MULTIMEDIA COMMUNICATIONS TECHNICAL COMMITTEE
http://www.comsoc.org/~mmc

MMTC Communications - Frontiers

Vol. 12, No. 5, September 2017

CONTENTS

Message from the MMTC Chair .................................................................................................................. 3
SPECIAL ISSUE ON Low Complexity Codecs for Power Constrained Video Communications .............................................................. 4

Guest Editors: Pedro A. A. Assunção1,2 and Carl James Debono3 .................................................. 4
1Instituto de Telecomunicações, 2Politécnico de Leiria, Leiria, Portugal ........................................... 4
3Dep. of Communications and Computer Engineering, University of Malta, Msida, Malta ................................................................. 4
{paassunc, c.debono}@ieee.org ........................................................................................................... 4

Fast and Low-Power Hardware Design for HEVC Fractional Motion Estimation .......... 6
Vladimir Afonso1,2, Ruhan Conceição3, Murilo Perleberg2 ............................................................ 6
Marcelo Porto2, Bruno Zatt4, Luciano Agostini2, Altamiro Susin1 .................................................. 6
1Graduate Program on Microelectronics, Federal University of Rio Grande do Sul, Porto Alegre, Brazil ................................................................. 6
2Video Technology Research Group, Federal University of Pelotas, Pelotas, Brazil ........ 6
{vafonso, radconceicao, mrperleberg, porto, zatt, agostini}@inf.ufpel.edu.br;
altamiro.susin@ufgrs.br .................................................................................................................. 6

Recent Advances in Complexity Reduction Methods for High Efficiency Video Coding ........................................................................ 12
Pedro Assuncao1,2, Luis Cruz1,3, Carl James Debono4 ..................................................................... 12
1Instituto de Telecomunicações, Portugal ......................................................................................... 12
2Instituto Politécnico de Leiria, Leiria, Portugal............................................................................. 12
3Dep. of Electrical and Computer Engineering, University of Coimbra, Coimbra, Portugal ............................................................................ 12
4Dep. of Communications and Computer Engineering, University of Malta, Msida, Malta ................................................................................. 12
paassunc@ieee.org; lcruz@deec.uc.pt; c.debono@ieee.org .......................................................... 12

Prediction of the Quad-tree Partitioning for Real-Time Low Energy HEVC Encoders .............................................................................. 20
Alexandre Mercat1, Florian Arrester1, Wassim Hamidouche1, Maxime Pelcat1, Daniel Menard .................................................................................. 20
1UBL, INSA Rennes, IETR CNRS UMR 6164, Rennes, France ........................................ 20
{amecart, farrester, mpelecat, whamidou, dmenard}@insa-rennes.fr ......................................... 20

Advanced Techniques for Power- and Energy-Efficient Design of Video Codecs in the Era of Internet-of-Things .................................................. 27
Muhammad Shafique1, Rehan Hafiz2, Semeen Rehman1, Jörg Henkel3 ........................................... 27

http://mmc.committees.comsoc.org

1/36

Vol.12, No.5, September 2017
IEEE COMSOC MMTC Communications – Frontiers

1Vienna University of Technology (TU Wien), Vienna, Austria ........................................... 27
2Information Technology University (ITU), Lahore, Pakistan ............................................. 27
3Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany ........................................... 27
{muhammad.shafique, semen.rehman}@tuwien.ac.at, rehan.hafiz@itu.edu.pk,
henkel@kit.edu ..................................................................................................................... 27

MMTC OFFICERS (Term 2016 — 2018) ............................................................................. 36
Dear MMTC colleagues and friends,

I would like to take this opportunity to give my thanks to MMTC Communications—Frontiers Director Guosen Yue and other three co-directors Danda Rawat, Hantao Liu and Dalei Wu, and all guest editors/editors for their great efforts and continuing to make MMTC Frontiers better. I also want to thank MMTC web chair Prof. Haixia Zhang for her outstanding work. MMTC Frontiers has been becoming an important platform that connects our community members and continues to make a positive impact to IEEE society. I encourage all of you to submit your work to MMTC Frontier.

In recent years, the significant growth of multimedia applications has raised many new research issues that we need to address, which also motivates further communications and collaborations among our community members. Therefore, there are strong needs to further expand our MMTC community and promote the communications. I am glad to see the continuous growth of our MMTC community. More and more members are being actively involved in the conferences such as IEEE ICME and ICC/Globecom, and IEEE journals such as IEEE TMM, TCSVT and IEEE Multimedia Magazine. I would like to encourage all our MMTC members to consider the opportunity of serving our community to enhance the communications. In addition, MMTC has created a newsletter, which is distributed weekly to our community through the MMTC mailing list. I would like to thank Prof. Mugen Peng for his outstanding leadership in organizing the newsletter. If there is any news, question or problem, please do not hesitate to contact with us.

Please enjoy this issue of MMTC Communications—Frontiers and let us know if you have any comments.

Have a wonderful holiday season. I wish you all the best!

Vice chair of MMTC Letters & Member Communications
Multimedia Communications TC of IEEE ComSoc

http://mmc.committees.comsoc.org
SPECIAL ISSUE ON Low Complexity Codecs for Power Constrained Video Communications

Guest Editors: Pedro A. A. Assunção¹,² and Carl James Debono³

¹Instituto de Telecomunicações, ²Politécnico de Leiria, Leiria, Portugal
³Dep. of Communications and Computer Engineering, University of Malta, Msida, Malta

(paassunc, c.debono)@ieee.org

Standard video codecs have been dramatically increasing their coding efficiency in the last decades, but this came at the expense of many-fold increase in computational complexity. Since the primary goal of video coding researchers and standardisation bodies was to reach high levels of coding efficiency, for some time the computational complexity was not the most important aspect of new algorithms and coding tools under investigation. However, after the target compression levels were achieved through intensive research and development, reducing and controlling the algorithmic complexity became an important research area to lower the computational power requirements and energy consumption of the High Efficiency Video Coders (HEVC). Since then, many different research approaches were followed for such purpose. The main challenge has been the optimisation of the trade-off between low computational power and high coding efficiency, by maximising the reduction of the former while minimising the loss in the latter. Recent research has been dealing with both software and hardware solutions, either separately or jointly to maximise the overall performance. In general, software-based methods are focused on algorithmic optimisation to speed-up the various coding decision processes and parameters by using several different approaches to reach near-optimal decisions in a rate-distortion sense, without searching the full space of possible options. In the case of hardware-oriented methods, the aim is also to simplify the most critical processing modules in order to obtain lower energy consumption, power-efficient memory architectures and specialised hardware accelerators to achieve lower energy consumption along with faster data processing. This special issue addresses relevant topics associated with video coding computational complexity and power/energy efficient techniques, through the following articles.

The first article, by V. Afonso et al., entitled “Fast and Low-Power Hardware Design for HEVC Fractional Motion Estimation”, presents a hardware design for fast motion estimation, which is able to process UHD 2160p videos at 60 fps with low-energy consumption. The hardware architecture is based on a complexity-reduction strategy for the motion estimation function of HEVC. The authors developed a method where the number of prediction unit (PU) sizes evaluated in the motion estimation process is constrained to the four square-shaped sizes, based on statistical analysis of the block-sizes distribution and comparison of different test cases using the constrained PU sizes. Better results than other previous works published in the literature are reported in terms of power and coding efficiency.

The second article by P. Assuncao et al., entitled “Recent Advances in Complexity Reduction Methods for High Efficiency Video Coding” provides an overview of methods to reduce the computational complexity of HEVC encoding through fast coding mode decisions. Two different approaches are addressed, classified as (i) Statistical methods based on spatio-temporal correlations of coding parameters related with algorithmic tools and processing; and (ii) Methods based on Machine Learning approaches, mostly used to achieve fast coding decisions based on different types of classifier models. Recent works, techniques and relevant results are presented and discussed.

The third article by A. Mercat et al, entitled “Prediction of the Quad-tree Partitioning for Real-Time Low Energy HEVC Encoders” describes two methods for reducing the energy consumption of HEVC Intra encoder by limiting the recursive rate-distortion optimization process. The first one exploits the correlation between coding tree unit (CTU) partitioning and the luminance variance of coding unit (CU) samples to devise a probabilistic CTU partitioning prediction method, while the second one follows a machine learning approach to predict the CTU partitioning. The authors focus this work on energy consumption for real-time encoding.

Finally, the fourth article by M. Shafique et al., entitled “Advanced Techniques for Power- and Energy-Efficient Design of Video Coders in the Era of Internet-of-Things” presents a quite comprehensive review and analysis of the HEVC encoding systems and optimized hardware/software solutions for power-energy-efficient design of video codecs. Research directions and emerging trends on software and hardware-based techniques to improve energy efficiency, by means of algorithmic techniques, hardware accelerator design and dense parallel processing are also presented and discussed. This article consolidates the previous ones by providing a complementary analysis.
IEEE COMSOC MMTC Communications - Frontiers

supported by relevant bibliographic references and hints for further research.

Although this Special Issue does not aim to cover the whole state of the art of low complexity video codecs for power efficient video communications, the four papers selected for publication present a quite comprehensive overview of this research area and also highlight different promising research directions. We hope that these papers provide relevant technical content for the interested audience working in the field and valuable insight for persisting challenges and future research.

Our special thanks go to all authors for their precious contributions to this Special Issue. We would also like to acknowledge the support from the Board of MMTC Communications - Frontiers.

Pedro Assuncao (M’99, SM’14) received the Licenciado and M.Sc. degrees in Electrical Engineering from the University of Coimbra, in 1988 and 1993, respectively, and the Ph.D. in Electronic Systems Engineering from the University of Essex, in 1998. He is currently professor of Electrical Engineering and Multimedia Communication Systems at the Polytechnic Institute of Leiria and researcher at the Institute for Telecommunications, Portugal. He is author/co-author of more than one hundred scientific/technical papers in conferences and journals, one book, seven book chapters and four US patents. His current research interests include high efficiency and 3D video and light field coding, multiple description and robust coding, error concealment and quality evaluation. He was the Chair of EU COST Action IC1105 (3D-ConTourNet) and senior member of the IEEE.

Carl James Debono (S’97, M’01, SM’07) received his B.Eng. (Hons.) degree in Electrical Engineering from the University of Malta, Malta, in 1997 and the Ph.D. degree in Electronics and Computer Engineering from the University of Pavia, Italy, in 2000. Between 1997 and 2001 he was employed as a Research Engineer in the area of Integrated Circuit Design with the Department of Microelectronics at the University of Malta. In 2000 he was also engaged as a Research Associate with Texas A&M University, Texas. In 2001 he was appointed Lecturer with the Department of Communications and Computer Engineering at the University of Malta and is now an Associate Professor. He also served as Deputy Dean of the Faculty of ICT between November 2009 and September 2015. He is currently the Head of the Department of Communications and Computer Engineering at the University of Malta. Prof. Debono is a senior member of the IEE. He was a member of the management committee of the COST Action IC1105 - 3D Content Creation, Coding and Transmission over Future Media Networks (3D-ConTourNet) where he chaired the 3D Media Coding Working Group. He currently chairs the IEEE MMTC Media Processing for Communications Interest Group (MPCIG) and is a board member of the IEEE MMTC Communications - Review. His research interests are in multi-view video coding, resilient multimedia transmission, and wireless systems design and applications.
IEEE COMSOC MMTC Communications - Frontiers

Fast and Low-Power Hardware Design for HEVC Fractional Motion Estimation

Vladimir Afonso\textsuperscript{1,2}, Ruhan Conceição\textsuperscript{3}, Murilo Perleberg\textsuperscript{3}, Marcelo Porto\textsuperscript{2}, Bruno Zatt\textsuperscript{2}, Luciano Agostini\textsuperscript{2}, Altamiro Susin\textsuperscript{1}

\textsuperscript{1}Graduate Program on Microelectronics, Federal University of Rio Grande do Sul, Porto Alegre, Brazil
\textsuperscript{2}Video Technology Research Group, Federal University of Pelotas, Pelotas, Brazil
\textsuperscript{3}vafonso, radconceicao, mrperleberg, porto, zatt, agostini}@inf.ufrgs.edu.br; altamiro.susin@ufrgs.br

1. Introduction

High-definition digital videos have become part of people’s daily life pushed forward by the increasing capability of streaming services over the internet and by the popularization of this media in smartphones and tablets. The huge amount of data required to represent high and ultra-high resolution videos leverages the need for more efficient video encoders. In this context, the High Efficiency Video Coding (HEVC) [1] is considered the state-of-the-art among available video coding standards [2].

In HEVC, Motion Estimation (ME) [3] – main step of the Inter-frames prediction – is the most computationally intensive coding tool, which leads the encoder to achieve significant gains regarding compression efficiency [4]. An HEVC-based video encoder divides each frame into smaller square-shaped blocks called Coding Three Units (CTUs), that, typically, has a size of 64x64 pixels. Besides, the CTUs are also divided in a quadtree structure of Coding Units (CUs). Each CU can be divides once more into smaller blocks called Prediction Units (PUs) with sizes that range from 8x4 or 4x8 up to 64x64 samples, totaling 24 different PU sizes [3]. On these PUs, the ME (and other prediction tools) is applied using a block matching algorithm (BMA) to find similar blocks within reference frames (previously processed frames), enabling data transmission/storage reduction. The HEVC ME may evaluate 24 different PU sizes whereas a whole encoding process for each possibility should be tested to choose the best one in terms of rate-distortion efficiency. These features make the ME responsible for a huge computational effort when compared to other coding tools. According to [4], the Inter-frames prediction is responsible for increasing by up to 1,567% the total encoding time depending on the video characteristics and encoder parameters.

