GPU Based Sample Adaptive Offset Parameter Decision and Perceptual Optimization for HEVC

Falei Luo*†, Shanshe Wang†, Nan Zhang§, Siwei Ma†‡, and Wen Gao†‡
Email: falei.luo@vipl.ict.ac.cn sswang@jdl.ac.cn zhangnan@ccum.edu.cn swma@pku.edu.cn wgao@pku.edu.cn
*Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
†Institute of Digital Media & Cooperative Medianet Innovation Center, Peking University, Beijing, China
‡Peking University Shenzhen Graduate School, Shenzhen, China
§School of Biomedical Engineering, Capital Medical University, Beijing, China

***Abstract*—In this paper, a graphics processing unit (GPU) based sample adaptive offset (SAO) parameters decision scheme is proposed for High Efficiency Video Coding (HEVC). Then, in order to further improve the performance of SAO, a perceptual based optimization scheme is provided according to the adjustment of Lagrange multiplier aiming to improve the subjective performance of SAO. Experimental results demonstrate that the proposed GPU based SAO parameter decision scheme can achieve average 0.76% and 0.78% BD-rate gain in terms of PSNR (Peak Signal to Noise Ratio) and SSIM (Structure Similarity) respectively. Combined with the perceptual optimization scheme, the maximum BD-rate gain in terms of PSNR and SSIM can be up to 1.77% and 3.3% with the average as 1.23% and 1.37%. Moreover, much computation complexity of SAO can be distributed to GPU.**

Keywords*—HEVC, SAO, GPU, Lagrange multiplier, perceptual optimization

I. Introduction

High Efficiency Video Coding (HEVC) significantly improves the coding efficiency compared to the previous coding standard H.264/AVC [1]. It is stated that in the same subjective quality level, half more bits can be saved [2]. Nonetheless, the traditional block based prediction and transform coding scheme is utilized, thus artifacts like blur and blocking artifacts still exist. In order to abate such artifacts, sample adaptive offset (SAO) was creatively proposed in the HEVC standardization process [3]. By applying independent offsets to samples in different categories, distortion can be evidently reduced and the coding performance can be significantly improved with SAO. However, the traditional implementation of SAO is based on objective rate distortion optimization (RDO). Since the final video receiver is human eyes, perceptual based RDO procedure can be expected to further improve the coding performance.

In the video coding optimization, perceptual based optimization is always a hot research topic for the optimization of video coding standard. In the early video coding standards, vision model has been considered [4]. However, human visual system (HVS) is so complicated that it cannot be represented accurately only by a simple model. In the actual video coding, a practicable scheme is to incorporate the perceptual distortion into the rate-distortion (R-D) cost for mode decision. In the previous researches, many works focused on the formulation of perceptual distortion. In [5], Structure Similarity (SSIM) is proposed to be an effective subjective video quality measurement. Based on SSIM, the perceptual distortion is described as (1-SSIM), and then perceptual RDO can be achieved [6][7]. In [8], Yeo et al. utilizes 1/SSIM as the perceptual distortion, and propose a perceptual optimization scheme based on the Lagrange multiplier (denoted as λ) for HEVC. In recent researches, the divisive normalization theory [6][9] is deeply investigated for the perceptual optimization. It is stated that it can reflect the visual characteristics of human eye at a certain extent [10]. In the image/video processing, the divisive normalization is widely utilized [11][12]. However, the calculation scheme for divisive normalized factor (DNF) is somewhat different, e.g. the local characteristic based scheme [12] and distribution model based scheme [10]. In [13], a Lagrange multiplier based perceptual optimization scheme is proposed to optimize HEVC coding, and much perceptual improvement is achieved.

However, the perceptual based RDO procedure would increase the computing complexity of HEVC encoder since additional computation is necessary, which may not be suitable for practical applications. In recent years, with the rapid development of the Graphic Processing Unit (GPU), many researchers tries to use General Purpose GPU (GPGPU) for video codec optimization [14][15]. It can be attributed to that modules like motion estimation is time consuming but can be conducted in parallel, which is suitable for GPU with strong computation ability for large scale concurrent computation.

