Future Generation Computer Systems 78 (2018) 316–329

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

FIPIP: A novel fine-grained parallel partition based intra-frame prediction on heterogeneous many-core systems

Wenbin Jiang[a,*], Min Long[a], Laurence T. Yang[a,d], Xiaobai Liu[b], Hai Jin[a], Alan L. Yuille[c,e,f], Ye Chi[a]

[a] School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China
[b] Department Computer Science, College of Sciences, San Diego State University, San Diego, CA, United States
[c] Department of Statistics, University of California, Los Angeles, Los Angeles, CA 90095, United States
[d] Department of Computer Science, St. Francis Xavier University, Antigonish, Canada
[e] Department of Cognitive Science, Johns Hopkins University, United States
[f] Department of Computer Science, Johns Hopkins University, United States

HIGHLIGHTS

- A fine-grained parallelism for intra prediction based on GPU is proposed.
- It is the first to promote intra prediction to pixel-level parallelism based on GPU.
- A new regular prediction formula is presented for parallelism.
- Two optimized encoding orders are adopted for multi-levels parallelism.
- An efficient self-synchronizing method is presented for task scheduling.

ARTICLE INFO

Article history:
Received 2 December 2015
Received in revised form
27 March 2016
Accepted 5 May 2016
Available online 12 May 2016

Keywords:
Parallelism
Fine-grained partition
Intra-frame prediction
Fast mode decision
GPU
H.264/AVC

ABSTRACT

Intra-frame prediction is an important time-consuming component of the widely used H.264/AVC encoder. To speed up prediction, one promising direction is to introduce parallelism and there have been many heterogeneous many-core based approaches proposed. But most of these approaches are limited by their use of highly irregular prediction formulas, which require significant amount of branch instructions. They only use coarse-grained parallel partition, which considers blocks or sub-region of images as parallel processing units. In this paper, by contrast, we propose a *fine-grained intra-frame prediction approach based on parallel partition* (FIPIP) and implement it on *Graphics Processing Unit* (GPU) based heterogeneous many-core systems. The approach is characterized by the following aspects. First, our approach takes individual pixels as parallel processing units, instead of blocks. Imposing pixel-level parallelism is capable of fully exploiting the computational power of heterogeneous GPU-based systems and hence tremendously reduces the encoding time. Second, we unify irregular prediction formulas in intra-frame prediction into a well-designed uniform one, and propose a table-lookup method to efficiently perform intra-frame prediction. Our formula can eliminate unnecessary branch instructions by using a unified predictor array, which improves the efficiency of the fine-grained parallel partition significantly. Third, two optimized encoding orders assisted by an improved combined frame strategy are adopted to implement multi-level parallelism. Finally, an efficient self-synchronizing method is realized for fine-grained task scheduling on heterogeneous CPU–GPU architecture. We apply FIPIP to encode a set of benchmark videos under varying conditions and compare it with other popular intra-frame prediction methods. Results show that FIPIP outperforms existing state-of-the-art work with speedups factor of 2–6.

[*] Corresponding author.
E-mail address: wenbinjiang@hust.edu.cn (W. Jiang).

http://dx.doi.org/10.1016/j.future.2016.05.009
0167-739X/© 2016 Elsevier B.V. All rights reserved.

1. Introduction

H.264/MPEG-4 Part 10 or *Advanced Video Coding* (AVC) [1] is the most widely used standard for video compression. It was proposed by the ITU-T *Video Coding Experts Group* (VCEG) together with the ISO/IEC JTC1 *Moving Picture Experts Group* (MPEG). The encoder is characterized by the use of multiple references frames, improved motion estimation, and pixel-wise *intra-frame prediction* (intra prediction for short), etc. As the key component of H.264/AVC, intra prediction is one of most time-consuming steps in the whole pipeline of H.264/AVC. The software *Joint Model* (JM) [2] suggested by H.264 employs *rate-distortion optimization* [3] (RDO) to decide the optimal intra mode. Intra prediction with RDO can achieve good trade-off between encoding rate and compression distortion. As a result, the complexity and computation load increase drastically, because the encoder needs to calculate all prediction modes exhaustively to find the best one for a 4×4 pixels block.