The HEVC, as other current video-coding standards, employs the Fractional Motion Estimation (FME) to extend the ME coding efficiency since the motion observed from temporal-neighbor frames is not limited to integer positions. On the one hand, FME is responsible for increasing by 59.44% the total encoding time (using Random Access configuration). On the other hand, FME provides a significant BD-Rate decrease of 9.63%, on average [4]. BD-rate measures the percentage variation in the bit rate for the same image quality [5]. Thus, negative BD-rate values are desirable whereas positive values are undesirable.

Energy-constrained portable devices, such as smartphones, camcorders, and tablets, are also expected to process high-resolution videos in real time demanding hardware implementations of current video encoders modules as ME. Therefore, in this paper we propose a FME hardware design able to process UHD 2160p videos at 60 fps with low-energy consumption. The developed architecture is based on a HEVC ME complexity-reduction strategy, also proposed in this work. The number of PU sizes evaluated along with the ME process is limited to the four square-shaped sizes based on a statistical analysis of block-sizes distribution (see Section 3) and a comparison of different test cases when limiting the PU sizes (see Section 4).

2. HEVC FME Background

As previously mentioned, the HEVC ME is applied at the level of PUs that can assume 24 distinct sizes (square-shaped, symmetric rectangular-shaped and asymmetric rectangular-shaped) where the PU that minimizes the rate-distortion cost (evaluating compression rate and image quality) is selected for encoding [3]. First, the Integer Motion Estimation (IME) is applied employing a BMA algorithm and then the FME generates new blocks at fractional positions for comparison with the best IME result.

FME has two main processes, that consist of (a) an interpolation process to generate sub-pixel samples around the integer samples of the IME best block; and (b) a search and comparison process to compare the best IME result with the blocks formed in step (a). HEVC defines FIR filters with 7-taps or 8-taps for the interpolation of luminance samples with quarter-pixel precision before a search-and-comparison process [1]. The HM (HEVC Model) Reference Software [6] implements the search using the fractional samples in two steps. First, a search comparing the eight fractional blocks at half-pixel positions is performed around the block with the best IME result. Then, a search with eight quarter-pixel fractional blocks is performed around the best matching half-pixel block.

http://www.comsoc.org/~mmc/ 6/36 Vol.12, No.5, September 2017
IEEE COMSOC MMTC Communications - Frontiers

Figure 1 illustrates the integer-position samples (green squares) in addition to the fractional samples (non-green squares) for the FME interpolation process. Figure 1-b represents an 8x4 block. After the interpolation, 48 new fractional blocks are formed for new comparisons (independently of the PU size), as can also be seen in Figure 1. In Figure 1-a, the first sample of each new fractional block is represented. The gray squares represent the half-pixel position samples while the white squares represent the quarter-pixel position samples. Once the FME filters have 7-taps or 8-taps, samples around the block that is being interpolated are needed to calculate the fractional samples.

![Figure 1](image)

**Figure 1.** An 8x4 block representation: (a) First samples of the 48 fractional blocks, (b) 8x4 block (green squares).

### 3. HEVC ME Evaluations

Evaluations using the HM software [6] were performed to support an effective HEVC ME complexity-reduction strategy and an efficient FME hardware design. The experiments were done to identify the PU sizes most frequently selected during the encoding and their representativeness in the frames, i.e., the PU sizes that encode more pixels, on average. Next subsection presents some important considerations about the test conditions.

#### 3.1. Experimental Setup

The experimental setup is based on the Common Test Conditions (CTC) document [7] recommended by the Joint Collaborative Team on Video Coding (JCT-VC). This document defines 24 video sequences, divided into classes according to their resolutions and features. In general, each class has video sequences at a specific resolution ranging from 416x240-pixels resolution up to WQXGA resolution (2560x1600 pixels). Class F is the only class that has video sequences at different resolutions, but all those are screen content videos, presenting different characteristics in comparison with other classes. The CTC document also defines four Quantization Parameters (QP) [7] to be used in experiments (QP=22, 27, 32, and 37). All evaluations were performed through the HM 16.15 version [6] using a configuration that combines the low complexity profile (Main Profile) with the Random Access (RA) temporal configuration considering all video sequences and QP values. The number of frames to be encoded was stipulated to be equivalent to two seconds of each sequence, based on the frame rate of the video sequences.

#### 3.2. PU-size Occurrences and Representativeness Analyses

As previously mentioned, the HEVC ME computational-effort is pushed up mainly by the decision of which encoding methods and PU sizes must be used, given that 24 PU sizes must be evaluated during the encoding process. Therefore, knowing the behavior of the PU sizes distribution is primordial. Figure 2-a shows the average percentage of occurrences of the PU sizes in the Inter-frames prediction (disregarding skip mode). The values are presented separately for each square-shaped PU size (64x64, 32x32, 16x16, and 8x8) and the remaining PU sizes (non-square shaped). Notice that the “Non-square PUs” bar denotes the average of the 20 remaining PU sizes. The overlapped lines in the “Non-square PUs” represent the range that the other PU sizes can reach.

Averagely, the 8x8 PU size presents the higher occurrence followed by the 16x16. Note that the 32x32 and 64x64 PU sizes are seldom selected in comparison with other sizes. The percentage selection of PU sizes suggests that larger PU sizes, as the 64x64, have lower importance for coding process. However, one may notice that larger PUs cover larger areas in the image. Thus, even presenting a lower occurrence, they may be more relevant for the coding.

IEEE COMSOC MMTC Communications - Frontiers

process. This way, we evaluated the representativeness of the PU sizes. Here, the concept of representativeness depicts the percentage of pixels that were encoded employing each PU size, as illustrated in Figure 2-b. Note that larger PUs sizes, as 64x64 and 32x32, are more representative in video sequences regardless the lower absolute occurrence.

This experiment also shows that square-shaped PUs are both frequent and representative when compared to the non-square PUs. Note that 8x8 PU size is the most frequent and the 16x16 PU size is the second most frequent, whereas the square-shaped sizes (64x64, 32x32, 16x16, and 8x8) are the most representative sizes. Based on these observations, several scenarios that limit the PU sizes in the ME were investigated targeting a complexity reduction that could support the FME hardware design. These new evaluations are presented in the next section.

![Figure 2. PU-size analyses: a) Occurrences of the PU sizes; b) Representativeness of the PU sizes.](image)

4. HEVC ME Complexity-Reduction Strategy

Since square-shaped sizes are the most representative PUs, a simple strategy to reduce the computational effort is limiting the PU sizes in the ME process. However, this approach impacts in the image quality and compression results so that any decision must be carefully analyzed. This way, a new set of experiments was performed to verify the impact in terms of rate-distortion when the number of available PU sizes is limited during the encoding process (using the same experimental setup described in Section 3).

Several test cases (TCs) were analyzed considering combinations with the square-shaped PU sizes. Despite the use of only one square-shaped PU size being the most-attractive scenario targeting hardware design due to its higher complexity reduction, this decision could impact in severe coding-efficiency degradation, as evaluated in [4]. Therefore, we decided to avoid these situations and only consider the 11 possible combinations using two, three, or four sizes among the four square-shaped sizes to measure the losses regarding rate-distortion efficiency.

Table 1 shows the obtained results for each TC (the supported block sizes are presented above each TC) applied in the Inter-frames prediction (disregarding the skip mode). The results presented in Table 1 consider the BD-Rate metric, comparing the coding efficiency when limiting the number of PU sizes with a regular ME flow (24 PU sizes). By fixing the number of possible PU sizes to two (TC$_{1}$–TC$_{3}$), the average compression losses are from 5.94% (TC$_{1}$) up to 17.33% (TC$_{3}$). Considering the use of three PU sizes (TC$_{4}$–TC$_{6}$), the impacts in BD-Rate vary form 3.76% up to 6.85% (TC$_{6}$). Considering the four square-shaped PU sizes (TC$_{10}$), the compression losses are minimized with a BD-rate increase of 3.1%, on average, varying from 2.56% to 4.8% according to the sequence class. In this last scenario, the total encoding time is reduced by 59% [4]. Therefore, we have adopted the TC$_{10}$ as the ideal operation point when jointly considering rate-distortion performance and encoding time reduction.

<table>
<thead>
<tr>
<th>Sequence Classes</th>
<th>Test Cases (TC$<em>{1}$–TC$</em>{10}$) according to the supported square-shaped PU sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class A - 2560x1600</td>
<td>TC$_{1}$ (S16)</td>
</tr>
<tr>
<td>Class B - 1920x1080</td>
<td>6.48%</td>
</tr>
<tr>
<td>Class C - 832x480</td>
<td>7.81%</td>
</tr>
<tr>
<td>Class D - 416x240</td>
<td>5.74%</td>
</tr>
<tr>
<td>Class E - 1280x720</td>
<td>5.97%</td>
</tr>
<tr>
<td>Class F - Several</td>
<td>7.83%</td>
</tr>
<tr>
<td>Average</td>
<td>4.73%</td>
</tr>
</tbody>
</table>
5. HEVC FME Hardware Design

The FME hardware design was developed based on the HEVC Main Profile and performs the FME over the PU size that presented the best IME result. The proposed design processes the supported PU size by breaking them into 8x8 blocks. This approach allows an efficient and scalable hardware design since one module designed for an 8x8 PU size can be reused to process the bigger square-shaped PU sizes (16x16, 32x32, and 64x64), allowing a better tradeoff between energy consumption and throughput.

As previously mentioned, the FME can be divided into two main processes: (a) the interpolation, and (b) the search and comparison. As can be seen in Figure 3, the interpolation module is composed of FIR filters to interpolate the luminance samples, a buffer to store some generated samples that are reused to interpolate other samples, and a Clip operator. Since interpolation filters have an expressive cost regarding hardware usage, some optimizations were implemented. The fifteen equations employed to generate the values at fractional positions [1] have some similarities. We exploited the similarities among these fifteen equations by employing algebraic manipulations, common sub-expressions sharing, and replacing constant multiplications by shift-adds. Due to the similarities among equations, which share the same multiplications by constants in some cases, only two different hardware architectures are needed for the filters, called here of Type-A and Type-B. The filter architectures were designed using three pipeline stages.

To interpolate the samples, a scheme able to perform the calculation of an entire line or column of fractional samples per cycle was adopted. Therefore, 27 units of filters (18 Type-A and 9 Type-B) were used to allow the calculation of 27 fractional samples per cycle, considering each 8x8 block. In Figure 3, a multiplexer is used to select the samples that must be connected to the filter inputs provided from reference frames stored in memory (integer positions with eight bits each) or from the Fractional Buffer, since some calculated fractional samples are reused. The Search and Comparison module receives all fractional samples with eight bits. Hence, a clip operation is needed. This clip operation cannot be performed before the samples are stored in the buffer, since this fact would cause an accumulative error. As 27 fractional samples are calculated per cycle, 51 cycles are needed to process an 8x8 block (considering that the pipeline of the filters is filled). In this work, we adopted the comparison of all 48 new fractional blocks formed during the interpolation process to guarantee optimal result and avoid data dependencies during search and comparison step (i.e., half and quarter-pixel) blocks are processed in parallel.

The Search and Comparison employs the SAD (Sum of Absolute Differences) as the similarity criterion. This module is composed of SAD Trees, SAD Accumulators, and one SAD Comparator, as can be seen in Figure 3. The SAD Trees module allows the SAD calculation for all new fractional blocks formed from the interpolation, and it has 12 SAD tree units featuring four pipeline stages. Each SAD tree unit, detailed in Figure 4-a, can calculate the SAD of one fractional block. The SAD tree unit obtains the absolute differences between the fractional samples of the reference frames (Rn-Rr) and the integer samples of the current block (Cn-Cr), for each position, and sum them up. One SAD tree processes an entire line or column (depending on the fractional samples) of the 8x8 block per cycle, since the unit processes the SAD of eight samples in parallel. As the SAD Trees module has 12 SAD tree units, this module can calculate 12 lines of 12 fractional blocks simultaneously.

After the latency of the SAD Tree units, the SAD results are accumulated in the SAD Accumulators since each block has eight lines or columns. Twelve outputs of the SAD trees are connected to 48 accumulators as presented in Figure 4-b so that 12 accumulators are selected every eight clock cycles. This way, after eight cycles, the FME
module has the SAD of 12 fractional blocks. The accumulator outputs are 20 bit-wide since the SAD of the bigger square-shaped PUs, as the 64x64 PU, can be calculated from 8x8 blocks without overflow. After the calculation of
the SAD values for all 48 fractional blocks, these values, and their respective motion vectors, are sent to the SAD Comparator, as can be seen in Figure 3. The SAD comparator uses 48 simplified comparators, as presented in Figure 4-d, distributed in six pipeline stages. This module is responsible for comparing all blocks simultaneously, two by
two. Then, the SAD Comparator delivers the SAD value and the motion vector of the block that presents the best
result among all fractional blocks and the best IME result.

The developed FME architecture was described in VHDL, and the synthesis results were generated targeting an
ASIC technology using the Cadence RTL Compiler [8] tool. The ASIC hardware results, obtained with the 45nm
Nangate standard-cells technology [9], consider a supply voltage of 0.95V. The developed architecture uses 148,410
gates to implement the complete FME architecture. This design reaches real-time processing of HD 1080p videos at
30 frames per second with low-power dissipation of 4.96mW when running at 49.6MHz. For real-time processing of
UHD 2160p at 60 frames per second, the architecture dissipates 15.85mW when operating at 396.8 MHz. When
compared with the related work [10], that uses the same operation frequency, this work reduces the hardware-
resource usage and the power dissipation about 40.4% and 38.8% to process 1080p videos at 30fps. Considering
2160p@60fps, this work reduces in three times the power dissipation obtained by [10]. The related work [11]
reaches real-time processing for UHD resolution (2160p@60fps) with a lower frequency than that reached in this
work due to its higher parallelism. However, the architecture presented in [11] uses an area eight times larger and a
power dissipation three times higher than reached by this work considering 2160p@60fps.

