Applied Mathematical Sciences, Vol. 9, 2015, no. 140, 6985 - 6995
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ams.2015.510655

Modified Intra Prediction Unit Size Selection Algorithm for H.265/HEVC Compression Systems

Oleg G. Ponomarev[1,2], Maxim P. Sharabayko[3], Dmitry Yu. Tyo[3] and Sergei E. Strelnikov[1]

[1]Tomsk State University of Control Systems and Radioelectronics
40 Lenin Prospekt, 634050 Tomsk, Russia
[2]Tomsk State University, 36 Lenin Prospekt, 634050 Tomsk, Russia
[3]Tomsk Polytechnic University, 30 Lenin Prospekt, 634050 Tomsk, Russia

Abstract

With the increased computational complexity of H.265/HEVC video compression fast decision on prediction unit size is essential for real-time coding applications. In this paper we provide an overview of existing intra prediction unit size decision algorithms and present our modification of one of the selection algorithms with improved compression performance.

Mathematics Subject Classification: 94A08

Keywords: intra prediction unit size selection, video coding decisions, H.265/HEVC

1 Introduction

The latest international video compression standard H.265/HEVC utilizes block-based hybrid video coding techniques just like its predecessors H.264/AVC, H.263, MPEG-2, etc. A set of non-overlapping pixels blocks divides each frame of a video sequence into smaller regions, for which spatial (intra) and temporal (inter) prediction is applied. Spatial prediction extrapolates pixels values

within a block by the adjacent pixels values of neighboring blocks. Temporal prediction utilizes pixels values from some region of the previously coded video frame. Subtraction of a predicted pixel values from the original values produces residual signal (residue). The residual signal is mapped into a frequency domain by two-dimensional discrete Fourier transform. Further quantization of transformed signal makes it possible to reduce the accuracy of representation of the transformed coefficients, thus providing an opportunity to loose less relevant data and reduce information redundancy. The transformed and quantized residue and supplemental information of compression algorithms used is a subject for entropy coding.

While the general compression data-flow is fixed by the H.265/HEVC standard, much more freedom is granted to the encoder in the process of prediction of block pixels values. Generally, the encoder is to decide how the frame is to be partitioned on blocks and how the pixels of a separate block are to be predicted. The increased number of available block partitionings and intra prediction modes increases computational complexity of HEVC compression process. The main computational bottleneck is described in Section 2 of this paper.

As the result of the increased encoder complexity there is a research field of developing fast unit size selection algorithms based on image features or some statistical information. In Section 3 we overview and compare different approaches to fast intra prediction unit size selection. Based on the comparison we choose the most perspective approach and in Section 4 we elaborate on its modification to further improve compression speed and then we provide experimental results on the obtained compression performance of our modified algorithm. The paper is concluded in Section 5.

2 Rate-Distortion Optimized Coding Decisions

The general compression data-flow of HEVC standard consists of partitioning of each frame of a video sequence on a series of coding tree units or CTUs. Each CTU covers a 64×64 pixels region of a video frame and is basically just a coding unit (CU). The adaptive quadtree partitioning of CTU, introduced in HEVC, makes it possible to optionally partition $2N \times 2N$ pixels CU on four $N \times N$ sub-CUs. The partitioning can continue until the minimum CU size is reached thus forming a quadtree-like partitioning structure. Each CU, produced as a result of a partitioning, can be coded using a plenty of coding options. Basically, each coding option corresponds to the prediction option used.

For a video encoder to properly select partitioning for a CTU and a coding option for each CU, rate-distortion optimization is involved. The technique is based on estimation of the compression rate (R) and the compression distortion

(D). Block compression options are compared with each other through the Lagrangian RDO cost [1]:

$$RD_{cost} = D + \lambda R, \tag{1}$$

where λ is the Lagrangian multiplier. Its value is usually determined empirically like [2, 3, 4]:

$$\lambda = 0.85 \cdot 2^{(QP-12)/3},$$

where QP is the quantization parameter used in compression data-flow. The RDO estimates the trade-off between compression rate and compression distortion thus providing a criterion of choosing the best coding option.

