The Impacts of Equirectangular 360-degrees Videos in the Intra-Frame Prediction of HEVC

Iago Storch[1], Luis A. da Silva Cruz[2], Luciano Agostini[1], Bruno Zatt[1], Daniel Palomino[1]

[1] Video Technology Research Group (ViTech), Federal University of Pelotas (UFPel), Pelotas, Brazil
[2] Instituto de Telecomunicações, Dep. of Elect. and Comp. Eng., University of Coimbra, Coimbra, Portugal
e-mail: [1] {icstorch, Agostini, zatt, dpalomino}@inf.ufpel.edu.br, [2]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. In order to explore the existing infrastructure for video coding, 360-degrees videos are pre-processed and then encoded by conventional video coding standards. However, the flattened version of 360-degrees videos present some peculiarities which are not present in conventional videos, and therefore, may not be properly exploited by conventional video coders. Aiming to find evidence that conventional video encoders can be adapted to perform better over 360-degrees 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-degrees videos present spatial properties that make some regions of the frame likely to be encoded using a reduced set of prediction modes and block sizes. This behavior could be used in the development of fast decision and energy saving algorithms by evaluating a reduced set of prediction modes and block sizes depending on the regions of the frame being encoded, which could improve the performance of modern video coding standards and contribute to the development of dedicated algorithms for encoding 360-degrees videos in future video coding standards.**

***Index Terms*— 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 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 or spherical 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 scene.

Nowadays, 360-degrees videos are being used mostly for entertainment purposes, such as broadcasting of sports events, concerts, and movies. However, given the level of immersion provided by this technology, 360-degrees videos can also be used for simulation of real environments. Some applications include virtual tours for real state and tourism agencies, surveillance applications, virtual labs for chemistry classes and training of new employees, military training, among others.

Considering that in 360-degrees 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-degrees video with the same quality as a conventional 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 properties 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].

Instead of being encoded in the spherical representation, 360-degrees videos are pre-processed in order to be represented in a rectangular planar domain, known as projection. More details regarding the projection process are given further in Section II. Since the projection generates a rectangular video, it is possible to encode the projected 360-degrees video with any conventional video coding standard, such as it is done with conventional videos.

Although it is possible to encode a projected 360-degrees video as a conventional video, 360-degrees videos present characteristics which are not present in conventional videos. Some of these characteristics are (1) the fact that for most applications, the user will watch a reduced portion of the video at a time (a point of view), (2) there are multiple manners to perform the projection, which can influence the encoding process differently, (3) there may be a recommended point of view for each video, among others.

Aiming to account for such characteristics during encoding and transmission whereas still maintaining interoperability among different content providers, industry and academia are working together towards the standardization of 360-degrees video processing framework.

The Joint Collaborative Team on Video Coding (JCT-VC), the organization responsible for the development of HEVC standard, started standardizing the signaling of 360-degrees video with messages indicating the used projection and preferred point of view, for instance [4].

The Joint Video Experts Team (JVET) is responsible for the development of the Versatile Video Coding (VVC) [5], the successor of the HEVC standard. The VVC standard is

Digital Object Identifier 10.29292/jics.v14i1.46

still under development, and several modifications of encoding tools have been evaluated to verify their performance when encoding 360-degrees videos. The work [6] presents a report on the findings of modifying in-loop filters and the interframe prediction, for instance.

Although future video coding standards are being developed aiming to support 360-degrees videos efficiently, current consumer market demands 360-degrees video, and since current video coding standards were not designed to deal with 360-degrees videos, they must be optimized for encoding such content while the next generation of video coding standards is not commercially available.

Along the years, many optimizations were 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 [7][8], explore parallel systems more properly [9][10], optimize memory usage [11], among others.

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

II. Overview of the 360-degrees Video Processing Chain

Since 360-degrees videos represent a spherical surface, its processing chain includes some peculiarities when compared to conventional videos. A general representation of the processing chain for 360-degrees videos is presented in Fig. 1, whereas each step is detailed as follows.

During the **capture**, it is not possible to use a single standard camera, such as used in conventional videos. Instead, the video acquisition is performed using special cameras equipped with multiple wide-angle lenses targeting different directions, therefore, several views of the same scene are captured into several conventional videos.

Following the acquisition is the **stitching** step. Since the acquisition generates several views of the same scene, now it is necessary to transform this set of conventional videos into a spherical video. Thus, the stitching is responsible for sewing all these individual videos together and performing the corrections to create a smooth, seamless spherical video representing the camera's surroundings.

The next step is responsible for the **projection**. Since conventional video coders are designed to encode 2D rectangular videos, the spherical videos must be projected into a flat rectangular surface to be properly handled by the encoders. The projection can be performed in many different ways. The most commonly used projection is the equirectangular projection (ERP) [12], in which each parallel of the sphere is translated into an entire row of a rectangle. This projection presents the property that it does not demand much computational power to be performed, however, it presents great distortion close to the poles. Another way to perform the projection is according to the Craster Parabolic Projection (CPP) [12], which is performed in a similar manner to the ERP projection, and the main difference is that it does not stretch the polar regions to fill the whole rectangle. The advantage of this approach is that it is not as distorted as the ERP projection, however, the area beyond the projected video is padded with inactive samples to form a rectangle, creating sharp edges in the image.

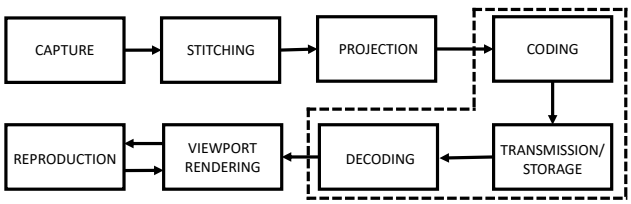


Fig. 1 Processing chain of 360-degrees videos

Another common projection is the cubemap (CMP) [12] 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. The CMP projection is interesting because the distortion presents itself mainly in the edges of each face, presenting a small distortion and representing the video in a more compact form when compared to the ERP projection. Also, there is the octahedral [12] projection, which is performed similarly to the CMP projection. In this projection, the sphere is projected into the inner faces of an octahedron which is then dismantled, and its faces rearranged. As an octahedron presents more faces than a cube, it presents smaller distortion. The ERP, CPP, CMP and octahedral projections for the AerialCity video are presented in Fig. 2 (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 transformed into a sphere again and a **viewport** is **rendered** from it. A viewport can be considered a point of view of the video from within the center of the sphere. Since 360-degrees 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, thus, watching one viewport at a time.