In SAO parameter decision, the main aim is to determine a suitable offset for each coding tree unit (CTU) in order to improve the reconstruction quality. The final selected modes and offsets are mainly determined by the reconstruction and original picture, while other coding information like coding unit partition has little effect. It inspired us to implement the decision process on GPU platform to save the computation burden of CPU. In this paper, a parallel mechanism for SAO parameter decision on GPU is firstly designed based on the Compute Unified Device Architecture (CUDA)[16], so the computing complexity is reduced and time saving is achieved. As numbers of threads can be provided on GPU, a more accurate slice control for SAO is suitable to be conducted and the coding performance can be improved. Furthermore, a perceptual based optimization scheme which can be applied in parallel is proposed based on perceptual R-D cost calculation and more coding performance improvement is obtained.

The remaining of this paper is organized as follows. In Section II, a brief introduction to SAO and its parameter decision process in HEVC reference software is provided.

978-1-4799-5341-7/16/$31.00 ©2016 IEEE

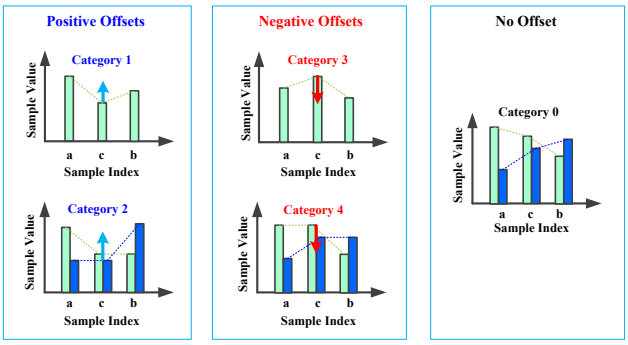


Fig. 1. Sample categories for EO in SAO

Then in Section III, the proposed implementation on GPU is detailed. And the experimental results are shown in Section IV. Eventually, Section V summarizes the paper.

II. Sample Adaptive Offset in HEVC

Two types of offset, edge offset (EO) and band offset (BO) are used in SAO of HEVC [3]. In EO, as shown in Fig. 1, samples are classified into five categories based on the variation tendency from one direction. In BO, the sample value range is divided into 32 bands evenly, and the category of one sample is determined by its value. Therefore, the computational complexity mainly focus on the judgment of SAO category and offset calculation.

In order to reduce the computational complexity, a fast parameter decision scheme is conducted in HEVC test model (HM) as follows [3]. Firstly the distortion increment between post SAO and pre SAO of samples with category i is represented as follows,

$$\Delta D = \sum_{i=1}^{4} \left(N_i {h_i}^2 - 2 h_i E_i \right), \tag{1}$$

where N_i is the number of samples in category i, h_i is an offset value, and E_i is the sum of difference between the original samples and the pre-SAO samples for a specific category. Then R-D cost increment ΔJ can be obtained by

$$\Delta J = \Delta D + \lambda \cdot R, \tag{2}$$

where λ is the Lagrange multiplier, and R is the estimated bits of coding the SAO parameters. Specially, when SAO is disabled for one coding tree unit (CTU), the distortion increment ΔD is zero and R-D cost increment is determined by the bits needed for coding the essential syntaxes. By comparing the R-D cost increment of each class and their offset values, the best SAO type and offset values are finally selected.

In order to improve the coding efficiency of SAO, HEVC adopts a switch to determine whether to execute SAO process for a slice. A fast slice level SAO decision algorithm is proposed [17][18] as follows. For a slice with none-zero depth N_D in a group of picture (GOP), a previous slice with depth $N_D - 1$ is set as the prediction slice. When more than 75% CTUs of the prediction slice disable SAO in luma or chroma component, SAO for corresponding component would be disabled for the current slice.