Each calculation of RD_{cost} equals to real compression of a block even in terms of computational complexity as soon as rate R estimation involves an entropy coder. Often the entropy coder in RDO is emulated [5, 6] as soon as the bits themselves are not required: only their number is the subject of rate estimation. Still the more there are coding options, the more RD_{cost} calculations are involved, the higher computational complexity is. Given the opportunity to skip RD_{cost} calculation for block partitions that are unlikely to become the optimal one, high computational load may be reduced.

3 Review of the Approaches

There are a lot of research papers, presenting the approaches to select intra prediction unit size faster. In this section we review several [7, 8, 9, 10, 11] that we find the most perspective.

One way to reduce intra PU search is to introduce some threshold for RD_{cost} for early split termination. In [7] the authors show that compression rate has a linear dependency upon a number of coding units that should have been split, but were not, while compression time has much higher impact. The authors propose to keep 5% false split termination hit rate and determined threshold empirically based on the size N of the PU and the QP value:

$$T_{64} = 962.7 \cdot e^{0.126 \cdot QP},$$

$$T_{32} = 164.6 \cdot e^{0.148 \cdot QP},$$

$$T_{16} = 19.75 \cdot e^{0.187 \cdot QP},$$

$$T_{8} = 1.054 \cdot e^{0.254 \cdot QP},$$

where T_{64}, T_{32}, T_{16}, T_8 are the thresholds for the corresponding block sizes $N = \{64, 32, 16, 8\}$.

Similar approach is offered in [8], where authors replace RD_{cost} calculation with the simplified LRD_{cost} (Low Complexity Rate-Distortion Cost). The threshold values T are chosen empirically. Once $LRD_{cost} < T$, further splitting does not happen, otherwise LRD_{cost} is calculated for sub-blocks like:

$$LRD_{cost} = \sum_{i,j} |A_{i,j}| + \lambda_m \cdot B_m,$$

where $A_{i,j}$ is the result of discrete Hadamard transform of the residual signal after prediction; B_m – number of bits to describe intra prediction mode in the bitstream, λ_m – Lagrangian multiplier. The threshold value T is determined by the PU size N and the QP value.

In [9] the authors noticed high spatial correlation of CU partitioning depth. They estimate partitioning depth by d_p equal to weighted sum of partitioning depths of previously coded blocks:

$$d_p = \sum_{i=0}^{3} \alpha_i \cdot d_i,$$

where d_p is an estimation of partitioning depth for current CU, d_i - partitioning depths of four neighbor CUs, α_i – weighting coefficients: $\alpha_0 = 0.3$, $\alpha_1 = 0.2$, $\alpha_2 = 0.3$, $\alpha_3 = 0.2$. The corresponding numbering of the neighbor blocks is illustrated on Fig. 1. Depending on the d_p value the CU being estimated is attributed to one of four groups each having its own range of partitioning depths shown in Table 1. Encoder chooses a certain partitioning depth within a given range depending on RD_{cost} values.

Figure 1: Positioning and numbering of neighboring CUs

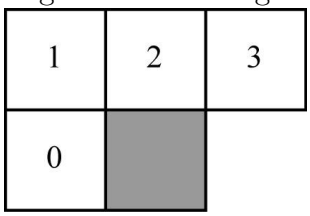

Table 1: Range of partitioning depths depending on sum of previous partitioning depth d_p

d_p	$d_p \leqslant 0.5$	$0.5 < d_p \leqslant 1.5$	$1.5 < d_p \leqslant 2.5$	$2.5 < d_p$
Partitioning depth	0–1	0–2	1–3	2–3

