LETTER

Fast Coding Unit Size Decision Based on Probabilistic Graphical Model in High Efficiency Video Coding Inter Prediction

Xiantao JIANG[†a)], ***Nonmember***, **Tian SONG**[††b)], ***Member***, **Wen SHI**[††], **Takafumi KATAYAMA**[††], ***Nonmembers***, **Takashi SHIMAMOTO**[††], ***Member***, ***and*** **Lisheng WANG**[†], ***Nonmember***

SUMMARY In this work, a high efficiency coding unit (CU) size decision algorithm is proposed for high efficiency video coding (HEVC) inter coding. The CU splitting or non-splitting is modeled as a binary classification problem based on probability graphical model (PGM). This method incorporates two sub-methods: CU size termination decision and CU size skip decision. This method focuses on the trade-off between encoding efficiency and encoding complexity, and it has a good performance. Particularly in the high resolution application, simulation results demonstrate that the proposed algorithm can reduce encoding time by 53.62%–57.54%, while the increased BD-rate are only 1.27%–1.65%, compared to the HEVC software model.

key words: *HEVC, CU size, encoding complexity, probability graphical model*

1. Introduction

High efficiency video coding (HEVC) is the latest video coding standard that is recommended in 2013 [1]. It is a hybrid coding model, and it achieves about 50% bitrate saving compared with H.264, while the computational complexity has been increased significantly. The CU size is from 64×64 to 8 × 8, and CU depth is [0, 3]. The cost is the computational complexity of CU size decision in comparison to the sub-CUs. Moreover, it is hard to implement real-time processing in encoder side.

Some works pay attention to reducing the computational complexity [2]–[7]. Gweon et al. proposed an early CU termination algorithm to reduce computational [2]. If the current mode satisfies the skip condition, the current CU is terminated to compute the rate distortion (RD) costs of the remaining CUs. Yang et al. proposed an early CU SKIP algorithm to reduce the computing complexity [3]. If the early SKIP condition is satisfied, the current depth CU mode selection is skipped, and go to the next depth of the CU. Xiong et al. proposed two fast CU decision methods to speed up the inter encoding [4], [5], and these methods are based on the Markov Random Fields and pyramid motion divergence, respectively. However, the two methods can reduce about 50% encoding time, while the encoding efficiency decreased more than 2%. Shen et al. proposed a fast CU size pruning method based on Bayes rule [6]. Zhang et al. propose fast CU depth decision algorithm based on support vector machine (SVM) [7]. However, the implementation cost of these methods is unacceptable in hardware design.

Hence, this work has a good performance particularly in high-resolution applications. Furthermore, it pays attention to the trade-off between encoding efficiency and encoding complexity in HEVC encoder. Moreover, the implementation cost of the proposed method is acceptable.

2. Proposed CU Size Decision in Inter Prediction

In HEVC inter coding, motion estimation for each inter prediction block is done based on the minimize low-complexity RD-cost function. In residual coding, a coded block flag (CBF) is transmitted to indicate whether transform unit has non-zero transformed coefficients or not. Moreover, when the CBF of the current CU is zero, it means that the texture of current CU tends to be smooth, and the probability of CU non-splitting is more than CU splitting. When the CBF is one, it means that the texture of current CU tends to be complex, and the probability of CU splitting is more than CU non-splitting. In this work, the low-complexity RD-costs and CBF in the transform coding are imported to determine the CU size.

The CU splitting or non-splitting is formulated as classification problem w^C, which is defined as

$$\begin{cases} w^C \in w_n^C, & \text{CU non-splitting,} \\ w^C \in w_s^C, & \text{CU splitting.} \end{cases}$$

The set of feature vectors is computed for the classifier $f = [f_1, f_2, \ldots, f_n]$. The posteriori probability $p(w_n^C|f)$ and $p(w_s^C|f)$ are the probability of observing CU belonging to w_n^C and w_s^C given feature vectors f. Thus, if the given features are conditionally independent, the posterior probability of w^C can be calculated based on Naive Bayes (NB) theorem

$$p(w^C|f) = \frac{p(f|w^C)p(w^C)}{p(f)} = \frac{\left(\prod_{i=1}^{n} p(f_i|w^C)\right)p(w^C)}{p(f)} \tag{1}$$

Actually, in order to make a prediction, the maximum a posteriori (MAP) probability $h(w^C)$ can be written as

Manuscript received November 19, 2015.
Manuscript revised May 26, 2016.
Manuscript publicized August 8, 2016.