6. Conclusions

This work presented a fast and low-power hardware design for the HEVC FME able to process high resolution
videos in real-time focusing on battery-powered devices. A detailed analysis to determine an efficient IME/FME
complexity-reduction strategy was performed based on evaluations with the HEVC Reference Software. From this
analysis, the FME architecture was designed to support only the four square-shaped PU sizes, avoiding the
evaluation of other 20 PU sizes. This approach reduces in 59% the total encoding time with only 3.1% increase on
BD-Rate. The synthesis results for ASIC 45nm technology show that the developed architecture can process HD
1080p videos at 30 frames per second with very low-power dissipation of 4.96mW when running at 49.6MHz. For
real-time processing of UHD 2160p at 60 frames per second, this architecture dissipates 15.85mW operating at
396.8 MHz. The reached power dissipation and used silicon area are lower than those reached by related works.

Acknowledgements

We have a special acknowledgement to CNPq, CAPES and FAPERGS Brazilian agencies to support this work.

References


http://www.comsoc.org/~mmc/ 10/36 Vol.12, No.5, September 2017
IEEE COMSOC MMTC Communications - Frontiers


Vladimir Afonso received the B.S. degree in industrial automation from the Sul-rio-grandense Federal Institute (IFSul), Brazil, in 2008 and the M.S. degree in Computer Science from the Federal University of Pelotas (UFPEl), in 2013. He is currently pursuing the Ph.D. degree in Microelectronics at the Federal University of Rio Grande do Sul, Brazil. He is a Professor since 2009 at the IFSul and a researcher with the Video Technology Research Group (ViTech) at the UFPEl. His research interests include 2D/3D video coding algorithms, FPGA-based and VLSI design.

Ruhan Conceição is a Ph.D. candidate in Computer Science at the Federal University of Pelotas (UFPEl), Brazil. He received his B.S. degree in Computer Engineering and M.S. degree in Computer Science from UFPEl in 2015 and 2016, respectively. He is a member of the Video Technology Research Group (ViTech) at the UFPEl. His current research interests include hardware and software solutions development focusing on HEVC standard and its 3D extension, emerging video coding standards, and light fields.

Murilo Perleberg is currently pursuing the B.S. degree in Computer Engineering at the Federal University of Pelotas (UFPEl), Brazil. He is a researcher with the Video Technology Research Group (ViTech) at the UFPEl. His research interests include hardware architectures for video coding and motion estimation algorithms.

Marcelo Porto received the B.S. degree in Computer Science from the Federal University of Pelotas (UFPEl), RS, Brazil, in 2006 and the M.S. and Ph.D. degrees, also in Computer Science, from the Federal University of Rio Grande do Sul (UFRGS), Brazil, in 2008 and 2012, respectively. He is a Professor since 2009 at the UFPEl, where he is member of the Video Technology Research Group (ViTech). His research interests include video coding, motion-estimation algorithms, FPGA-based design, and VLSI design for video coding.

Bruno Zatt received his B.S. and M.S. in Computer Engineering from the Federal University of Rio Grande do Sul (UFRGS), Brazil, in 2006 and 2008, respectively. He has received his Ph.D. degree in Microelectronics at the same university in 2012 with "summa cum laude" distinction. Currently, he is a Professor at the Federal University of Pelotas (UFPEl), Brazil, and a member of the Video Technology Research Group (ViTech).

Luciano Agostini is a Professor since 2002 at Federal University of Pelotas (UFPEl), Brazil, where he leads the Video Technology Research Group (ViTech). He received the M.S. and Ph.D. degrees in Computer Science from Federal University of Rio Grande do Sul (UFRGS), Brazil, in 2002 and 2007, respectively, and the B.S. degree in Informatics from UFPEl in 1998. His research interests include 2D and 3D video coding, algorithmic optimization, arithmetic circuits, FPGA-based design and microelectronics.

Altamiro Susin received the B.S. degree in Electrical Engineering and the M.S. degree in Computer Science from the Federal University of Rio Grande do Sul (UFRGS), Brazil, in 1972 and 1977, respectively, and the Ph.D. degree in Microelectronics from the Polytechnique National Institute of Grenoble, Grenoble, France, in 1981. Since 1975, he is Full Professor with the Electrical Engineering Department at the Federal University of Rio Grande do Sul (UFRGS). His main research interests include integrated circuit architecture, system-on-chip design, and signal processing.
Recent Advances in Complexity Reduction Methods for High Efficiency Video Coding

Pedro Assunção\textsuperscript{1,2}, Luís Cruz\textsuperscript{3,1}, Carl James Debono\textsuperscript{4}

\textsuperscript{1}Instituto de Telecomunicações, Portugal
\textsuperscript{2}Instituto Politécnico de Leiria, Leiria, Portugal
\textsuperscript{3}Dep. of Electrical and Computer Engineering, University of Coimbra, Coimbra, Portugal
\textsuperscript{4}Dep. of Communications and Computer Engineering, University of Malta, Msida, Malta

\texttt{paassunc@ieee.org; lcruz@deec.uc.pt; c.debono@ieee.org}

1. Introduction

The High Efficiency Video Coding (HEVC) standard supports applications and services with demanding requirements in quality and compression ratio, such as Ultra High Definition Television (e.g., 4K, 8K), multi-view video, high dynamic range and panoramic video. The higher coding efficiency achieved by HEVC comes with a significantly increased computational complexity of encoding in comparison with previous standards [1]. The nested coding data structures and the huge multidimensional search spaces used to find optimal coding decisions and motion information, are the main causes for the high coding efficiency and also complexity. Reducing the algorithmic complexity is a crucial factor to achieve power-efficient video encoding in multimedia devices, especially in those with constrained computational resources and limited battery life.

In the recent past, various techniques have been proposed to reduce the computational complexity of HEVC encoding using different approaches. To achieve fast coding mode decisions, such approaches can be classified as (i) Statistical methods based on spatio-temporal correlations of coding parameters related with algorithmic tools and processing; and (ii) Methods based on Machine Learning approaches, mostly used to achieve fast coding decisions based different types of classifier models. Recently, fast motion estimation techniques have also been developed to speed up the process of finding the best motion vectors for optimum prediction [2]. Statistical methods based on spatio-temporal correlations exploit spatio-temporal similarities in video coding data attempting to reuse coding decisions from spatial neighbors in the current frame and also from co-located regions in temporally adjacent frames. In general, these techniques try to estimate the optimum block type, prediction mode and motion vectors of coding units (CU) and prediction units (PU), based on either static or dynamic decision thresholds, computed from the statistics of coding data. The other type of methods, based on machine learning, use diverse features extracted from the coding data and also from the video data to train a specific classifier model. Different types of classifiers have been used to speed up the encoding process required to produce HEVC compliant streams and to achieve power-efficient video encoding by means of such algorithmic complexity reduction.

2. Statistical methods

Statistical methods are commonly based on correlations between different types of spatial and temporal parameters in adjacent CUs and/or frames. In general, such correlations allow faster coding decisions due to the prior information brought by correlated parameters, which contribute to reduce the search spaces for optimization. The computational complexity of high efficiency video encoders is also significantly dependent on the encoding configurations, as found in a previous study with results presented in Figure 1, which shows the normalized encoding complexity of 16 different configurations with respect to CFG1 (configuration 1) [3]. It is worthwhile to point out that all video sequences exhibit a similar monotonic increase of the encoding complexity from CFG 1 to CFG 15, but, from CFG 10 to CFG 15 the slope is very small and the normalized computational complexity is roughly constant. Since these configurations correspond to different functional modes of motion estimation/compensation, these results show that the choice of more accurate motion estimation (ME) modes is responsible for most of the computational complexity increases. The comprehensive study and results about encoding complexity is presented in [3].

A previous study on statistical dependence regarding PU splitting modes and its relevance in the design of computational reduction methods can be found in [4]. A significant correlation between the partition mode of the current PU and further splitting decisions into smaller CUs was found, based on the high probability that a CU predicted with large-size PUs (e.g., a single 2N×2N PU) does not need being split into smaller CUs. Statistics that support this relationship are presented in Figure 2(a) for 32×32 CUs. The chart shows that 78.5% of the CUs encoded as a 2N×2N PU were not split into sub-CUs. Moreover, an average of 86.7% of CUs encoded with the remaining PU splitting modes were divided into sub-CUs.

http://www.comsoc.org/~mmc/
In Figure 2(b), the Neigh_Depth parameter is related with the coding tree depths used in neighboring CTUs already encoded and calculated as follows. First, for each neighboring CTU, the average depth of all CUs is computed. Then, the average of averages is computed among all neighbors available for the current CTU. The top, left, top-left and top-right CTUs in the current frame, as well as the co-located CTUs in the first frames of both reference lists (list 0 and list 1) are considered as neighbors. Figure 2(b) shows the distribution of Neigh_Depth for 32×32 CUs. The curves show that there is a clear relationship between the distribution of Neigh_Depth and the CU splitting decision. CUs which are not split into smaller CUs have Neigh_Depth values that cluster towards low magnitudes, while for CUs which are split into smaller CUs the opposite occurs. Since the two distributions do not fully overlap it is possible to determine an optimal decision threshold to determine whether a split or no-split decision should be taken.

Figure 2. Frequency of occurrence of CUs split and not split into smaller CUs: (a) PU splitting mode chosen for the current CU and (b) average of CU depths in neighboring CTUs [4].

2.1 HEVC computational complexity reduction based on statistical methods

Previous research studies based on statistical methods have been following different ideas with the common goal of speeding-up the encoding algorithms and to reduce the computational power. The work reported in [5,6], propose methods for fast CU size decision and adaptive inter-mode decision. In [5] the authors propose to find a CU depth range for each CU and skip those depth levels that are rarely used in spatially adjacent CUs and also in temporally co-located ones in the previous frame. The complexity reduction also benefits from early termination methods based on motion homogeneity. The use of fixed thresholds does not allow adaptability to content characteristics. In [6] the same authors further analyze the distribution of prediction modes at each depth level and the correlations between adjacent CUs to devise adaptive inter-mode decision strategies. This correlation-based approach can achieve more than 50% of complexity reduction in HEVC. Another method for early CU skip detection and fast CU split decision, based on spatio-temporal encoding parameters, is described in [7]. Sample-adaptive-offset (SAO) parameters are used to estimate the texture complexity of CUs and motion vectors are used to estimate the temporal complexity. Since the spatio-temporal parameters used by the algorithm are byproducts of the encoding process, no additional computation is required to reach complexity reductions between 42% to 49%. The texture of each CU was used in [8].
IEEE COMSOC MMTC Communications - Frontiers

to model the CU partitioning through the use of the Sobel operator to optimize the partition decision based on edge features. The inter-mode decision algorithm devised by the authors presents an average speed up gain of 45% in comparison with the reference H265.

Since the first spatio-temporal methods for HEVC appeared in the literature, further improvements were developed and more recently more than 50% of complexity reduction can be achieved. For instance, in [9] the authors propose a method based on a Laplacian transparent composite model, for discrete cosine transform (DCT) coefficients of an original image and for residual signals of an inter-coded frame. In such model, the flat tail of the DCT coefficient distribution is modeled as uniform distribution while the main body is a truncated Laplacian distribution. For the Low Delay configuration, the method is reported to achieve 60% of complexity reduction. A different algorithm was presented in [10] for UHD video, which extends the CU size and classifies it according to the information collected from previous frames. A probability model is used in the Split/No Split decision process of CUs. The results report an average complexity reduction of 62%. An optimal selection model of CU depth is devised in [11] to estimate the range of candidate CU depth by exploiting the temporal correlation of CU depth among the current CU and temporally co-located CUs. This strategy is capable of achieving 56% of complexity reduction. More recently the depth correlation between the current and the co-located CU was investigated to avoid irrelevant CU modes and truncate some PU predictions. A model for the relationship between the RD costs after the Merge/SKIP prediction was investigated for adaptive termination of CU search. In addition, the relationship of the RD costs after the Merge/SKIP prediction and 2N×2N mode is also explored to achieve an average encoding time reduction of 52%.

In general, the methods based on statistics have been consistently improving in recent years, mainly due to better modeling of correlations among the various signal characteristics and coding parameters that can be computed with low complexity. While fixed models present limited accuracy due to the non-stationary characteristics of video signals, adaptive statistical modeling techniques still have room for improvement by considering increased number of parameters and look-ahead strategies. Overall, further accuracy in the fast coding decision processes is also dependent on adaptive models for better estimation of the statistical properties and correlations of both global and localized features in the spatio-temporal domain.

3. Machine learning methods

The difficulty of making sense of statistical information regarding HEVC encoding mode choice and inferring relationships between video data and HEVC encoder past status and the optimal encoding mode choices, as required to define encoding acceleration heuristics, can be sidestepped by the use of automated data analysis tools and decision methods grounded in machine learning.

In the last few years several works have been published that rely on data mining and machine learning techniques to design algorithms enabling fast mode decision for application in the different stages of HEVC encoding like CU partitioning depth decision and PU and TU partition choices. Most of the solutions published so far take the form of a classifier where each class is associated with one of the possible mode choices, e.g. in the case of CU partitioning, a binary classifier is designed which is then incorporated into the HEVC encoder processing flow to decide if a given CU should be further partitioned or not. Different types of classifiers can be used and popular choices are decision trees (DT), random forests (RF), artificial neural networks (ANN), support vector machines (SVM) and more recently convolutional neural networks (CNN). In all cases the classifiers require input information which represents features of the data that are relevant to the decision problem at hand. The majority of the works use features which in essence represent previous states of the encoder (like the partitioning decision of a collocated CU in a previously encoded video frame) and features obtained by processing intermediate encoding results, for instance energies of prediction residues after motion estimation. Only recently have features based on the processing of pixel data to be encoded started being used, in most cases through application of CNNs. The methods proposed can be further divided into off-line trained algorithms, where a dataset of training data is used to build a classifier which is then used without modification during encoding, and on-line methods which have the ability of being continuously adaptable to video data characteristics through periodic retraining or parameter adjustment.

