# A HARDWARE-FRIENDLY ALGORITHM AND ITS ARCHITECTURE FOR ANGULAR INTRA PREDICTION OF VVC

VINICIUS BORGES[1]; MURILO PERLEBERG[2];
MARCELO PORTO[3]; LUCIANO AGOSTINI[4]

[1]*Universidade Federal de Pelotas – vdaborges@inf.ufpel.edu.br*
[2]*Universidade Federal de Pelotas – mrperleberg@inf.ufpel.edu.br*
[3]*Universidade Federal de Pelotas – porto@inf.ufpel.edu.br*
[4]*Universidade Federal de Pelotas – agostini@inf.ufpel.edu.br*

## 1. INTRODUCTION

The digital video consumption on the internet is increasing expressively in the last years. One of the main motivations is the current COVID-19 pandemic. Great part of the population, avoiding the pandemic, stay at home. Consequentially, they consume more digital information, such as video streaming services, provided by Netflix, YouTube, Amazon Prime Video, and so many others. The giant companies were forced to reduce the video quality to support the demand.

For this purpose, Joint Video Experts Team (JVET) was created in a collaboration between ISO Moving Picture Experts Group (MPEG) and ITU-T Video Coding Experts Group (VCEG) to develop the Versatile Video Coding (VVC) standard (ITU-T AND ISO/IEC, 2020), which was established as a Final Draft International Standard (FDIS) in July 2020. VVC has several improvements to increase the encoding efficiency, including block partitioning with the increase in the maximum size of the Coding Tree Unit (CTU) to 128x128 and allowing a flexible partitioning of blocks using a Quadtree with nested Multi-type Tree (QTMT) structure. Some improvements are related with the intra prediction, such as, the increase of the number of angular prediction modes and the creation of new intra prediction tools, like Multiple Reference Line (MRL), Matrix-based Intra Prediction (MIP) and Intra Subpartition (SALDANHA, 2021).

The focus of this work is the VVC block partitioning and angular intra-frame prediction. All new tools added in VVC improved the encoding efficiency but raised the encoder computational effort expressively. Aiming the reduction of this computational effort, this work proposes a hardware-friendly strategy by removing part of the less used angular modes and removing the new block sizes for the angular intra-frame prediction. Also, the work presents an architecture design for the proposed strategy. The results of the algorithm and hardware architecture will be presented in the next sections.

## 2. METHODOLOGY

The VVC inherited HEVC angular prediction modes used in intra prediction and extended them, increasing the number of angular modes to 67. To evaluate the encoder behavior, the usage from the intra prediction modes was extracted from the VVC Test Model (VTM) (SALDANHA, 2021). The Planar and DC modes have the higher use, reaching 38% and 5% of use, respectively. The other modes have less than 5%, where the angular modes 18 and 50 were used in 3.48% and 4.55% of the cases, respectively. The rest of the angular modes were used from 0.5% to 1% of the cases, with a few exceptions which reach a little more than 1%. Another important point is related to the use of different block shapes supported on VVC. Even

with the new rectangular block sizes added on VVC, the most used block sizes tend to be the quadratic ones inherited from HEVC (SALDANHA, 2021).

Based on these facts, this work adopts a strategy aiming reduce the computational effort of the VTM, based on the block sizes and intra prediction modes most used. The first computational effort reduction strategy was to remove the new rectangular block sizes adopted on VVC intra prediction. Therefore, only 4x4, 8x8, 16x16, 32x32 and 64x64 block sizes were used.

The second computational effort reduction strategy was to reduce the number of intra prediction modes evaluated. Then, only the modes with an usage higher than or equal to 1% were selected. Therefore, initially 14 modes were selected to be supported by our dedicated hardware: Planar, DC, 2, 3, 10, 18, 33, 34, 35, 43, 46, 49, 50 and 54. In addition to these modes, the angular modes 7, 23, 26 and 30 were also considered to fulfill the highest gaps among the previously selected modes. Thus, this second strategy reduced the total intra prediction modes from 67 to only 18 modes.

### A. Experimental Setup and Evaluation Results

The experiments performed to evaluate the algorithm were based on the Common Test Conditions (CTC) document (BOSSEN, 2019) recommended by the Joint Video Experts Team (JVET). The experiments were conducted using version 14.0 of the VTM reference software in All-Intra (AI) temporal configuration (BOSSEN, 2019). Five Full HD (1920×1080) video sequences were used: MarketPlace, RitualDance, Cactus, BasketballDrive, and BQTerrace. From each video sequence were encoded from 100 to 120 frames, representing two seconds of video from each evaluated sequence. Finally, the Quantization Parameters (QPs) defined in the CTC were used for the experiments: 22, 27, 32, and 37.

The efficiency of the proposed low-complexity angular algorithm was evaluated based on the Bjontegaard Delta Rate (BD-Rate) metric. The BD-Rate value represents the bitrate increase to maintain the same image quality. In comparison with the original VTM algorithm, the developed solution has an average increase of 8.92% in the BD-rate for AI configuration. The computational cost reduction (CR) to encode each video sequence was also evaluated. In comparison with the original algorithm, the developed solution has an average decrease of 73.34% in computational effort.

### B. Dedicated Hardware Design

A dedicated architecture of the proposed low-complexity angular algorithm was designed targeting a low-power and high-performance solution. The high-level architecture manages the control of the machines and provides signals for the other units. The architectural design was divided in nine unities: multi-mode unit and eight directional units. The multi-mode unit has two parallel modules to perform both planar and DC calculations. The directional units are responsible to process two angular modes in parallel, totalizing 16 angular modes.

