Geodesic Disparity Compensation for Inter-View Prediction in VR180

Kruthika Koratti Sivakumar, Bharath Vishwanath, Kenneth Rose
Department of Electrical and Computer Engineering
University of California, Santa Barbara
Santa Barbara, CA, 93106
{ kruthika, bharathvishwanath, kenrose } @ucsb.edu

***Abstract*—The VR180 format is gaining considerable traction among the various promising immersive multimedia formats that will arguably dominate future multimedia consumption applications. VR180 enables stereo viewing of a hemisphere about the user. The increased field of view and the stereo setting result in extensive volumes of data that strongly motivate the pursuit of novel efficient compression tools tailored to this format. This paper's focus is on the critical inter-view prediction module that exploits correlations between camera views. Existing approaches mainly consist of projection to a plane where traditional multi-view coders are applied, and disparity compensation employs simple block translation in the plane. However, warping due to the projection renders such compensation highly suboptimal. The proposed approach circumvents this shortcoming by performing *geodesic disparity compensation on the sphere*. It leverages the observation that, as an observer moves from one view point to the other, all points on surrounding objects are perceived to move along respective geodesics on the sphere, which all intersect at the two points where the axis connecting the two view points pierces the sphere. Thus, the proposed method performs inter-view prediction on the sphere by moving pixels along their predefined respective geodesics, and accurately captures the perceived deformations. Experimental results show significant bit-rate savings and evidence the efficacy of the proposed approach.**

***Index Terms*—inter-view prediction, VR180, immersive video, HEVC, virtual reality**

I. Introduction

Immersive virtual reality technologies have been "booming" in recent years with diverse application fields ranging from automotive through healthcare, education to the entertainment industries. Video formats in these applications either capture a 360-degree field-of-view, which enables users to view in all directions, or a 180-degree field-of-view, which captures the front half of the user's surroundings, but in both cases do not capture natural depth information. This unnatural aspect of the video is undesirable and may be a contributing factor in the so-called virtual reality sickness, with symptoms ranging from user fatigue to nausea. Consequently, users are drawn towards 3D (stereoscopic) videos, that resemble the binocular human vision system, as they offer a sense of depth perception by capturing two synchronous views of the same scene and provide better user experience. In this paper, we focus on Google's 3D-180° video format known as VR180 [1]. In VR180, two cameras are positioned with a fixed inter-pupillary distance between them, so that they synchronously capture a 180° field of view of the surroundings. VR180 generates large amounts of data due to its increased field of view and its stereo setting. Thus, there is compelling strong motivation for the development of efficient compression tools tailored to this scenario.

Most existing approaches for stereo and multi-view coding of immersive content project the spherical videos from camera views onto planes via one of several known projection geometries such as equirectangualar, or cube-map projections [2], to be processed by current (2D) multi-view coders (see e.g., [3]–[5]). For conventional 2D videos, the multi-view extension of High Efficiency Video Coding (HEVC) [6] employs inter-view prediction (also referred to as disparity compensated prediction) to exploit correlations between camera views. Specifically, reconstructed frames of adjacent view/s are used as additional reference frames, enabling inter-view prediction along with standard temporal prediction. Both temporal and inter-view predictions employ a simple block translation model to perform motion or disparity compensation. However, projection from a spherical video induces unintended warping that severely compromises the efficacy of the translation model in both (temporal and inter-view) prediction scenarios. Moreover, the motion vectors in the projected domain have no natural/physical meaning. Motivated by recognition of this shortcoming, recent work proposed approaches to model motion on the sphere for temporal prediction, albeit in the context of monocular spherical videos [7]–[9]. To the best of our knowledge, there is no work that has focused on overcoming this drawback in the setting of disparity compensation for inter-view prediction, despite the realization [6] that inter-view prediction is critical to efficient stereo and multi-view coding. Thus, there is a strong motivation for a physically sound model that can accurately capture the perceived disparity on the sphere as the observer translates between view points.

