D world of the object depicted by the pixels of the block $\mathcal{B}$ (see Fig.1)
- the *object motion*: the type of motion done by the object in the 3D world (see Fig.2)

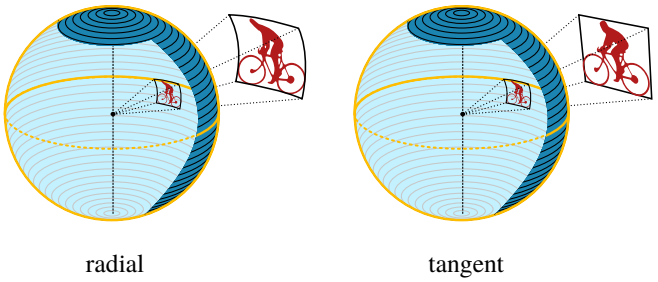
2.2. Object shape modeling

The color of a pixel $\mathbf{p}$ corresponds to the color of a 3D point that is on the line $(O\mathbf{p})$, where O is the center of the sphere. The position of the object that is actually captured at pixel position $\mathbf{p}$ can thus be anywhere on this line. The object shape modeling states how the pixels of a block $\mathcal{B}$ are mapped into the 3D world. We categorize them into two types: radial (*e.g.,* [12]) and tangent (*e.g.,* [11]).

The *radial* object shape model maps each pixel $\mathbf{p}$ to a 3D point $\mathbf{q}$ that is on the line $(O\mathbf{p})$ at a constant distance ρ. The exact value of ρ does not have any importance, since all pixels of the block are mapped with the same ρ:

$$\forall \mathbf{p} \in \mathcal{B}, \quad \mathbf{q} = \rho \mathbf{p}, \tag{3}$$

considering that $\mathbf{p}$ is expressed in terms of a 3D coordinate on the sphere (*i.e.*, $\mathbf{p} = [p_x, p_y, p_z]^\top$). This simple model gives a curved shape of the object as depicted in Fig.1.



radial tangent

Fig. 1: The pixels of the spherical blocks are considered as point in the 3D world. Two approaches exist in the literature: radial (*e.g.,* [12]) and tangent (*e.g.,* [11]).

The *tangent* object shape model maps each pixel $\mathbf{p}$ to a point $\mathbf{q}$ belonging to the tangent plane that is orthogonal to the line $(O\mathbf{p}_0)$ and at a distance of h from O, where $\mathbf{p}_0$ is the center of the block $\mathcal{B}$. The point $\mathbf{q}$ is given by:

$$\forall \mathbf{p} \in \mathcal{B}, \quad \mathbf{q} = h \begin{bmatrix} p'_x/p'_z \\ p'_y/p'_z \\ 1 \end{bmatrix} \quad \text{with } \mathbf{p}' = \mathbf{R}_{\mathbf{p}_0}\mathbf{p}, \tag{4}$$

where $\mathbf{R}_{\mathbf{p}_0}$ is the rotation matrix to point the z axis towards the pixel $\mathbf{p}_0$. As the radial shape, the distance h does not have impact. This model gives a planar shape to the object as depicted in Fig.1.

2.3. Object motion modeling

Once all pixels of a block have been modeled as points in the 3D world, the motion model then consists in displacing these points. This is done using the rigid transformation model:

