# An Assessment of HEVC Intra-Frame Prediction over 360-degrees Videos

Iago Storch
*Video Technology Research Group (ViTech)*
*Federal University of Pelotas*
Pelotas, Brazil
icstorch@inf.ufpel.edu.br

Daniel Palomino
*Video Technology Research Group (ViTech)*
*Federal University of Pelotas*
Pelotas, Brazil
dpalomino@inf.ufpel.edu.br

Luciano Agostini
*Video Technology Research Group (ViTech)*
*Federal University of Pelotas*
Pelotas, Brazil
agostini@inf.ufpel.edu.br

Bruno Zatt
*Video Technology Research Group (ViTech)*
*Federal University of Pelotas*
Pelotas, Brazil
zatt@inf.ufpel.edu.br

Luís Cruz
*Instituto de Telecomunicações*
*Universidade de Coimbra*
Coimbra, Portugal
lcruz@deec.uc.pt

***Abstract*—Recent technological advancements allowed videos to come from a simple sequence of 2D images to be displayed in a flat screen display into spherical representations of one's surroundings, capable of creating a realistic immersive experience when allied to head-mounted displays. Although 360-degrees videos present different characteristics from conventional videos, they are encoded using the same tools, which may not yield the best results. Aiming to find evidence that conventional video encoders can be adapted to perform better over 360-videos, this work performs an evaluation on the intra-frame prediction performed by the High Efficiency Video Coding over 360-degrees videos in the equirectangular projection. Experimental results point that 360-videos present spatial features that make it likely to be encoded using a reduced set of prediction modes, which could be used in the development of fast decision and energy saving algorithms.**

***Keywords—spherical video, equirectangular projection, intra-frame prediction evaluation***

## I. INTRODUCTION

Digital videos are very popular nowadays and can be found in a variety of scenarios, such as entertainment, video calls, e-learning, outdoor advertisement, among others. Their presence in most people lives is such that it is expected that by the year 2021 digital videos will account for 82% of total consumer internet traffic [1]. With the constant technological advances in video acquisition, display and processing, the industry is investing in both higher resolutions and interactive/immersive approaches for digital videos.

Among these approaches are 3D videos, which have been a subject of study for decades and many technologies have surfaced in the meantime. Another approach is 360-degrees videos, also known as immersive, spherical or simply 360 videos, which is a relatively new technology and there is still work to be done towards its popularization. In such videos, the scene is captured in all directions from within one point using special cameras, and during playback, the user can freely look around as if it were in the middle of the scene.

Considering that in 360 videos it is necessary to represent a whole sphere instead of a single point-of-view of it, the amount of data required to represent a 360 video with the same quality as a standard video is considerably higher. The standard videos themselves require an enormous amount of data to be represented, and its popularization was only possible due to the video coding standards. These video coding standards exploit features such as spatial and temporal redundancies to achieve high compression rates. The High Efficiency Video Coding (HEVC) standard [2] is the current state of the art, and surpassed its predecessor presenting a coding efficiency of about 50% higher [3].

Given that video coding standards have been studied since the 80's, nowadays there is a robust infrastructure for video coding. Aiming to exploit such infrastructure, instead of creating a 360-degrees-specific video encoding standard, these videos are pre-processed to be represented in a flat fashion, and then encoded by conventional video encoders.

Along the years many optimizations have been proposed to video coding standards, and such optimizations vary in a myriad of ways. Among such ways, there are proposals to reduce the complexity and/or energy consumption of video encoding and decoding [4][5], explore parallel systems more properly [6], optimize memory usage [7], among others.

However, both the coding standards and most of its improvements were developed aiming to exploit features of conventional videos, and may not present the best results when applied to 360 videos in their flat form. Considering it, this work aims to perform a study on the features of 360 videos and how they can be used to improve their encoding.

## II. OVERVIEW OF THE 360 VIDEO PROCESSING CHAIN

Since 360 videos represent a spherical surface, its processing chain includes some peculiarities when compared to conventional videos.

During the **acquisition**, it is not possible to use a single standard camera, as in conventional videos. Instead, the video acquisition is performed using multiple wide-angle lenses targeting different directions.

Following the acquisition is the **stitching** step. Since multiple lenses are used to perform the capture, the end result is multiple videos facing different directions of the scene. The stitching is responsible for sewing all these individual videos together and performing the corrections to create a smooth, seamless spherical video.