The proposed method is inspired by our work in [8], where we propose a geodesic motion model for temporal prediction in monocular spherical videos dominated by camera motion. The approach builds on a straightforward but highly useful observation that straight lines in 3D space map to geodesics on the sphere. In this paper, we extend the model to perform

This work was partly supported by Google Inc.

geodesic inter-view prediction that can effectively and accurately capture the disparity on the sphere as we move from one camera view to the other in VR180. The cameras capture the same scene synchronously, implying that all the surrounding objects, captured at the same time instant, are spatially 'stationary' between the views. Thus, the difference in apparent location between views stems solely from the separation of the cameras. In this setting, as we move between the view points, surrounding stationary objects exhibit *relative motion* along straight lines that are parallel to the axis connecting the view points (cameras), and this 3D motion maps to motion along geodesics on the sphere. All these geodesics intersect at the points where the inter-camera axis pierces the sphere. This observation opens the door to performing optimal disparity compensated prediction on the sphere. Specifically, we map a given block of pixels in the current view onto the sphere, move each pixel along its respective geodesic and map the geodesic-translated pixels back to the reference frame of the adjacent view to derive the prediction signal. It is worthwhile to note that in our earlier work in the context of a moving 360^o camera, we further had to account for (unknown) actual object motion between frames. However, in the context of VR180 inter-view prediction, where surrounding objects are stationary in space between the two synchronous frames, the geodesic model is unbeatable in that it perfectly captures the perceived disparity of the objects across the views.

Additional benefits, beyond the accurate capture of disparity on the sphere, include the fact that the disparity vectors are now one-dimensional, as we only have to signal the amount of geodesic translation to the decoder. This amounts to significant bit-rate savings in side-information. Moreover, the model operates entirely on the sphere, making it agnostic of the projection geometry and easily applicable to any new projection format. Experimental results demonstrate significant bit-rate savings validating the above mentioned benefits.

The rest of the paper is organized as follows. Section II provides an overview of equirectangular (ERP) and equatorial cylindrical projection (ECP). The proposed approach is described in section III. Section IV summarizes the experimental results, followed by conclusions in section V.

II. Overview of two projection formats

A. Equirectangular Projection (ERP)

This projection format is depicted in Fig. 1. It maps meridians to vertical straight lines of uniform sampling density in the projected plane. Latitudes are mapped to horizontal straight lines. Thus, any point p on the sphere, with θ denoting the elevation (pitch) and ϕ denoting the azimuth (yaw), is mapped to the position obtained on the 2D grid as the intersection of the vertical and horizontal lines corresponding to its latitude and longitude on the sphere. ERP maintains constant vertical sampling density, whereas the horizontal sampling density increases as we move towards the poles. The mathematical foundation and practical procedures for ERP mappings between sphere and plane are available in [10].

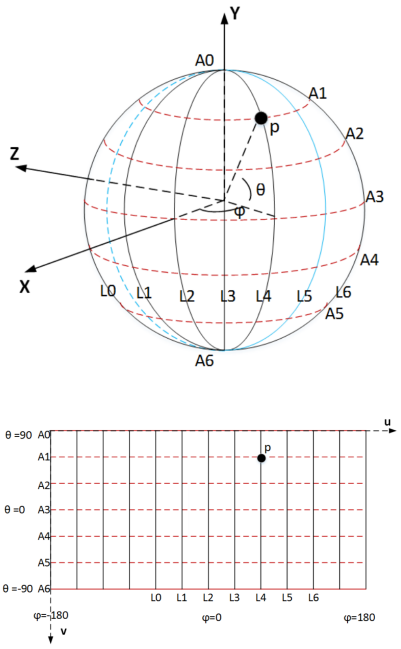

Fig. 1: ERP Sampling: on the sphere (top) and in two-dimensions (bottom)

B. Equatorial Cylindrical Projection (ECP)

ECP is obtained by projecting the equatorial region of the sphere using the Lambert cylindrical equal-area projection [2] and the two polar regions of the sphere onto square faces. The area of the sphere covered by the equatorial region and each polar region are 2/3 and 1/6 of the total sphere area, respectively. Geometry mappings for ECP are more involved and their exact definitions have been skipped here for brevity. A detailed description is provided in [11].