In [10, 11] the authors use various measures of homogeneity as a criteria to choose partitioning size for intra PU. For example, in [11] the authors loop

through partitioning upwards from bottom depth: four neighboring blocks of minimal size $N = 4$ are combined into one block of $N = 8$ once at least three of those blocks are considered homogeneous. The same rule is applied for larger blocks. The mean absolute values of four directional derivatives g_0, g_{90}, g_{45} and g_{135} are used to estimate image homogeneity:

$$g_0 = \sum_{k=0}^{\frac{N}{4}-1} \sum_{l=0}^{\frac{N}{4}-1} \sum_{i=0}^{1} \sum_{j=0}^{1} \left| I_{4k+i,4l+2j} - I_{4k+i+2,4l+2j} \right|,$$

$$g_{90} = \sum_{k=0}^{\frac{N}{4}-1} \sum_{l=0}^{\frac{N}{4}-1} \sum_{i=0}^{1} \sum_{j=0}^{1} \left| I_{4k+2i,4l+j} - I_{4k+2i,4l+j+2} \right|,$$

$$g_{45} = \sum_{k=0}^{\frac{N}{4}-1} \sum_{l=0}^{\frac{N}{4}-1} \sum_{i=0}^{1} \sum_{j=0}^{1} \left| I_{4k+i,4l+j} - I_{4k+i+2,4l+j+2} \right|,$$

$$g_{135} = \sum_{k=0}^{\frac{N}{4}-1} \sum_{l=0}^{\frac{N}{4}-1} \sum_{i=0}^{1} \sum_{j=0}^{1} \left| I_{4k+i+2,4l+j} - I_{4k+i,4l+j+2} \right|,$$

where $I_{x,y}$ is the intensity of a pixel in the position (x, y) within a coding block; N is the size of the coding block. The homogeneity measure of a block is

$$C = \left| g_{min} - g_{ort} \right|,$$

where $g_{min} = min\left\{g_0,\ g_{45},\ g_{90},\ g_{135}\right\}$, and g_{ort} is the value of the derivative in a orthogonal direction to g_{min}. The block is considered homogeneous when $C < T$, while the threshold value T is determined by block size N and quantization parameter QP:

$$T = QP \cdot N.$$

In [10] the authors suggest to use eight homogeneity measurements, each called either a local or a global complexity of an image. The four measurements of global complexities within a separate direction are calculated the following way:

$$G_0 = \sum_{i=0}^{N-1} \sum_{j=0}^{\frac{N}{2}-1} \left| I_{i,j} - \bar{I} \right| - \sum_{i=0}^{N-1} \sum_{j=\frac{N}{2}}^{N-1} \left| I_{i,j} - \bar{I} \right|,$$

$$G_{90} = \sum_{i=0}^{\frac{N}{2}-1} \sum_{j=0}^{N-1} \left| I_{i,j} - \bar{I} \right| - \sum_{i=\frac{N}{2}}^{N-1} \sum_{j=0}^{N-1} \left| I_{i,j} - \bar{I} \right|,$$

$$G_{45} = \sum_{i=0}^{N-1} \sum_{j=0}^{i} \left| I_{i,j} - \bar{I} \right| - \sum_{i=0}^{N-1} \sum_{j=i}^{N-1} \left| I_{i,j} - \bar{I} \right|,$$

$$G_{45} = \sum_{i=0}^{N-1} \sum_{j=0}^{N-i-1} \left| I_{i,j} - \bar{I} \right| - \sum_{i=0}^{N-1} \sum_{j=N-i-1}^{N-1} \left| I_{i,j} - \bar{I} \right|,$$

where $I_{x,y}$ is the intensity of the pixel of a coded block; N - size of a coded block; $\bar{I}$ - the average intensity of pixels within a coded block. Local complexities for the same directions are calculated as:

$$L_{ang} = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \left| I_{i,j}^{ang} - \bar{I}^{ang} \right|,$$

where $ang = 0,\ 90,\ 45,\ 135$; $I_{i,j}^{ang}$ - the value of derivative of pixel intensities in the direction given by the value of ang: $I_{i,j}^{0} = I_{i-1,j} - I_{i+1,j}$, $I_{i,j}^{90} = I_{i,j-1} - I_{i,j+1}$, $I_{i,j}^{45} = I_{i-1,j-1} - I_{i+1,j+1}$, $I_{i,j}^{135} = I_{i+1,j-1} - I_{i-1,j+1}$; and $\bar{I}^{ang}$ is the average value of a derivative of pixels intensities in a given direction ang.