$$\mathbf{q}_{\text{motion}} = \mathbf{R}\mathbf{q} + \mathbf{t}, \tag{5}$$

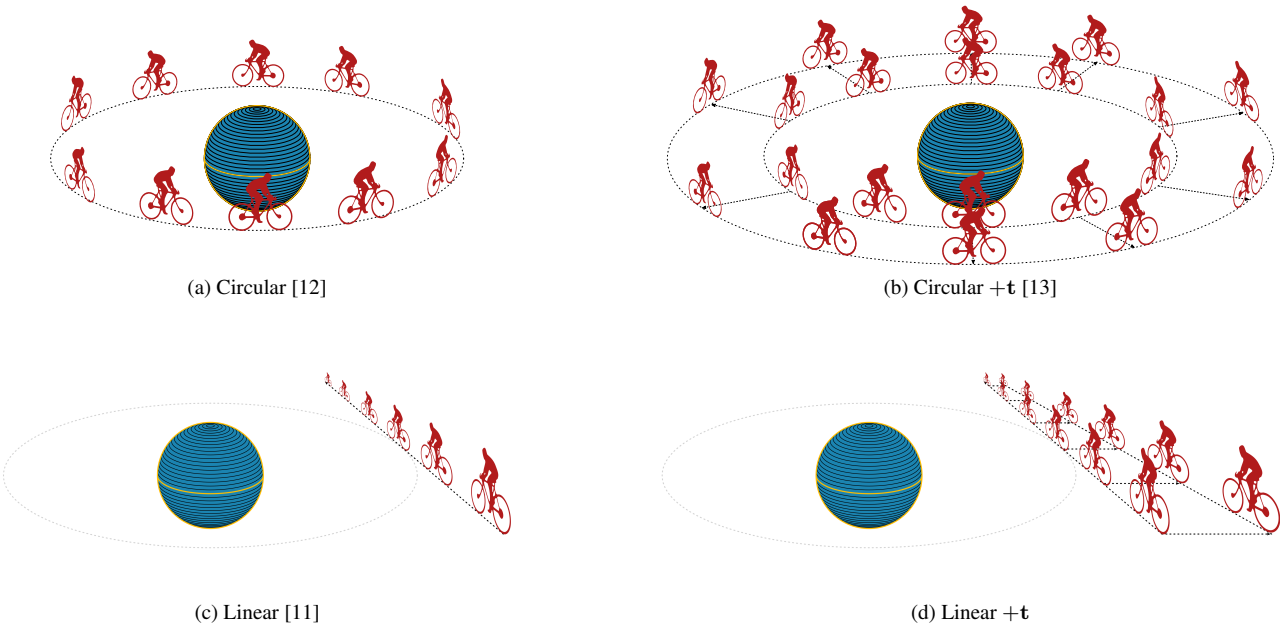
where $\mathbf{R}$ is a rotation matrix driven by three rotation angles, $(\theta_{\text{yaw}}, \theta_{\text{pitch}}, \theta_{\text{roll}})$, and where $\mathbf{t} = [t_x, t_y, t_z]^\top$ is a translation vector.

The parameters of the rigid transformation are denoted $\mathcal{H}$ (already introduced in (1) and (2)). Depending on the type of motion model, the transformation in (5) can be only rotational (*i.e.*, $\mathbf{t} = \mathbf{0}$), or only translational (*i.e.*, $\mathbf{R} = \mathbf{I}_3$), or more generally constrained to fewer *degrees of freedom* (less than $dim(\mathcal{H})$). Indeed, the parameters of the transformation have to be transmitted to the decoder. There is naturally a trade-off between the accuracy of the predicted block and the number of parameters to transmit.

Finally, the position on the sphere of the pixel corresponding to the displaced object is given by

$$\tau_{\mathcal{H}}(\mathbf{p}) = \frac{\mathbf{q}_{\text{motion}}}{||\mathbf{q}_{\text{motion}}||}. \tag{6}$$

There exist two families of methods for the object motion: *circular* and *linear*.



(a) Circular [12]

(b) Circular $+\mathbf{t}$ [13]

(c) Linear [11]

(d) Linear $+\mathbf{t}$

Fig. 2: The 4 object motion models considered in this study.

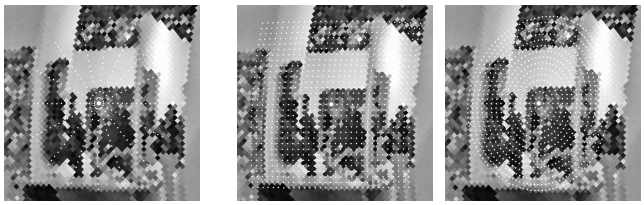
The *circular* motion consists in rotation around the center of the spherical camera. In the 2 DoF model in [12], the circular motion considers that the rotation is performed with the same radius, *i.e.,* $\mathbf{t} = \mathbf{0}$ as illustrated in Fig. 2a. In the 3 DoF circular model [13], a translation with an additional DoF α is allowed. It corresponds to a change of radius in the rotational motion, *i.e,* $\mathbf{t} = \alpha\mathbf{R}\mathbf{q}$ as illustrated in Fig. 2b.