III. Proposed approach for inter-view prediction

We first illustrate the perceived disparity on the sphere as the observer moves from one view point (camera) to the other, and then propose a disparity compensation procedure for VR180.

A. Observation: the Perceived Disparity on the Sphere

Consider the setup shown in Fig. 2, with the left camera of a VR180 rig located at point A and the right camera located at point B. Although the cameras in VR180 capture hemispherical views, for ease of illustration, we include full spherical views centered at A and B. Point P in space is seen by the left camera (located at point A) as its projection point S on the sphere. At a given time instant, let us move the "observer" in the direction of the vector v, towards the right camera (located at point B). As the observer moves, point P is perceived as moving in the opposite direction, resulting in a displacement to the new point P′, which is seen by the right camera as its projection point S′ on the sphere. Thus, the perceived displacement P-P′ is seen as the projected arc S-S′ on the sphere, which is observed to be a portion of a geodesic that connects the two points of intersection of an axis, along

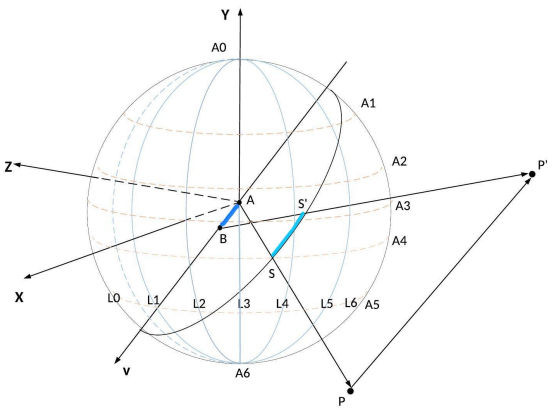

Fig. 2: Perceived disparity due to positioning of the left and right cameras

vector v, with the sphere. To re-emphasize, at a given time instant, as we translate from one view to the other, we make the following observations:

- All surrounding objects are effectively stationary in space.
- A projected point S in one spherical view is perceived to have displaced to projected point S′ in the other view, and the trajectory is along a well defined geodesic passing through point S.
- All geodesics (for perceived trajectories), corresponding to different points in space, intersect at the two points where the axis (along vector v) connecting the two cameras intersects the sphere.

This observation paves the way for an optimal inter-view prediction paradigm. Given the reconstruction from one camera view, we can predict the other camera view by a geodesic translation of pixels on the sphere along their respective geodesics. Thus, we next introduce the proposed geodesic model for disparity compensated prediction.

B. Geodesic Model for Inter-view Prediction

Existing methods for inter-view prediction of stereoscopic videos project the videos to planes using one of the different projection geometries, perform disparity compensation in the projected plane, before finally performing inverse projection back to the spherical domain. As observed earlier, this results in suboptimal performance since disparity compensation in the projected domain fails to account for perceived disparity on the sphere. The proposed geodesic motion model determines the exact relationship between adjacent views based on the camera geometry. Given the reconstructions of one camera view, we perform disparity compensated prediction for the other camera view by the following steps:

- Sphere Mapping: Given a block of pixels in the projected domain that is to be inter-view predicted, map the pixels in the block onto the sphere. This steps enables capturing the disparity directly on the sphere and renders the method completely agnostic of the projection geometry. For ease of presentation (and calculation), replace the original spherical coordinate system of the video with spherical coordinates with respect to a polar axis aligned with vector v, i.e., the axis connecting the two cameras. Let (ϕ_{ij}, θ_{ij}) be the spherical coordinates with respect to vector v as the polar axis.
- Geodesic translation: Given a disparity motion vector (m, n), translate the pixels along their respective geodesics to obtain the new spherical coordinates $(\phi'_{ij}, \theta'_{ij})$ as,

$$\phi'_{ij} = \phi_{ij} + m\Delta\phi_s \quad \theta'_{ij} = \theta_{ij} + n\Delta\theta_s$$

  where $\Delta\phi_s$ and $\Delta\theta_s$ are predefined step sizes. Note that we expect change in elevation and no change in azimuth, i.e, m should be zero. Although one-dimensional motion is expected, one may still include a (zero) second component for compliance with standard codecs.
- Projection and interpolation: The reference frame is in the projected domain. Map the spherical coordinates of translated pixels, $(\phi'_{ij}, \theta'_{ij})$, to the plane. As the projected pixels may not fall on the sampling grid of the reference frame, perform interpolation as needed to derive the prediction signal.