The spherical video is then **projected** onto a flat surface to be encoded by conventional video coding standards. The projection can be performed in many different ways. The most commonly used projection is the equirectangular projection (ERP) [8], in which each parallel of the sphere is translated into an entire row of a rectangle. Another common projection is the cubemap (CMP) [8] projection, in

which the sphere is put inside of a cube and its surface is projected into the 6 inner faces of said cube. After this, the cube is dismantled and its faces rearranged to form a rectangle. Other projections such as octahedral [8] and truncated square pyramid [8] are performed similarly to the CMP projection, however in these cases, the sphere is projected into an octahedron and a flat-topped square pyramid, respectively, which are then dismantled and their faces rearranged. The ERP, CMP, octahedral and truncated square pyramid projections for the *AerialCity* video are presented in Figure 1 (a), (b), (c), and (d), respectively.

Once the video is projected onto a flat surface, it is **encoded**, **transmitted/stored** and **decoded** following the same process as conventional videos.

In the user-end of the processing chain, once the video is decoded it is projected into a sphere again and a **viewport** is **rendered** from it. Since 360 videos are intended to increase the immersion, the user watches a reduced portion of the sphere at a time, as if it were looking that way from within the sphere, and this portion is called a viewport.

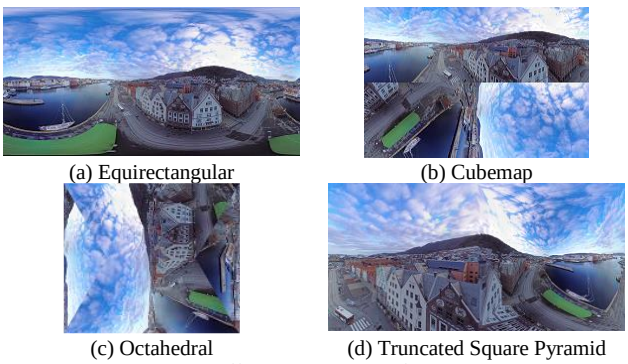
Finally, when 360 videos are **reproduced** in a flat screen the viewport can be selected with a mouse or joystick, whereas when reproducing it in a head-mounted display it is preferable to perform motion tracking.

Once this processing chain is analyzed, it is possible to notice that stitching several videos into a spherical video and then project it into a flat surface creates distortions and artifacts nonexistent in conventional videos. Apart from that, as presented in Figure 1, it is visible that different projections create distinct distortions in the image, therefore, it is not possible to determine the impact of all the projections during the encoding at once. Considering that the ERP projection is the most commonly used projection, this will be the projection considered during this work.

## III. Evaluation Aspects and Methodology

In the ERP projection, each parallel is transformed into a row of the rectangle, therefore there is no vertical distortion. On the other hand, since the parallels radius decrease as we approach the polar regions, there is less information to be represented in the same row length. In the most extreme case the north/south pole, which represent a single point, will be translated into a complete row of the rectangle. Considering this, it is visible that the central region of the ERP video will have a faithful representation of the original 360 video, whereas the polar regions will present a heavily distorted video due the stretching, as visible in Figure 1 (a).

Considering that the ERP projection creates high redundancy in horizontally neighbor samples as much as close to the poles, it is possible that the HEVC intra-frame prediction presents a different behavior when processing such videos since the intra-frame prediction exploits spatial redundancies. Considering this, this work will focus on the intra-frame prediction behavior over ERP videos.



(a) Equirectangular (b) Cubemap

(c) Octahedral (d) Truncated Square Pyramid

Figure 1. Different projections of 360 videos

### A. The HEVC intra-frame prediction

One aspect that led the HEVC standard to achieve such performance improvement over previous standards is its highly flexible partitioning structure. Each frame is composed of several Coding Tree Units (CTUs), which are the basic partitioning structure of the HEVC and have a standard size of 64×64 samples. Each CTU can be divided into 4 equal-sized Coding Units (CUs) in a quadtree structure, which can be recursively divided into more CUs until they reach dimensions of 8×8 samples. In addition, when performing intra-frame prediction each CU comprises either 1 or 4 square-shaped Prediction Units (PUs): CUs with dimensions from 16×16 to 64×64 necessarily comprise a single PU, whereas 8×8 CUs can either comprise a single PU or be split into a quadtree comprising 4 PUs of size 4×4 [2].

Since the intra-frame prediction aims to exploit spatial redundancy in the image, the prediction is performed by representing the current PU samples using samples of neighboring PUs. The HEVC standard has 35 intra-prediction modes to exploit different texture orientations, which is another great advancement over its predecessor, H.264, which used only 9 modes [9]. From these 35 modes, 33 of them are angular whereas 2 are non-angular [2]. The HEVC intra-prediction modes are depicted in Figure 2.