The following paragraphs present a chronological abridged review of the most recent and relevant HEVC encoding complexity reduction methods based on these ideas.

In [12] Shen and Yu propose a fast CU decision method where CU partitioning is modeled as a binary classification problem, which is solved using an SVM. The authors use a weighted SVM to reduce the rate-distortion loss that occurs during misclassifications when a CU that should have been partitioned is prematurely defined as not needing

http://www.comsoc.org/~mmc/ 14/36  Vol.12, No.5, September 2017
further partitioning (the other type of misclassification increases computation but does not degrade rate-distortion performance). The authors study the usefulness of features based on prediction-error values like the sum of absolute values of transformed prediction residuals (SATD), information about partitioning depth of temporally and spatially adjacent CUs, pixel-domain luminance gradient computed on the CU to be encoded, a motion consistency feature and several others related to RD costs. The results listed show that the proposed algorithm achieves about 45% complexity reduction on average with 1.4% BD-BR increase under the Random Access (RA) configuration.

Corrêa et al. took a different approach in [4] choosing decision tree based classifiers to decide if a CU should be further split, to choose one from the different PU partition possibilities and also to decide the transform unit (TU) partitioning to be used in coding the residues. The decision trees designed are tied to specific block sizes and constructed using the C4.5 algorithm as implemented in the Weka framework. This suite of tools take as input training data expressed in Attribute-Relation File Format (ARFF) (as illustrated in Figure 3(a)) and can be used to invoke different tools for data analysis to mine important attributes and classifier design tools. Previous to the classifier design an in-depth study of several tens of possible features was performed and only the features with a high enough information gain were retained for the training of the decision trees. Among the features found to be useful one counts neighboring CUs partitioning depth, ratios of RD costs obtained at the previous level when using 2N×2N and MSM modes, as well as several binary variables indicating whether skip or merge modes were selected in neighboring CUs. After training the decision trees, similar to the one presented in Figure 3(b), were embedded into the HEVC reference encoder and the performance of the proposed method evaluated, with the CU, PU and TU partitioning prediction algorithms taken independently or operating jointly. For the case of the CU partitioning early termination method, operating alone, an average complexity reduction of 36.7% was achieved with an increase in bitrate of 0.284% as measured by the Bjontegaard Delta (BD-Rate). The PU early termination method, also taken alone, resulted in a 49.6% computational complexity reduction with a 0.562% BD-Rate penalty. The TU partitioning decision acceleration method had less impressive results, and when used alone it allowed computation savings of about 7% with a BD-Rate degradation of 0.055%. Joint use of the three partitioning decision methods (CU, PU and TU) resulted in a computational saving of about 65% and 1.35% degradation of BD-Rate.

![Figure 3. Extract of ARFF file with attributes and class information (b) Example decision tree after training [4].](image)

Unlike the two previous works that involved off-line training without the possibility of retraining during encoding, Momcilovic et al. describe in [13] a fast CU partitioning method based on ANN which uses on-line learning to adapt the neural network to the changing video characteristics. The method operates by having different computing threads, one devoted to (modified) HEVC encoding and some others dedicated to the retraining of the ANN classifier whenever necessary. The entire process is controlled by a finite-state machine, which decides when the ANN needs to be retrained. Figure 4 shows an example of the operations occurring in parallel in the various threads, where \( i \) stands for (thread) initiation, \( r \) for training, \( v \) for validation and \( a \) for application (of the model trained).

Concerning the ANN, the main input features are depths and RD values of neighboring CUs as well as a MergeFlag binary feature. The results reported list a complexity reduction of 47.5% for a BD-Rate increase of 1.17%. However, these results do not consider the extra computation involved in the retraining of the ANN by the secondary threads, which in all likelihood contribute a non-negligible amount of complexity to the overall encoding procedure.

http://www.comsoc.org/~mmc/ 15/36 Vol.12, No.5, September 2017
The authors of [14] describe a CU partitioning early termination algorithm to reduce HEVC encoding complexity based on a three-level binary hierarchical classifier where each stage binary classifier is a SVM. The features used include intermediate coding features like status of the SKIP flag, motion vector information, and neighboring CUs depth. Two different modes are discussed based on on-line and off-line training as illustrated in Figure 5, but after showing that through on-line learning it is possible to build classifiers that adapt to the scene characteristics, the authors ended up opting for off-line training with optimized classifier parameterization.

This method proposed reduces the computational complexity by an average of 51.45% with an accompanying 1.98% coding efficiency (BD-Rate) loss.

In [15] Liu et al. address the complexity reduction of HEVC intra encoding. While the method is not general enough as it is not directly applicable to inter-frame encoding, it is nonetheless noteworthy as it uses a CNN-based solution and it is especially adapted to hardware parallel implementations. The entire intra encoder with CNN processor is shown in Figure 6.

The objective of the algorithm proposed is the reduction of computation via CU and PU fast mode decision. The algorithm tries to decide which 8x8 blocks should undergo splitting and entire RD optimization step of HEVC intra-coding, employing a CNN-based classifier to do so. As shown in Figure 6(b) the classifier is applied to 8x8 pixel blocks applying first a convolution layer with 6 filter kernels of size 3x3. This convolutional layer is followed by a subsampling layer and two hidden fully connected perceptron layers with two final outputs which indicated the probability of splitting or not splitting the block. Although the method is aimed at hardware implementations the authors performed some empirical evaluations comparing their algorithm with the HEVC reference implementation version 12, achieving 61.7% computational savings with a 2.67% BD-Rate increase, for intra-only encoding.
Also based on SVMs is the fast CU and PU partitioning algorithm introduced by Zhu et al. in [16]. This proposal uses several SVM classifiers with input features such as pixel-domain sum of absolute differences between the current block and co-located block in the previous frame, size of current block, some RD costs and binary features like Merge mode and Skip mode flags, and other information like the quantization parameter (QP) in use. The system uses three different classifiers according to a “multiple reviewers” principle, where one of the classifiers acts as a higher ranked tie-breaker which is called to solve conflicting decisions issued by the other classifiers as shown in Figure 7.

The algorithm uses both off-line training and an on-line learning mode, where one of the lower ranked classifiers is trained off-line and the other low-ranked and the high-ranked classifiers are adjusted with recourse to on-line training, using a predefined schedule mechanism. The authors report overall computational complexity savings in the order of 68%, but with a very significant BD-Rate penalty of about 4.2%, for the low-delay P coding configuration.

More recently Xu et al. propose in [17] a CTU partition map prediction method based on CNN and Long Short Term Memory (LSTM) networks applied to intra and inter frame encoding. In this work the authors proposed to

http://www.comsoc.org/~mmc/
IEEE COMSOC MMTC Communications - Frontiers

predict a partition map through direct computation from the pixels with recourse to deep classifiers having in excess of nine layers and about 1.2 million coefficients for the CNN networks and 750 thousand for the LSTM case. The networks are trained, validated and tested using about 110 million CTUs with corresponding partition maps for the intra case and about 300 million CTUs for the inter case. Although the networks are exceedingly complex to train, the authors maintain that their use at encode time does not add significantly to the total complexity. Using these structures for the intra modes, complexity reductions in the range of 57% to 67% are reported for several QPs with a corresponding average BD-Rate increase of about 2.25%. In inter coding cases the complexity savings documented go from 44% to 63% with a BD-Rate increase of about 1.5%

Although the results reported in the literature show that machine learning methods can reduce the encoding complexity by significant amounts with small degradation in encoding performance, these methods are not without limitations. Probably the biggest drawback of these methods is the need to spend a considerable amount of time and ingenuity choosing the best performing features, both for off-line and on-line trained algorithms. In the case of off-line methods, collecting the training data and training the classifiers can also require significant effort, a problem compounded by the possibility of occurrence of overfitting effects typical of badly designed classifiers. Furthermore, some classifier structures are not easily implementable in hardware, due to their complexity or irregular flow of processing operations. With respect to this type of issue, decision tree based fast encoding algorithms are probably the friendlier towards hardware implementations as they are structurally simple. The recently introduced CNN-based solutions solve some of the problems identified above, for instance relieving the designer of the trouble of designing the features, but at a cost of training procedures several orders of magnitude more complex than competing methods and, more importantly, a much more demanding computational structure that has to include multiple convolutional layers where many filtering operations are applied to the input video data, not to mention the following neural network layers necessary to obtain the final mode estimates.

5. Conclusion

This paper presented a short summary of recent research works addressing the problem of computational complexity reduction in HEVC. These are methods aimed at reducing the algorithmic complexity, which implicitly reduce the computational power requirements and energy consumption of high efficiency video encoders. Since the ever increasing coding efficiency has been mostly accomplished at the cost of a very high increase in computational complexity, its reduction and control is expected to remain an open field for research and innovation in the coming years.

Acknowledgement

This work is part of the project FAVIETOL supported by Instituto de Telecomunicações, FCT UID/EEA/50008/2013, Portugal.

References


http://www.comsoc.org/~mmc/ 18/36 Vol.12, No.5, September 2017
IEEE COMSOC MMTC Communications - Frontiers


Pedro Assuncao (M’99, SM’14) received the Licenciado and M.Sc. degrees in Electrical Engineering from the University of Coimbra, in 1988 and 1993, respectively, and the Ph.D. in Electronic Systems Engineering from the University of Essex, in 1998. He is currently professor of Electrical Engineering and Multimedia Communication Systems at the Polytechnic Institute of Leiria and researcher at the Institute for Telecommunications, Portugal. He is author/co-author of more than one hundred scientific/technical papers in conferences and journals, one book, seven book chapters and four US patents. His current research interests include high efficiency and 3D video and light field coding, multiple description and robust coding, error concealment and quality evaluation. He was the Chair of EU COST Action IC1105 (3D-ConTourNet) and senior member of the IEEE.

Luis Cruz (M’12) received the Licenciado and M.Sc. degrees in Electrical Engineering from the University of Coimbra, Portugal, in 1989 and 1993 respectively. He also holds an MSc degree in Mathematics and a Ph.D. degree in Electrical Computer and Systems Engineering from Rensselaer Polytechnic Institute (RPI), Troy, NY, US granted in 1997 and 2000 respectively. He has been with the Department of Electrical and Computer Engineering of the University of Coimbra in Portugal since 1990 first as a Teaching Assistant and as an Assistant Professor since 2000. He is a researcher of the Institute for Telecommunications of Coimbra where he works on video processing and coding, medical image processing and wireless communications. He is a member of the EURASIP, SPIE and IEEE technical societies.

Carl James Debono (S’97, M’01, SM’07) received his B.Eng. (Hons.) degree in Electrical Engineering from the University of Malta, Malta, in 1997 and the Ph.D. degree in Electronics and Computer Engineering from the University of Pavia, Italy, in 2000. Between 1997 and 2001 he was employed as a Research Engineer in the area of Integrated Circuit Design with the Department of Microelectronics at the University of Malta. In 2000 he was also engaged as a Research Associate with Texas A&M University, Texas. In 2001 he was appointed Lecturer with the Department of Communications and Computer Engineering at the University of Malta and is now an Associate Professor. He also served as Deputy Dean of the Faculty of ICT between November 2009 and September 2015. He is currently the Head of the Department of Communications and Computer Engineering at the University of Malta. Prof. Debono is a senior member of the IEEE. He was a member of the management committee of the COST Action IC1105 - 3D Content Creation, Coding and Transmission over Future Media Networks (3D-ConTourNet) where he chaired the 3D Media Coding Working Group. He is a board member of the IEEE Multimedia Communications Technical Committee Communications - Review. His research interests are in multi-view video coding, resilient multimedia transmission, and wireless systems design and applications.

http://www.comsoc.org/~mmc/ 19/36  Vol.12, No.5, September 2017
IEEE COMSOC MMTC Communications - Frontiers

Prediction of the Quad-tree Partitioning for Real-Time Low Energy HEVC Encoders
Alexandre Mercat\textsuperscript{1}, Florian Arrebi\textsuperscript{1}, Wassim Hamidouche\textsuperscript{1}, Maxime Pelcat\textsuperscript{1}, Daniel Menard\textsuperscript{1}
\textsuperscript{1}ULB, INSA Rennes, IETR CNRS UMR 6164, Rennes, France
\{amecart, farrebi, mpeclat, whamidou, dmenard\}@insa-rennes.fr

1. Introduction

With the progress of microelectronics, many embedded applications now encode and stream live video contents. The High Efficiency Video Coding (HEVC) \cite{1} \cite{2} \cite{3} standard represents the state-of-the-art in video coding. When compared with the previous MPEG AVC, HEVC Main profile reduces the bit rate by 60\% on average for a similar objective video quality \cite{4} \cite{5} at the expense of a significant computational complexity increase. On the other hand, the main limitation of recent encoded systems, particularly in terms of computational performance, comes from the bounded energy density of batteries. Therefore, energy consumption represents a serious challenge for embedded HEVC real-time encoders. For both hardware and software codecs, a solution to reduce energy consumption is to decrease the computational complexity while controlling compression quality losses.

An HEVC encoder is based on a classical hybrid video encoder structure that combines Inter and Intra predictions. While encoding in HEVC, each frame is split into equally-sized blocks named Coding Tree Units (CTUs) as represented in Figure 1. Each CTU is then divided into Coding Units (CUs), themselves nodes in a quad-tree. In HEVC, the size of CUs is equal to \(2^N \times 2^N\) with \(N \in \{3,2,1,0,1,2,3,4\}\). Then, CUs may be split into Prediction Units (PUs) of smaller size to be predicted. In intra prediction mode, PUs are square and can take the size of \(2^N \times 2^N\) (or \(N \times N\) only when \(N = 4\)), which can be associated to a quad-tree depth range \(d \in \{0,1,2,3,4\}\) as illustrated in Figure 1. In the rest of this paper, PU of size \(4 \times 4\) will be associated to CU \(4 \times 4\). The HEVC Intra-frame prediction is complex and supports in total 35 modes performed at the level of PU including planar (surface fitting) mode, DC (flat) mode and 33 angular modes \cite{2}. The Quantization Parameter (QP) impacts the RDO process to balance the encoding between distortion and quality. To achieve the best Rate-Distortion (RD) performance, the encoder performs an exhaustive search process, named Rate-Distortion Optimization (RDO), testing every possible combination of partitioning structures with the 35 Intra prediction modes. This exhaustive search tests 341 different decompositions which considerably increases the coding complexity.