The *linear* motion consists in translation in the 2D plane that is tangent to the sphere (of radius h in (4)) at the center of the block to be displaced. In the 2DoF model [11], $\mathbf{R} = \mathbf{I}_3$ and the translation is characterized by a horizontal and a vertical displacement in the tangent plane, see Fig. 2c. For the sake of completeness, we consider in this paper an extended linear motion model where a third DoF β is allowed in the translation, namely $\mathbf{t} = \beta\mathbf{q}_0$ the direction orthogonal to the tangent plane, as illustrated in Fig. 2d.

3. RD-DRIVEN MOTION ESTIMATION

When looking for the best motion parameters $\mathcal{H}$ for a block $\mathcal{B}$ as in (2), the underlying question that is raised is: *what is the best motion model?* While the state-of-the-art papers [12, 11, 13] consider a unique motion model for all the blocks, we propose here to optimally choose among different motion models based on a rate-distortion criterion. This is motivated by the fact that objects' motion in the real world are of different types (see Fig. 2), and considering one for the whole frame is naturally suboptimal.

Since the different models may require a different amount of bits to describe the motion parameters (*e.g.,* depending on the DoF), we rewrite (2) as the minimization of a rate term



(a) (b) (c)

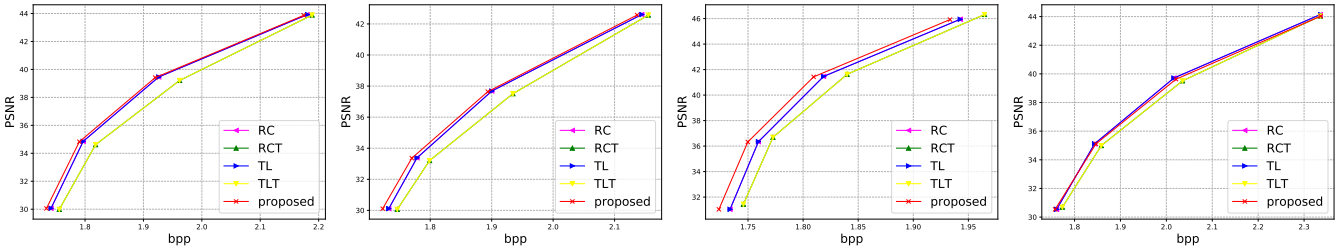
Fig. 3: Search patterns from (a) [12], (b) [11], (c) our SU.

under distortion constraints to take into account the cost to encode the model parameters. The proposed motion estimation algorithm is the following:

$$\mathcal{H} = \text{argmin}_{\mathcal{H}' \in \{H_1,\ldots,H_M\}} R(\mathcal{H}') + R(\tilde{I}_t) \quad \text{s.t.} D < D_{\max}, \tag{7}$$

where $R(\mathcal{H}')$ is the rate needed to code the parameter $\mathcal{H}$ (coded with fixed-length code in this study) and $R(\tilde{I}_t)$ is the rate needed to code the quantized residual $Q(I_t(\mathcal{B}) - \tilde{I}_t(\mathcal{B}))$ under the constraint that the resulting distortion is lower than $D_{\max}$. In that equation, the H_m are the search spaces corresponding to the different motion models. In our algorithm, we consider the following motion models:

- *Radial-circular (RC)* [12]: the object shape model is radial and the considered motion is circular. In that case $dim(\mathcal{H}) = 2$.
- *Radial-circular+t (RCT)* [13]: the object shape model is radial and the displacement is circular with a one degree of freedom translation that is allowed. In that case $dim(\mathcal{H}) = 3$.



