SPHERICAL VIDEO CODING WITH GEOMETRY AND REGION ADAPTIVE TRANSFORM DOMAIN TEMPORAL PREDICTION

Bharath Vishwanath and Kenneth Rose

Department of Electrical and Computer Engineering, University of California Santa Barbara, CA 93106
{bharathv, rose}@ece.ucsb.edu

ABSTRACT

Many virtual and augmented reality applications depend critically on efficient compression of spherical videos. Current approaches apply a projection geometry to map a spherical video onto the plane(s), wherein a standard codec can be used for compression. Video coders employ simple pixel copying from reference frames for inter-prediction, which ignores underlying spatial correlations, and is hence suboptimal. A novel paradigm of transform domain temporal prediction (TDTP) was developed previously in our lab to effectively overcome this suboptimality of standard video coding. This paper is motivated by the observation that projected spherical videos exhibit significantly more statistical variation due to *i)* the choice of projection geometry and *ii)* position of the block on the sphere, which reflect variations in sampling density and various statistical features. To account for such variations, we propose geometry and region adaptive TDTP that is tailored to spherical videos. For a given geometry, the sphere is divided into regions, according to expected signal statistics, and prediction filters are designed for each region. Experimental results show significant performance gains as evidence for the efficacy of TDTP in spherical video coding.

Index Terms— spherical video coding, virtual reality, motion compensation, temporal prediction

1. INTRODUCTION

Spherical video or 360-degree video offers an immersive experience for users by capturing the surroundings on a sphere enclosing the user who can then view in any desired direction. With its increased field of view, spherical video generates enormous amounts of data and necessitates efficient compression. Current approaches simply project a spherical video onto planes via different projections such as the equirectangular projection, cubemap, etc., [1]. This facilitates the use of standard coders to compress the projected video. The projected videos are sampled uniformly in the plane, thus inducing non-uniform sampling on the sphere.

Modern video coders perform motion compensated prediction to exploit temporal redundancies. A translation motion model is employed to identify a matching reference block in a previously reconstructed frame, which is used as the prediction signal. In the case of spherical video, advanced motion models were proposed to perform motion compensation on the sphere [2–4]. All these approaches perform pixel domain copying for prediction, which largely ignores underlying spatial correlations, rendering them suboptimal. Many approaches that account for spatial correlations in standard video coding including multi-tap filtering [5, 6] and three-dimensional subband coding [7,8] often result in high encoder complexity. In an earlier work in our lab, a fundamentally different approach of performing temporal prediction in transform domain was proposed [9]. The core idea was to first spatially decorrelate the block by a transform such as the discrete cosine transform (DCT), and model the temporal (inter-frame) evolution of each transform coefficients as a first order auto-regressive process. TDTP offers two fold benefits by: *i)* achieving both spatial and temporal decorrelation and *ii)* capturing the variations in the temporal correlations of the DCT coefficients which is otherwise hidden in the pixel domain.

In earlier work, we have established the benefits of TDTP in regular video coding [10, 11]. Projected spherical videos exhibit signal statistics that differ considerably from standard videos. Clearly, the statistics vary across different projection geometries. Further, uniform sampling in the projection format induces varying sampling density on the sphere. Thus, for a given geometry, signal statistics vary significantly for different regions of the sphere. For instance, let us consider temporal correlations of DCT coefficients for blocks along their motion trajectories in equatorial versus polar regions in ERP. Such correlations, extracted from a sample sequence, are tabulated in Table 1 and clearly illustrate that they vary significantly for different regions on the sphere. Motivated by these observations, we propose to design TDTP filters that adapt to these variations in statistics. TDTP filters are tailored separately for different geometries. Further, for each geometry, the sphere is divided into regions of similar sampling density (or expected statistics) and TDTP filters are designed for each region. Note that the proposed approach achieves geometry and spatial adaptivity without incurring any additional cost in side-information.

A major challenge, in TDTP filter design, is design insta-

This work was supported in part by a Google Faculty Research Award.