There are three possible partitioning decisions based on the values of local and global complexities:

1. block is not to be partitioned (the further search is terminated);
2. block is to be partitioned (no further search on current partitioning depth is performed);
3. no decision on block partitioning can be made (split decision is based on RD_{cost}).

The first case takes place when global and local complexities in any direction are less than a predefined threshold for current CU and all its subblocks. The second case takes place when all the complexities are higher than the threshold values. The third condition takes place when none of the above conditions are true. The value of the threshold for local complexities was found empirically and conforms to all the directions and quantization parameters:

$$T_{loc} = 5120 \cdot \left(\frac{3}{4}\right)^{d},$$

where $d = 0,\ 1,\ 2,\ 3,\ 4$ is the current partitioning depth, that corresponds to block sizes $N = 64,\ 32,\ 16,\ 8,\ 4$. The threshold value for global complexity depends on the size of a CU and on the quantization parameter QP:

$$T_{glb} = C(QP) \cdot \left(\frac{3}{4}\right)^{d},$$

where $C(22) = 448$, $C(27) = 704$, $C(32) = 832$, $C(37) = 1216$, and for the rest values of QP the values for $C(QP)$ are found upon interpolation.

In [12] the homogeneity of an image is measured upon the entropy of the quantized values of pixels intensities. Quantization step is equal to 8, thus the quantized values are located within a range from 0 to 31 for 8-bit images. Shannon entropy is calculated the following way:

$$H = -\sum_{i=0}^{31} p_i \cdot log_2 p_i,$$

where $p_i = \frac{n_i}{N^2}$ is a relative frequency of value i among the quantized pixels values, n_i is the number of quantized values that equal to i, and N is the size of coding unit.

The work proposes three conditions to make partitioning decisions:

1. if the entropy of the quantized values of pixel intensities within a block $H < 1.2$, then block is not partitioned;
2. if $H > 3.5$, the block is partitioned into four subblocks;
3. if $H_{avg} - 0.15 < H < H_{avg} + 0.15$, where H_{avg} is the average entropy values for subblocks of all possible sizes, then the block is not split.

In case none of the conditions is true the CU is partitioned in four subCUs.

The results presented in [7, 8, 9, 10, 11, 12] that describe the efficiency of the proposed approaches are summarized in Table 2. The first column highlights the first author and references corresponding paper. The second column provides results in terms of BD-Rate [13]. The BD-Rate value shows the average bitrate overhead with respect to the reference data set that is usually the default reference compression algorithm. The third column shows average values of $\Delta T = \frac{T_{ref} - T}{T_{ref}} \cdot 100\%$, where T and T_{ref} are compression times for the modified and reference algorithms correspondingly.

Table 2: The efficiency of the proposed algorithms

Approach	BD-Rate, %	ΔT, %
Kim [7]	0.8	23.8
Cho [8]	2.0	55.8
Shen [9]	1.7	21.1
Yongfei Zhang [11]	4.8	56.7
Min [10]	0.8	52.3
Mengmeng Zhang [12]	3.7	62.0

The algorithm is considered the best if it provides the least BD-Rate overhead with the highest ΔT time savings. The least BD-Rate overhead among the ones described have Kim [7] and Min [10]. The first one provides only 23.8% time savings, while the second one provides 52.3% time savings. Therefore, the Min algorithm is used for further improvements.