As shown in \cite{6}, in real-time software HEVC Intra encoder, two specific parts of the encoding algorithm have the bigger opportunities of energy reduction; the Intra prediction level offers at best 30\% of energy reduction whereas the CTU quad-tree partitioning level has a potential of energy reduction up to 78\%. Authors in \cite{7} present an Intra CU size classifier based on data-mining with an offline classifier training. The classifier is a three-node decision tree using mean and variance of CUs and sub-CUs as features. This algorithm reduces the coding time of 52\% at the expense of bit rate increase of 2.67\%. Works in \cite{8} \cite{9} \cite{10} \cite{11} \cite{12} use CTU texture complexities to predict the quad-tree partitioning. Min et al. \cite{8} propose to decide if a CU has to be split, non-split or if it is undetermined using the global and local edge complexities according to four directions (horizontal, vertical, 45\(^{\circ}\) and 135\(^{\circ}\) diagonals) of CU and sub-CUs. This method provides a computational complexity reduction around 52\% for a Bjontegaard Delta Bit Rate (BD-BR) increase around 0.8\%. Feng et al. \cite{9} uses information entropy of CUs and sub-CUs saliency maps to predict the CUs size. The method reduces the complexity of 37.9\% for a BD-BR increase of 0.62\%.

We propose in this paper a fair comparison between two proposed methods to predict the CTU partitioning in order to budget the energy consumption of a real-time HEVC encoder. The two proposed methods limit the recursive RDO process, which can drastically reduce the energy consumption of HEVC Intra encoder. The first proposed method is a probabilistic CTU partitioning prediction technique which exploits the correlation between the CTU partitioning and the variance of the CU luminance samples. The second methods proposed in this paper aims to predict the CTU partitioning using Machine Learning. Compared to previous solutions, our works focuses on energy consumption for real-time encoding.

The rest of this paper is organized as follows. Section 2 presents the Probabilistic CTU partitioning prediction technique. Section 3 details the second proposed technique based on Machine Learning CTU partitioning prediction. The performance comparison of the two proposed energy reduction techniques are presented in Section 4. Finally, Section 5 concludes the paper.
2. Probabilistic CTU Partitioning Prediction

It has already been shown that the CTU partitioning during the RDO process is highly correlated to the QP value and the texture complexity which can be statistically represented by the variance of blocks [11] [8] [10] in Intra coding.

To study how to predict the quad-tree partitioning from the variances of CU luminance samples, two populations of CUs at a current depth \( d \) are defined: Merge (M) and Non Merge (NM). The CUs belong to the Non Merge population when the RDO process chooses to encode the CUs at the current depth \( d \), while the CUs belong to the Merge population when the RDO process chooses to encode the CUs at a new depth \( d' \) with \( d' \leq d \). With a bottom-up approach (i.e. \( d \) from 4 to 1), all CUs of the quad-tree decomposition of all CTUs can be classified in one of these two populations.

2.1 Variance-Based Decision for Quad-Tree Partitioning

To classify each CU in one of the two populations, the Cumulative Distribution Function (CDF) of the Non Merge is used. In our case, the CDF defines the probability of the variance population of a given CU size being less or equal to a given value. Figure 2 shows the CDF of CU variances depending on CU size for the sixth frame of the BasketballDrive video sequence. The CDF curves show that the probability for a CU size to be selected during the RDO process decreases when the variance of the CU increases. In other terms, it is rare for a CU selected during the RDO process to have a variance greater than a certain threshold. From this observation, a variance threshold \( v_{\text{th}}(\Delta, d) \) for each depth \( d \) can be extracted from the CDF for a specific probability \( \Delta \). The main limitation of this criteria is that the thresholds being highly dependent of the video content, they cannot be computed offline and thus have to be determined during encoding, using a Learning Frame.

![Figure 1. Quad-tree structure of CTU divided into CU and PU (dimensions in luminance pixels).](image1)

![Figure 2. CDFs of the Merged population depending on CU size for the sixth frame of the sequence BasketballDrive.](image2)

The thresholds can be approximated directly from Non Merge population moments: the mean \( \mu_v(d) \) and the standard deviation \( \sigma_v(d) \) using a linear relation:

\[
v_{\text{th}}(\Delta, d) = a(\Delta, d) \cdot \mu_v(d) + b(\Delta, d) \cdot \sigma_v(d) + c(\Delta, d)
\]

where \( a(\Delta, d), b(\Delta, d) \) and \( c(\Delta, d) \) are coefficients allowing to model the threshold according to the probability \( \Delta \) and for each depth \( d \). The coefficients are computed offline for each \( \Delta \) and \( d \) values using a linear regression on all frames of BasketballDrive, BQTerrace, Cactus, ParkScene, PeopleOnStreet and Traffic for 4 values of QP: 22, 27, 32 and 37.

2.2 Prediction Algorithm for CTU Partitioning

A description of the CTU partitioning is needed to explicitly depict the prediction of the quad-tree and then force the encoder to just encode this specific decomposition. A matrix 8 by 8, called CTU Depth Map (CDM), is chosen as representation of the CTU partitioning. Each element of the matrix represents the depth \( d \) of an 8 by 8 square sample of the CTU. A depth of 4 in the CDM corresponds to 4 CU 4 \( \times \) 4 in the CTU decomposition.

First of all, the full CDM is initialized with the depth value 4 and all the variances of the CU 4 \( \times \) 4 are computed. Then, the algorithm explores the CTU decomposition with a bottom-up approach: from \( d = 4 \) to \( d = 1 \). For the current depth \( d \), the algorithm browses the CDM and tests if the 4 neighbours blocks in the Z-scan order have the same depth \( d \). If the previous condition is true, then the algorithm tests whether the blocks have to be merged or not using the variance criteria previously detailed in Section 2.1. If the 4 blocks variances are lower than the given
IEEE COMSOC MMTC Communications - Frontiers

threshold \( v_{th}(\Delta, d) \) then the blocks are merged and the corresponding elements is the CDM are set at \( d - 1 \) and the variance of the merged block is calculated.

2.3 Variance-Aware Algorithm Scheme

Figure 3 presents a high-level diagram of our overall variance-aware CTU partitioning prediction technique. Firstly, the video sequence is split into GOFs.

The first frame of a Group of Frame (GOF), called learning frame is encoded with a full RDO process. From this encoding, the variances are extracted according to the depth \( d \in \{0,1,2,3,4\} \) of the CUs selected during the full RDO process. Then, the two following statistical moments according the depth \( d \) are computed: the means \( m_d(d) \) and the standard deviations \( \sigma_d(d) \) of the variance populations. According to the parameter \( \Delta \), the set of thresholds \( v_{th}(\Delta, d) \) are calculated using the coefficients \( a(\Delta, d), b(\Delta, d) \) and \( c(\Delta, d) \) computed off-line (cf. Section 2.1).

The other frames of the GOF, called constrained frames \( F_c \), are encoded with a limited RDO process. For each CTU, an according CDM is generated by applying the algorithm described in the previous section using the sets of thresholds previously computed from the learning frame \( F_l \). However, the CDM generated is very restrictive for the RDO process as it allows only one depth to be searched for each CU of a CTU. To increase the accuracy of the depth map prediction with limited impact on the complexity, we propose to refine the CDM by generating a second CDM, called Refined CTU Depth Map (RCDM), and force the HEVC encoder to only apply the RDO process between the two CDMs. The RCDM is generated from the CDM by merging all the group of four neighbouring blocks (in the Z-scan order) with the same depth in the input CDM.

To conclude this section, our proposed energy reduction scheme takes as input the parameter \( \Delta \) to generate two CDMs. The CDMs are predicted from the variance of its samples and refined to generate the RCDM. Then, the HEVC encoder is forced to only apply the RDO process between the CDM and the RCDM to limit the exploration of the quad-tree decomposition of a CTU.

3. Machine Learning CTU Partitioning Prediction

The previous Section shows a probabilistic approach to predict the CTU partitioning from the variance information of the samples. The state-of-the-art shows that several attributes may also be used to predict the CTU partitioning as explain in the introduction [11] [7]. However, manually applying a probabilistic approach with such large attributes vector will grow exponentially in complexity and thus it is well suited for a machine learning approach.

The Waikato Environment for Knowledge Analysis (WEKA) [13] is used to generate the Machine Learning classifiers based on offline training. WEKA is a free, university open-source, Machine Learning tool that includes several machine learning algorithms for classification, regression, clustering, association rules, and visualization.

3.1 Quad-Tree Partitioning Based on Machine Learning

Like the probabilistic approach, the machine learning approach aims to classify each CU in one of the two population Merge (M) and Non Merge (NM). The first step is the feature selection which will be used by the Machine Learning decision trees. This step aims to select the most relevant attribute to the predictive modeling problem, in our case, to the CU classification. Numerous features can be extracted from a CU samples to describe its content. The 12 following features are extracted from the state-of-the-art and selected for this study:

- The variance of the current CU samples (1 feature)
IEEE COMSOC MMTC Communications - Frontiers

- The variances of the 4 sub-CU samples (4 features)
- The variance of the upper CU samples (1 feature)
- The variance of the 3 neighboring CU samples in the Z-scan order (3 features)
- The variance of the variance of the 4 sub-CU samples (1 feature)
- The variance of the mean of the 4 sub-CU samples (1 feature)
- The QP value (1 feature)

The training of the decision trees is performed with the C4.5 classifier [14] on a training video sequence set composed by: Traffic (2560x1600), Cactus (1920x1080), BQMall (832x480), BasketballPass (416x240) and Johnny (1280x720) which being different in term of resolutions and video content characteristics (both Spatial and Temporal information). The training dataset is composed by 80000 instances equally extracted from 4 QP values (22, 27, 32, 37). At each depth $d$, two different types of tree are trained:

- the Merge Tree which predicts if the CU at the depth $d$ have to be merge in CU at the depth $d-1$,
- the Split Tree which predicts if the CU at the depth $d$ have to be split in 4 CUs at the depth $d+1$.

The measure of the trained decision tree accuracy is carry out by the 10-fold cross-validation process of WEKA. The Table 1 show the percentage of correctly classified instances for the 8 trees:

<table>
<thead>
<tr>
<th>Depth $d$</th>
<th>Merge Tree</th>
<th>Split Tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-</td>
<td>80.93%</td>
</tr>
<tr>
<td>1</td>
<td>81.26%</td>
<td>80.87%</td>
</tr>
<tr>
<td>2</td>
<td>80.19%</td>
<td>80.90%</td>
</tr>
<tr>
<td>3</td>
<td>80.64%</td>
<td>82.24%</td>
</tr>
<tr>
<td>4</td>
<td>81.39%</td>
<td>-</td>
</tr>
</tbody>
</table>

The results of the training phase show that all the decision trees have a percentage of correctly classified instances up to 80%. The next Section presents the algorithm of CTU partitioning prediction using the Machine Learning decision trees.

3.2 Prediction Algorithm for CTU Partitioning based on Machine Learning

Similarly to the previous probabilistic algorithm for CTU partitioning, the Machine Learning strategy follows a bottom-up approach from depth $d = 4$ to $d = 1$. However, to decide if a group of 4 neighboring CUs have to be merged, the Machine Learning approach gives 5 answers: 4 from the Merge Tree applied on the 4 neighboring CUs of the depth $d$ and 1 from the Split Tree applied on the equivalent CU (in term of samples) at the depth $d - 1$.

The Machine Learning algorithm starts by initializing the CDM with the depth value 4 and all the features of the CUs $4 \times 4$ are computed. Then, the algorithm explores the CTU decomposition from $d = 4$ to $d = 1$. For the current depth $d$, the algorithm browses the CDM and applies the Merge Tree on the 4 neighbours CUs in the Z-scan order in the depth $d$ and applies the Split Tree on the equivalent CU at the depth $d - 1$. If all the decision trees predict that the neighbours CUs have to be merged, then the CUs are merged and the corresponding elements is the CDM are set at $d - 1$ and the features of the merged CU are calculated.

Figure 4 presents a high-level diagram of our overall Machine Learning CTU partitioning prediction technique. For each CTU of each frame of the video sequence, the previous algorithm is firstly applied to generate the CDM. Then, the RCDM is computed from the CDM applying the Depth Map Refinement algorithm described in Section 2.2 to increase the accuracy of the depth map prediction with a limited impact on the complexity. To finish, the RDO process of the HEVC encoder is processed between the CDM and the RCDM.

http://www.comsoc.org/~mmc/
IEEE COMSOC MMTC Communications - Frontiers

4. Performance Evaluation and Comparison

The following section gives the experimental setup and a comparison of the results obtained for our two proposed energy reduction schemes: the Probabilistic and the Machine Learning approaches on a real-time HEVC encoder.

4.1 Experimental Set-Up

All experiments are performed on one core of the embedded EmETXe-187M0 platform from Arbor Technologies based on an Intel Core i5-4402E processor at 1.6 GHz. The studied HEVC software encoder is Kzaaar [15] in All Intra configuration. Since the configuration aims to be real-time, from the study of [6], the RDOQ and the Intra transform skipping are disabled. Each sequence is encoded with 4 different QP values: 22, 27, 32, 37.

To measure the energy consumed by the platform, Intel Running Average Power Limit (RAPL) interfaces are used to get the energy of the CPU package, which includes cores, I/Os, DRAM and integrated graphic chip set. As shown in [16], RAPL power measurements are coherent with external measurements.

