Extremely Randomised Trees for Computational Complexity Reduction of Omnidirectional Intra Video Coding

Jose N. Filipe[1]
jose.filipe@av.it.pt
J. Carreria[1,2]
jcarreira@co.it.pt
Luis M. N. Tavora[2]
luis.tavora@ipleiria.pt
Sergio M. M. Faria[1,2]
sergio.faria@co.it.pt
Antonio Navarro[1,3]
navarro@av.it.pt
Pedro A. A. Assuncao[1,2]
amado@co.it.pt

[1] Instituto de Telecomunicações,
Campus Universitário de Santiago
Aveiro, PT

[2] Politecnico de Leiria,
ESTG, Morro do Lena - Alto do Vieiro,
Leiria, PT

[3] Universidade de Aveiro,
Campus Universitário de Santiago,
Aveiro, PT

Abstract

This paper presents a novel method to reduce the computational complexity of intra-coded 360° video in Equirectangular Projection (ERP) format. The proposed method is based on three Extremely Randomised Trees models to predict the maximum partition depth that should be used for intra-coding of the complex nested data structures (Quad Tree, Binary Tree and Ternary Tree) used in the forthcoming video coding standard (Versatile Video Coding). The results show that an average complexity reduction of 31.35% is achieved, with a negligible loss of coding efficiency of just 0.42%, outperforming other state of the art low complexity solutions.

1 Introduction

Image and video data represent the majority of all internet traffic, due to new applications and emerging services using mixed reality, namely 4K and 8K resolutions, cloud gaming, self-driving vehicles, smart surveillance systems, among others. These advanced services and applications require very high resolutions and complex compression algorithms. Thus, standardisation efforts specifically targeting emerging video formats, such as 360° video [5] are in place. Specifically, the Joint Video Exploration Team (JVET) is developing the forthcoming video compression standard, named Versatile Video Coding (VVC), to face the challenging requirements posed by higher resolutions and new visual representation formats. This new standard greatly increases the coding efficiency of its predecessor High Efficiency Video Coding (HEVC). However, such improvement is achieved at the cost of a great deal of additional computational complexity. As the VVC encoder is 7 to 9 times more complex than HEVC [10], fast computational methods are of utmost importance to ease adoption of the this standard and to meet implementation constraints.

Historically, previously proposed methods to reduce the complexity of video encoders focus on reducing the number of tests performed by the Rate-Distortion Optimisation (RDO) method, or by replacing such process by a fast decision algorithm that avoids the computation of each block coding cost, aiming to reduce the number of methods used to partition the Coding Tree Units (CTUs) into Coding Units (CUs). This process is even more complex in VVC than in HEVC, given that besides Quadtree (QT) partitions, two more partition types have been added, namely Binary Tree (BT) and Ternary Tree (TT) partitions. To tackle this problem, Na Tang *et al*. leverage the Canny Edge detector to preform early termination if the CU is uniform enough, or select horizontal or vertical partitions, depending on the ratio between the number of horizontal and vertical detected edges [8]. Jing Cui *et al*. base their decision upon the gradients of the CU, to choose what partition type should be applied to the CU [2]. Thomas Amestoy *et al*., take advantage of a number of features fed into a set of Random Forests models trained for each partition depth, to decide whether QT or BT should be applied [1]. Genwei Tang *et al*., on the other hand, propose a Split/No-Split approach where a shape-adaptive Convolutional Neural Network replaces the RDO process and to decide whether a given CU should be or not further split [7].

In this paper, we propose an off-loop early termination method, that leverages three Extremely Randomised Trees (ERT) models to predict the maximum partition depth per partition type (QT, BT and TT), that can be achieved in a given CTU. Once the maximum partition depth for a given partition type is achieved, no further partition of the same type is performed. The remainder of the paper is organised as follows: section 2 presents a detailed description of the proposed method, section 3 presents the achieved results and, finally, some conclusions are drawn in section 4.

Table 1: Summary of the used features.