Table 1. Transform domain temporal correlations for blocks in ERP in :

(a) Equatorial region

0.99	0.96	0.92	0.91	0.89	0.84	0.79	0.67
0.97	0.95	0.91	0.87	0.83	0.78	0.73	0.58
0.96	0.93	0.88	0.86	0.84	0.75	0.69	0.6
0.93	0.88	0.88	0.84	0.79	0.72	0.64	0.58
0.89	0.90	0.90	0.84	0.75	0.66	0.62	0.46
0.83	0.89	0.84	0.83	0.70	0.58	0.54	0.44
0.83	0.81	0.82	0.74	0.62	0.53	0.49	0.4
0.77	0.71	0.62	0.66	0.58	0.45	0.39	0.38

(b) Polar region

0.99	0.94	0.89	0.71	0.6	0.37	0.37	0.08
0.97	0.98	0.90	0.69	0.62	0.34	0.3	0.23
0.98	1.0	0.94	0.79	0.66	0.49	0.21	0.32
0.96	0.96	0.95	0.77	0.53	0.33	0.3	0.22
0.94	0.92	0.92	0.79	0.43	0.23	0.13	0.32
0.9	0.88	0.84	0.75	0.52	0.24	0.27	-0.04
0.82	0.78	0.82	0.72	0.49	0.11	0.3	0.11
0.71	0.78	0.62	0.61	0.48	0.05	0.32	0.23

bility due to the closed loop nature of the coders. The prediction filters are applied to reconstructed samples, which in-turn depend on the prediction filters. This gives a glimpse of the closed loop conundrum. Standard closed loop design often suffers from significant (and sometimes catastrophic) design instability due to error propagation in the prediction loop. An effective remedy, called asymptotic closed loop (ACL) design, was proposed in [12]. In ACL, prediction is performed iteratively in an open loop fashion, ensuring design stability, but such that upon convergence, reconstructed samples remain unchanged from iteration to iteration, so that the resulting system is optimal for closed loop operation. In this paper, we use ACL as a stable platform for designing TDTP filters.

2. BACKGROUND

2.1. Transform Domain Temporal Prediction (TDTP)

TDTP models the temporal evolution of DCT coefficients as a first order AR process. Let x_n be a particular DCT coefficient in a given block in frame n, along a motion trajectory. The evolution of x_n is thus modeled as,

$$x_n = \rho \hat{x}_{n-1} + e_n \tag{1}$$

where $\hat{x}_{n-1}$ is the corresponding DCT coefficient of the block in the reconstructed frame $n-1$ along the motion trajectory and e_n is the innovation sequence. The optimal prediction coefficient that minimizes the mean square prediction error is given by,

$$\rho = \frac{\sum x_n \hat{x}_{n-1}}{\sum \hat{x}_n^2} \tag{2}$$

By performing temporal prediction in DCT domain, TDTP effectively achieves both temporal and spatial decorrelation. Further, TDTP captures the variations in temporal correlations for different frequencies by optimizing ρ for each DCT coefficient.

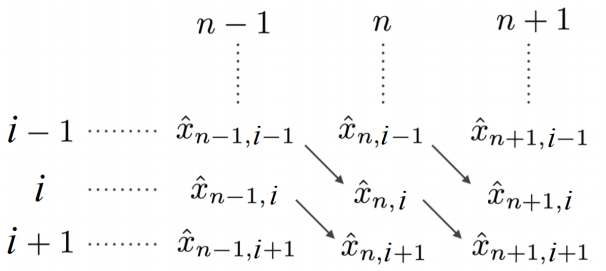

Fig. 1. Asymptotic closed loop design (arrows depict the prediction direction)

2.2. Asymptotic Closed Loop Design

The optimal prediction filter depends on the reconstructions, as is evident from (2). These reconstructions further depend on the prediction coefficient. This inter-dependency poses a major challenge in designing prediction filters. In the standard closed-loop technique [13], predictors are designed iteratively. The predictor designed for the reconstructed sequence in iteration $i-1$ is applied to the reconstructions in the next iteration i. The resulting error, due to statistical mismatch, propagates in the prediction loop causing a growing design instability. ACL design effectively resolves this issue by updating the reconstructions in an open-loop fashion as illustrated in Fig. 1. Note that predictors are applied to the same set of reconstructions they were designed for, ensuring better reconstructions and hence better predictions over iterations. On convergence, the reconstructed sequence remains essentially unchanged. Therefore, predicting from the previous iteration's reconstructions mimics predicting from the current iteration, i.e., it effectively operates in closed loop. Thus, ACL asymptotically optimizes the predictors for closed loop operation.