The $\Delta$ parameter of the Probabilistic approach is fixed at 0.6 and the GOF size is fixed at 50.

4.2 Experiment Results

The performance of our proposed energy reduction scheme was evaluated by measuring the trade-off between Energy Reduction (ER) in % and RD efficiency using the BD-BR and BD-PSNR. ER is defined by the following equation:

$$ ER = \frac{100}{4} \sum_{QP_t \in \{22, 27, 32, 37\}} \frac{E_{\text{Ref}}(QP_t) - E_{\text{Red}}(QP_t)}{E_{\text{Ref}}(QP_t)} $$

where $E_{\text{Ref}}(QP_t)$ is the energy using to encode the video sequence without constraint and $E_{\text{Red}}(QP_t)$ the energy to encode the same sequence with our proposed energy reduction scheme, both with $QP = QP_t$.

Table 2 details the performance of the two proposed energy reduction schemes: the Probabilistic and the Machine Learning approaches in terms of BD-BR, BD-PSNR and ER for 18 different sequences belonging to the 5 classes A, B, C, D and E each one corresponding to a specific resolution or video content. The energy values include the energy overhead due to the entire energy reduction scheme as the variance computation.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>BD-BR (in %)</th>
<th>BD-PSNR (in dB)</th>
<th>ER (in %)</th>
<th>BD-BR (in %)</th>
<th>BD-PSNR (in dB)</th>
<th>ER (in %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Traffic</td>
<td>4.59</td>
<td>-0.24</td>
<td>60.16</td>
<td>4.05</td>
<td>-0.21</td>
<td>58.69</td>
</tr>
<tr>
<td>A PeopleOnStreet</td>
<td>4.28</td>
<td>-0.24</td>
<td>59.06</td>
<td>3.73</td>
<td>-0.21</td>
<td>57.09</td>
</tr>
<tr>
<td>B Kimono</td>
<td>13.28</td>
<td>-0.43</td>
<td>52.59</td>
<td>9.51</td>
<td>-0.31</td>
<td>61.33</td>
</tr>
<tr>
<td>B ParkScene</td>
<td>4.29</td>
<td>-0.19</td>
<td>55.84</td>
<td>3.86</td>
<td>-0.17</td>
<td>60.28</td>
</tr>
<tr>
<td>B Cactus</td>
<td>3.69</td>
<td>-0.11</td>
<td>60.75</td>
<td>4.65</td>
<td>-0.13</td>
<td>60.16</td>
</tr>
<tr>
<td>B BQTerrace</td>
<td>3.56</td>
<td>-0.13</td>
<td>60.40</td>
<td>3.79</td>
<td>-0.14</td>
<td>61.34</td>
</tr>
<tr>
<td>B BasketballDrive</td>
<td>2.25</td>
<td>-0.14</td>
<td>62.06</td>
<td>1.99</td>
<td>-0.15</td>
<td>58.93</td>
</tr>
<tr>
<td>C RaceHorses</td>
<td>3.11</td>
<td>-0.18</td>
<td>59.68</td>
<td>2.95</td>
<td>-0.17</td>
<td>60.37</td>
</tr>
<tr>
<td>C BQ Mall</td>
<td>1.88</td>
<td>-0.13</td>
<td>59.48</td>
<td>1.10</td>
<td>-0.08</td>
<td>57.10</td>
</tr>
<tr>
<td>C PartyScene</td>
<td>2.74</td>
<td>-0.13</td>
<td>58.33</td>
<td>4.97</td>
<td>-0.24</td>
<td>62.27</td>
</tr>
<tr>
<td>C BasketballDrill</td>
<td>3.46</td>
<td>-0.19</td>
<td>57.54</td>
<td>3.16</td>
<td>-0.17</td>
<td>57.13</td>
</tr>
<tr>
<td>D RaceHorses</td>
<td>2.47</td>
<td>-0.15</td>
<td>59.07</td>
<td>1.93</td>
<td>-0.12</td>
<td>58.57</td>
</tr>
<tr>
<td>D BQSquare</td>
<td>2.73</td>
<td>-0.21</td>
<td>57.81</td>
<td>1.00</td>
<td>-0.08</td>
<td>55.51</td>
</tr>
<tr>
<td>D BlowingBubbles</td>
<td>1.45</td>
<td>-0.10</td>
<td>55.25</td>
<td>1.24</td>
<td>-0.08</td>
<td>53.54</td>
</tr>
<tr>
<td>D BasketballPass</td>
<td>2.39</td>
<td>-0.14</td>
<td>59.20</td>
<td>2.31</td>
<td>-0.14</td>
<td>58.01</td>
</tr>
<tr>
<td>E FourPeople</td>
<td>4.78</td>
<td>-0.27</td>
<td>55.70</td>
<td>4.51</td>
<td>-0.25</td>
<td>55.13</td>
</tr>
<tr>
<td>E Johnny</td>
<td>5.76</td>
<td>-0.23</td>
<td>51.63</td>
<td>5.49</td>
<td>-0.22</td>
<td>52.26</td>
</tr>
<tr>
<td>E KristenAndSara</td>
<td>4.10</td>
<td>-0.21</td>
<td>54.24</td>
<td>4.62</td>
<td>-0.23</td>
<td>53.39</td>
</tr>
<tr>
<td>Average</td>
<td>3.93</td>
<td>-0.19</td>
<td>57.71</td>
<td>3.60</td>
<td>-0.17</td>
<td>57.84</td>
</tr>
</tbody>
</table>

http://www.comsoc.org/~mmc/ 24/36 Vol.12, No.5, September 2017
IEEE COMSOC MMTC Communications - Frontiers

The results show that the both energy reduction techniques achieve an average of 58% of energy reduction. Indeed, the overhead due to the unconstrained Learning Frames and the variance computations of the Probabilistic approach is approximately equal to the overhead of the features computations of the Machine Learning approach. However, even if the Probabilistic approach does not constrain all the frames (only 49 every 50 frames), this approach causes more encoding degradations: +0.33% of BD-BR and -0.02dB of BD-PSNR, than the Machine Learning approach.

It is noticeable in the Table REF than the Kimono sequence has more degradations than the other sequences: 13.28% of BD-BR increasing with the Probabilistic approach and 9.51% of BD-BR increasing with the Machine Learning approach. This can be explained by the texture specificity of the Kimono video sequence which is composed by a traveling of trees and vegetation in the background. This video sequence has the highest Spatial Information due to the details. Nevertheless, the results show that the Machine Learning approach reduces the degradation of 3.77% of BD-BR compare to the Probabilistic approach.

To conclude, the Machine Learning approach achieved overall better results than the Probabilistic approach while releases the learning frames needed to predict the quad-tree partitioning. Not using learning frames allows to constrain every frame and make easier the use of this energy reduction technique to control the energy consumption of the HEVC encoding process.

5. Conclusion

This paper presents a fair comparison between two proposed methods of CTU partitioning prediction in order to reduce the energy consumption of a real-time HEVC encoder. The two CTU partitioning prediction techniques limit the recursive RDO process and thus reduce the energy consumption of the encoding process. The first energy reduction technique proposed exploits the correlation between the CTU partitioning and the texture complexity and variance of the CU luminance samples with a Probabilistic approach. Experimental results show that the proposed overall algorithm reduces the energy consumption by 58% on average for a bit rate increase of 3.93%. The second energy reduction technique uses Machine Learning to predict the CTU partitioning. This technique achieved overall better results than the Probabilistic approach with an average bit rate increase of 3.6% for the same energy reduction of 58%. Future work will use the Machine Learning CTU partitioning prediction technique to control the energy consumption of the HEVC Intra encoder for a given energy consumption budget.

Acknowledgement

This work is partially supported by the French ANR ARTEFaCT project, by COVBIE project funded by Brittany region and by the European Celtic-Plus project 4KREPROSYS funded by Finland, Flanders, France, and Switzerland.

References


http://www.comsoc.org/~mmc/
Alexandre Mercat received the Engineering degree in Electronics and Computer Engineering from the National Institute of Applied Sciences (INSA), Rennes in 2015. Now, he is pursuing the Ph.D. degree in Electrical and Computer Engineering from the Institute of Electronics and Telecommunications of Rennes (IETR) laboratory. His research interests include implementation of image and signal processing applications in many core embedded systems, video coding algorithms, complexity-aware video coding, approximate computing, power consumption and digital systems design.

Florian Arrestier has received the Engineering degree in Electronics and Computer Engineering from the National Institute of Applied Sciences in Rennes (INSA) in 2017. He is currently pursuing a PhD degree in Electrical and Computer Engineering from the Institute of Electronics and Telecommunications of Rennes (IETR) laboratory.

Wassim Hamidouche received the Engineering Degree in Computer Science from the University of Sciences and Technologies of Algiers, Algeria, in 2006 and the Master Degree in Electrical Engineering from the University of Poitiers, France, in 2007. He received a Ph. D. Degree in Signal and Image Processing from the University of Poitiers, France in 2010. From 2011 to 2012 he has been a Research Engineer with Canon Research Centre, Rennes, France, where he worked on video compression standard High Efficiency Video Coding (HEVC) and its scalable extension SHVC. From 2013 to 2015 he was a research engineer with the IETR laboratory, IMAGE group, Rennes, France. Since 2015 he has been an associate professor at INSA Rennes. His research interests focus on efficient real time and parallel architectures for the new generation video coding standard, multimedia transmission over heterogeneous networks, and multimedia content security.

Maxime Pelcat is an Associate Professor at the INSA in Rennes. He holds a joint research appointment at IETR in Rennes and at Institut Pascal in Clermont Ferrand, two CNRS research units. Maxime Pelcat obtained his Ph.D. in signal processing from INSA Rennes in 2010, thesis resulting from a collaboration of Texas Instruments, Nice and INSA Rennes. Previously, after one year in the Audio and Multimedia department at Fraunhofer Institute IIS in Erlangen, Germany, he worked as a contractor at France Telecom Research and Development until 2006. He is an author of 50+ peer reviewed publications since 2009 in the domains of models of computation, energy efficiency, multimedia and telecommunication processing, and programming of parallel embedded systems. Maxime Pelcat has served as Guest Editor for the Springer Journal of Signal Processing Systems and is an author of the best paper award at DASIP 2014, the best demo awards at ICME 2015 and EDERC 2014 and of the book “Physical Layer Multi-Core Prototyping” Springer, 2012.

Daniel Menard is Professor in the Electronics and Computer Engineering (ECE) department at INSA-Rennes (graduate engineering school). He is also member of the IETR/CNRS laboratory. He received the Ph.D. and HDR (habilitation to conduct researches) degrees in Signal Processing and Telecommunications from the University of Rennes, respectively in 2002 and 2011. From 2003 to 2012 he was an Associate Professor at the ECE department of the University of Rennes engineering school, ENSSAT. He was also member of the IRISA/INRIA laboratory. He is also author of more than 90 international papers distributed in the areas of embedded systems, computer-aided design, arithmetic and signal processing. His research interests include implementation of image and signal processing applications in embedded systems, approximate computing, fixed- point arithmetic, low power systems for video compression.
Advanced Techniques for Power- and Energy-Efficient Design of Video CODECs in the Era of Internet-of-Things
Muhammad Shafique, Rehan Hafiz, Sameen Rehman, Jörg Henkel
Vienna University of Technology (TU Wien), Vienna, Austria
Information Technology University (ITU), Lahore, Pakistan
Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
{muhammad.shafique, sameen.rehman}@tuwien.ac.at, rehan.hafiz@itu.edu.pk, henkel@kit.edu

1. Introduction
In today’s era of smart Internet-of-Things (IoTs) and Cyber Physical Systems (CPS), video coding based applications are ubiquitous, ranging from mobile devices, smart homes/buildings/cities, smart hospitals, and industrial monitoring, to surveillance, automotive, entertainment, Multimedia Sensing as a Service (MaaS) [1], etc. This also comes with the big data problem due to the increasing video resolutions (ultra-HD to Super-Vision) and gigantic rates of video content generation and transmission over internet (ca. 80% of global consumer traffic [2]). To address such big data challenges, several video coding standards have emerged like High Efficiency Video Coding (HEVC) [3] and Google’s VP9 [4]. The HEVC provides 1.6x–2x improved coding efficiency compared to the H.264/AVC standard [5] (see Figure 1a) by leveraging several new coding tools like Coding-Tree Block or Unit (CTB, CTU) with recursive partitioning into multiple Coding Units (CU) of sizes 64x64, 32x32, down to 4x4 pixels (see Figure 1d-e), variable-sized Transform and Prediction Units (TU, PU), and numerous prediction modes for intra- and inter-prediction [3][6][7][8]. A CU can be encoded using inter- or intra-prediction. In case of inter-prediction, Motion Estimation (ME) is employed to search a matching block within a so-called search window, in a given reconstructed reference frame, using a matching metric like Sum of Absolute Differences (SAD); see Figure 1f-g. ME can be performed for each possible CU inside a CTU. To determine the best coding mode for achieving a high compression efficiency, HEVC employs an extensive Rate-Distortion Optimized (RDO) mode decision process, which evaluates all possible combinations of CU, PU, and TU to determine the best Quad-Tree partitioning. However, design space exploration of these CU, PU and TU modes, increased Motion Estimation (ME) effort due to so many inter-prediction modes, and numerous intra-prediction modes lead to significantly increased execution time, power/energy consumption, and memory requirements compared to the H.264 encoder [7]-[15] (also see Figure 1a-c), which also leads to elevated on-chip thermal profiles [16][17] (see Fig. 2). Besides computations, a majority of energy portion is spent in memory transfers during the ME process [18]-[21]. Moreover, parallelizing HEVC using the built-in video tile feature also aggravate the memory pressure due to concurrent memory access and overlapping data access across different tile boundaries [20]. Therefore, there is a need to investigate advanced power/energy reduction techniques at both hardware and software levels for 2D/3D video codecs, while also accounting for the escalating thermal and reliability issues in the emerging nano-scale fabrication technologies. Such techniques need to jointly optimize for different computational components and memory sub-systems to enable highly energy-efficient video coding systems for IoT and CPS.