Finally, when 360-degrees 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. Since the reproduction device generally controls the viewport selection, there is a bi-directional communication among the

(a) Equirectangular (b) Craster Parabolic

(c) Cubemap (d) Octahedral

Fig. 2 Different projections of 360-degrees videos

reproduction and viewport rendering steps.

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 Fig. 2, it is visible that different projections create distinct distortions in the image.

The ERP projection presents an overstretching on the polar regions whereas the central region is represented in a more faithful way. The CPP projection does not present such overstretching, however, it is composed of several inactive samples which do not carry any useful information and must be encoded nonetheless, apart from creating sharp edges in the active-inactive area frontier. The CMP and octahedral projections present an average distortion, however, they also present sharp edges in the faces junction. Apart from these projections, there are several other projections which present distinct distortions.

Since different projections lead to different distortions, it is not possible to determine the impact of every projection during the encoding at once. Given that and considering that the ERP projection is the most commonly used projection, this will be the projection considered during this work.

III. Evaluation Methodology

The ERP projection can be simplistically defined as transforming each parallel of the sphere 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, the parallels closer to the equator present more information than the parallels closer to the poles, therefore, most parallels must be stretched in order to fill the whole rectangle width. In the most extreme case the north/south pole, which represent a single point in the sphere, will be translated into a complete row of the rectangle. This stretching is performed interpolating neighboring samples. Considering this, it is visible that the central region of the ERP video will have a faithful representation of the original 360-degrees video, whereas the polar regions will present a heavily distorted video due to the stretching.

The strength of the stretching can be more accurately assessed utilizing the Tissot's indicatrix, which is a method to characterize local distortions in projections [13]. The indicatrix is performed by projecting a circle from the spherical video into the projection plane. If the projection poses no distortion to the spherical image, the circles will present the same area in the projection. However, in case the projection poses any distortion, the circles will present different areas and/or be transformed into ellipses.

The Tissot's indicatrix for the ERP projection is presented in Fig. 3 in which an Earth map is used as an example of projection. When analyzing Fig. 3, it is visible that the row in the middle of the frame is composed of perfect circles. As the rows farther from the equator are analyzed, it is visible that they are composed of ellipses with the same height as the circles (characterizing the lack of vertical distortion), whereas the ellipses width increase as we approach the poles (characterizing the horizontal stretching). On the top and bottom of the frame, the red line represents a circle which has been severely stretched to fill the entire row.

The stretching caused far from the equator can be seen as a redundancy, since the new samples are created through interpolation of the original samples. Therefore, it is possible that the HEVC intra 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 behavior of the HEVC intra-frame prediction when applied to ERP 360-degrees 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 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 [14]. From these 35 modes, 33 of them are angular whereas 2 are non-angular [2]. The HEVC intra prediction modes are depicted in Fig. 4.

Each arrow in Fig. 4 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 a combination of samples from the immediate left, upper-left, and above PUs. Whereas the angular modes (that is, modes from 2 to 34) are intended to predict PUs with highly orientated textures, the non-angular modes 0 (Planar) and 1 (DC) are intended to predict PUs with smooth surfaces or which do not present a highly oriented texture. When employing the DC mode for predicting a

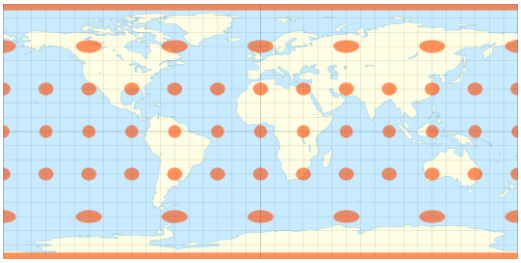


Fig. 3 Tissot's indicatrix of distortion for ERP projection

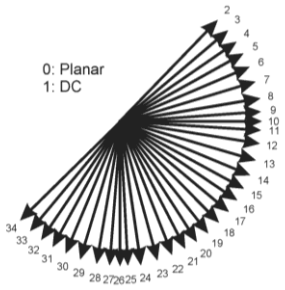


Fig. 4 Intra prediction modes of HEVC

given PU, for instance, the entire area of such PU is going to be represented with samples with the same value, which is the average from the samples of the original PU. When employing the Planar mode, on the other hand, the PU is represented using the average of two linear predictions, which use four corners as references. The planar mode is intended to prevent discontinuities in the boundaries of the PUs.

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. However, performing a precise evaluation for all combinations of prediction modes and PU sizes poses great computational costs. Aiming to evaluate all prediction modes whereas maintaining a reasonable computational complexity, the HEVC reference software (HEVC Test Model 16.16 – HM-16.16) [15] employs a three-steps decision scheme, as presented in Fig. 5.

Firstly, when a PU is going to be encoded, all 35 prediction modes are evaluated accordingly to the Rough Mode Decision (RMD). During this process, a Hadamard Transform is applied to the prediction residue and the transform coefficients are summed into the Sum of Absolute Hadamard Transformed Coefficients (SATD). In this scenario, the SATD acts as an approximation of the coding efficiency, in which the lower the SATD, the greater is the probability of such mode being the best prediction mode for the current PU. After performing this process to all the prediction modes, the modes with the lowest SATD are added to a list of better prediction candidates. The number of modes added to such list is 8, 8, 3, 3, and 3 for PUs of size 4×4, 8×8, 16×16, 32×32, and 64×64, respectively.