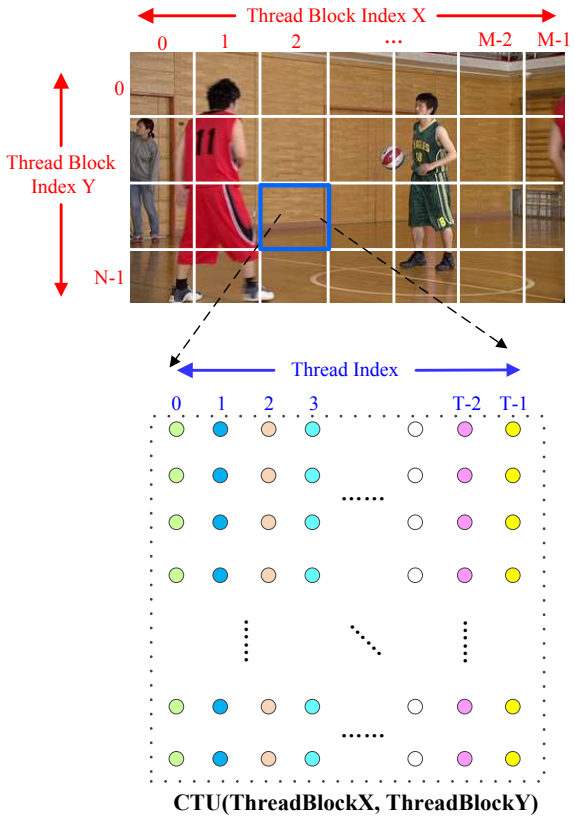


Fig. 2. Thread allocation for the SAO variables calculation

III. Proposed SAO Parameter Decision on GPU

In this section, the proposed GPU based SAO parameters determination scheme and the proposed perceptual optimization scheme are elaborately illustrated.

A. GPU Based SAO Parameter Decision

The SAO parameter decision includes two aspects, the CTU level decision and the slice level control decision. Both decisions are proposed to be implemented on GPU, thus time complexity reduction and coding efficiency improvement can be expected.

1) CTU Level SAO Parameter Decision: As mentioned in Section II, the decision of SAO parameters depends on many corresponding variables, i.e. N_i, h_i and E_i in (1). And the CTU level decision consists of two steps, i.e. the calculation of the above variables and the RDO process for SAO parameter decision of each CTU.

In the first step, the calculation of the above variables for all CTUs in one frame can be conducted at the same instant. In order to obtain more acceleration, much data is stored on the high speed shared memory on GPU, thus the number of threads in one thread block N_T is limited. It is mainly restricted as follows,

$$N_T \times S_T < S_{MAX}, \tag{3}$$

where S_T denotes the size of shared memory needed per thread, and S_{MAX} is the maximum size of shared memory for any thread block on a CUDA enabled GPU.

During the calculation of N_i and E_i for each CTU in EO class, 40 bytes of memory is needed for each thread since a sample may be classified into 5 categories and 8 bytes are needed to store the corresponding N_i and E_i pair. During the calculation of variables for BO class, 32×8 bytes are needed since a sample can be classed into any of the 32 bands. Therefore, at least 256 bytes are needed for each thread on GPU. While S_{MAX} is 48 KB for each thread block on CUDA devices with capacity 2.0, so N_T is limited to 192 for such devices. Due to such limitation, using one thread for one sample is not realistic while utilizing one thread for one column of samples is a better trade-off. As shown in Fig. 2, in the proposed implementation, each thread block is allocated for one CTU, and T threads are used for each thread block. To decrease latency caused by synchronization in one CTU, T and N_T are both set to 32, same as the warp size of GPU. Thus $(M \cdot N \cdot T)$ times speedup is expected in this step, where M and N are the number of CTUs in width and height.