The multi-mode unit is responsible to sequentially process the predicted block (PB) of Planar and DC modes. After obtaining the PB for each of those modes, a multiplexer in the multi-mode unit selects the output of the mode being processed to obtain the residual block. To compute the predicted block of the Planar mode, each sample of the PB is computed from the weighted average of its four neighboring samples, located at the edge of the PB. For the DC mode, the predicted block

is represented by the DC value, which is the average value of all neighboring samples from the processed PB.

Each directional unit is responsible for processing only the Angular modes. Then, the directional unit processes the angular modes 2, 3, 7, 10, 18, 23, 26, 30, 33, 34, 35, 43, 46, 49, 50 and 54. The computation of these 16 modes is divided into eight instances of the directional unit, where each one is responsible to process two angular modes in order. As example, the first directional unit has angular modes 2 and 3, the second directional unit has modes 7 and 10, and so on. All these units operate in parallel filled by the same neighboring input samples.

To process each mode, each sample from the predicted block is generated based on the weighted average of some specific neighbor samples, according to the direction of the angular mode processed. Then, each angular unit adopts the strategy of deriving a reference vector from the neighboring samples before predicting the samples of the PB using that vector. This enables the sharing of the weights between all the lines or between all the columns from the predicted block.

## 3. RESULTS AND DISCUSSION

The designed architecture was described in VHDL and synthesized targeting the TSMC 40nm technology. The Cadence RTL compiler tool was used to perform this synthesis. Table I summarizes the reached results and presents the results of some related works.

The synthesis results show that the developed hardware requires an area of 1,453k gates. The designed architecture requires 4,991 cycles to process all block sizes smaller than a 64×64. The synthesis results shows that when operating at 75.8 MHz, the designed architecture dissipates 91.65 mW to process Full HD videos at 30 frames per second.

The designed architecture was compared with the two related works, as presented in Table I. This comparison was done even with the difficulties to do a fair comparison due to the use of distinct standards, different technologies, and the different methodology used in these works to measure the coding efficiency. These related works propose hardware solutions for the HEVC, supporting all HEVC intra prediction. However, it is important to highlight that VVC standard has more encoding tools and reaches higher coding efficiency than HEVC, which makes unfair the comparisons about the BD-Rate results.

The works of (HUANG, 2016) and (ZHANG, 2019) are focused on 55nm and 90nm technologies, whereas our work is focused on 40nm technology. The comparison about the coding efficiency was done considering the reported BD-Rate results in relation with the target standard, and through a normalized BD-Rate. Considering that VVC doubled the coding efficiency in relation to HEVC, then it is plausible to do a normalization using the HEVC as basis. This normalization showed that our solution, in fact, has a much higher coding efficiency than the related works, as presented in Table I (negative results are gains in BD-Rate).

When considering area and power, our architecture surpasses the related works. Our area is 7.5% smaller than (HUANG, 2016) and 36.5% smaller than (ZHANG, 2019). Our power dissipation is about 52.75% lower than (HUANG, 2016) and 61.15% lower than (ZHANG, 2019). However, it is hard to fairly compare the power results of these works because each work focus on a different technology, which can explain great part of the difference.

Table I: Synthesis Results and Comparisons.

| | This work | Huang X. | Zhang |
|---|---|---|---|
| Standard | VVC | HEVC | HEVC |
| Modes | Planar, DC, 2, 3, 10, 18, 23, 26, 30, 33, 34, 35, 43, 46, 49, 50 and 54 | Planar, DC and 33 angular modes | Planar, DC and 33 angular modes |
| BD-Rate (%) for Full HD | 8.92% in relation to original VVC | 4.3% in relation to original HEVC | 3.02% in relation to original HEVC |
| Normalized BD-Rate (%) for Full HD | -91.08% | 4.3% | 3.02% |
| Computational Effort Reduction | 73.34% | 41.6% | 27% |
| Technology | 40nm | 55nm | 90nm |
| Area (k gates) | 1,453 | 1,571.7 | 2,288 |
| Frequency (MHz) | 75.8 | 294 | 320 |
| Power | 91.65 mW (1080p@30fps) | 194mW (1080p@60fps) | 236mW (2160p@30fps) |

## 4. CONCLUSIONS

This paper presented a hardware architecture for the VVC angular intra prediction tool. This dedicated architecture implements a new low-complexity algorithm for angular that was also proposed in this work. This algorithm reduces the number of evaluated modes from 67 to 18. Moreover, the algorithm also reduces the PB sizes supported to only quadratic ones. These simplifications allowed an expressive computational effort reduction of 73.34% in average, when compared with the original VTM implementation, at a cost of 8.92% in coding efficiency (BD-Rate). Synthesis results showed that the designed architecture can process Full HD videos at 30 frames per second, with a power dissipation of 91.65 mW. When compared with related works, the presented solution reached the best results in all compared axes.

## 5. REFERENCES

ITU-T and ISO/IEC, Versatile Video Coding, ITU-T Rec. H.266 and ISO/IEC 23090-3, 2020.

Saldanha M. et al. Performance analysis of VVC intra coding. **Journal of Visual Comm. and image rep.,** V. 79, 2021.

Huang, X.; Jia, H.; Cai, B.; Zhu, C.; Liu, J.; Yang, M.; Xie, D.; Gao, W. Fast algorithms and VLSI architecture design for HEVC intra-mode decision. **J. Real-Time Image Process**. p. 285–302, 2016.

Zhang, Y.; Lu, C. Efficient Algorithm Adaptations and Fully Parallel Hardware Architecture of H.265/HEVC Intra Encoder. **IEEE Transactions on Circuits and Systems for Video Technology**, v. 29, n. 11, p. 3415-3429, 2019.

Bossen F.; Boyce J.; Suehring X.; Li X.; Seregin V. JVET common test conditions and software reference configurations for SDR video. **JVET 14th Meeting**, JVETN1010, 2019.