In this paper, we investigate how to speed up intra prediction for encoding generic videos by using heterogeneous GPU-based many-core systems. In the literature, there have been many efforts in this direction. For example, Milani proposed to use spatial correlation between adjacent blocks to select a reduced set of prediction modes according to their probabilities, which were estimated adopting a belief-propagation procedure, and finally speed up the mode decision progress [4]. Pan et al. and Tsai et al. instead used the relationship between textures and prediction modes [5,6]. Other researchers implemented intra prediction on *Graphics Processing Units* (GPUs) [7,8] with the *Compute Unified Device Architecture* (CUDA) [9]. Diversified research works [7,9,10] have been done for coarse-grained intra prediction parallelism at frame-level, slice-level or block-level, which have obtained considerable speedups. However, the advantages of many-core resources such as GPUs are still not fully taken into account, because these methods usually employ a single thread to process one unit (such as a 4×4 block, a slice formed of several blocks, or even a frame). Note that a thread is the basic parallel unit of CUDA. All threads in the same grid execute the same kernel code simultaneously. To improve the parallelism degree further, a fine-grained parallel partition over pixel-level, instead of block-level, is potentially a good choice. This fine-grained parallelism, however, is likely to generate large amount of branches, and hence requires complex design. Other concerns include that several kinds of constraints over threads posed by the H.264 standard must be satisfied during encoding, which challenges the model of *Single Instruction Multiple Threads* (SIMT) of CUDA.

To address the aforementioned issues, we propose a *fine-grained intra-frame prediction approach based on parallel partition* (FIPIP) which aims to parallelize intra prediction at pixel-level on GPU-based many-core systems. To implement FIPIP, we present a unified formula to describe the prediction modes. We also propose a table-lookup based algorithm to efficiently eliminate the branches, which makes fine-grained parallelism much more efficient. Moreover, two encoding orders are presented to enable block-level parallelism to improve the parallelism degree further. Meanwhile, we introduce a combined frame strategy to sew some frames together to make a bigger frame for frame-level parallelism. Moreover, an efficient self-synchronizing method is realized for fine-grained task scheduling on heterogeneous CPU–GPU architecture.

By focusing on fine-grained parallelism, while cooperating with the other aforementioned methods, the proposed intra prediction method is capable of achieving high performance on heterogeneous GPU-based systems. The presented work is based on our previous work [11] and extends it by the following additions: (1) a 5-step sorting algorithm for the existing fast mode decision; (2) a selective encoding order neglecting intra modes 3 and 7; (3) more discussion about self-synchronizing task scheduling; (4) a more detailed evaluation on two more platforms, on frame combination, detailed speed-ups, and detailed PSNR and bit rate losses; (5) more introduction about the related work.

The rest of this paper is organized as follows. We first discuss the related work in Section 2, and introduce the background of intra prediction and CUDA architecture in Section 3. Section 4 analyzes the problems existing in multi-level parallelism methods. The proposed fine-grained parallel intra prediction is discussed in Section 5 in detail. The presented approach is evaluated in Section 6. Last, we conclude this paper and remark the future works in Section 7.

2. Related work

H.264 was approved as the video coding standard in 2003. Since then, lots of ways to accelerate the processing speed have emerged [7,12,13]. These solutions fall into two categories.

The first category reduces the computational overhead of the intra prediction, and is called fast mode decision. Pan et al. [5] built an edge map and a local edge direction histogram for each block by Sobel edge operators and chose the candidate modes according to their statistics. In a similar way, Tsai et al. [6] introduced an intensity gradient to select a subset of prediction modes. In [4], Milani S. proposed a method to determine the candidate modes based on their probabilities, by using belief-propagation. Researchers have also studied the transform domain as well as the pixel domain. Chen et al. [14] converted intra prediction from the pixel domain to the transform domain by matrix manipulation in order to obtain the transform domain predictions for various intra modes. Furthermore, [15] adopted the differences in the low frequency region to represent the total distortion of one 4×4 block. In [16], the authors even applied Support Vector Machines for Regression to intra prediction to improve the performance of H.264. Although the fast mode decision based solutions mentioned above have achieved good performance, they are limited to the serial computational process over CPU.

The second category for accelerating intra prediction is called parallel intra prediction. It is emerging in recent years and performs much faster than the ones with the fast mode decision mentioned above, as more and more parallelism-specific hardware devices (such as GPUs, stream processors and VLSI (*Very Large Scale Integration*) architectures) and related technologies spring up.

Fig. 2(d) shows a parallel intra prediction method by using an improved 10-step encoding order presented in [7,8]. It breaks through the restriction of a macroblock, which improves the GPU-based parallelism degree of the intra prediction. Fig. 2(c) depicts a 7-step order presented and used in [7,9]. These works promote the intra prediction to a higher parallelism degree by dividing a frame into a number of slices. Inside a macroblock, the 7-step order can improve the parallelism obviously. However, some prediction modes are reduced in this method, which decreases the quality of service to some extent. Ren et al. [10] presented a new real-time encoding method by applying stream processing model. Meanwhile, the 7-step encoding order is also used for high parallelism. It focuses on video sequences with 1080p resolution.

The aforementioned researches apply some revised encoding orders, which result in some loss of video quality. Moreover, the parallelism strategies used are still coarse-grained, which limits the speedup.

Aside from GPUs, some other many-core devices are also applied for the acceleration of the intra prediction. VLSI processors and stream processors are typical representatives. Wu et al. [12] implemented intra prediction on various stream processors by