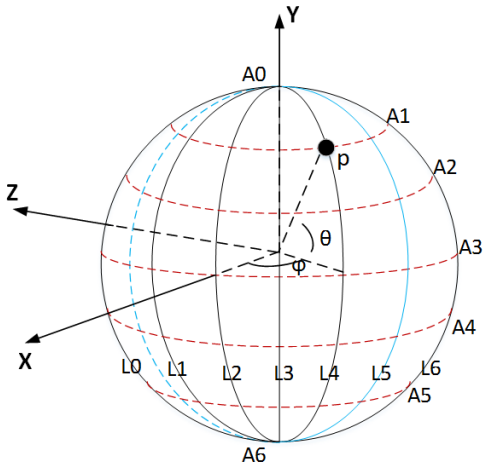

Fig. 2. Sphere sampling pattern for equirectangular projection

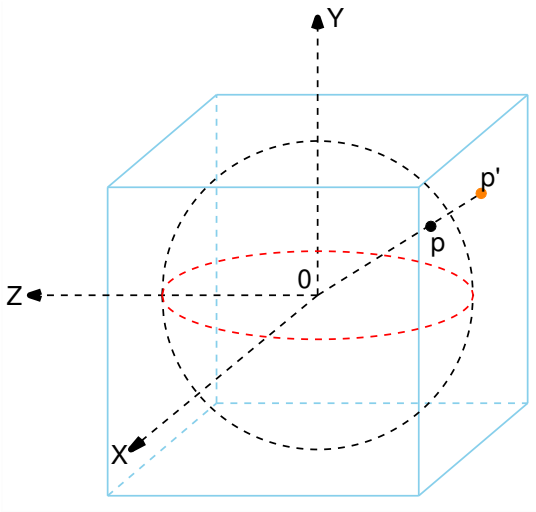

Fig. 3. Sphere mapping for Equi-Angular Cubemap

3. PROPOSED METHOD

Our earlier work on TDTP focused on learning TDTP filters in the standard video coding scenario. In this paper, we focus on spherical videos and address the challenges therein. Since different projection formats induce different sampling on the sphere, we propose to have different TDTP filters/modes for different projection formats. As discussed earlier, for a given geometry, the sampling density varies for different regions on the sphere. We first consider different projection formats and define regions of similar signal statistics on the sphere and then discuss the overall encoding paradigm. Although for now regions are defined heuristically based on analysis of the geometry, future work will develop data-based algorithms to directly optimize the regions.

3.1. Defining regions on the sphere

3.1.1. Equirectangular Projection (ERP)

The sphere sampling for ERP is shown in Fig. 2. For ERP, the vertical sampling density is constant. However, the horizontal sampling density increases with respect to the elevation θ as $sec\theta$. Thus, for ERP, we define regions based on the elevation of the center of the block on the sphere. Let θ_c be the elevation of the center of the current prediction unit. The region r is defined as,

$$\begin{aligned} r &= 1 \text{ if } |\theta_c| \leq sin^{-1}(\frac{1}{3}) \\ &= 2 \text{ if } sin^{-1}(\frac{1}{3}) < |\theta_c| \leq sin^{-1}(\frac{2}{3}) \\ &= 3 \text{ otherwise} \end{aligned} \quad (3)$$

The partitions are chosen such that the area of each region is same on the sphere.

3.1.2. Equi-angular Cubemap

Equi-angular cubemap is shown in Fig. 3. In a traditional cubemap, the sphere is enclosed in a cube and each face of the cube is uniformly sampled, resulting in non-uniform sampling on the sphere. However, sampling in EAC is done such that it results in near uniform sampling on the sphere [14]. Although, EAC has similar sampling density throughout the sphere, we nevertheless define regions as in (3) based on the expectation that object motion would exhibit different characteristics in each region, resulting in different signal statistics. Recall that data-based optimization of EAC regions is currently being investigated.