In the next step, the table-based bit estimation for CABAC in HEVC is utilized for rate calculation [19]. Like the SAO decision process in HM, only SAO syntaxes (merge flags, types and offset values) are taken into account, thus the number of contexts on GPU can be cut down to 2. At the same time, best offset calculation for non-merge modes can be conducted in advance since they do not depend on the neighboring data and entropy coding status, which in turn brings in speedup for the RDO procedure. Furthermore, the parameter decision of CTUs in different CTU rows can be conducted at the same time on GPU like wavefront parallel processing (WPP) [2], and some times speedup can be expected.

2) Slice Level SAO Parameter Decision: Based on the sum of R-D cost increment for all CTUs, an improved scheme to determine whether to utilize SAO for a slice or not is proposed. Firstly, the R-D cost increment for enabling only luma or chroma and enabling both components are calculated respectively as follows,

$$\Delta J_{luma} = \sum_{i=1}^{MN} \Delta D_{luma}(i) + \lambda_{luma} R_{luma}(i), \quad (4)$$

$$\Delta J_{chroma} = \sum_{i=1}^{MN} \Delta D_{chroma}(i) + \lambda_{chroma} R_{chroma}(i), \quad (5)$$

$$\Delta J_{all} = \sum_{i=1}^{MN} \Delta D_{all}(i) + \lambda_{all} R_{all}(i), \quad (6)$$

$$\Delta J_{off} = 0, \quad (7)$$

where ΔJ_{luma} is the R-D cost increment of the slice when executing SAO just on luma component, and it is achieved from the SAO parameter decision resulting from the former SAO parameter decision process. ΔJ_{chroma} is the R-D cost increment of the slice when applying SAO just for chroma component. And ΔJ_{all} is the R-D cost increment when SAO for all components are enabled.

When the sum R-D cost increments as in (4), (5) and (6) are all generated, the minimum R-D cost increment of a slice is determined. If the minimum is negative, it is suitable to utilize SAO for the slice. Then corresponding achieved parameters should be encoded.

B. Lagrange Multiplier Based Perceptual Optimization

The shortcoming of traditional RDO as shown in (2) is that the distortion D cannot reflect the characteristics of HVS. However, it is intuitive that samples in some CTUs would have more impact on the subjective quality and others may have less. Thus, it is suitable to determine whether to turn on SAO and determine the offset values on a subjective perspective.

In [13], a perceptual based RDO for a CU is presented as,

$$J_p = D + \lambda_p \cdot R, \quad (8)$$

$$\lambda_p = \lambda \cdot f^2, \quad (9)$$

where J_p denotes the perceptual based rate distortion cost, f is relevant with the DNF. The detailed description of DNF can be referred to our previous work [13]. The scheme in (8) cannot be directly utilized for SAO since the processing unit of SAO in HEVC is CTU.

In this paper, based on the scheme as in (8), a perceptual based scheme for the decision of SAO parameters is proposed. The calculation of f for a CTU is improved as follows,

$$f = 1.4 - 0.8^{f'}, \quad (10)$$

$$f' = \frac{\sum_{i=1}^{l} F_i}{l \cdot E(F_i)}, \quad (11)$$

$$F_i = \sqrt{\frac{\sum_{k=1}^{N-1} \left(2 \cdot I_i^g(k)^2\right)}{N-1} + C}, \quad (12)$$

where f' is the DNF [6], denotes the Gaussian filtered coefficients and $E(F_i)$ means mathematic expectation of F_i. From (10), it can be seen that f is restricted in [0.6, 1.4]. Then the rate distortion increment can be represented as

$$\Delta J_p = \Delta D + \lambda_{p,CTU} \cdot R, \quad (13)$$

where ΔJ_p denotes perceptual based rate distortion increment for one SAO class and $\lambda_{p,CTU}$ is the perceptual optimized Lagrange multiplier for current CTU. When the minimum ΔJ_p of all the SAO types is negative, it is suitable to utilize SAO for the CTU.