After the RMD, the Most Probable Modes (MPMs) are selected for the current PU. The MPM step consists of extracting the prediction modes selected for two neighbor PUs, and based on these modes, three modes are derived as being the most probable modes for the current PU. These three prediction modes are then added to the list of prediction candidates. If a given mode was discarded during the RMD and selected during the MPM, it is going to be added to the candidates list. If a given mode was already selected during RMD and is selected again during the MPM, it is kept in the list. The MPM is intended to exploit the spatial redundancy among neighboring PUs in the frame.

Lastly, the list of prediction candidates, which is composed of 3 up to 11 modes depending on the PU size and redundancy in the RMD and MPM modes selections, is evaluated accordingly to the rate-distortion optimization (RDO). The RDO consists of performing all further steps on the

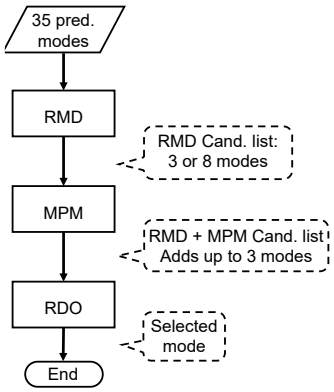


Fig. 5 HEVC intra prediction mode decision scheme

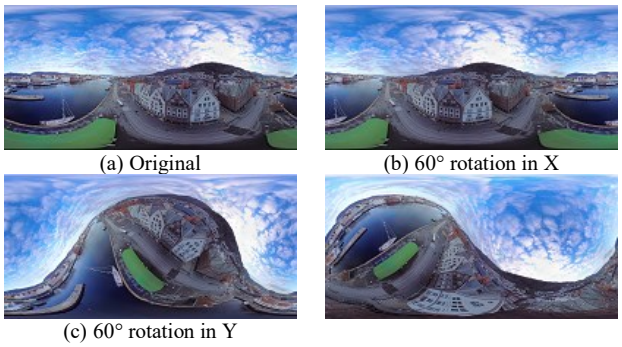
encoding, i.e., transforming and quantizing the prediction residue, performing entropy coding over the quantized coefficients and decoding the PU. With this process, the encoder is able to **O**ptimize the necessary bitrate (**R**ate) and the quality (**D**istortion) provided by each prediction mode and select the one with the best RD tradeoff [16].

As the HEVC intra prediction mode decision scheme was developed aiming to deal with the properties of conventional videos, it is important to evaluate the behavior of this tool when ERP 360-degrees videos are being encoded.

B. Evaluation Setup

A set of videos are encoded using the intra prediction of the HM-16.16 [15] along with Lib360 [12]. All the encodings are set according to the Common Test Conditions and Evaluation Procedures for 360° Video Coding (CTCs) [17]. During these encodings, the prediction mode and the size selected for each PU is extracted to perform the evaluation of the intra prediction over 360-degrees videos. The evaluation was performed over the video sequences AerialCity, Broadway, PoleVault and SkateboardInLot since these videos compose both stationary and moving camera, and low and high movement sequences. However, most of the 360-degrees 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 impact of the ERP projection distortion in the intra prediction, encoding a set of videos with such similar contents in co-located regions of the video sequences could yield results which are due to the content of the videos, and not to the distortion caused by the ERP projection

Since 360-degrees videos represent a sphere, they have no specific orientation and can be projected considering any central position. In this case, different central positions are equivalent to different camera orientations during the recording. Aiming to avoid getting content biased results, the 360-degrees videos are projected considering several central positions, more specifically, the evaluated video sequences are rotated in multiples of 30º from 0º up 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 Fig. 6, in which (a) represents the original frame and (b), (c) and (d) represent the frame rotated 60º in the X, Y and Z axis, respectively.



(a) Original (b) 60° rotation in X

(c) 60° rotation in Y

Fig. 6 Original video and rotations along X, Y and Z axis

When analyzing Fig. 6, at a first glance the images may represent different scenes. However, as stated before, they represent the same scene recorded with different camera orientations. In Fig. 6 (a), it can be said that the camera is facing the horizon behind the houses and the hill, whereas in Fig. 6 (b) the camera is facing the horizon behind the houses and the hill, but 60° to the right. In Fig. 6 (c) the camera has turned 60° down and now it is facing the road, whereas in Fig. 6 (d) the camera has turned 60° counterclockwise around its own axis. Using these rotations allows to evaluate the impact of the ERP distortion disregarding the contents of the video sequences: although the rotations change the contents distribution within the frame, all rotations represent the same scene and the ERP distortion is applied to all of them in the same manner, i.e., the closer to the poles, the stronger is the horizontal stretching. Therefore, using several rotated sequences makes for a better test set when evaluating the impacts of ERP 360-degrees videos in the intra prediction.

IV. Evaluation Results

Once the videos were encoded and the intra prediction mode and size of each PU extracted, the data was processed to perform the evaluation.

A. Prediction mode evaluation

Since ERP videos present different degrees of distortion in different regions, the first step is performing 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 mode selected for each co-located sample is averaged based on the results for all the videos and rotations. However, since the Planar and DC modes do not represent directions, their contribution to the average value would not present mathematical meaning and, therefore, they are removed from this evaluation.

The result from this average mode evaluation is presented as a heatmap in Fig. 7, where the darkest blue represents mode 2 whereas the darkest red represents mode 34, according to the prediction modes in Fig. 4.

When analyzing the average intra prediction modes depicted in Fig. 7, 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 in this area, there is a chance that all the modes occur with the same frequency. However, when the samples towards the upper and lower edges of the frame are analyzed, it is visible that the average 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 in this area, it is probable that some modes (around the horizontal mode 10) are more likely to occur than others. Since the evaluated videos are rotated into several positions along the X, Y and Z axis, all the content from the videos is encoded into different regions of the frame, therefore, the results depicted in Fig. 7 are due to the ERP distortion and not the content of the videos themselves.