4 Modified Intra Prediction Unit Size Selection Criteria

Let us describe the approach presented in [10]. After the computation of the global and local complexities the split decision is made within several stages. First, image complexity calculated within each of the four directions is compared with the threshold value. If there is at least one direction having image complexity below the threshold, the block is not partitioned. Otherwise there is the second stage where the minimum values for the local and global complexities within each direction are compared to the threshold values. If the minimum local or minimum global complexity is above the threshold value, the block is partitioned. The third stage comes otherwise where the split decision is made in a conventional way by the minimal RD_{cost} value.

The obvious way to modify the algorithm is to utilize additional criteria on the third stage to increase time savings. The proposed modification of Min algorithm is to utilize the criteria suggested in [7] on the third stage of decision making algorithm. This criteria is applyed only to those CUs that passed the first two stages of the Min algorithm. Obviously the CUs that passed the forst two stages are to have specific features, that is why the threshold values presented in [7] are to be updated. In our experiments on the JCT-VC test video sequences set [14] we estimated a dependency between Kim threshold values and the QP value in our modification of intra prediction unit size selection algorithm. The dependency is proved to be exponential. However, the dependency has parameters different to [7]. We obtained the minimum bitrate overhead on the test sequences with the following threshold values:

$$T_{64} = 1265.00 \cdot e^{0.086 \cdot QP},$$

$$T_{32} = 179.50 \cdot e^{0.1329 \cdot QP},$$

$$T_{16} = 26.41 \cdot e^{0.1678 \cdot QP},$$

$$T_{8} = 1.054 \cdot e^{0.254 \cdot QP},$$

The results of the proposed modification of intra prediction unit size selection algorithm are provided in Table 3. The experiments are held on the JCT-VC test video sequences with respect to the common test conditions [14]. The test video sequences set is divided into six classes A–F. The sequences within each class share similar or close video resolution and similar nature. The results are obtained on the HM reference encoder v. 14.0 [15]. The BD-Rate values characterize the average bitrate overhead of the modified HM encoder with respect to the original compression algorithm, while ΔT describes the average time savings provided by the proposed algorithm. As seen from the

Table 3, on average the proposed algorithm provides 55.90% time savings at the cost of only 2.36% bitrate overhead.

Table 3: Compression efficiency loss of the proposed rate estimation algorithm

Class	Sequence	Resolution	BD-RATE, %	ΔT, %
A	Traffic	2560×1600	1.71	54.51
	PeopleOnStreet		1.75	52.29
	Nebuta		7.05	41.18
	SteamLocomotive		8.55	50.41
B	Kimono	1920×1080	6.57	70.21
	ParkScene		1.41	57.32
	Cactus		1.64	58.37
	BQTerrace		1.02	52.95
	BasketballDrive		2.20	68.42
C	RaceHorses (C)	832×480	1.45	50.35
	BQMall		2.10	48.03
	PartyScene		0.26	40.96
	BasketballDrill		1.89	56.17
D	RaceHorses (D)	416×240	0.82	44.60
	BQSquare		0.40	46.65
	BlowingBubbles		0.73	44.05
	BasketballPass		1.01	53.12
E	Vidyo1	1280×720	2.30	69.08
	Vidyo3		5.59	63.33
	Vidyo4		2.57	69.08
F	BaskeballDrillText	832×480	1.49	54.51
	ChinaSpeed	1024×768	1.45	60.32
	SlideEditing	1280×720	0.45	53.85
	SlideShow		2.31	81.75
Average			**2.36**	**55.90**

5 Conclusion

In this paper we proposed new modification of a known fast algorithm for intra prediction unit size selection for H.265/HEVC compression. Based on the review of the existing papers we highlighted the approach of [10] with the most prominent results and observed a possible way to further improve this algorithm introducing additional PU size decision criteria, presented in [7]. The proposed algorithm was implemented in the HM reference software. Exper-

imental results show that the approach provides performance characteristics superior to the original approach.