Each arrow in Figure 2 represents a different texture orientation direction, and the modes are represented by numbers. The mode 26, for instance, represents a vertical-oriented texture, therefore if a given PU is predicted using such mode it will be represented using the samples from the directly above PU. On the other hand, when using mode 18 the PU will be represented using the samples from the immediate left, upper-left, and above PUs. Whereas the modes from 2 to 34 are intended to predict PUs with orientated textures, the modes 0 (Planar) and 1 (DC) are intended to predict PUs with smooth surfaces or which do not present a highly oriented texture.

During the encoding, each CTU is partitioned into all previously mentioned structures, from 64×64 down to 4×4 PUs, and each PU is predicted with the 35 intra-prediction modes available in order to determine which combination yields the best coding efficiency. Since there are many possible combinations, the HEVC employs a two-step evaluation: for each combination of PU size and mode, firstly a rough mode decision (RMD) is performed to

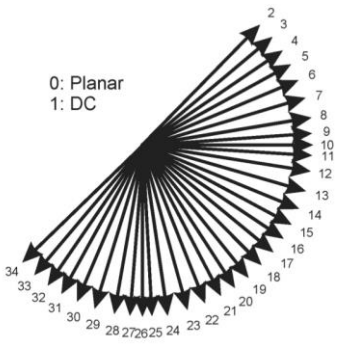



Figure 2. Intra prediction modes of HEVC

separate some modes with probable better Rate-Distortion (RD), and then this reduced group is passed through the proper RD-cost evaluation. During the RMD evaluation, the cost is calculated using the Sum of Absolute Hadamard Transformed Differences (SATD), which is a low complexity approximation of the RD-cost evaluation [2].

Considering this, a set of videos are encoded using the intra-prediction of the HEVC Test Model 16.6 (HM-16.6) [10] along with Lib360 [8]. During these encodings, the prediction mode selected for each PU is extracted to perform the evaluation of the intra-prediction over 360 videos.

### B. *Evaluation methodology*

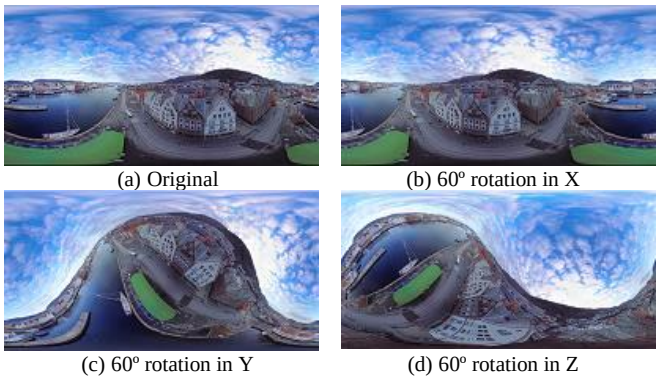
The evaluation was performed over the videos *AerialCity*, *Broadway*, *PoleVault* and *SkateboardInLot* since they compose both stationary and moving camera, and low and high movement videos. However, most of the 360 videos have simple and stationary textures in the polar regions, which are mostly composed by the ground floor, ceilings or the sky. Since this work aims to evaluate the intra-prediction, such similar contents in the videos could bias the results towards the video contents instead of the ERP distortion.

Since 360 videos represent a sphere, they have no specific orientation and can be projected considering any central position. Aiming to avoid getting content biased results, the 360 videos were projected considering several central positions, more specifically, the evaluated videos were rotated in multiples of 30º from 0º to 330º in the X, Y, and Z axis. As such, each video turns into 34 videos (original video plus 11 rotations in each axis) with the same content but differently distributed throughout the frame. The first frame of *AerialCity* is depicted in Figure 3, in which (a) is the original frame and (b), (c) and (d) is the frame rotated 60º in the X, Y and Z axis, respectively.

## IV. Evaluation Results

Once the videos were encoded and the intra prediction mode of each PU extracted, the data was processed to perform the evaluation. Since ERP videos present different degrees of distortion in different regions, the first step is a spatial evaluation aiming to find evidence that there is a tendency for certain modes in particular regions of the frame. To achieve such, the average selected mode for each sample was calculated based on the results of all videos and rotations. However, since Planar and DC modes do not represent directions, they were removed from this evaluation.

The resulting mode heatmap is presented in Figure 4, where the darkest blue represents mode 2 whereas the darkest red represents mode 34, according to Figure 2.