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

Afterward, a similar spatial evaluation is performed considering only the non-angular modes. In this evaluation, the occurrence rate of the Planar (mode 0) and DC (mode 1) modes are extracted and individually compared against the remaining modes. Fig. 8 and Fig. 9 present the results for the Planar and DC modes, respectively.

In Fig. 8, the color is normalized to represent the occurrence rate of the Planar mode when compared to the remaining modes, that is, a sample with value 0.2 represents that in this sample the Planar mode is selected 20% of the times, whereas in 80% of the times the remaining modes (33 angular modes plus DC mode) are selected. In Fig. 9 the same method is employed, however, the value of a sample represents the occurrence rate of the DC mode. When analyzing Fig. 8 and Fig. 9, it is clear that the Planar mode is selected more often than the DC mode. However, it is not possible to draw clear conclusions regarding the occurrences per frame region, since the entire frame is covered with a similar tone both for the Planar and DC mode. This evaluation reveals

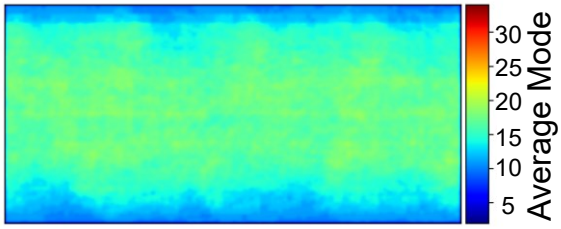


Fig. 7 Average angular modes in 360-degrees videos

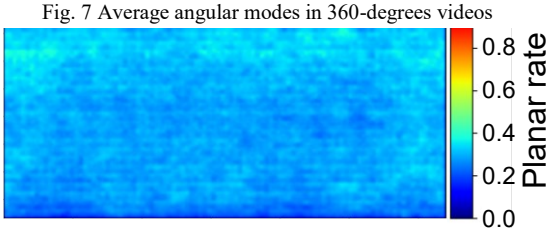


Fig. 8 Planar mode occurrence rate

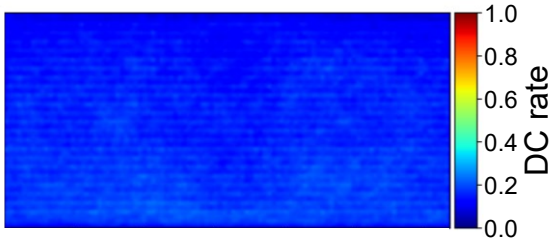


Fig. 9 DC mode occurrence rate

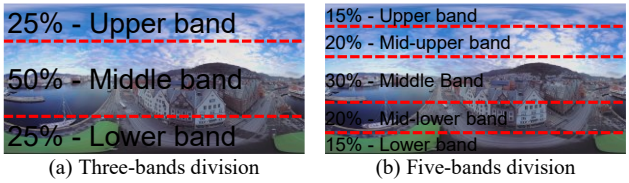
that, differently from the angular modes, the non-angular modes do not present a clear spatial behavior in videos in the ERP projection.

Aiming to perform a more accurate assessment of the intra modes per region of the frame, 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 frame, 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. Fig. 10 (a) presents this three-bands division over the first frame of *AerialCity* video sequence.

When performing such division, the PUs belonging to each band are evaluated separately considering their size and intra prediction mode, and finally used to create a distribution of prediction mode per PU size per frame region. The results for the lower, middle and upper bands are presented as histograms in Fig. 11 (a), (b) and (c), respectively, where the bars heights represent the occurrence rate of each 'Prediction mode' in the respective 'PU size' and frame region. Aiming to improve the visibility, the Planar and DC modes are represented by green and blue bars, respectively, the horizontal mode is represented by red bars, the vertical mode is represented by yellow bars, whereas the remaining modes are represented by bars in alternating gray tones.

When analyzing these distributions, it is noticeable that the non-angular modes are responsible for a significant part of the selected modes independently of 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.

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 closer modes decreases; (4) the occurrence of the vertical mode is reduced when compared to the middle band, whereas its closer modes rarely occur.



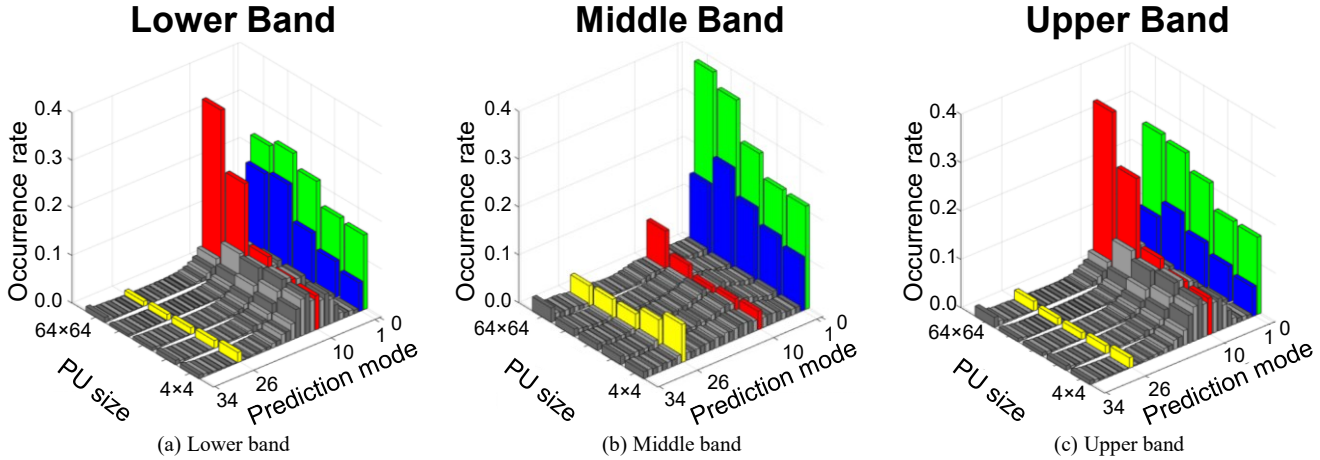
(a) Three-bands division (b) Five-bands division

Fig. 10 Examples of three- and five-bands division

As pointed out by the spatial evaluation, this behavior is due to 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.

