Complexity Control of HEVC Based on Region-of-Interest Attention Model

Xin Deng*, Mai Xu , Shengxi Li , Zulin Wang

School of Electronic and Information Engineering, Beihang University
Xueyuan Road, Beijing, 100191, China
**Corresponding to: cindydeng@ee.buaa.edu.cn*

***Abstract*—In this paper, we present a novel complexity control method of HEVC to adjust its encoding complexity. First, a region-of-interest (ROI) attention model is established, which defines different weights for various regions according to their importance. Then, the complexity control algorithm is proposed with a distortion-complexity optimization model, to determine the maximum depth of the largest coding units (LCUs) according to their weights. We can reduce the encoding complexity to a given target level at the cost of little distortion loss. Finally, the experimental results show that the encoding complexity can drop to a pre-defined target complexity as low as 20% with bias less than 7%. Meanwhile, our method is verified to preserve the quality of ROI better than another state-of-the-art approach.**

***Index Terms*—Complexity control, region of interest, HEVC.**

I. Introduction

Recently, high-resolution videos and large-sized screens are flooding into lives of humans, which brings about perfect visual enjoyment but at the same time huge challenge on bandwidth of communication channels. It is labored for the once-booming H.264/AVC to complete this challenge with its available coding efficiency, thus motivating the birth of a new video coding standard, called High Efficiency Video Coding (HEVC). The draft of HEVC was issued in January, 2013 and has gone through more than ten modified versions to constantly improve its coding performance.

Compared with H.264/AVC, HEVC can save bit rates by 50%, but at the cost of higher complexity [1]. However, most of multimedia-ready devices, such as portable computers and smartphones, do not have the ability to sustain such massive complexity due to limited communication technology and computational power. To solve this problem, many methods [2][3][4][5][6][7][8] have been proposed to decrease the encoding complexity of HEVC. Leng *et al.* [4] proposed an early coding unit (CU) depth determination approach in frame level. The basic idea of this approach is to skip some specific depth which is rarely used in the precious frame for CUs in current frame to save encoding time. Shen *et al.* [5] developed a fast CU size decision method based on the Bayesian decision rule. It employs important and computational-friendly features to help making a precise and fast selection on CU size by minimizing the Bayesian risk. Corrêa *et al.* [6] reduced the complexity of rate distortion optimization (RDO) process using a maximum CU depth pre-specified scheme. It divided all frames into two categories: unconstrained frames (*Fu*) and constrained frames (*Fc*). The *Fu* frames are encoded normally in which the RDO process needs to test all possible CU depths for optimization. In contrast, for *Fc* frames, the maximum depth of CUs is constrained on the basis of the history CU depths in *Fu* frames. In addition, some early prediction unit (PU) mode decision methods were used in [7] [8] to speed up the PU mode selection by reducing the optional prediction modes. However, all these methods do not take the subjective factors into consideration when performing the complexity control, thus to some extent jeopardizing video quality.

In this paper, we propose a HEVC complexity control method based on an ROI attention model. In such an attention model, each pixel is endowed with one weight according to its visual importance. The maximum depth of largest coding unit (LCU) is then determined by a proposed depth decision algorithm, on the basis of its average weight and the target complexity. For the LCU with smaller weight, its maximum depth tends to be set smaller, thus simplifying the RDO process to reduce the encoding complexity. The maximum depth of LCU grows along with its importance, thus preserving the quality of regions with higher importance well.

This paper is organized as follows. In Section II, we give an overview of the quadtree-based coding tree unit (CTU) block partitioning structure in HEVC. Section III introduces the ROI attention model and details our method for complexity control. Then, the experimental results are shown in Section IV and Section V concludes this paper.