Acknowledgements. The results were obtained at Tomsk State University of Control Systems and Radioelectronics as part of the complex project 'Provision of multimedia broadcasting services in Internet public networks, based on peer-to-peer network technology and adaptive data streaming' with the financial support of the Ministry of Education and Science of the Russian Federation.

References

[1] T. Berger, *Rate Distortion Theory: Mathematical Basis for Data Compression, (Prentice-Hall series in information and system sciences)*, Prentice-Hall: Endlewood Cliffs, New Jersey, 1971.

[2] T. Stockhammer, D. Kontopodis, T. Wiegand, Rate-distortion optimization for JVT/H.26L video coding in packet loss environment, *In International Packet Video Workshop on Packet Loss Environment*, Munich University of Technology, Munich, Germany, 2002, 1-12.

[3] G.J. Sullivan, T. Wiegand, Rate-distortion optimization for video compression, *IEEE Signal Processing Magazine,* **15** (1998), 74-90. http://dx.doi.org/10.1109/79.733497

[4] T. Wiegand, B. Girod, Lagrange multiplier selection in hybrid video coder control, *Proceedings 2001 International Conference on Image Processing*, IEEE: Thessaloniki, Greece, **3** (2001), 542-545. http://dx.doi.org/10.1109/icip.2001.958171

[5] F. Bossen, CE1: Table-based bit estimation for CABAC, *Document of ITU-T Q.6/SG16 JCTVC-G763*, ITU-T, Geneva, CH, 2011.

[6] M.P. Sharabayko, O.G. Ponomarev Fast rate estimation for RDO mode decision in HEVC, *Entropy*, **16** (2014), no. 12, 6667-6685. http://dx.doi.org/10.3390/e16126667

[7] J. Kim, Y. Choe, Y.-G. Kim, Fast coding unit size decision algorithm for intra coding in HEVC, *2013 IEEE International Conference on Consumer Electronics (ICCE)*, (2013), 637-638. http://dx.doi.org/10.1109/icce.2013.6487050

[8] S. Cho, M. Kim, Fast CU splitting and pruning for suboptimal CU partitioning in HEVC intra coding, *IEEE Transactions on*

Circuits and Systems for Video Technology, **23** (2013), 1555-1564. http://dx.doi.org/10.1109/tcsvt.2013.2249017

[9] L. Shen, Z. Zhang, P. An, Fast CU size decision and mode decision algorithm for HEVC intra coding, *IEEE Transactions on Consumer Electronics*, **59** (2013), no. 1, 207-213. http://dx.doi.org/10.1109/tce.2013.6490261

[10] B. Min, R.C.C. Cheung, A fast CU size decision algorithm for the HEVC intra encoder, *IEEE Transactions on Circuits and Systems for Video Technology*, **25** (2015), no. 5, 892-896. http://dx.doi.org/10.1109/tcsvt.2014.2363739

[11] Y. Zhang, Z. Li, Bo Li, Gradient-based fast decision for intra prediction in HEVC, *Visual Communications and Image Processing (VCIP)*, (2012), 1-6. http://dx.doi.org/10.1109/vcip.2012.6410739

[12] M. Zhang, J. Qu, H. Bai, Entropy-based fast largest coding unit partition algorithm in high-efficiency video coding, *Entropy*, **15** (2013), no. 6, 2277-2287. http://dx.doi.org/10.3390/e15062277

[13] G. Bjontegaard, Improvements of the BD-PSNR model, Input Document to ITU-T Q.6/SG16 JCTVC-G323, ITU-T: Berling, Germany, 2008.

[14] F. Bossen, Common test conditions and software reference configurations. In *Document of ITU-T Q.6/SG16 JCTVC-K1100*, ITU-T: Shanghai, CN, 2012.

[15] HEVC software repository, `https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/`

Received: November 1, 2015; Published: December 3, 2015