ID	Feature
1	Latitude of the centre point of the CTU
2	Secant of the latitude of the centre point of the CTU
3	Spatial Information
4	Std. Dev. of Sobel filtered CTU along x
5	Std. Dev. of Sobel filtered CTU along y
6	Std. Dev. of Sobel filtered bottom left fourth of CTU, along y

2 Proposed Method

The proposed method extracts off-loop features from a given CTU, feeds them into tree ERT models (one for each type of partition, *i.e*. QT, BT, and TT), that predicts the maximum depth (per partition type). This limits the partition depth that is going to be tested by the RDO process. For example, if the models predict that the maximum depth of the QT, BT, and TT partitions are 2, 2, and 2, respectively, it means that only two depths of each partition type will to be tested, greatly reducing the total number of hypothesis to be tested using RDO, and thus reducing the overall complexity of the encoding process.

It is worthwhile to notice that not all CUs resulting from the RDO process present the maximum estimated complexity, since larger CUs may be more suitable to certain regions of the CTU. However, in order to limit the impact of this early termination method in the coding efficiency, only the maximum depths are estimated. In other words, if the predictive models had 100% accuracy, this method would have had absolutely no impact on the coding efficiency. Therefore, all coding efficiency losses are directly caused by the models mis-classification.

2.1 Features

Initially, a set of 56 features was extracted from the CTUs of the training dataset. Then, ERT models were used to preform Recursive Feature Elimination (RFE).

Features 1 and 2 from Table 1 take advantage of the geometric characteristics of the Equirectangular Projection (ERP) [6]. As noticed in [4], most of the coding complexity related to the ERP format is clustered near the equator, while regions near the poles tend to require lower complexity. Furthermore, it can be demonstrated that the ERP distortion that originates regions of lower complexities has a direct relationship with the $secant(l)$ of the latitude (l). Therefore, Feature 1 discriminates the vertical position of a given CTU, while Feature 2 is related to the distortion that

spawns the low complexity regions. Feature 3 is the Spatial Information [9], while Features 4 and 5 discriminate between the horizontal and vertical directions, respectively. Finally, to capture finer detail, the CTUs were split into four smaller squares of 64 by 64 pixels, and the same spatial features were computed to each of these squares. The standard deviation of the bottom left square along the y direction was deemed relevant to predict the maximum CTU partition depth, by the RFE process.

2.2 ERT Model

The advantages of ERT over Random Forests, arise from the method used to choose attributes and cut-points while cutting a tree node. In Random Forest this is done by finding the local optimum cut-point for each feature, using a metric such as Information Gain. In ERT, a set of cut-points is randomly generated for each feature. Then, the cut-point from the set that yields the best accuracy is selected. Furthermore, in ERT each decision tree is trained over the entire training dataset.

The three ERT models were trained to predict the maximum partition depth for each of the partition schemes (QT, BT, and TT). To achieve this, a training/testing dataset was generated by encoding the 10 ERP sequences recommended by [3], in all intra configuration and $QP = 22$, and then registering the maximum depth achieved by each partition type for all CTUs in the frame. Furthermore, a set of 6 features mentioned in Section 2.1 was extracted for each CTU.

Finally, the ERT models were trained and tested using Cross-Validation, such that the models were trained using data from 9 sequences and then tested against the remaining one. Using this methodology, the QT model achieved an Average Accuracy on the test dataset of $71 \pm 6\%$, the BT model $67 \pm 12\%$, and the TT model $92 \pm 7\%$. Leveraging the 0-1 loss metric, we can approximate the bias and the variance of each of the 3 models. The QT model presents an estimated bias of 29% and an estimated variance of 6%, the BT model bias of 33% and variance of 12%, and the TT model bias of 8% and variance of 7%. This shows that the TT model presents low levels of both bias and variance, as desired, indicating that this model present neither over nor underfitting. Regarding the QT and BT models, a relatively low variance and higher bias is presented, indicating some level of underfitting.

After the three models have been trained, they were implemented within the VVC encoder and the respective functions are called each time a new CTU is encoded. Then, the partition depth limits are updated according to the prediction of the models. Some constrains were implemented in the encoder, so that no partition depths above the predicted maximum are evaluated by the RDO process. It is worthwhile to note that this off-loop approach has the advantaged that the models are required to run only once per CTU, resulting in negligible complexity overhead. In-loop approaches often have to compensate the introduced overhead, since their functions are typically called several times during the encoding of a single CTU.

3 Results

The proposed method was evaluated by measuring the processing time required to encode each of the 10 sequences, and then comparing the time and coding efficiency with the same sequences encoded using the standard VVC reference software (VTM 8.0). All sequences have a 4432 by 2216 resolution, and were encoded using all intra configuration, next profile, and a set of 4 QPs (22, 27, 32, and 37), in order to compute Bjontegaard Delta Rate (BD-Rate). This metric was used to evaluate the coding efficiency, while the complexity was evaluated by computing the average across the 4 QPs of the difference between the encoding time using the proposed method and using the reference VVC, normalised to the encoding time of the latter.