II. CTU block partitioning structure in HEVC

The CTU block partitioning scheme is regarded as one of the most significant contributions to the coding efficiency of HEVC. Compared with the fixed 16×16 macroblocks in H.264/AVC, the flexible size of CTU can range from 64×64 down to 8×8. Furthermore, each CTU can be split into smaller CUs according to a quadtree structure [1]. Owing to the flexibility of CU sizes, HEVC allows the encoder to match different characteristics of either high or low resolutions video content. Especially for the large flat regions in high-resolution videos, larger CUs can represent the regions with smaller amounts of bits, thus saving bit rates tremendously. However, the searching for final CU sizes is quite a complex work, which

978-1-4799-6139-9/14/$31.00

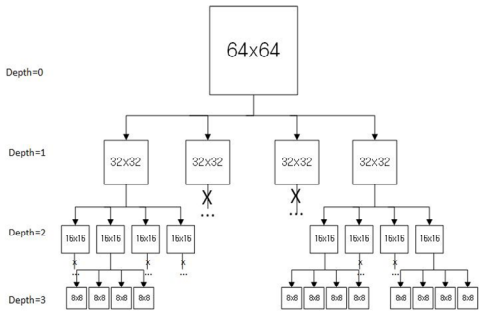

Fig. 1. An example of CTU partitioning structure.

is responsible for most of the encoding complexity. We show an example of CTU partitioning structure in Figure 1. HEVC adopts the RDO technique to decide whether current CU is split or not. When the rate-distortion (RD) cost of root CU is larger than that of sum of its leaf node CUs, the splitting of root CU is implemented. Otherwise, the root CU is not allowed to be split. The RDO process needs to compare all possible CU sizes allowed by the encoder and finally chooses the best one. Assuming that C_n is the total encoding complexity when the maximum depth of LCUs is n, then we have [9]

$$C_n = \sum_{i=0}^{n} 4^i \times \Phi_i, \tag{1}$$

where i stands for the actual depth and Φ_i is the complexity of CU at i-th depth. According to (1), C_n increases sharply as n grows. In other words, the maximum depth of LCU plays a dominant role in the encoding complexity.

III. Proposed Method

A. ROI attention model

The ROI attention model is built on the basis of human visual system (HVS) [10]. According to HVS, when a person watches a generic video, he/she may not pay attention to the whole scene, but only a small region around a point of fixation, called ROI region [11]. In other words, different regions have hierarchical importance for HVS.

For the establishment of ROI attention model, regions with distinct significance need to be split apart from each other. Then, they are sorted according to their relative importance. Regions with highest importance are assigned with greatest weight. Take conversational videos for example, facial regions are much more important than background. Among the facial regions, facial features (e.g., eyes, mouth, and nose) seem more significant than others, bringing about greater importance. Therefore, the facial features should have the greatest weight, followed by the facial regions and background. The detailed information of the weight allocation can be seen in [12]. For other non-conversational videos, the build-up process is analogous, except for the extraction of ROI regions. Therefore, we just discuss the conversational videos in the rest of paper due to the limited space.

B. Complexity control algorithm

The main objective of encoding complexity control is to optimize distortion D of compressed videos given a target complexity T_c,

$$\min(D) \quad \text{s.t.} \quad \frac{1}{J}\sum_{j=1}^{J} T_j = T_c, \tag{2}$$

where T_j is the encoding complexity of j-th LCU, and J is the total number of LCUs in each frame. Since HEVC encodes through CTU partitioning scheme, we thus regard the complexity sum of encoding all LCUs as the total complexity.

Complexity estimation. Upon the distortion-complexity model (2), the foremost problem is how to decrease T_j to make the sum of T_j equivalent to T_c. As we have mentioned in Section II, the maximum LCU depth has a significant effect on the encoding complexity. This provides some cues for us to manipulate the complexity. T_j can be reduced by means of decreasing the maximum LCU depth.