(a) Original (b) 60º rotation in X

(c) 60º rotation in Y (d) 60º rotation in Z

Figure 3. Original video and rotations along X, Y and Z axis

When analyzing the average intra-prediction modes depicted in Figure 4, it is visible that the area in the center of the frame is filled with a yellowish-blue tone – the center of the scale – which points that this area has an equiprobable intra modes distribution. However, when the samples towards the upper and lower edges of the frame are analyzed, it is visible that the color gets within the range centered in mode 10. A color range centered in mode 10 – which is not the center of the scale – points that this area is more likely to be encoded with modes around the mode 10, that is, modes which are horizontally oriented. Since the evaluated videos are rotated into several positions along the X, Y and Z axis, all their content is encoded into different regions of the frame, therefore, the results depicted in Figure 4 are due to the ERP distortion and not the content of the videos.

Considering the presented results, it is clear that the polar regions of the video present a higher tendency to be encoded using horizontally oriented modes, whereas the central area has a more evenly distributed mode selection.

Aiming to perform a more accurate assessment of the intra modes per region of video, the extracted encoding parameters are divided according to three regions: lower, upper and middle bands. The upper band comprises the top 25% samples of the video, the middle band comprises the 50% central samples, whereas the lower band comprises the bottom 25% samples. The upper and lower bands combined are called polar bands and comprise 50% of the video.

When performing such division, the PUs belonging to each band are evaluated separately considering their size and intra mode, and finally used to create a distribution of prediction mode per PU size per frame region. These results are presented as histograms in Figure 5, where the bars heights represent the occurrence rate of such mode in the respective PU size and frame region. Aiming to improve the visibility, the Planar and DC modes are represented by red and blue bars, respectively, the horizontal mode is represented by green bars, the vertical mode is represented by purple bars, whereas the remaining modes are represented by grayish-blue bars.

When analyzing these distributions, it is noticeable that the non-angular modes are responsible for a significant part of the selected modes whether the frame region, and the bigger the PU size the more probable is the selection of such modes: considering the Planar mode alone, it is responsible for at least 15% of the total occurrences in the smaller PUs, reaching more than 25% in the bigger PUs. The angular modes, however, present a distinct behavior in the polar and middle bands. When analyzing the middle band it is visible that the horizontal and vertical modes stand out with a slightly higher occurrence rate, whereas the remaining modes present a low and similar occurrence rate.

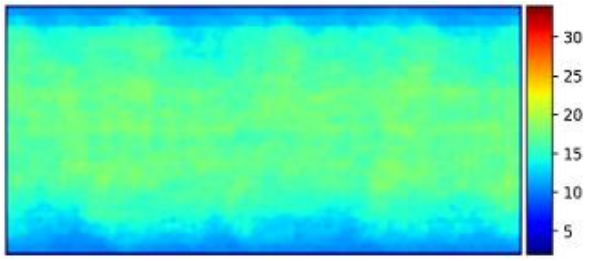



Figure 4. Average angular modes in 360 videos

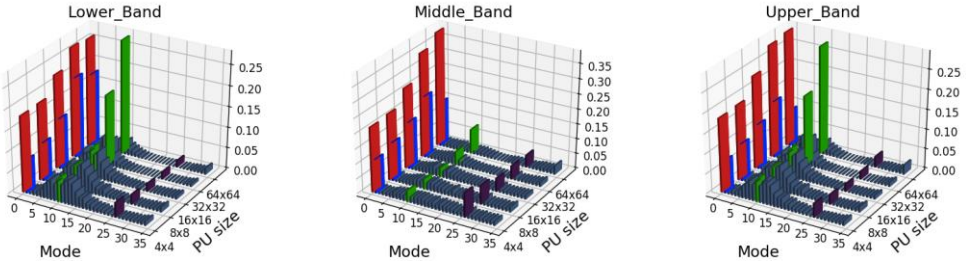



Figure 5. Occurrence rate of intra modes in ERP videos

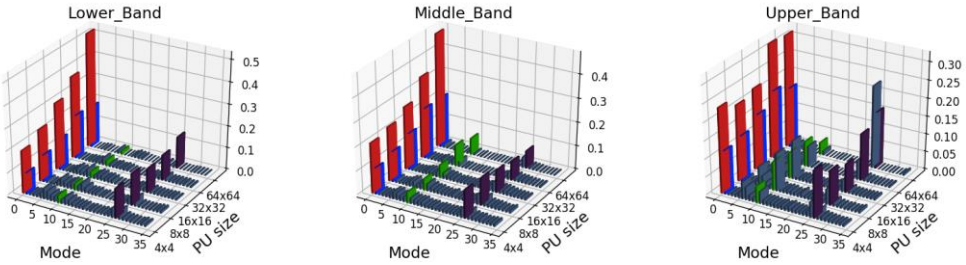



Figure 6. Occurrence rate of intra modes in conventional videos

Furthermore, when analyzing the distribution for the polar bands the behavior is quite different: (1) the horizontal mode is highly probable, in some cases more than the non-angular modes; (2) there is a high occurrence rate of modes close to the horizontal mode; (3) as the PU size increases, the occurrence of the horizontal mode increases whereas the occurrence of its neighboring modes decreases; (4) the occurrence of the vertical mode is reduced when compared to the middle band, whereas its neighbors rarely occur.