Table 2 shows these results for all 10 sequences. In all cases, the proposed method presents significantly reduced complexity, when compared to the unaltered implementation of VVC, with a negligible loss of coding efficiency, that is less than 1% for 9 out of the 10 cases. In fact, the proposed method presents on average 31.35% complexity reduction, with an average increase in bitrate for a given visual quality of about 0.42%.

Moreover, if one divides the average complexity reduction by the BD-Rate, to determine the percentage of complexity reduction that a given method can achieve per each 1% of BD-Rate loss, we can conclude that the proposed method achieves a ratio of 74.30, outperforming other state of the art methods, such as [8] (23.39), [7] (33.75), [2] (50.00), and [1] (52.63).

Table 2: Results for the proposed method.

Sequence	BD-Rate (%)	Avg. Complex. Reduction (%)
Harbor	0.79	−26.00
KiteFlite	0.42	−28.46
Balboa	0.36	−31.43
BranCastle	0.24	−35.71
Broadway	0.36	−31.23
Landing2	0.30	−35.88
SkateBoardInLot	0.96	−36.45
ChailiftRide	1.06	−35.92
Trolley	0.42	−31.27
Gaslamp	0.72	−25.69
Average	0.42	−31.35

4 Conclusions

In this paper we propose a novel algorithm, that leverages 3 ERT models to predict the maximum partition depth of each partition type (QT, BT, and TT) for every CTU in intra frames of 360° video sequences in ERP format. The prediction is used in a modified version of the VVC, to limit the RDO process to depths smaller than the maximum partition depths predicted by the models. The proposed method achieves an average complexity reduction of 31.35%, with a negligible average coding efficiency loss of just 0.42%. Additionally, the proposed method is able to outperform other state of the art low complexity solutions, such as [7, 8].

Acknowledgements

This work was supported by Programa Operacional Regional do Centro, project ARoundVision CENTRO-01-0145-FEDER-030652 and by FCT/MCTES through national funds and when applicable co-funded EU funds under the project UIDB/EEA/50008/2020, Portugal.

References

[1] T. Amestoy, A. Mercat, W. Hamidouche, C. Bergeron, and D. Menard. Random forest oriented fast qtbt frame partitioning. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1837–1841, 2019.

[2] J. Cui, T. Zhang, C. Gu, X. Zhang, and S. Ma. Gradient-based early termination of cu partition in vvc intra coding. In *2020 Data Compression Conference (DCC)*, pages 103–112, 2020.

[3] P. Hanhart, J. Boyce, K. Choi, and J.-L. Lin. L1012: JVET common test conditions and evaluation procedures for 360° video. Technical report, Joint Video Experts Team (JVET), 12th Meeting: Macau, CH, October 2018.

[4] B. Ray, J. Jung, and M. Larabi. A low-complexity video encoder for equirectangular projected 360 video content. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1723–1727, April 2018. doi: 10.1109/ICASSP.2018.8462368.

[5] R. Skupin, Y. Sanchez, Y. . Wang, M. M. Hannuksela, J. Boyce, and M. Wien. Standardization status of 360 degree video coding and delivery. In *IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4, December 2017. doi: 10.1109/VCIP.2017.8305083.

[6] John P. Snyder. Map projections: A working manual. Technical report, U.S. Government Printing Office, 1987.

[7] G. Tang, M. Jing, X. Zeng, and Y. Fan. Adaptive cu split decision with pooling-variable cnn for vvc intra encoding. In *2019 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4, 2019.

[8] N. Tang, J. Cao, F. Liang, J. Wang, H. Liu, X. Wang, and X. Du. Fast ctu partition decision algorithm for vvc intra and inter coding. In *2019 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*, pages 361–364, 2019.

[9] H. Yu and S. Winkler. Image complexity and spatial information. In *Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 12–17, July 2013. doi: 10.1109/QoMEX.2013.6603194.

[10] Fan Zhang, Angeliki V. Katsenou, Mariana Afonso, Goce Dimitrov, and David R. Bull. Comparing vvc, hevc and av1 using objective and subjective assessments, 2020.