IV. Experimental Results

To evaluate the coding performance of the proposed scheme, experiments are conducted on HEVC test model HM16.2. The experiment was conducted on a desktop with Intel i5 760 core in 2.80 GHz, 4 GB RAM and GTX 480 which has a CUDA capacity version 2.0 and clock rate 1.40 GHz. The operating system is Windows 7. During the test, low delay P configuration with asymmetric motion partition (AMP) off and 32 as largest coding unit size is adopted.

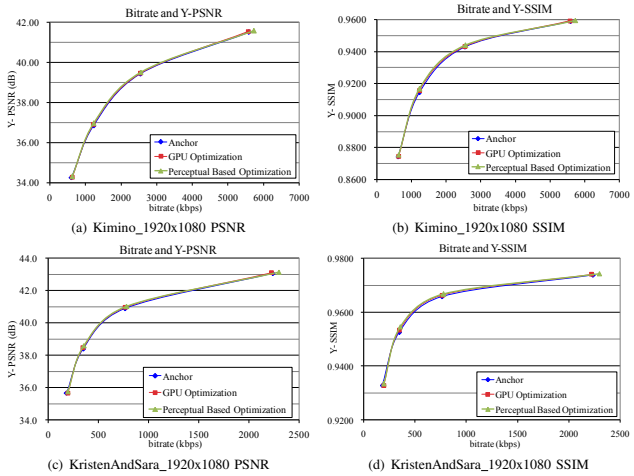
BD-rate [20] in terms of SSIM and PSNR are utilized to assess subjective and objective performance respectively. Speedup of SAO process is calculated as

$$Speedup = \frac{T_{anchor}}{T_{pro}}, \quad (14)$$

where T_{anchor} and T_{pro} respectively represent the average time consumption of SAO decision for one slice in HM16.2 and the

TABLE I. CODING PERFORMANCE OF THE PROPOSED SCHEME

Sequence	GPU Optimization			Perceptual based Optimization		
	BD-RATE PSNR	BD-RATE SSIM	*Speedup*	BD-RATE PSNR	BD-RATE SSIM	*Speedup*
Class A	-0.78%	-0.84%	2.56	-1.45%	-1.65%	1.80
Class B	-0.90%	-0.83%	2.47	-1.74%	-1.72%	1.79
Class C	-0.90%	-0.93%	2.16	-0.88%	-1.06%	1.54
Class D	-0.62%	-0.22%	1.13	-0.42%	0.36%	0.77
Class E	-0.54%	-1.08%	2.56	-1.77%	-3.32%	1.75
Average	-0.78%	-0.76%	2.18	-1.23%	-1.37%	1.53



(a) Kimino_1920x1080 PSNR (b) Kimino_1920x1080 SSIM

(c) KristenAndSara_1920x1080 PSNR (d) KristenAndSara_1920x1080 SSIM

Fig. 3. R-D curves comparison of the proposed scheme with HM

proposed implementation, including GPU execution time and overhead for data transferring between CPU and GPU.

The coding performance is tabulated in TABLE I. It can be observed that the proposed GPU implementation brings in 0.78% and 0.76% performance improvement on objective and subjective quality. This can be attributed to the traversal of all possible slice level switches. Although much speedup is expected, only 2.18 times speedup can be achieved on average, the computing capacity of a single GPU thread during RDO and data transferring cost can be the limitation.

With the proposed perceptual based scheme, more objective and subjective quality improvements can be achieved. As shown in TABLE I, the objective coding gain is 1.23% and the subjective coding gain in is 1.37% on average. Especially for sequences in Class E, over 3.3% subjective coding gain can be achieved, since more bits would be allocated in the regions of interests by the proposed scheme. The R-D curves of the proposed implementation are illustrated as shown in Fig. 3.

V. CONCLUSION

In this paper, a GPU based SAO parameter decision algorithm is proposed. Moreover, an optimization scheme is provided according to the divisive normalization to further improve the coding efficiency of SAO. Experimental results show that the proposed scheme can efficiently reduce the time complexity of SAO and significantly improve the coding performance.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China (61322106, 61571017), National Basic Research Program of China (973 Program, 2015CB351800), and Shenzhen Peacock Plan, which are gratefully acknowledged.