(a) *Touvet* (b) *Turtle* (c) *Teatro* (d) *UnderwaterPark*

Fig. 4: Rate-distortion comparison. The BD-rate gains of our proposed approach with respect to the best of the single-model approaches are respectively: -0.28%, -0.35%, -0.51% and 0.25% for the four test images.

- *Tangent-linear (TL)* [11]: the object shape model is tangent and only 2D translations in the tangent plane is allowed. In that case $dim(\mathcal{H}) = 2$.
- *Tangent-linear+t (TLT)* (proposed here for the completeness of the study): the object shape model is tangent and 3D translations are allowed. In that case $dim(\mathcal{H}) = 3$.

For the motion models with 2 DoF, the search space can be described by a set of points on the sphere corresponding to candidate positions for the center of motion compensated block. In order to compare all the models fairly, we propose a common search pattern, called *spherical-uniform (SU)* that is presented in Fig. 3. Contrary to the patterns proposed in [12, 11], the proposed pattern has a search window (*i.e.,* the external boundary of the search space) that is circular, and that is uniformly sampled. It consists of a set of concentric circles for which the number of candidate points at each circle is proportional to the radius. For the model with 3 DoF, the same search pattern is used, while for each central block position different values for the translation parameter that are tested.

4. EXPERIMENTAL COMPARISON

In this section, we compare all motion compensation methods. For a fair comparison, the same pixelization, namely HEALPix [8] is used in all methods. This has several advantages. First, the sphere is uniformly sampled on the sphere so neighboring pixels/blocks represent more informative data than other projections. Second, since it is uniform, the rate-distortion optimization for mode selection in motion estimation is more accurate. Third, due to the uniformity of the pixelization, the PSNR calculation in rate-distortion evaluation is closer to the Spherical PSNR [18].

Experiments have been conducted on four 360° video sequences from [19] entitled *Touvet*, *Turtle*, *Teatro*, *UnderwaterPark*. We consider a GOP size of 4. The RD comparison is shown in Fig. 4. We see that the proposed RD-driven approach performs better than existing motion estimation techniques consisting in choosing a unique motion model for the whole frame. This is further confirmed by the vector field label shown in Fig. 5. In this image, we show, for each block, which method has been chosen by our proposed approach. We can clearly see the heterogeneity of the motion vector field, demonstrating the interest of choosing the best motion model locally. We also show the proportion of the chosen motion models at low and high bitrate in Table 1. This demonstrates that complex motion models (with DoF of 3) are chosen more often at high bitrate, *i.e.,* when the additional parameter bit overhead becomes acceptable compared to the total bit budget.

Table 1: Percentage of blocks for which each motion model is chosen for *UnderwaterPark* sequence.

	no motion	RC	RCT	TL	TLT
Low bitrate	82.01	0.06	0.04	17.45	0.44
High bitrate	43.01	3.4	3.12	32.99	17.48

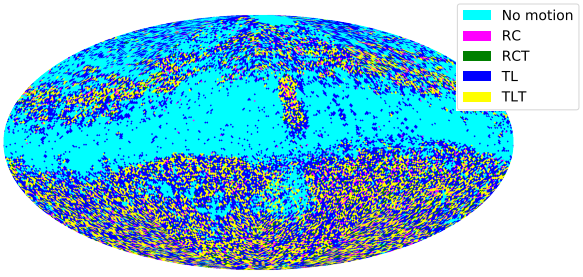


Fig. 5: Chosen motion model for each block for *UnderwaterPark* sequence.

5. CONCLUSION

This paper proposed a motion compensation algorithm to achieve on-the-sphere video compression. In particular, a novel motion model and a novel search pattern have been proposed to characterize a large set of motions in omnidirectional data. To avoid increasing the cost to encode the motion model, we also proposed a rate-distortion optimization of the motion estimation.