For accurate complexity control, we need to explore the relationship between T_j and LCU maximum depth d_j (from 0 to 3). Similar to [6], we calculate the average encoding complexity at various d_j, by doing statistical analysis on encoding time. Specially, the training sequences we used are standard conferencing videos *KristenAndSara* and *Vidyo3* from Class E. The encoding time is recorded using a 64-bit and 3.4 GHz processor on HM 14.0 platform. Assuming that T_j is normalized to 1 when d_j is 3, the results of T_j for d_j=0, 1, and 2 are 0.18, 0.37, and 0.68.

After exploring the relationship between T_j and d_j, next is to reasonably adjust d_j, to make the sum of T_j equal to T_c. Then, the complexity control can be reduced to LCU maximum depth decision algorithm. First, every pixel is endowed with a weight based on ROI attention model. Then, before compressing j-th LCU, its average weight w_j is computed by averaging the weights of its inclusive pixels. There are three thresholds of w_j: λ_1, λ_2, and λ_3, employed to define d_j. Note that λ_1, λ_2, and λ_3 are not fixed values, since they depend on target complexity T_c. Given a settled T_c, Figure 2 shows an example of the relationship between $\boldsymbol{\lambda}$=(λ_1, λ_2, λ_3), w_j, and d_j. When $w_j \geq \lambda_3$, d_j is set to be 3. When $\lambda_2 \leq w_j < \lambda_3$, d_j is defined to 2. If $\lambda_1 \leq w_j < \lambda_2$, d_j is 1. Otherwise, d_j is equal to 0. Since T_j is determined by d_j which depends on $\boldsymbol{\lambda}$ and w_j, $T_j(\boldsymbol{\lambda}, w_j)$ can be further defined according to aforementioned statistical analysis results of T_j and d_j,

$$T_j(\boldsymbol{\lambda}, w_j) = \begin{cases} 1 & \text{if } w_j \geq \lambda_3 \\ 0.68 & \text{if } \lambda_2 \leq w_j < \lambda_3 \\ 0.37 & \text{if } \lambda_1 \leq w_j < \lambda_2 \\ 0.18 & \text{if } w_j < \lambda_1. \end{cases} \tag{3}$$

Distortion Optimization. The distortion of each LCU is related to its maximum depth. Assuming that the distortion of LCUs at the same depth are the same, we use D_i to represent the distortion of LCU with maximum depth being i. The whole distortion D can be defined as,

$$D = \sum_{J_0} D_0 + \sum_{J_1} D_1 + \sum_{J_2} D_2 + \sum_{J_3} D_3, \tag{4}$$

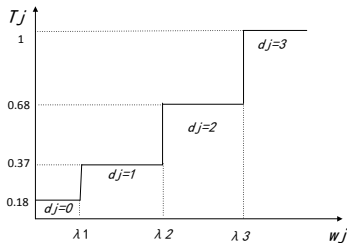

Fig. 2. An example of the correlation of T_j, w_j, $\boldsymbol{\lambda}$ and d_j.

where J_i means the number of LCUs with maximum depth being i, and $J_0 + J_1 + J_2 + J_3 = J$. Then, combining (3) and (4), the distortion-complexity model (2) turns to

$$\min \left(\sum_{J_0} D_0 + \sum_{J_1} D_1 + \sum_{J_2} D_2 + \sum_{J_3} D_3 \right) \quad \text{s.t.} \quad \frac{1}{J} \sum_{j=1}^{J} T_j(\boldsymbol{\lambda}, w_j) = T_c. \tag{5}$$