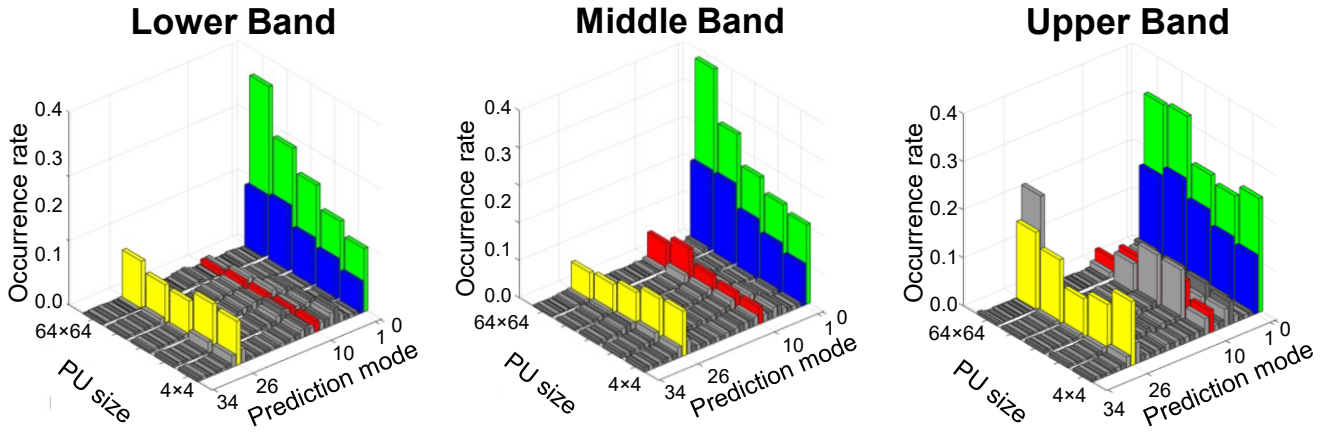
This statistical evaluation considering three-bands division is also performed over a set of conventional videos aiming to confirm that the results are due the ERP projection and not a common behavior of the HEVC intra prediction. During this assessment, the videos *BasketballDrive*, *BQTerrace*, *Cactus*, *Kimono*, and *ParkScene* are evaluated, and the distribution of prediction modes for the lower, middle and upper band of these videos are presented in Fig. 12 (a), (b) and (c), respectively, respecting the same color scheme as Fig. 11. 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 three evaluated bands since the remaining modes are approximately uniformly distributed. The only exception is the upper band, in which the contents of the videos made mode 9 and mode 25 to present an occurrence rate higher than mode 10 and mode 26, respectively. When comparing these results to the ones of 360-degrees ERP videos, it is clear that the behavior in Fig. 11 is characteristic of ERP videos and not a common behavior of the HEVC intra prediction.

Although the three-bands division evaluation points that the distortion caused by the ERP projection impacts the



(a) Lower band (b) Middle band (c) Upper band

Fig. 11 Occurrence rate of intra modes in ERP videos, considering three-bands division



(a) Lower band (b) Middle band (c) Upper band

Fig. 12 Occurrence rate of intra modes in conventional videos, considering three-bands division

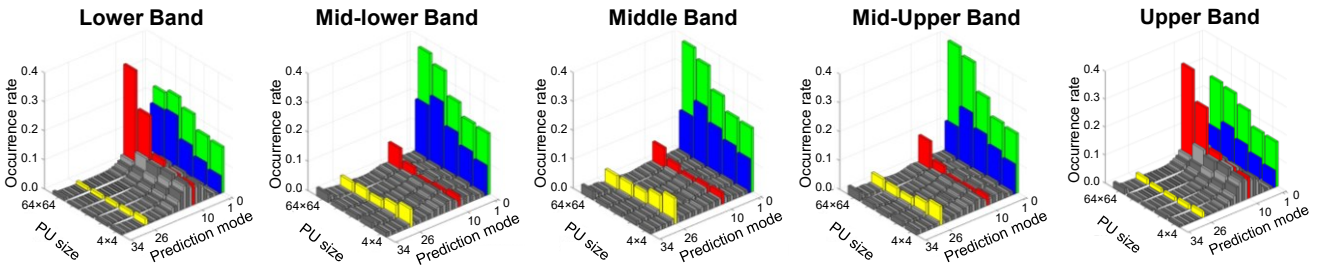
behavior of the intra prediction, the frame cannot be simply divided into "distorted" and "non-distorted" regions, as the three-bands division suggests. Aiming to account for the transition between the highly distorted areas and the non-distorted ones, the frames are also evaluated according to a five-bands division. When employing this division, the videos are divided into upper, mid-upper, middle, mid-lower, and lower bands. The upper and lower bands comprise the top and bottom 15% of the frame, respectively. The middle band comprises the middle 30% of the frame. Whereas the mid-upper and mid-lower bands comprise the region between the upper and middle band, and the region between the middle and lower band, respectively. Such division is depicted in Fig. 10 (b). The mid-upper and mid-lower bands combined are also called mid-polar bands. After performing this division, the PUs belonging to each band are separately evaluated to create a distribution of prediction modes per PU size per band. This evaluation is presented in Fig. 13, respecting the same color scheme as before.

When analyzing the results, it is visible that the non-angular modes still represent a significant part of the selected modes independently of the frame region, as expected. Also, the middle band (which is reduced when compared to the three-bands division) still presents a similar occurrence rate for all the angular modes, except for the vertical and horizontal modes. When analyzing the upper and lower bands, which now comprise a smaller region and thus a better representation of the poles themselves, it is visible that the behavior observed before (with three bands) is accentuated: when employing five bands, the concentration of occurrences around the horizontal mode is increased, that is, a smaller set of modes is responsible for a greater part of the selections. This behavior presents itself as a narrower and higher peak centered in mode 10 when compared to the three-bands division. When analyzing the mid-polar bands, which represent the transition between the highly distorted polar bands and the less distorted middle band, a different scenario can be seen: although their distribution is not as uniform as in the middle band, it is also far less concentrated than the polar bands. This behavior is probably due to the variety of distortion affecting this region, since it is composed by PUs close to the polar bands and close to the middle band as well, resulting in a hybrid region.