[†] The authors are with the Department of Electronics and Information Engineering, Tongji University, Shanghai, China.
[††] The authors are with the Department of Electrical and Electronics Engineering, The University of Tokushima, Tokushima-shi, 770–8506 Japan.
a) E-mail: xiantao@ee.tokushima-u.ac.jp
b) E-mail: tiansong@ee.tokushima-u.ac.jp

DOI: 10.1587/transinf.2015EDL8237

$$\begin{aligned} h(w^C) &= arg\max\ \ln p(w^C|f) \\ &= arg\max\ \ln p(f|w^C)p(w^C) \\ &= arg\max\ [\ln p(f|w^C) + \ln p(w^C)] \\ &= arg\max \left[\ln\left(\prod_{i=1}^{n} p(f_i|w^C)\right) + \ln p(w^C)\right] \end{aligned} \tag{2}$$

The CU size decision rule is follow:

$$\begin{cases} h(w_n^C) > h(w_s^C), & w^C \in w_n^C \\ h(w_n^C) < h(w_s^C), & w^C \in w_s^C \end{cases}$$

Where the features f_i are the coded block flag (CBF) and RD cost of partition $2N \times 2N$, denoted as f_1 and f_2. The prior probability function $p(f_1|w^C)$ is modeled using a discrete Bernoulli (ϕ) distribution. The prior probability function $p(f_2|w^C)$ obeys the Gaussian distribution [4]. Thus, the prior probability of $p(f_2|w^C)$ are defined as:

$$p(f_2|w_n^C) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left\{-\frac{(f_2-\mu_0)^2}{2\sigma_0^2}\right\} \tag{3}$$

$$p(f_2|w_s^C) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{-\frac{(f_2-\mu_1)^2}{2\sigma_1^2}\right\} \tag{4}$$

where the parameters (μ_0, σ_0), (μ_1, σ_1) are mean vectors and covariance of CU non-splitting and splitting, respectively.

In HEVC, there is correlation between current CU and neighborhood CU. In order to utilize the spatio-temporal correlation, a set of neighborhood system, Ω is defined as

$$\Omega = \{CU_1, CU_2, CU_3, CU_4, CU_5\} \tag{5}$$

As Fig. 1 shown, CU_1, CU_2, CU_3, CU_4 denote the spatially adjacent CUs of the current CU_0, and CU_5 denote the temporally adjacent CU of the current CU_0. Whereas, the prior distribution $p(w^C)$ can be confirmed by the probabilistic graphical model: Markov Random Fields (MRF) [8].

$$p(w^C) = \frac{1}{Z}\exp\left(-\sum_{j\in\Omega} V_j(X_j)\right) \tag{6}$$

From the physicists, this is the Gibbs distribution with interaction potential $\{V_j, j \in \Omega\}$, energy $U = \sum_j V_j$, and partition function of parameters Z. Configurations of lower energies are the more likely, whereas high energies correspond to low probabilities. The CU size decision is a binary classification problem, and the binary problem can be modeled by a simple ISING-MRF model [8]:

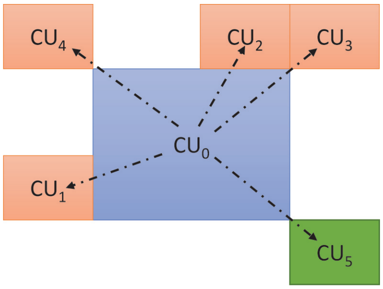

Fig. 1 The neighborhood system of the current CU.

$$V_j(X_j) = -\beta \times (X_0 \times X_j) \tag{7}$$

where the X_0 denotes the flag of the CU_0 splitting or non-splitting, X_j denotes the flag of the CU_j splitting or non-splitting in neighborhood system Ω, if $X_j = X_0$, $V_j(X_j) = 1$, else $V_j(X_j) = 0$. β is coupling factor, which indicates the strength of CU correlation with neighborhood system (in this work $\beta = 0.7$). So $p(w^C)$ exhibits a factorized form

$$p(w^C) \propto \exp\left(-\sum_{j\in\Omega} -\beta \times (X_0 \times X_j)\right) \tag{8}$$

Then when CU_0 neighborhood system Ω is valid, the prior distribution $h(w^C)$ can be rewritten as:

$$h(w^C) = arg\max \left[C_1 - \frac{1}{2\sigma_i^2}(f_2-\mu_i)^2 - \sum_{j\in\Omega} -\beta \times (X_0 \times X_j)\right] \tag{9}$$

where constant $C_1 = \ln p(f_1|w^C)$, $i = 0, 1$. However, when CU_0 neighborhood system Ω is invalid, the $p(w^C)$ can be confirmed by Bernoulli (ψ) distribution. $h(w^C)$ can be rewritten as:

$$h(w^C) = arg\max \left[C_1 - \frac{1}{2\sigma_i^2}(f_2-\mu_i)^2 + \ln p(w^C)\right] \tag{10}$$