As pointed out by the spatial evaluation, this behavior is due the polar overstretching caused by ERP projection. As explained in Section 3 the amount of useful data in polar regions is reduced, notwithstanding this data is horizontally stretched to fit in a rectangle. This stretching makes horizontal samples highly similar since many of them are created through interpolation of the others, therefore predicting such samples with horizontal modes performs well. In addition, it is known that CTUs with complex texture tend to be partitioned into small PUs whereas CTUs with simple/homogeneous texture tend to be encoded with bigger PUs [11]. Allying this with the fact that the more stretching a CTU suffers the less information it will contain (and more uniform the texture will be), it is expectable that overstretched regions tend to be encoded with 64×64 PUs and either with a non-angular or horizontal mode, as visible by the green bars standing out in polar bands of 64×64 PUs.

Finally, the same evaluation is done to a set of conventional videos aiming to confirm that the results are due the ERP projection and not a common behavior of the HEVC intra prediction. During this assessment, the videos *BasketballDrive*, *BQTerrace*, *Cactus*, *Kimono*, and *ParkScene* are evaluated, and the histogram representing the distribution of intra-prediction modes for these videos is presented in Figure 6. In this distribution, it is clear that the non-angular modes represent a significant part of the occurrences, and the vertical and horizontal modes stand out from the others. However, there is no clear behavior in the 3 evaluated bands since the remaining modes are approximately randomly and/or uniformly distributed.

## V. Conclusion

This work presented evidence that ERP videos cause the intra-prediction of the HEVC standard to behave differently from conventional videos. The evaluation results point that due to overstretching in the polar areas, the intra-prediction of ERP videos have a high probability of choosing horizontally oriented modes in polar areas whereas the central area behaves similarly to conventional videos. These observations can be exploited to develop fast-selection algorithms for intra-frame prediction or even disabling some modes depending on the frame region, which can lead to considerable complexity reduction and power saving in embedded devices, with marginal coding efficiency penalty.

## Acknowledgment

The authors would like to acknowledge CAPES, CNPq, and FAPERGS for supporting this work.

## References

[1] Cisco, “Cisco Visual Networking Index: Forecast and Methodology 2016-2021”, 2017.

[2] High Efficiency Video Coding, ITU-T Rec. H.265 and ISO/IEC 23008-2, February 2018.

[3] J. R. Ohm *et al.*, "Comparison of the Coding Efficiency of Video Coding Standards-Including High Efficiency Video Coding (HEVC)," in *IEEE Transactions on Circuits and Systems for Video Technology.*, vol. 22, no. 12, pp. 1669-1684, Dec. 2012.

[4] G. Correa *et al.*, "Fast HEVC Encoding Decisions Using Data Mining," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 25, no. 4, pp. 660-673, April 2015.

[5] W. Penny *et al.*, "Pareto-based energy control for the HEVC encoder," *2016 IEEE Int. Conf. on Image Processing*, Phoenix, 2016, pp. 814-818.

[6] I. Storch *et al.*, "Speedup-aware history-based tiling algorithm for the HEVC standard," *2016 IEEE Int. Conf. on Image Processing*, Phoenix, 2016, pp. 824-828.

[7] A. Martins *et al.*, "Cache Memory Energy Efficiency Exploration for the HEVC Motion Estimation," *2017 VII Brazilian Symposium on Computing Systems Engineering (SBESC)*, Curitiba, 2017, pp. 31-38.

[8] Y. Ye *et al.*, “Algorithm descriptions of projection format conversion and video quality metrics in 360Lib”, JVET Doc. JVET-E1003, 2017.

[9] Advanced Video Coding for Generic Audio-Visual Services, ITU-T Rec. H.264 and ISO/IEC 14496-10 (AVC), April 2017.

[10] High Efficiency Video Coding Test Model 16.6, available: <https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/tags/HM-16.6>. Last access: June 2018.

[11] M. U. K. Khan *et al.*, "An adaptive complexity reduction scheme with fast prediction unit decision for HEVC intra encoding," 2013 IEEE Int. Conf. on Image Processing, Melbourne, VIC, 2013, pp. 1578-158.