REFERENCES

[1] I. Draft, "Recommendation and final draft international standard of joint video specification (ITU-T Rec. H. 264— ISO/IEC 14496-10 AVC)," *Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, JVTG050*, vol. 33, 2003.

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

[3] C.-M. Fu, E. Alshina, A. Alshin, Y.-W. Huang, C.-Y. Chen, C.-Y. Tsai, C.-W. Hsu, S.-M. Lei, J.-H. Park, and W.-J. Han, "Sample adaptive offset in the HEVC standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1755–1764, 2012.

[4] I. Dalgiç and F. A. Tobagi, "A Constant Quality MPEG-1 Video Encoding Scheme and its Tra c Characterization," in *International Picture Coding Symposium*, 1996, pp. 105–110.

[5] Z. Wang, A. C. Bovik, H. R. Sheikh, S. Member, E. P. Simoncelli, and S. Member, "Image Quality Assessment : From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[6] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, "Perceptual video coding based on SSIM-inspired divisive normalization," *IEEE Transactions on Image Processing*, vol. 22, no. 4, pp. 1418–1429, 2013.

[7] Y. H. Huang, T. S. Ou, P. Y. Su, and H. H. Chen, "Perceptual rate-distortion optimization using structural similarity index as quality metric," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 11, pp. 1614–1624, 2010.

[8] C. Yeo, H. L. Tan, and Y. H. Tan, "SSIM-based adaptive quantization in HEVC," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 1690–1694.

[9] S. Wang, "Study on video coding techniques based on visual characteristics," Ph.D. dissertation, Peking University, 2013.

[10] D. J. Heeger, "Normalization of cell responses in cat striate cortex," in *Visual neuroscience*, vol. 9, no. 2, 1992, pp. 181–97.

[11] A. Rehman and R. Zhou, "Reduced-reference image quality assessment by structural similarity estimation." *IEEE transactions on image processing*, vol. 21, no. 8, pp. 3378–89, 2012.

[12] M. J. Wainwright and E. P. Simoncelli, "Scale mixtures of Gaussians and the statistics of natural images," *Advances in neural information processing systems*, vol. 12, no. 1, pp. 855–861, 2000.

[13] S. Wang, D. Zhao, and W. Gao, "Lagrange Multiplier Based Perceptual Optimization for High Efficiency Video Coding," in *Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)*, 2014, pp. 1–4.

[14] W. Xiao, B. Li, J. Xu, G. Shi, F. Wu, and G. Shi, "HEVC Encoding Optimization Using Multi-core CPUs and GPUs," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, pp. 1–14, 2015.

[15] F. Luo, S. Ma, J. Ma, H. Qi, L. Su, and W. Gao, "Multiple Layer Parallel Motion Estimation on GPU for High Efficiency Video Coding (HEVC)," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2015, pp. 1122–1125.

[16] N. Corporation, "Compute unified device architecture programming guide," 2007.

[17] G. Laroche, P. T., and O. P., "Non-CE1: Encoder modication for SAO interleaving mode," ITU-T/ISO/IEC Joint Collaborative Team on Video Coding (JCTVC) document JCTVC-I0184, April 2012.

[18] E. Alshina, A. A., and J. H. Park, "Encoder modication for SAO," ITU-T/ISO/IEC Joint Collaborative Team on Video Coding (JCTVC) document JCTVC-J0044, July 2012.

[19] F. Bossen, "CE1: Table-based bit estimation for CABAC," ITU-T/ISO/IEC Joint Collaborative Team on Video Coding (JCTVC) document JCTVC-G763, November 2011.

[20] G. Bjntegaard, "Calculation of average psnr differences between RD-Curves," VCEG-M33, April 2001.