Through the above analysis, CU size decision algorithm based on PGM model is proposed as algorithm 1, which includes CU termination decision and CU skip decision. CU termination decision rule is: CBF is zero, and

Algorithm 1: CU Size Decision Algorithm Based PGM

```
Perfrom CU inter prediction with motion estimation
The statistical parameters estimation with online learning
for $depth = 0$ to 3 do
    Calculate the probability $h(w_n^C)$ and $h(w_s^C)$
    if CBF is 0 then
        if $h(w_n^C) > h(w_s^C)$ then
            CU tremination process is made
        else
            CU splitting 4 sub-CUs, and $depth$++
    if CBF is 1 then
        if $h(w_n^C) < h(w_s^C)$ then
            CU skip process is made, and $depth$++
        else
            Check the other modes in current depth
Determine the best CU size
```

$h(w_n^C) > h(w_s^C)$. CU skip decision rule is: CBF is one, and $h(w_n^C) < h(w_s^C)$.

3. Statistical Parameters Update

Owing to the different video characteristics, therefore, the online learning method is used to estimate the statistical parameters: $\{(\mu_0, \sigma_0),\ (\mu_1, \sigma_1),\ \phi,\ \psi\}$. In each group of picture (GOP), the first frame is used for the training picture and the estimated parameters are stored in a lookup table (LUT), while the rest of the frames are coded with the fast CU size decision method.

4. Experiment Results

The proposed algorithm is implemented and verified based on HEVC test model HM12.0. The test conditions are set to evaluate the performance of the proposed algorithm at *lowdelay* (LD) and *randaccess* (RA) profiles, The quantization parameters (QP_i) are set to 22, 27, 32 and 37, respectively.

The performance of the proposed algorithm is evaluated Bjøntegaard Delta Bit Rate (BR) according to [9], and the average time saving (TS) is defined as

$$TS(\%) = \frac{1}{4}\sum_{i=1}^{i=4} \frac{T_{HM}(QP_i) - T_{pro}(QP_i)}{T_{HM}(QP_i)} \times 100\% \quad (11)$$

where $T_{HM}(QP_i)$ and $T_{pro}(QP_i)$ are the encoding time by using the HEVC reference software and the proposed method with different QP_i.

The performance of the fast CU size decision are shown as Table 1. From the table, we can see that the proposed method can reduce significantly the encoding time. The BR and TS are (1.11%, 48.24%), and (1.40%, 53.38%) in the LD and RA profiles.

Table 1 The results of proposed CU size decision.

		LD		RA	
Class	**Sequence**	**BR(%)**	**TS(%)**	**BR(%)**	**TS(%)**
2560 × 1600	**Traffic**	1.20	53.93	1.15	57.88
	SteamLocomotive	0.68	50.51	1.19	54.83
1920 × 1080	**Kimono**	1.75	42.01	2.51	47.04
	ParkScene	1.16	53.09	1.17	57.64
	Cactus	1.29	48.26	1.63	52.67
	BasketballDrive	2.33	46.91	3.74	51.13
	BQTerrace	0.67	56.44	1.10	59.14
1280 × 720	**Vidyo1**	1.06	63.05	1.41	66.54
	Vidyo3	1.50	60.50	1.16	63.57
	Vidyo4	1.09	61.63	1.44	64.95
High Res.	**Average**	**1.27**	**53.62**	**1.65**	**57.54**
832 × 480	**BasketballDrill**	1.01	40.30	0.99	48.27
	BQMall	0.98	46.25	1.22	52.42
	PartyScene	0.59	37.95	0.77	45.14
	RaceHorses	0.98	34.91	1.54	39.50
416 × 240	**BasketballPass**	0.76	48.85	0.88	53.68
	BQSquare	0.36	39.94	0.48	49.63
	BlowingBubbles	1.55	35.69	1.54	43.46
Low Res.	**Average**	**0.89**	**35.69**	**1.06**	**47.44**
Average		1.11	48.24	1.40	53.38

Compared with previous work, the results are shown as Table 2. My proposed method has good performance, particllarly in the high resolution (Res.), and the BR and TS are (1.27%, 53.6%), and (1.65%, 57.54%) in the LD and RA profile, while (2.41%, 62.59%) in Zhang's work, and (3.30%, 69.24%) in the Xiong's work. The benefit is from the adaptive online learning method. Moreover, the proposed method achieves a good tradeoff between encoding efficiency and encoding complexity. Furthermore, the implementation cost of the proposed method is more efficiency in hardware design.