Recall that the geodesics are determined by an axis aligned with vector v. In the setting of VR180, v is known from the placement of the two cameras, which therefore defines the fixed set of spherical coordinates one may conveniently use to perform disparity compensation. It is worth re-emphasizing that the disparity vectors in the proposed method are one-dimensional, which nets significant savings in side-information, a clear gain from compression perspective.

IV. Experimental results

The geodesic model was implemented with HM-16.15 [12] as the video codec. Geometry and sample rate conversion between source and coding formats were performed using the projection tool 360Lib-3.0 [10]. The VR180 sequences used in the experiment were obtained from the VR Gallery available on the website of Humaneyes Technologies[1] [13]. The 180° videos from the camera views were grey padded to generate full spherical videos. These videos are then projected to low-resolution projection formats. In our testing, we chose ERP and ECP as the projection formats. For each sequence, without loss of generality, the left view video was first encoded using standard HEVC temporal prediction in low-delay P profile. To show the full-potential of the proposed method, right-view is restricted to predict either from intra prediction or from inter-view prediction, i.e, temporal prediction is disabled for the right view. The competitors for right view compression are: i) codec using block based translation model for inter-view prediction and ii) codec using the proposed geodesic model for inter-view prediction. For the geodesic model with ERP, the step sizes $\Delta\phi_s$ and $\Delta\theta_s$ are chosen to be $\frac{\pi}{H}$, where H is the height of the ERP video. The corresponding step

[1]The authors would like to thank Humaneyes Technologies for providing permission to use these VR180 sequences in this research.

sizes for ECP with face-width W are chosen to be $\frac{\pi}{2W}$ since a face-width of W corresponds to a field of view of $\frac{\pi}{2}$rad. For geodesic model, we use sinc interpolation at $\frac{1}{64}^{th}$ pixel accuracy to derive prediction signal from the reference frame. R-D points were obtained by encoding at QP values of 22, 27, 32, and 37. We measured the distortion in terms of end-to-end weighted spherical PSNR [14], as recommended in [15]. Average bit-rate reduction is calculated using the Bjontegaard function, as per [16]. Table I shows the bit-rate savings (in %) of the proposed method over standard inter-view prediction using block translation model for right view. Note that for one sequence under one projection we incur a small loss. This can be explained by the fact that the HEVC framework was not (as yet) optimized for the new prediction technique, which would impact motion vector prediction and entropy coding, etc. Fig. 3 shows the rate-distortion curves for the *TelAviv* sequence for QP values 22, 27 and 32 in both projection formats. This sequence had constant-rate PSNR gains of up to 0.32dB for ERP and 0.23dB for ECP. It is clear from the overall gains presented in the table and the RD curves that our proposed method, which accounts for the geometry of the camera setup while performing inter-view prediction, substantially outperforms the standard translational model.

TABLE I: Bit-rate savings in % from the proposed geodesic model over the block-based translation method in HEVC (Y-component)

Geometry	Sequence	Bit-rate Savings
ERP	TelAviv	13.99
	HET	21.37
	Travel	1.97
	Average	**12.44**
ECP	TelAviv	12.36
	HET	2.82
	Travel	-0.54
	Average	**4.88**

V. Conclusions and Future Work

This paper proposes a novel geodesic model for disparity compensated prediction of VR180 videos. The proposed model perfectly captures the perceived disparity on the sphere. The model is agnostic of the projection format and can be easily extended to new geometries. Significant bit-rate savings in the experiments demonstrate the efficacy of the proposed model. Future work will focus on optimally combining the temporal and inter-view predictions.