We aim to reduce the encoding complexity to a pre-defined level with minimal video distortion. In (4), we assume that $D_3 \ll D_2 \ll D_1 \ll D_0$, since larger depth usually leads to better video quality with less distortion. Therefore, for minimal distortion, we just assign LCUs with two largest depth, according to T_c. Take T_c=0.6 for example, since $0.68 \leq T_c < 1$, we only set d_j= 2 and 3 to LCUs, without d_j= 0 and 1, guaranteeing less distortion. Thus, D in (5) can be minimized to $D(\boldsymbol{\lambda}, w_j)$ on the basis of T_c,

$$D(\boldsymbol{\lambda}, w_j) = \begin{cases} \sum_{J_3} D_3 + \sum_{J_2} D_2, & J_3 = \{j | w_j \geq \lambda_3\}, \quad 0.68 \leq T_c < 1 \\ \sum_{J_2} D_2 + \sum_{J_1} D_1, & J_2 = \{j | w_j \geq \lambda_2\}, \quad 0.37 \leq T_c < 0.68 \\ \sum_{J_1} D_1 + \sum_{J_0} D_0, & J_1 = \{j | w_j \geq \lambda_1\}, \quad 0.18 \leq T_c < 0.37 \\ \sum_{J_0} D_0, & J_0 = \{j | w_j < \lambda_1\}, \quad T_c < 0.18. \end{cases} \tag{6}$$

Note that for each condition in (6), the sum of J_i is always J. For example, when $T_c \geq 0.68$, J_3+J_2=J. The specific value of J_i in each condition is determined by $\boldsymbol{\lambda}$, given a T_c. We can see from (6) that LCUs with higher w_j are assigned with larger d_j for smaller distortion, thus guaranteeing that the distortion reduction of ROI regions is less than that of non-ROI regions. Therefore, subjective video quality can be optimized.

Solution. According to (6), $\boldsymbol{\lambda}$=$(\lambda_1, \lambda_2, \lambda_3)$ should be correspondingly set to,

$$\boldsymbol{\lambda} = \begin{cases} (0, 0, \lambda_3) & \text{if } T_c \geq 0.68 \\ (0, \lambda_2, \mathrm{W}) & \text{if } 0.37 \leq T_c < 0.68 \\ (\lambda_1, \mathrm{W}, \mathrm{W}) & \text{if } 0.18 \leq T_c < 0.37 \\ (\mathrm{W}, \mathrm{W}, \mathrm{W}) & \text{if } T_c < 0.18, \end{cases} \tag{7}$$

where W is the maximum value of w_j among each frame and may change from frame to frame. Finally, $\boldsymbol{\lambda}$ is calculated by solving

$$\frac{1}{J} \sum_{j=1}^{J} T_j(\boldsymbol{\lambda}, w_j) = T_c. \tag{8}$$

TABLE I
THE COMPLEXITY CONTROL APPROACH

- **Input:** The target complexity T_c.
- **Output:** $\boldsymbol{\lambda}$=$(\lambda_1, \lambda_2, \lambda_3)$.
- Initialize J to the total number of LCUs in each frame.
- Initialize θ_2, θ_1, and θ_0 to be 0.68, 0.37, and 0.18.
- **For:** $j < J$
  1. Calculate w_j by averaging the pixel weights in j-th LCU
  2. Obtain the maximum weight of w_j W=max$\{w_j\}_{j=0}^{J}$.
  3. **Case** T_c

  Case 1: $\theta_2 < T_c < 1$
  Set λ_1 and λ_2 to zero, and calculate λ_3 through

  $$\frac{1}{J} \sum_{j=1}^{J} T_j(\lambda_3, w_j) = T_c.$$

  Case 2: $\theta_1 < T_c \leq \theta_2$
  Set λ_1 to zero and λ_3 to W, calculate λ_2 through

  $$\frac{1}{J} \sum_{j=1}^{J} T_j(\lambda_2, w_j) = T_c.$$

  Case 3: $\theta_0 < T_c \leq \theta_1$
  Set λ_2 and λ_3 to W, calculate λ_1 through

  $$\frac{1}{J} \sum_{j=1}^{J} T_j(\lambda_1, w_j) = T_c.$$

  Case 4: $T_c \leq \theta_0$
  Set all λ_1, λ_2 and λ_3 to W.

  End
- Save the $\boldsymbol{\lambda}$=$(\lambda_1, \lambda_2, \lambda_3)$ and use it to determine the maximum depth of LCUs when encoding.