3.1.3. Equatorial Cylindrical Cubemap

Equatorial Cylindrical cube-map (ECP) was proposed in [15]. In ECP, the equatorial region corresponding to $\{-sin^{-1}\frac{2}{3} \leq \theta \leq sin^{-1}\frac{2}{3}\}$ is mapped to four faces of the cube via Lambert equi-area sampling [1]. Each polar region is mapped to a circular region in a plane and then stretched to fit the face of a cubemap. Even with ECP, the sampling pattern changes with respect to the elevation of the block on the sphere. We thus define three regions similar to ERP as in (3).

3.2. Overall Design Paradigm

The overall design follows a two loop ACL design similar to [10]. Inner-loop involves designing TDTP filters for a fixed encoder decision and in the outer loop, the encoder is given the new set of TDTP filters to update various decisions such as quad-tree partition, motion field etc. In the inner loop, with the encoder decisions held fixed, TDTP filters are designed for each region r in ACL fashion. In an iteration i of ACL, $\rho^i_{k,l,r}$, the prediction filter for $(k,l)^{th}$ DCT coefficient is given by,

$$\rho^{i}_{k,l,r} = \frac{E\{(\hat{x}^{i}_{k,l,n-1,r} x_{k,l,n,r})\}}{E\{(\hat{x}^{i}_{k,l,n,r})^2\}} \quad (4)$$

where, $E\{\}$ is the expectation operator, $x_{k,l,n,r}$ is the DCT coefficient of a block in region r in frame n of the source video and $\hat{x}^{i}_{k,l,n,r}$ is the reconstructed DCT coefficient of reference block in frame $n-1$. These filters are used to update the reconstructions in open loop fashion as,

$$\hat{x}^{i+1}_{k,l,n,r} = \rho^{i}_{k,l,r} \hat{x}^{i}_{k,l,n-1,r} + \hat{e}^{i}_{k,l,n,r} \quad (5)$$

where $\hat{e}^{i}_{k,l,n,r}$ is the quantized prediction error. Upon convergence in the inner loop, various encoder decisions are updated in the outer loop with the encoder using the TDTP filters learnt in the inner-loop. The overall design paradigm is illustrated in Algorithm 1.

Define regions as in (3);
Get initial closed loop encoder reconstruction ;
while *outer_iter* $<$ *max_outer_iter* **do**
  Fix encoder decisions;
  while *MSE decreases* **do**
    (a) learn TDTP filters for each region on sphere;
    (b) Update reconstructions in ACL fashion ;
  end
  Update encoder decisions with new TDTP filters and get new closed loop reconstruction;
end

Algorithm 1: Overall design approach

4. EXPERIMENTAL RESULTS

To obtain experimental results, region-adaptive TDTP is implemented within HM-14.0 [16]. The geometry mappings are done using the projection conversion tool of [17]. We chose the low delay P profile in HEVC for experiments. To simplify the experiments, we only use the previous frame as the reference frame. The projection formats were ERP, EAC and ECP. We encoded 30 frames of five video sequences over four quantization parameter (QP) values of 22, 27, 32 and 37. Since statistics vary for different QP, we design different filters for different QPs. In order to illustrate the full potential of proposed approach, we learn different TDTP filters for each sequence and for each region in a given projection format. The resolution for ERP was 2K and the face-width for EAC and ECP was 512. We measured the distortion in terms of end-to-end weighted spherical PSNR [18], as recommended in [19]. Bitrate reduction is calculated in the standard manner in terms of BD-rate [20]. The bit-rate savings for each projection format are tabulated in Table 2. It is evident that region adaptive TDTP consistently outperforms HEVC across different projection formats. The RD curve for *bicyclist* sequence with EAC projection format coding is shown in Fig. 4.