![Figure 1: (a) HEVC time and bitrate normalized to H.264, (b) Memory bandwidth requirement for HEVC and H.264, (c) Energy distribution [%] of the HEVC encoder for the “BasketballDrive” Fall-HD (1920x1080) sequence using 5 reference frames and CTU size of 64x64, (d) CTU structure with recursive partitioning, (e) CTU partitioning for the Basketball sequence, (f) Motion estimation/search process, (g) SAD accelerator [9][23].](http://www.comsoc.org/~mmc/)

This paper discusses the following key research directions to develop power-/energy-efficient video coding systems:
1) **Software-level techniques for energy efficiency** that include workload balancing, partitioning of video frames into multiple tiles, adaptive complexity reduction, hardware/software co-design, content-driven memory power management, content-driven thermal management, application-level approximate computing, etc.

2) **Hardware-level techniques for energy efficiency** that include specialized hardware accelerators, approximate accelerator design, power-efficient architectures for on-chip video memories, memory power/energy reduction considering both on-chip and off-chip memories, approximate memories, etc.

3) **Power-efficient reliability** for video codecs considering content properties.

4) **Emerging trends** in computing hardware and complexity reduction for video codecs.

2. **Analysis of HEVC Encoding for Processing, Memory, and Temperature Profiles**

In the following, we will present analysis of HEVC encoder to derive some key observations to enable efficient hardware/software solutions.

**Complexity Analysis of ME**: Fig. 2a shows the number of CTUs, CUs, and search candidates evaluated per frame using a full-search and fast adaptive ME (FME, AME) for different video resolutions. Since one SAD computation for a 32x32 CU requires ca. 2K arithmetic operations, the FME processing load for real-time (30fps) processing of HD1080p and CIF videos is ca. 50x10^12 and 160x10^9 operations, respectively (1.8x10^12 and 5.4x10^10 operations for AME considering on average 80%-90% reduction in the search complexity [22]). Therefore, besides employing specialized hardware accelerators for SAD computations, the emerging concept of approximate computing can also be exploited to obtain high-energy savings as demonstrated in [23]-[30][34].

**Early CU/PU Prediction**: Fig. 2b shows that the contribution of larger CU/PU increases with an increasing value of QP. Typically, image areas with smooth texture (low variance, homogenous texture) and slow motion can easily be captured by large-sized CUs/PUs, see Figure. 1c. Therefore, early CU/PU size prediction by exploiting the knowledge of texture and motion in the spatial and temporal proximity (also inter-view correlation in case of 3D videos), can provide significant reduction in the processing and energy requirements.

![Diagram](image)

**Early CU/PU Prediction**: Fig. 2c shows the increased memory requirements of HEVC compared to that of H.264. Fig. 2d shows the percentage of data used within the search window during the ME of various blocks. This reveals an interesting observation that most of data in the search window is not even accessed by the ME due to its adaptive nature and refinement process towards the global minima. Since search windows are typically stored in on-chip memories, such behavior leads to area wastage as well as power wastage in on-chip memory leakage and off-chip memory accesses. Therefore, such an analysis and observations can be exploited to develop efficient on-chip video memory architectures, adaptive search window re-sizing, intelligent data prefetching, frame compression, content-driven power management of memories, approximate memories, as demonstrated in previous works [10], [18]-[21], [31]-[35].

**Thermal Analysis**: Fig. 2e-f show two scenarios of HEVC encoding and on-chip thermal profiles (measured using an IR-Thermal camera; see details of the setup and measurement process in [16][17]) for two different video sequences with diverse texture and motion content. Fig. 2e illustrates the tradeoff between video content, frequency scaling, and the resulting temperature. For a given deadline, the frequency of the video sequence with relatively low content/motion content is throttled down, which results in a lower peak temperature. Fig. 2f shows the tradeoff between video content, adaptive core count scaling, and the resulting temperature. For a given deadline, the video

---

http://www.comsoc.org/~mmc/
sequence with relatively high texture/motion content requires two cores to be simultaneously powered-on, processing two video tiles in parallel. This leads to a high peak temperature. This analysis shows that video content properties can be leveraged to scale frequency, voltage, and number of active cores in order efficiently manage the temperature of a video coding system. Such studies can be found in [16][17].

3. Key Research Directions for Power-/Energy-Efficient Video Coding Systems

In the quest to achieve a high energy-efficiency for video coding systems (based on HEVC or H.264), we have explored various methods at both software (SW) and hardware (HW) levels for the computational and memory components. Several methods at the HW/SW levels operate jointly in a collaborative way to achieve higher energy savings.

At the software layer, first, the input video is pre-processed to statistically characterize different regions w.r.t. their texture and motion properties [22][36]. This is used to optimize the software architecture of a given video encoder [37]-[39], and then to perform workload budgeting for different tiles and CUs considering the underlying hardware platform and the user-provided QoS constraints [13][14]. Afterwards, following key steps are taken to determine the computational demands of a video codec from the underlying hardware platform.

1) Partitioning of a video frame into multiple tiles such that their workloads are balanced across different cores to reduce their voltage/frequency for energy minimization [37][43][44].

2) Adaptive computational complexity reduction to eliminate improbable coding modes (CU/PU) and to select a small subset of coding modes for RDO mode decision, while keep the video quality loss to minimal/imperceptible [11][12][36][40]-[42].

3) Hardware/software partitioning and hardware/software co-design techniques, i.e. how to trade-off hardware and software for energy efficiency [57].

At the hardware layer, the key to get high energy-efficiency and performance is specialized hardware accelerators and coprocessors [11][21][45]-[47] and memory sub-system [10], [18]-[21], [31]-[35] for the compute-intensive kernels of the video codec as well as their efficient run-time management depending upon the varying properties of the input videos and software-level decisions (like mode decision and ME algorithm). Both the accelerators and memories can be of exact or approximate in nature depending upon the tolerable trade-offs between the output quality degradation and required energy savings [23][54]. In depth details of these works can be found in the mentioned references. In the following, we briefly explain the key concepts, followed by a short discussion of the emerging trends in achieving extreme energy efficiency.

3.1 Adaptive Complexity Reduction [11][12][22][36][40]-[42]: Reducing the computational complexity of the HEVC coding algorithms (like RDO mode decision and motion estimation) not only directly lower the energy requirements due to reduced number of operations, but also facilitates reducing the voltage and frequency of the underlying processing platform that directly impacts the resulting power/energy consumption. The key challenge here is to proactively eliminate the set of improbable coding modes and identify the set of most efficient coding modes considering the input video characteristics and the available spatial/temporal correlation, such that, the loss in video quality is minimal when compared to the exhaustive RDO-based mode decision. Fast coding mode decisions involve prediction of the near-optimal CU, PU, and TU sizes as well as fast motion estimation and intra-mode decisions. Towards this end, exploitation of the variance and motion properties of different video blocks as well as previously selected modes for the spatial/temporal neighboring blocks can play an important role. For instance, a large-sized CU/PU can be predicted for a homogeneous block with small-to-medium motion, while small-sized CU/PU would be amenable high-texture blocks with high motion. Care has to be taken when determining the thresholds for different fast decision policy. For this, statistical analysis of video properties and their relationship to the optimal coding modes need to be established while accounting for the impact of Quantization Parameter (QP) on the optimal decisions obtained through exhaustive RDO mode decision. Our studies have shown that considering the above aspects leads to significant complexity reduction ca. 60% with minimal quality loss to up to 80%-90% when a certain amount of quality degradation is allowed, for instance, in case the battery status of a device is very low.

3.2 Workload Balancing and Video Tiling for Multi-/Many-Core Processors [37][43][44]: The HEVC standard allows breaking a video frame into multiple video tiles that can be independently and concurrently processed on different cores available in a multi-/many-core processor. Under different system constraints, different types of mapping policies may be adopted [48]. Since the workload of each tile can vary depending upon the texture and motion properties of the objects occurring in that tile, a challenging problem is “tile formation”, i.e., how many and which CTUs should form a particular tile such that their workload is balanced across all the cores while minimizing the quality loss due to tiling. This is important to reduce the voltage and frequency of all cores, or a set of core in a particular voltage/frequency-island, which will lead to significant power/energy reduction. Alternatively, several

http://www.comsoc.org/~mmc/  29/36  Vol.12, No.5, September 2017
IEEE COMSOC MMTC Communications - Frontiers

tiles can be mapped to a single core to increase the number of idle cores, which can then be power-gated to additionally save leakage power. This would require solving a bin-packing problem. If the workload is very high such that the available cores cannot meet the throughput constraints, coordination with the software-level complexity reduction schemes will be required to further curtail the processing requirements. Therefore, such a “tile formation” decision has to account for the characteristics of the input videos, the underlying processing platform (e.g., number of available cores, types of cores, available pairs of voltage-frequency, different power-gating modes, etc.) and the QoS/system constraints as well as the “tile formation” decisions for the previously encoded frames/group-of-pictures as they may exhibit high correlation. During the encoding, the “tile formation” may need to be dynamically adapted due to scene cuts or sudden variations in the video content, which can lead to a sudden workload imbalance. Our proposed methods for power-efficient “tile formation” have shown power savings of up to 53%. In [9][11], we developed a multithreaded version of the HEVC Intra-only encoder in C++, which allows threading capabilities at GOP-, slice- and video tile-level. Its single thread is ≈13.2x faster than the HM-9.2 reference software.

3.3 Exact and Approximate Hardware Accelerators and Coprocessors [23][24][28][29]: The computational kernels (like motion estimation and intra prediction) of HEVC are accelerated using dedicated HW accelerators, which can be of exact (i.e. complying to the precise Boolean equivalence between specification and implementation) or approximate nature. Approximate computing an emerging area that relaxes the bounds of Boolean equivalence of a computation to achieve high-energy savings. For instance, Figure. 3(a-b) illustrates various designs and area/power/energy results of different approximate 1-bit full adders (FA) [25][26] used to develop the multi-bit approximate adders (see open-source library of approximate modules in [27]-[29]), and their impact on the quality of a low-pass image processing filter. Due to the short circuit of input to output, the AppxAdd3 design offers the best area, power, and latency results but with the highest error rate. On the contrary, the AppxAdd1 and AppxAdd2 designs provide good tradeoff between power and error rate. In a multi-bit adder design, only the selected number of LSBs can be approximated using a particular 1-bit approximate adder considering the given constraints on the tolerable error rate and magnitude; Figure. 3(c). It can be seen in Figure. 3(a) that for the image filtering application, AppxAdd3 leads to high quality loss with yet recognizable objects, and provides very high power savings. However, despite Peak Signal to Noise Ratio (PSNR) loss, the AppxAdd2 design produces similar subjective quality results as of the accurate design while providing high power savings, and therefore is a good design option for the approximate image filtering.

![Figure 3](image)

Resilience Analysis of the Motion Estimation Process: An important question is: Can different kernels in HEVC be approximated to achieve high-energy savings, while still providing correct output stream. The answer is yes! The SAD (Sum of Absolute Difference) computations during the motion estimation (ME) process are highly amenable to approximate computing without affecting the correctness of the HEVC encoder, as shown in Figure. 3(d). In this particular example even the best candidate “S3” remains unchanged when replacing the accurate adders/subtractors with the approximate ones in the SAD accelerators. In general, the error surface for different motion search candidates shift in roughly the same direction, i.e., following the same trend as shown in Figure. 3(d). Therefore, in most of the cases, the global minima found by the motion search strategy is unchanged or near-optimal, thus resulting in slight increase in the prediction error leading to slightly higher bit rates compared to the accurate case. This illustrates the inherent resilience of the ME process to the approximation-induced errors, which can be leveraged to achieve high-energy savings through (relaxed or aggressive) approximations.

http://www.comsoc.org/~mmc/ 30/36 Vol.12, No.5, September 2017
IEEE COMSOC MMTC Communications - Frontiers

Approximate Architecture for the HEVC Motion Estimation – Figure 3(c): It supports heterogeneous approximate tiles of different SAD variants (of sizes 8x1, 8x8, 16x16, and 32x32), providing different energy vs. quality tradeoffs. A subset of approximate/accurate SAD tiles can be selected at run time depending upon the input video types, user constraints (e.g., tolerable error), and run-time available energy budgets, while others are kept power-gated. The accelerator array receives the input data from the on-chip memory. The search strategy executes on the main processor core, which communicates with the SAD array through a coprocessor interface. We synthesized our ME coprocessor architecture using the ASIC design flow for a 45nm TSMC library. The approximate designs provide similar PSNR with a bit rate increase in the range of 2% to 20% for full-HD sequences, while providing an energy saving of more than 2x in SAD computations.

3.4 Accurate and Approximate Video Memories [10], [18]-[21], [31]-[35]: Video codecs, especially the HEVC, are a memory-intensive application, as can also be seen in Figure 1b and Fig. 2c. This is mainly due to the large storage required for multiple reference frames and repetitive accesses to the memory during the motion estimation (ME) and intra-prediction processes, especially for high resolutions like full-HD and beyond. Typically, during the ME process, large portions of the reference frames like search windows are brought to SRAM-based on-chip video memories from the DRAM-based off-chip memories. This leads to high DRAM access energy and high leakage energy of SRAMs. To mitigate these issues for both 2D and 3D video coding, we developed highly optimized on-chip video memory architectures and management policies in [10], [18]-[21], [31]-[35]. Key techniques that we proposed cover (1) low-power memory architectures with content-driven and application-specific power management policies for 3D video coding and HEVC [10][18]-[21][35]; (2) memory pressure balancing for 3D video coding [31][32]; (3) reference frame compression to reduce DRAM accesses [33]; (4) designing hybrid on-chip and off-chip video memories that combine the benefits of SRAM/DRAM and emerging non-volatile memories (NVMs) [10][35]; and (5) exploiting algorithm and content knowledge to approximate NVMs deployed in video codecs. Besides these, we have also introduced several other policies in the above papers like reducing off-chip memory accesses through intelligent data reuse, motion trajectory prediction, and search window re-sizing. Our techniques provide significant energy reductions in DRAM accesses as well as leakage and dynamic energy reduction for on-chip video memories compared to state-of-the-art. More details of these comparisons can be found in the above papers.