The specific procedure of our algorithm is presented in Table I. It includes a classification of T_c using thresholds θ_2, θ_1, and θ_0 being 0.68, 0.37, and 0.18, respectively. After the classification of T_c, $\boldsymbol{\lambda}$ is set and calculated according to (7) and (8). For example, when T_c is larger than 0.68, λ_1 and λ_2 are specified to be zero. We only need to calculate λ_3 by solving (8). The same holds for other cases.

IV. EXPERIMENTAL RESULTS

In this section, experiments were performed on three video sequences to validate the effectiveness of the proposed approach. Here, the method in [6] executed on HM 14.0 software is regarded as the reference approach.

A. Test video sequences and parameter selections

We carried out the experiments on another two standard conferencing videos from Class E: *Johnny* and *Vidyo4*. The proposed method was executed on HM 14.0 software. The typical parameter settings of HM 14.0 in our experiments are tabulated in Table II.

B. Complexity Control Performance

In this subsection, we evaluate the performance of our complexity control method. The actual running complexity

TABLE II
THE CONFIGURATION PARAMETERS OF HM 14.0

Frames to be encoded	240
Frame rate	60
GOP structure	IPPP
LCU size	64×64
Maximum CTU depth	3
SAO	1
FEN	1
Initial QP	32
Intra period	-1

TABLE III
R_c, BIT RATES, AND Y-PSNRS FOR EACH TARGET COMPLEXITY

Video Sequence	Tc[%]	Rc[%]	Y-PSNR [dB]	Bit rate [Kbps]
Johnny	100	100	41.66	1002.8
	80	84	41.65	1002.9
	60	64	41.63	1003.2
	40	44	41.56	1003.4
	20	24	41.26	1003.2
Vidyo4	100	100	40.61	1002.4
	80	76	40.58	1002.3
	60	67	40.56	1002.2
	40	45	40.39	1002.3
	20	24	**40.01**	1002.4

(R_c) is measured by the encoding time reduction as follows,

$$R_c[\%] = \frac{\text{Enc.Time(Proposed)}}{\text{Enc.Time(Original)}}. \quad (9)$$

Table III presents the results of R_c, Y-PSNRs, and bit rates under five target complexity level. From Table III, we can observe that the proposed method can control the complexity accurately for the two test videos. For the same video, bit rates are nearly unchanged for different complexity level. The Y-PSNRs are not affected so much, with reduction no more than 0.6 dB (*Vidyo4*, $Tc = 20\%$). The R_c keeps close to T_c, with the deviation no more than 7% (*Vidyo4*, $Tc = 60\%$), verifying the accuracy of our method.

Table IV compares the Y-PSNRs results between our and reference approaches [6]. For simplicity, $\mathrm{PSNR_a}$, $\mathrm{PSNR_f}$, and $\mathrm{PSNR_b}$ are used to represent Y-PSNRs of whole, ROI, and non-ROI regions. We can see that, for complexity level above 40%, the Y-PSNRs of our method are higher than those of the reference approach [6] under the same T_c, especially for the ROI regions. As presented in Table IV, our method can increase the Y-PSNRs of ROI regions up to 0.07 dB compared with the reference approach [6]. For target complexity level 20%, our $\mathrm{PSNR_a}$ is lower than that of the reference method. Because the actual running complexity of the reference method under 20% target is nearly 30%, while our running complexity is only 24%. Even under this condition, our method can still offer 0.03 dB Y-PSNR increasement in ROI regions, further verifying the effectiveness of our method.

V. CONCLUSIONS

In this paper, we have proposed an effective method for complexity control of HEVC. It employs an ROI attention model to assign different weights to LCUs according to their importance. The complexity control can be achieved by reasonably setting the maximum depth of each LCU in light of its weight. The experimental results show that the actual running complexity can follow the target complexity well, with bias no more than 7%, at the cost of Y-PSNRs reduction no more than 0.6 dB. Besides, our method can make the Y-PSNRs of ROI regions drop more slowly than the reference approach. The proposed method is useful for high-resolution video conferencing system to keep high video quality at low complexity, thus reducing the burden of encoding.