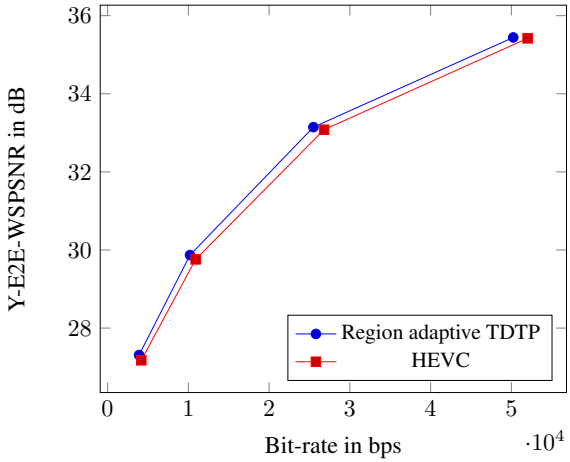

Fig. 4. RD curves for *bicyclist* sequence for equiangular cubemap

Table 2. Bit-rate savings (%) for Y component over HEVC with different projection formats

Geometry	Sequence	Bit-rate Reduction
ERP	Bicyclist	7.8
	Chair	9.2
	Balboa	7.2
	Broadway	7.3
	Glacier	5.1
	Average	**7.3**
EAC	Bicyclist	10.6
	Chair	8.4
	Balboa	7.1
	Broadway	8.6
	Glacier	4.8
	Average	**7.9**
ECP	Bicyclist	8.2
	Chair	8.1
	Balboa	7.0
	Broadway	10.1
	Glacier	5.2
	Average	**7.7**

5. CONCLUSIONS

In this paper, we proposed a transform domain temporal prediction paradigm for spherical videos that adapts for varying statistics on the sphere for different geometries. The issue of design instability was handled by an asymptotic closed-loop design approach. Significant gains demonstrate the potential of TDTP for spherical video coding, even in conjunction with heuristic definition of the regions. Future work will focus on developing data-based algorithms to optimize the regions on sphere.

6. REFERENCES

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

[2] B. Vishwanath, T. Nanjundaswamy, and K. 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.

[3] Y. Wang, D. Liu, S. Ma, F. Wu, and W. 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.

[4] 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.

[5] J. Kim and J. W. Woods, "Spatiotemporal adaptive 3-d kalman filter for video," *IEEE Transactions on Image Processing*, vol. 6, no. 3, pp. 414–424, 1997.

[6] T. Wedi, "Adaptive interpolation filter for motion and aliasing compensated prediction," in *Visual Communications and Image Processing 2002*. International Society for Optics and Photonics, 2002, vol. 4671, pp. 415–423.

[7] S.-J. Choi and J. W. Woods, "Motion-compensated 3-d subband coding of video," *IEEE Transactions on image processing*, vol. 8, no. 2, pp. 155–167, 1999.

[8] J.-R. Ohm, "Three-dimensional subband coding with motion compensation," *IEEE Transactions on image processing*, vol. 3, no. 5, pp. 559–571, 1994.

[9] J. Han, V. Melkote, and K. Rose, "Transform-domain temporal prediction in video coding: exploiting correlation variation across coefficients," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*. IEEE, 2010, pp. 953–956.

[10] S. Li, T. Nanjundaswamy, Y. Chen, and K. Rose, "Asymptotic closed-loop design for transform domain temporal prediction," in *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 4907–4911.

[11] B. Vishwanath, T. Nanjundaswamy, and K. Rose, "Deterministic annealing based transform domain temporal predictor design for adaptive video coding," in *2019 Data Compression Conference (DCC)*. IEEE, 2019, pp. 192–200.

[12] H. Khalil, K. Rose, and S. L. Regunathan, "The asymptotic closed-loop approach to predictive vector quantizer design with application in video coding," *IEEE transactions on image processing*, vol. 10, no. 1, pp. 15–23, 2001.

[13] P-C. Chang and R. Gray, "Gradient algorithms for designing predictive vector quantizers," *IEEE transactions on acoustics, speech, and signal processing*, vol. 34, no. 4, pp. 679–690, 1986.

[14] M. Zhou, "AHG8: A study on equi-angular cubemap projection," *Document JVET-G0056*, 2017.

[15] M. Zhou, "AHG8: A study on equi-angular cubemap projection," *Document JVET-G0056*, 2017.

[16] "High efficiency video coding test model, HM-14.0," *https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/tags/*, 2016.

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

[18] Y. Sun, A. Lu, and L. Yu, "AHG8: WS-PSNR for 360 video objective quality evaluation," *Document JVET-D0040*, 2016.

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

[20] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," *Doc. VCEG-M33 ITU-T Q6/16, Austin, TX, USA, 2-4 April*, 2001.