Figure 2 shows a typical core architecture of the HEVC encoder with the proposed CU pruning model. The proposed CU pruning model can help the inter prediction engine to select optimal CU size before rate-distortion optimization (RDO) process. For each $2N \times 2N$ CU, a probability calculator is used to decide CU splitting and non-splitting by using RD-cost and CBF. This probability calculator is based on probability graphical model and Gaussian distribution of RD cost. It supports all $2N \times 2N$ CU size for 64×64 to 8×8.

This CU pruning model can reduce redundant CU size directly which leads to very low power cost. Different from a fixed threshold, this proposed method can achieve efficient tradeoff between encoding efficiency and encoding complexity. By reducing the redundant RDO iteration the working clock frequency can be decreased and low cost and low power hardware architecture can be easily achieved.

5. Conclusion

In this paper, a fast CU size decision algorithm is proposed based on PGM. The proposed algorithm consists of CU termination and CU skip methods to reduce the redundant computing. Finally, simulation results demonstrate that the proposed algorithm can reduce 53.62%–57.54% encoding time in the high-resolution application.

Table 2 Performance compared with previous work.

		(BR, TS)		
Config	**Method**	**High Res.**	**Low Res.**	**Average**
LD	**Proposed**	**(1.27, 53.62)**	**(0.89, 35.69)**	**(1.11, 48.24)**
	Xiong's [5]	(2.59, 44.40)	(1.83, 36.26)	(2.21, 40.33)
	Zhang's [7]	(2.41, 62.59)	(1.55, 40.31)	(1.98, 51.45)
RA	**Proposed**	**(1.65, 57.54)**	**(1.06, 47.44)**	**(1.40, 53.38)**
	Xiong's [4]	(3.30, 69.24)	(2.18, 57.10)	(2.74, 63.17)
	Shen's [6]	(1.25, 51.05)	(1.33, 38.61)	(1.35, 44.7)

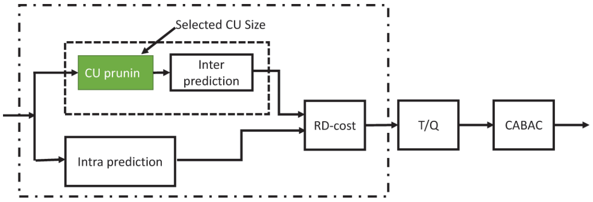

Fig. 2 Architecture design of HEVC encoder.

Acknowledgments

This research was supported by the National High Technology Research and Development Program of China (863 Program NO: 2013AA040302), and the Shanghai Economic and Information commission major technical equipment development project (NO: ZB-ZBYZ-03-12-1067-1). It also was supported by JSPS KAKENHI Grant Number 26330066.

References

[1] G.J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," IEEE Trans. Circuits Syst. Video Technol., vol.22, no.12, pp.1649–1668, 2012.

[2] R. Gweon, Y. Lee, and J. Lim, "Early termination of CU encoding to reduce HEVC complexity," JCTVC-F045, Torino, IT, July 2011.

[3] J. Yang, J. Kim, K. Won, H. Lee, and B. Jeon "Early skip detection for HEVC," JCTVC-G543, Geneva, CH, Nov. 2011.

[4] J. Xiong, H. Li, F. Meng, S. Zhu, Q. Wu, and B. Zeng, "MRF-based fast HEVC inter CU decision with the variance of absolution differences," IEEE Trans. Multimed., vol.16, no.8, pp.2141–2153, Dec. 2014.

[5] J. Xiong, H. Li, Q. Wu, and F. Meng, "A fast HEVC inter CU selection method based on pyramid motion divergence," IEEE Trans. Multimed., vol.16, no.2, pp.559–564, Feb. 2014.

[6] X. Shen, L. Yu, and J. Chen, "Fast coding unit size selection for HEVC based on Bayesian decision rule," Proc. Picture Coding Symp. (PCS), pp.453–456, May 2012.

[7] Y. Zhang, S. Kwong, X. Wang, H. Yuan, Z. Pan, and L. Xu, "Machine learning based coding unit depth decisions for flexible complexity allocation in high efficiency video coding," IEEE Trans. Image Process., vol.24, no.7, pp.2225–2238, July 2015.

[8] P. Patrick, Markov random fields and images, IRISA, 1998.

[9] G. Bjøntegaard, "Calculation of average PSNR differences between RD-curves," ITU-T SG16 Q.6, VCEG-M33, April 2001.