TABLE IV
Y-PSNRS COMPARISON (*Johnny*, 1 MBPS)

Tc[%]	Method	Rc[%]	$\mathrm{PSNR_a}$[dB]	$\mathrm{PSNR_f}$[dB]	$\mathrm{PSNR_b}$[dB]
100	-	100	41.66	39.28	41.76
80	Our	84	41.65	39.27	41.75
	Reference	78	41.64	39.25	41.74
	ΔPSNR	-	+0.01	+0.02	+0.01
60	Our	64	41.63	39.25	41.73
	Reference	65	41.56	39.18	41.67
	ΔPSNR	-	+0.07	**+0.07**	+0.06
40	Our	44	41.56	39.22	41.65
	Reference	47	41.53	39.17	41.63
	ΔPSNR	-	+0.03	+0.05	+0.02
20	Our	24	41.26	39.02	41.32
	Reference	29	41.31	38.99	41.38
	ΔPSNR	-	-0.05	+0.03	-0.06

VI. ACKNOWLEDGEMENT

This work was supported by the NSFC projects under Grants 61202139, and the China 973 program under Grant 2013CB329006, and MSRA visiting young faculty program.

REFERENCES

[1] G. J. Sullivan, J. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 22, no. 12, pp. 1649–1668, 2012.

[2] W. S. e. a. Ma S, Wang S, "Low complexity rate distortion optimization for HEVC," *Data Compression Conference (DCC)*, pp. 73–82, 2013.

[3] G. Correa, P. Assuncao, L. Agostini, and L. Cruz, "Coding tree depth estimation for complexity reduction of hevc," in *Data Compression Conference (DCC), 2013*. IEEE, 2013, pp. 43–52.

[4] J. Leng, L. Sun, T. Ikenaga, and S. Sakaida, "Content based hierarchical fast coding unit decision algorithm for HEVC," in *CMSP, 2011 International Conference on*, vol. 1. IEEE, 2011, pp. 56–59.

[5] X. Shen, L. Yu, and J. Chen, "Fast coding unit size selection for HEVC based on bayesian decision rule," in *Picture Coding Symposium (PCS), 2012*. IEEE, 2012, pp. 453–456.

[6] G. Corrêa, P. Assuncao, L. Agostini, and L. A. da Silva Cruz, "Complexity control of high efficiency video encoders for power-constrained devices," *Consumer Electronics, IEEE Transactions on*, vol. 57, no. 4, pp. 1866–1874, 2011.

[7] J. Kim, J. Yang, K. Won, and B. Jeon, "Early determination of mode decision for HEVC," in *Picture Coding Symposium (PCS), 2012*. IEEE, 2012, pp. 449–452.

[8] T. L. da Silva, L. V. Agostini, and L. A. da Silva Cruz, "Fast HEVC intra prediction mode decision based on edge direction information," in *EUSIPCO*. IEEE, 2012, pp. 1214–1218.

[9] K. Choi and E. S. Jang, "Fast coding unit decision method based on coding tree pruning for high efficiency video coding," *Optical Engineering*, vol. 51, no. 3, pp. 030 502–1, 2012.

[10] C. t. Blakemore and F. Campbell, "On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images," *The Journal of physiology*, vol. 203, no. 1, pp. 237–260, 1969.

[11] Z. Wang and A. C. Bovik, "Embedded foveation image coding," *Image Processing, IEEE Transactions on*, vol. 10, no. 10, pp. 1397–1410, 2001.

[12] M. Xu, X. Deng, S. Li, and Z. Wang, "Region-of-interest based conversational HEVC coding with hierarchical perception model of face," *IEEE JSTSP*, vol. 8, no. 3, 2014.