![Graphs showing PSNR, Bit Rate, and Temperature Reduction](image)

Figure 4: (a-d) Temperature, bit rate and PSNR for different parameters like QP, CU Size, number of Reference Frames and Search Area, (e-f) Configuration points for temperature reduction, (g-h) PSNR and Bit Rate results [16][17].

3.5 Dynamic Thermal Management for Video Coding [16][17]: In the nano-scale CMOS technology era, it is very important to efficiently manage on-chip temperature profiles, especially for compute and memory-intensive applications like video codecs. In this regard, exploitation of the application- and input content-specific properties can be vital. Towards this, we developed application-driven dynamic thermal management (DTM) policies for the HEVC encoders [16][17]. To understand the impact of several key encoding parameters like Quantization Parameter (QP), size of the Coding Units (CU), number of reference frames (RF) and the search area (SA) on the on-chip temperature and encoding quality, we performed an extensive analysis for the HEVC encoder; see Figure 4(a-d). For an efficient design of an application-aware DTM policy, we first perform a Pareto analysis of different configurations for encoding different test video sequences from different complexity classes mentioned by the standardization committee; see Figure 4(e-f). These Pareto-Optimal configuration points then serve as the input to our policy. By exploiting an application-level prediction of the temperature trend, our content-aware DTM policy adapts the above parameters at run time to select a Pareto-Optimal point under a given thermal constraint in order to keep the temperature within safe thermal limits while incurring minimum penalties in terms of bit rate and video quality; see Figure 4(g-h). In case of thermal emergencies, our policy triggers the voltage and frequency scaling to

http://www.comsoc.org/~mmc/
IEEE COMSOC MMTC Communications - Frontiers

rapidly cool down the core. Figure. 4(g-h) shows the impact of our policy on the video quality (in terms of PSNR and bit rate) compared to without thermal optimization for different video sequences. The quality degradation increases when decreasing the threshold voltage constraint which makes it tougher to process high complexity computations for HEVC. At a constraint of 54 °C, the average PSNR loss is of 0.007 dB while the bit rate slightly increases 0.99% on average for all sequences. Note that, for the medium-low complexity sequences like BQMall and BasketBallDrill, both the PSNR and bit rate degradation is relatively low compared to the medium-high sequences like PartyScene, RaceHorses and Keiba. It occurs, since for reducing the temperature of high complexity sequences, the encoder configuration provided by our technique needs to reduce the workload for low complexity sequences. Compared to state-of-the-art techniques that typically employ frame drops in such thermal emergencies leading to severe low in PSNR (typically in the range of tens of dB), our policy provides significantly improved output quality.

3.6 Power-Efficient Reliability for Video Codec [49]-[56]: The video coding systems fabricated in the nano-scale CMOS technologies are highly susceptible to various reliability threats like soft errors, aging and process variation, which result in transient or permanent errors. However, protecting against these reliability threats can potentially incur significant overhead in terms of area, power/energy, and latency. Therefore, there is a dire need for providing power-efficient reliability features for video coding systems. Soft errors manifest as transient bit flips in the logic and memory, and can corrupt different parts of the video coding process like header information, motion vectors, transformed coefficients, etc. In [49]-[51], we developed a fault-tolerant entropy coding hardware for the H.264/AVC encoder, which leverages the knowledge of both algorithm (like distributions of different syntax elements for different types of videos) and architecture to protect the entropy coding process in a power-efficient way. It also employs concepts like selective redundancy [49][53] to protect important coefficients, as well as partitioning of different VLC tables such that memories for the unused sub-tables with their parity hardware can be power-gated to reduce the power consumption. Our results showed a reduction of 2x in the area and performance overhead while providing significant power/energy gains compared to state of the art techniques. Aging fault typically result in threshold voltage and static-noise margin (SNM) degradation in the SRAM-based memories, which lead to access errors. In [54][55], we proposed a content-aware microarchitectural-level technique for mitigating aging of these SRAM-based memories, by efficiently altering the input and output data with very low overhead. Our technique performs power, area, and aging analysis for different aging balancing circuits integrated into an on-chip video memory architecture, and leverages this analysis to develop a power-efficient anti-aging memory architecture. It is coupled with a run-time anti-aging controller that exploits the data characteristics to efficiently take the following key decisions: “which aging balancing circuit to activate for which SRAM cells and at what time instant”. Our experiments illustrate significant aging improvements at a low power overhead while demonstrating power vs. reliability tradeoffs under different run-time scenarios. Furthermore, in the emerging Dark Silicon era, where not all transistors on a chip can simultaneously be powered-on at the full performance level due to high power densities, elevated thermal profiles and cost-constrained cooling technology, judiciously using the power budgets is crucial. Towards this, in [56], we proposed an adaptive technique to distribute power budgets and resources among concurrently executing threads of a video codec. Our technique achieved ca. 30% higher throughput in terms of frames-per-second compared to the state-of-the-art.

3.7 Emerging Trends and Open Research Problems: The recent advancements in brain-inspired computing and emerging hardware platforms (like approximate computing and neuromorphic hardware) can be leveraged to achieve extreme energy efficiency in next-generation video coding systems. In this context, we present the following perspectives and open research challenges for the video community.

1) **Light-weight machine learning techniques can be exploited to develop self-learning complexity reduction schemes** that can allow video coding systems to learn from the quality of their previous decisions to increase the probability of correctness for the future coding/mode decision while also accounting for the spatial and temporal correlation.

2) **Cross-layer approximate computing** concepts can be exploited at both software and hardware levels to synergistically steer heterogeneous approximation knobs for achieving a high degree of energy savings without incurring perceptible quality losses. This has to be done by carefully analyzing different spatial and temporal properties of the objects within a frame, across different frames, and even across different views in 3D-video coding systems.

3) **Massively parallel accelerator-abundant architectures with distributed on-chip video memories and multi-grained power- and thermal management features** can provide an attractive solution for developing highly energy-efficient video coding systems. These accelerators and memories can be of accurate/exact or approximate nature, where the approximation control needs to consider the run-time operating contexts and user constraints.
IEEE COMSOC MMTC Communications - Frontiers

4) **Self-aware concepts** that can allow a video coding system to learn and optimize its decisions on-the-fly to efficiently handle the changing video content and varying operating contexts.

5) **Emerging technologies** like Spintronic devices and memristors can be leveraged for building energy-efficient accelerators, as well as leveraging the emerging non-volatile memory (NVM) technologies like STT-MRAMs and PCMs (Phase-Change Memories) for on-chip and off-chip memories. An attractive solution could be replacing the complete memory hierarchy with NVMs, and even computing near to memory, or inside the memory, as discussed below.

6) **Emerging hardware platforms like Neuromorphic Processors and In-Memory Computing** can be leveraged to investigate new generation of extreme-energy efficient coding standards that go beyond the traditional block-based coding standards. In particular, the in-memory computing can serve as a powerful architectural means for significantly reducing the energy of extensive data transfers during the memory-intensive motion estimation and intra-prediction processes. For example, research at the IBM Research Zürich has shown experimental demonstrations of in-memory computing for unsupervised learning of temporal correlations and solution of linear equations, using up to a million phase-change memory devices [58]. Demonstrating such concepts for the next-generation video coding standards would be a great research direction.

4. Conclusion

In this paper, we have presented several trends and new research directions for developing power-energy-efficient video coding systems that exploit both software and hardware level techniques. The software-level technique range from efficient video tile sizing and workload balancing to adaptive complexity reduction. The hardware-level techniques cover design of application-specific hardware accelerators and specialized on-chip video memories. We also discussed how to handle new challenges like reliability in a power-efficient way as well as dynamic thermal management for video codecs. At the end, we discussed several emerging trends in advanced video coding systems ranging from brain-inspired computing concepts like approximate hardware and software, machine-learning based coding mode decisions, and massively parallel accelerator-abundant architectures with self-aware features.

Acknowledgement

This research was supported in parts by the German Research Foundation (DFG) as part of the GetSURE project under the priority program “Dependable Embedded Systems” (SPP 1500 - spp1500.iti.ekit.edu) [52] and as part of the Transregional Collaborative Research Centre “Invasive Computing” (SFB/TR 89); http://invasic.de.

References


IEEE COMSOC MMTC Communications - Frontiers


IEEE COMSOC MMTC Communications - Frontiers


Muhammad Shafique is a full professor at the Institute of Computer Engineering, Vienna University of Technology (TU Wien), Austria. He is directing the Group on Computer Architecture and Robust, Energy-Efficient Technologies (CARE-Tech). He received his Ph.D. in Computer Science from the Karlsruhe Institute of Technology (KIT), Germany in January 2011. Before, he was with Streaming Networks Pvt. Ltd. where he was involved in research and development of video coding systems for several years. His research interests are in computer architecture, power- & energy-efficient systems, robust computing covering various dependability & security aspects, emerging computing trends like neuromorphic and approximate computing, neurosciences, emerging technologies and nanosystems, self-learning and intelligent systems, FPGAs, MPSoCs, and embedded systems. His research has a special focus on cross-layer analysis, modeling, design, and optimization of computing and memory systems, and their integration into use-cases from IoT and CPS. Dr. Shafique received the prestigious 2015 ACM/SIGDA Outstanding New Faculty Award, six gold medals in educational career, and several best paper awards and nominations at prestigious conferences like DATE, DAC, ICCAD and CODES+ISSS, Best Master Thesis Award, and Best Lecturer Award. He holds one US patent and over 150 papers in premier journals and conferences.

Rehan Hafiz is currently an associate professor at the Electrical Engineering department at the Information Technology University (ITU), Lahore, Pakistan. He received his Ph.D. degree in EE from the University of Manchester, UK in 2008. He is the director and founder of Vision Processing Lab (ViSpro) at ITU and earlier at NUST. His lab has a focus on the design of energy-efficient architectures and systems for computer vision, image processing, machine learning, and multi-projector immersive display technologies. He holds several patents in US, South Korea and Pakistan patent office.

Sumeen Rehman is an assistant professor at the Vienna University of Technology (TU Wien), Austria, where she is heading the research group on “Dependable and Energy-Efficient Computing” at the Institute of Computer Technology (ICT), Faculty for Electrical Engineering and Information Technology (EETI). Earlier, she was a post-doc researcher at the TU Dresden and Karlsruhe Institute of Technology (KIT) for 2 years. She received her Ph.D. in computer science from KIT, Germany, in 2015. Her current research interests include cross-layer reliability modeling and optimization covering various system layers like compiler and run-time system, approximate computing, embedded systems, Internet-of-Things, and Complexity Sciences. Dr. Rehman received the CODES + ISSS 2011 and 2015 Best Paper Awards, DATE 2017 Best Paper Award Nomination, and several HiPEAC Paper Awards.

Jörg Henkel is currently with Karlsruhe Institute of Technology (KIT), Germany, where he is directing the Chair for Embedded Systems CES. Before, he was a Senior Research Staff Member at NEC Laboratories in Princeton, NJ. He received his PhD from Braunschweig University with “Summa cum Laude”. His research interests are in design and architectures for energy efficient and reliable on-chip systems. Prof. Henkel received several best paper awards and nominations from conferences like CODES+ISSS, DATE, DAC, and ICCAD. He has been a General Chair of various conferences in embedded systems and CAD including ESWEEK and ICCAD. He is the Chairman of the IEEE Computer Society, Germany Section, and was the Editor-in-Chief of the ACM TECs for two consecutive terms. He is an initiator and the coordinator of the German Research Foundation’s (DFG) program on “Dependable Embedded Systems” (SPP 1500). He is the site coordinator (Karlsruhe site) of the Three-University Collaborative Research Center on “Invasive Computing” (DFG TR89). He is the Editor-in-Chief of the IEEE Design & Test Magazine since January 2016. He holds ten US patent and is a Fellow of the IEEE.

http://www.comsoc.org/~mmc/
IEEE COMSOC MMTC Communications - Frontiers

MMTC OFFICERS (Term 2016 — 2018)

CHAIR
Shiwen Mao
Auburn University
USA

STEERING COMMITTEE CHAIR
Zhu Li
University of Missouri
USA

VICE CHAIRS
Sanjeev Mehrotra (North America)
Microsoft
USA

Fen Hou (Asia)
University of Macau
China

Christian Timmerer (Europe)
Alpen-Adria-Universität Klagenfurt
Austria

Honggang Wang (Letters&Member Communications)
UMass Dartmouth
USA

SECRETARY
Wanqing Li
University of Wollongong
Australia

STANDARDS LIAISON
Liang Zhou
Nanjing Univ. of Posts & Telecommunications
China

MMTC Communication-Frontier BOARD MEMBERS (Term 2016—2018)

Guosen Yue Director Huawei R&D USA USA
Danda Rawat Co-Director Howard University USA USA
Hantao Liu Co-Director Cardiff University UK
Dalei Wu Co-Director University of Tennessee at Chattanooga USA
Zheng Chang Editor University of Jyvaskyla Finland
Lei Chen Editor Georgia Southern University USA
Tasos Dagdiklas Editor London South Bank University UK
Melike Erol-Kantarci Editor Clarkson University USA
Kejie Lu Editor University of Puerto Rico at Mayaguez Puerto Rico
Nathalie Mitton Editor Inria Lille-Nord Europe France
Shaoen Wu Editor Ball State University USA
Kan Zheng Editor Beijing University of Posts & Telecommunications China