The analysis of Fig. 11 and Fig. 13 leads to the conclusion that such aggregate evaluation of the prediction modes, when an arbitrary contiguous region is considered, is able to present evidence regarding the tendency around some modes in specific regions of the video. However, in order to obtain a better insight into the behavior of the intra prediction when applied to ERP 360-degrees video sequences, a more fine-grained evaluation is performed.

To perform such evaluation, firstly the *Broadway* and *SkateboardInLot* video sequences, which are 6k and 8k, respectively, are resized to 4k aiming to have four videos of the same resolution (*AerialCity* and *PoleVault* are already on 4k resolution). Then these videos are encoded with the same parameters as before, including the rotations depicted in Section III – B, and the useful information is extracted.

With this setup, all the evaluated videos are encoded in the same resolution, therefore, all of them are composed by the same number of CTU rows. Furthermore, since the distortion is proportional to the distance from the equator, it is



(a) Lower band (b) Mid-lower band (c) Middle band (d) Mid-upper band (e) Upper band

Fig. 13 Occurrence rate of intra modes in ERP videos, considering five-bands division

possible to perform a fine evaluation of the distortion evaluating each CTU row separately. Therefore, another occurrences distribution evaluation is performed in which each CTU row is evaluated independently from each other.

In order to reduce the amount of data, the results for the four evaluated videos in all mentioned rotations are averaged, and all PU sizes are grouped in the same evaluation considering different weights. Since the minimum PU size is 4×4, this is considered the reference PU size, and each 4×4 PU counts as a single PU. Since 8×8 PUs are composed of 4 PUs of size 4×4, each 8×8 PU counts as 4 PUs. The same process is applied to the remaining sizes, with the 64×64 PUs counting as 256 PUs. This evaluation is presented as a boxplot in Fig. 14, where the horizontal axis represents the prediction modes and the vertical axis represents the CTU rows. It is interesting to highlight that, although the 4k videos are composed of 30 CTU rows (30 × 64 = 1920 vertical samples), the CTCs specify that 4k videos must be encoded into 3328×1664 resolution, which yields 26 rows of CTUs (from 0 up to 25), as depicted in the vertical axis [17]. Furthermore, since the objective of this evaluation is to provide a better insight on the concentration of occurrences around the horizontal mode based on the CTU row, only the horizontally-oriented modes are considered, that is, only modes from 2 up to 18 are considered in the horizontal axis of Fig. 14 and the remaining modes are discarded. In Fig. 14 the boxes represent the second and third quartiles, that is, from 25% up to 75% of all occurrences, the whisker represent the range between 5% and 95% of all occurrences, and the fliers are not represented. In addition, the solid orange line represents the median mode whereas the dashed green line represents the mean mode.

When analyzing the boxplots from Fig. 14, it is visible that the median of the horizontally-oriented modes is mode 10 for all the CTU rows, whereas the mean mode is very close to mode 10 in every row as well. When analyzing the evolution of the boxes throughout the rows, it is visible that the boxes start comprising only the mode 10 in the uppermost and lowermost rows and begin spamming more modes as they approach the central rows. For the group of rows centered in the middle of the frame, however, the boxes present a similar design, and therefore, the distribution of modes is similar. This behavior shows that on the most polar rows the horizontal stretching is such that, on row 0, only three modes (from mode 9 up to mode 11) are responsible for 90% of the total selections, whereas in row 25, only four modes (from mode 8 up to mode 11) are responsible for 90% of the selections. In rows closer to the middle of the frame, the horizontal stretching is alleviated, and it is necessary to employ modes with different angles in order to obtain a good prediction and therefore, the modes are more uniformly selected.

Considering the presented evaluations, it is clear that the ERP projection creates spatial properties which make some modes more likely to be selected than others during the encoding of specific regions of the frame. Furthermore, there is a high concentration of occurrences around the horizontal modes in the polar regions, and this concentration is dispersed as we approach the equator region. Considering this behavior, *it is possible to exploit the distortion caused by the ERP projection to evaluate a reduced set of prediction modes when encoding some areas of the frame, leading to significant complexity reduction whereas posing small harm to the coding efficiency.*