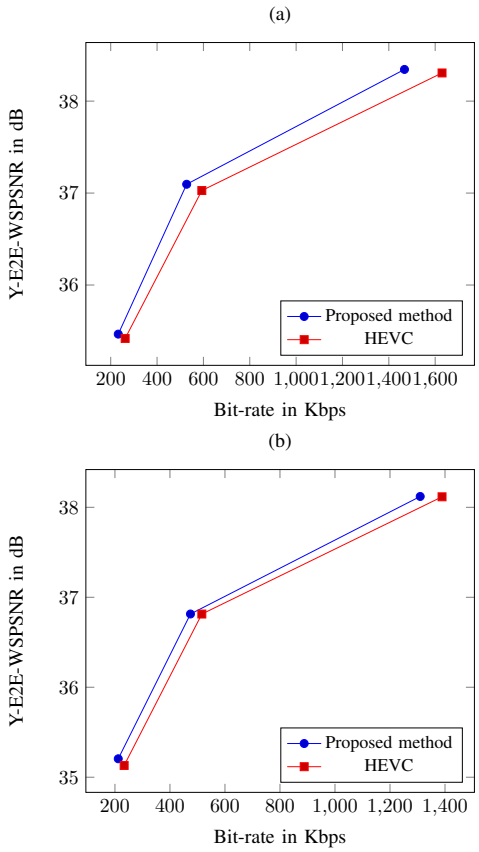

Fig. 3: RD curves for the *TelAviv* sequence with (a) ERP and (b) ECP as the projection format

References

[1] "VR-180," *https://arvr.google.com/vr180/*.

[2] J. P. Snyder, *Flattening the earth: two thousand years of map projections*, University of Chicago Press, 1997.

[3] M. Jamali, F. Golaghazadeh, S. Coulombe, A. Vakili, and C. Vazquez, "Comparison of 3D 360-degree video compression performance using different projections," in *2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)*. IEEE, 2019, pp. 1–6.

[4] J. Jung, F. Henry, A. Ouach, B. Ray, and P. Schwellenbach, "Stereoscopic 360 video compression with the next generation video codec," *JVET-G0064*, 2017.

[5] G. Lafruit, D. Bonatto, C. Tulvan, M. Preda, and L. Yu, "Understanding MPEG-I coding standardization in immersive VR/AR applications," *SMPTE Motion Imaging Journal*, vol. 128, no. 10, pp. 33–39, 2019.

[6] G. Tech, Y. Chen, K. Müller, J-R. Ohm, A. Vetro, and Y-K. Wang, "Overview of the multiview and 3D extensions of high efficiency video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 1, pp. 35–49, 2015.

[7] B. Vishwanath, K. Rose, Y. He, and Y. Ye, "Rotational motion compensated prediction in HEVC based omnidirectional video coding," in *Picture Coding Symposium (PCS)*, 2018, pp. 323–327.

[8] B. Vishwanath, T. Nanjundaswamy, and K. Rose, "Motion compensated prediction for translational camera motion in spherical video coding," in *International Workshop on Multimedia Signal Processing (MMSP)*, 2018, pp. 1–4.

[9] L. Li, Z. Li, X. Ma, H. Yang, and H. 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.

[10] Y. He, B. Vishwanath, X. Xiu, and Y. Ye, "AHG8: InterDigital's projection format conversion tool," *Document JVET-D0021*, 2016.

[11] G. Van der Auwera, M. Coban, and M. Karczewicz, "AHG8: Equatorial cylindrical projection for 360-degree video," *Document JVET-F0026*, 2017.

[12] "High efficiency video coding test model, HM-16.15," *https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/ tags/*, 2016.

[13] "Vuze VR gallery," *https://vuze.camera/vr-gallery/*.

[14] B Vishwanath, Y He, and Y Ye, "AHG8: Area weighted spherical PSNR for 360 video quality evaluation," *JVET-D0072, Chengdu, CN*, 2016.

[15] J. Boyce, E. Alshina, A. Abbas, and Y. Ye, "JVET common test conditions and evaluation procedures for 360-degree video," *JVET-F1030*, 2017.

[16] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," *Doc. VCEG-M33 ITU-T Q6/16, Austin, TX, USA, April*, 2001.