6. REFERENCES

[1] N. Mahmoudian Bidgoli, T. Maugey, and A. Roumy, "Intra-coding of 360-degree images on the sphere," in *2019 Picture Coding Symposium (PCS)*, 2019, pp. 1–5.

[2] Anke Berns, José Miguel Mota, Ivan Ruiz-Rube, and Juan Manuel Dodero, "Exploring the potential of a 360 video application for foreign language learning," in *Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality*, 2018, pp. 776–780.

[3] Christopher MP Russell, "360-degree videos: a new visualization technique for astrophysical simulations," *Proceedings of the International Astronomical Union*, vol. 12, no. S329, pp. 366–368, 2016.

[4] Xing Liu, Qingyang Xiao, Vijay Gopalakrishnan, Bo Han, Feng Qian, and Matteo Varvello, "360 innovations for panoramic video streaming," in *Proceedings of the 16th ACM Workshop on Hot Topics in Networks*, 2017, pp. 50–56.

[5] Marek Domański, Olgierd Stankiewicz, Krzysztof Wegner, and Tomasz Grajek, "Immersive visual mediampegi: 360 video, virtual navigation and beyond," in *2017 International Conference on Systems, Signals and Image Processing (IWSSIP)*. IEEE, 2017, pp. 1–9.

[6] Navid MahmoudianBidgoli, Thomas Maugey, and Aline Roumy, "Fine granularity access in interactive compression of 360-degree images based on rate-adaptive channel codes," *IEEE Transactions on Multimedia*, 2020.

[7] Lemonia Argyriou, Daphne Economou, and Vassiliki Bouki, "360-degree interactive video application for cultural heritage education," in *3rd Annual International Conference of the Immersive Learning Research Network*. Verlag der Technischen Universität Graz, 2017.

[8] K. M. Górski, E. Hivon, A. J. Banday, B. D. Wandelt, F. K. Hansen, M. Reinecke, and M. Bartelmann, "HEALPix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere," *The Astrophysical Journal*, vol. 622, no. 2, pp. 759–771, apr 2005.

[9] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 83–98, May 2013.

[10] I. Tosic, I. Bogdanova, P. Frossard, and P. Vandergheynst, "Multiresolution motion estimation for omnidirectional images," in *2005 13th European Signal Processing Conference*, 2005, pp. 1–4.

[11] Francesca De Simone, Pascal Frossard, Neil Birkbeck, and Balu Adsumilli, "Deformable block-based motion estimation in omnidirectional image sequences," in *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2017, pp. 1–6.

[12] Bharath Vishwanath, Tejaswi Nanjundaswamy, and Kenneth Rose, "Rotational motion model for temporal prediction in 360 video coding," in *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2017, pp. 1–6.

[13] Yefei Wang, Dong Liu, Siwei Ma, Feng Wu, and Wen Gao, "Spherical coordinates transform-based motion model for panoramic video coding," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 1, pp. 98–109, 2019.

[14] Li Li, Zhu Li, Madhukar Budagavi, and Houqiang Li, "Projection based advanced motion model for cubic mapping for 360-degree video," in *2017 IEEE International Conference on Image Processing (ICIP)*, Beijing, Sept. 2017, pp. 1427–1431, IEEE.

[15] Muhammed Zeyd Coban, Geert Van der Auwera, Fnu HENDRY, and Marta Karczewicz, "Motion compensation for cubemap packed frames," Aug. 2019.

[16] Li Li, Zhu Li, Xiang Ma, Haitao Yang, and Houqiang Li, "Advanced spherical motion model and local padding for 360 video compression," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2342–2356, 2018.

[17] Bharath Vishwanath and Kenneth Rose, "Spherical video coding with geometry and region adaptive transform domain temporal prediction," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 2043–2047.

[18] 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*, 2015, pp. 31–36.

[19] Erwan J David, Jesús Gutiérrez, Antoine Coutrot, Matthieu Perreira Da Silva, and Patrick Le Callet, "A dataset of head and eye movements for 360 videos," in *Proceedings of the 9th ACM Multimedia Systems Conference*, 2018, pp. 432–437.