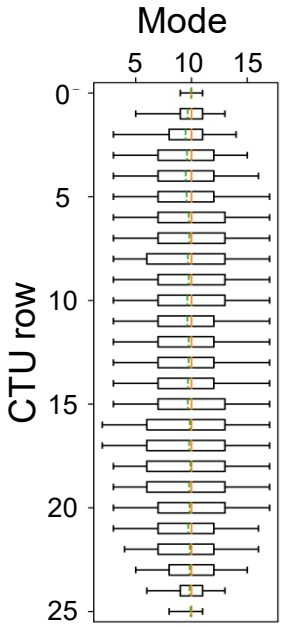


Fig. 14 Boxplot of horizontally-oriented prediction modes per CTU row

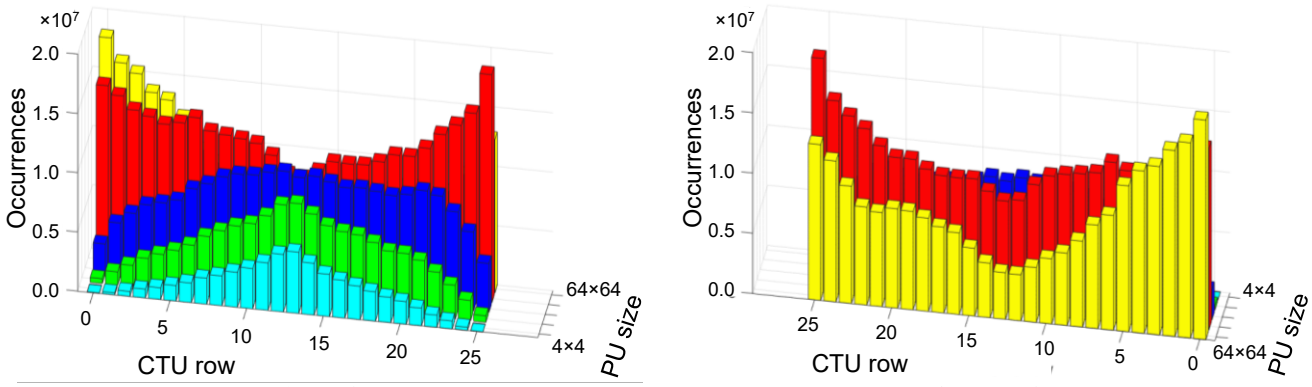
B. PU size evaluation

Considering 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 [18], it is possible that the PU sizes are influenced by the frame region of ERP 360-degrees videos as well, given that the polar areas present a stronger stretching and, therefore, tend to be composed of more homogeneous samples.

Aiming to evaluate the impact of the frame region on the selected PU size, the set composed of the original *AerialCity* and *PoleVault* sequences, plus the *Broadway* and *SkateboardInLot* sequences resized to 4k are evaluated considering all the mentioned rotations, their results are averaged and then, the occurrence rate of each PU size for each CTU row is evaluated. This evaluation is presented in Fig. 15, in which the bars heights represent the 'Occurrences' of each 'PU size' for each 'CTU row', normalizing the PUs according to their dimensions as in Section IV – A. In Fig. 15 the bars for PUs of size 4×4, 8×8, 16×16, 32×32 and 64×64 are colored cyan, green blue, red, and yellow, respectively. In addition, in Fig. 15 (a) the PUs 4×4 are presented in the first plane of the view and PUs 64×64 are presented in the last plane, whereas in Fig. 15 (b) the perspective is inverted and the PUs 64×64 are presented in the first plane and the PUs 4×4 are presented in the last plane. Notice that the rows ordering is also inverted in both figures.

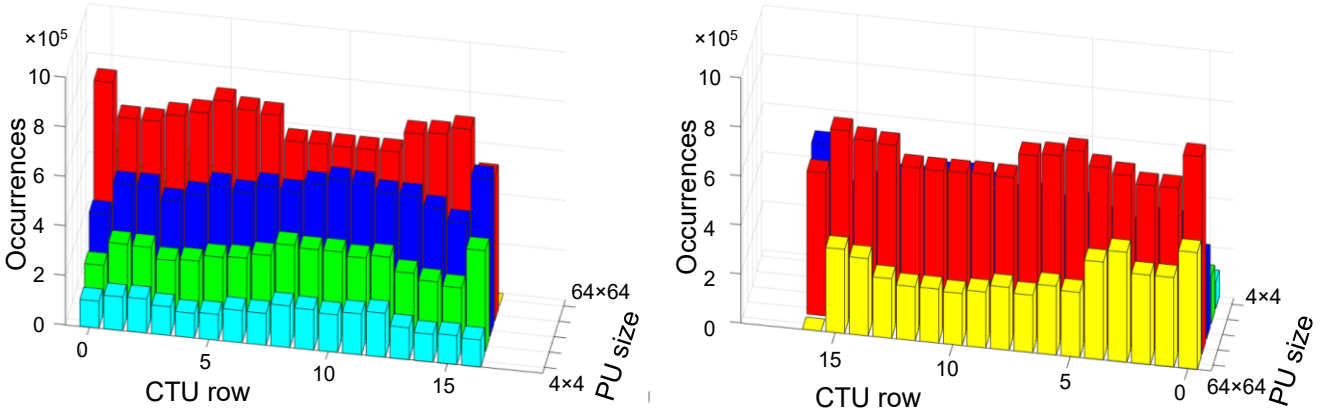
Since all CTU rows have the same number of samples, that is, the same number of normalized PUs, the sum of all PU sizes occurrences is the same CTU row yields the same value. That is, summing any aligned group of bars perpendicular to the 'CTU rows' axis results in the same value.

When analyzing Fig. 15, it is visible that different PU sizes present distinct behaviors. When analyzing the 64×64



(a) 4×4 view (b) 64×64 view

Fig. 15 Distribution of PU sizes per CTU row for ERP 360-degrees videos



(a) 4×4 view (b) 64×64 view

Fig. 16 Distribution of PU sizes per CTU row for conventional videos

PUs, it is clear that the occurrences are clustered in the polar rows, that is, rows 0 and 25, and the occurrences decrease towards the center. For the 32×32 PUs, although there is a higher occurrence rate in the polar rows, the difference is not as significant as for the 64×64 PUs. From the 16×16 PUs to the 4×4 PU, this behavior is inverted. For the 16×16 PUs, the rows around the center of the frame present a similar occurrence rate, whereas the polar rows present a smaller occurrence rate. For the 8×8 and 4×4 PUs, the rows in the center of the frame present a very high occurrence rate, whereas in the polar rows, these PUs rarely occur.

Comparing all the PU sizes, in row 0 the PUs 64×64 are responsible for more than 50% of the total selected normalized PUs, whereas the combination of PUs 64×64 and 32×32 are responsible for more than 90% of the total selections. The contribution of larger PUs decreases as we approach the middle rows. In row 5, for instance, the PUs 64×64 are responsible for around 37% of total selections, whereas the combination of 64×64 and 32×32 PUs corresponds to 71% of the total selections. The largest PU size presents the smallest number of occurrences in the two middle rows (rows 12 and 13), in which the PUs of size 64×64 are responsible for 11% of the total selections, and the PUs 4×4 are responsible for 13% and 14% of the total selections. This behavior shows that, in the polar rows, the PUs 64×64 and 32×32 are responsible for most of the selected PUs, whereas in the middle rows the different PU sizes present a similar occurrence rate.

The same evaluation is performed to the same set of conventional videos as in Section IV – A, for comparison. The results are presented in Fig. 16 (a) and (b), respecting the same color and ordering scheme as in Fig. 15. When analyzing Fig. 16, it is visible that different PU sizes present different occurrence rates, however, there is no clear behavior for any PU size along the CTU rows. The exception is row 16, in which PUs 64×64 never occur. This behavior is due to the resolution of the evaluated conventional videos. All evaluated conventional videos present 1920×1080 resolution, which does not result in an integer number of vertical CTUs (1080 / 64 = 16.875) and therefore, the last row must be partitioned in smaller CUs and PUs.

When comparing the results from Fig. 15 and Fig. 16, it is clear that ERP 360-degrees videos present specific spatial properties that make some PU sizes more likely to be selected in the polar regions, whereas in the middle region the different PU sizes are more uniformly selected. This behavior is due to the overstretching observed in the polar regions: since the stretching is more aggressive in polar regions, more samples are created through interpolation which makes these regions more likely to be encoded with larger PUs. In the middle region, the stretching is less aggressive or even nonexistent, therefore, the distribution of PU sizes is very similar between the conventional and ERP 360-degrees videos. Considering this behavior, *it is possible to exploit the specific spatial properties caused by the distortion of ERP projection to evaluate a reduced set of PU sizes when encoding some regions of the frame, leading to significant complexity reduction whereas posing small penalty to the coding efficiency.*

V. Conclusion

This work presented a thorough evaluation of the impact of the distortion caused by the ERP projection on 360-degrees videos when encoded by the intra-frame prediction of the HEVC standard. During the evaluations, both the prediction modes and PU sizes generated by the intra prediction process were evaluated, and the behavior of the intra prediction when encoding ERP 360-degrees videos was compared with the encoding of conventional videos.

The conducted experiments revealed that when encoding ERP 360-degrees videos, different regions of the frame presented different encoding characteristics. More precisely, it was revealed that in regions closer to the poles the intra prediction tends to select horizontally-oriented modes to perform the prediction more often than the remaining angular modes, and this tendency is reduced as we encode regions closer to the middle of the frame. When encoding PUs in the middle region of the frame, the intra prediction tends to behave similarly to when encoding conventional videos, i.e., it selects all prediction modes with a similar frequency.

The tendency observed when encoding PUs near the poles is due to the horizontal redundancies caused by the ERP projection in such regions. Since these regions are horizontally stretched during the projection, new samples are created through interpolation of existing samples, and therefore, there is an increase in the horizontal redundancy in these regions. Since there are high horizontal redundancies in these regions, horizontally-oriented prediction modes tend to be selected more often than others.

In addition, the performed evaluations made it clear that there are differences in the PU size selection when encoding ERP 360-degrees videos and conventional videos. The experiments showed that whereas in conventional videos the PU sizes present a similar occurrence rate for any region of the frame, when encoding ERP 360-degrees videos the PU sizes selection is heavily based on the frame region. When encoding ERP 360-degrees videos, the regions closer to the poles tend to be encoded using larger PUs, whereas when encoding regions in the middle of the frame, smaller PUs tend to be employed.

This behavior is due to the overstretching in the polar regions of ERP 360-degrees videos. It is known that homogeneous regions tend to be encoded with larger PUs whereas detailed regions tend to be encoded with smaller PUs. Since the horizontal stretching increases the redundancies in the polar regions of the frame, these regions tend to be encoded with larger PUs. Since the central regions present a less aggressive stretching, it presents redundancies similar to those observed in conventional videos, and therefore, a similar PU size distribution is observed.

Finally, the specific behavior observed when encoding ERP 360-degrees videos could be exploited to design fast intra prediction algorithms. A fast intra prediction algorithm designed to deal with ERP 360-degrees videos could evaluate a subset of prediction modes depending on the frame region which is being encoded. In addition, it is possible to discard the evaluation of some unlikely PU sizes according to the frame region being encoded. In doing so, it is possible to achieve expressive complexity reduction whereas posing small coding efficiency penalties, since for some regions a very small combination of prediction modes and PU sizes is responsible for most of the encoder decisions. A fast intra prediction algorithm exploiting the exposed findings could be used to reduce the runtime for current video coding standards, or be employed in the design of the next generation of video coding standards.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and Fundação de Amparo à pesquisa do Estado do Rio Grande do Sul Brasil (FAPERGS).

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] R. Skupin, Y. Sanchez, Y. -. Wang, M. M. Hannuksela, J. Boyce and M. Wien, "Standardization status of 360 degree video coding and delivery," 2017 IEEE Visual Communications and Image Processing (VCIP), St. Petersburg, FL, 2017, pp. 1-4.

[5] Versatile Video Coding, Online: https://jvet.hhi.fraunhofer.de/.

[6] P. Hanhart, J-L. Lin, C. Pujara, “CE13: Summary report on coding tools for 360° omnidirectional video”, JVET Doc. JVET-M0033, 2019.

[7] 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.

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

[9] 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.

[10] I. Storch *et al.*, “Speedup-Oriented History-Based Tiling Algorithm for the HEVC Standard Targeting an Efficient Parallelism Exploration,” in Journal of Integrated Circuits and Systems, vol. 13, no. 1, pp. 1-8, 2018.

[11] 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.

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

[13] K. Mulcahy *et al.*, “Symbolization of Map Projection Distortion: A Review,” in Cartography and geographic information science, vol. 28, no. 3, pp. 167-182, 2001.

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

[15] High Efficiency Video Coding Test Model 16.16, available: <https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/tags/HM-16.16>. Last access: November 2018.

[16] G. J. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," in IEEE Signal Processing Magazine, vol. 15, no. 6, pp. 74-90, Nov. 1998.

[17] J. Boyce *et al.*, “Common Test Conditions and Evaluation Procedures for 360° Video Coding,” JVET Doc. JVET-D1030, 2016.

[18] 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. 1578-1582.