摘 要

随着视频产业的快速发展，视频内容数量增长迅速，视频分辨率也越来越高。面对有限的存储容量和带宽，视频数据的高效压缩一直是学术界的重要研究课题。从上世纪八十年代以来，不断有国际编码标准推出，刷新编码性能的上限，包括广泛应用的MPEG-2、H.264/AVC 等。2013 年，高效视频编码标准 HEVC 作为新一代视频编码的国际标准发布，相比于 H.264/AVC 编码性能又提升了近一倍，这得益于大块划分结构、丰富的编码模式等新的编码技术。然而，在移动互联网时代，除了压缩性能，实时视频编码也成为了影响用户体验的主要因素，而 HEVC 提升压缩性能是以牺牲编码复杂度为代价的，这使得 HEVC 编码器离实时编码的市场需求还有一段距离。

近几年，随着 GPU 通用计算能力的提高，GPU 逐渐转变为协处理器，辅助 CPU 完成高性能并行计算。凭借低成本、低功耗的轻量处理器阵列，GPU 的整体并行计算能力优于 CPU。因此，基于 CPU-GPU 并行的计算框架广泛用于机器学习、数值分析、视频编码等数据密集型应用。

基于 HEVC 目前发展的瓶颈和 GPU 发展背景，本文从两个方向研究 HEVC 视频编码优化：编码本身的速度优化和并行编码器加速。在编码速度优化方面，本文分析HEVC 在编码复杂度上的优化空间，并提出高效的运动矢量优化算法。在编码器加速方面，研究基于 CPU-GPU 异构平台的 HEVC 编码器加速，并提出高效的 HEVC 帧间和帧内并行编码框架。

首先，针对编码器复杂度优化问题，本文从HEVC运动矢量编码入手，研究了HEVC运动矢量精度和编码复杂度与率失真性能的关系，提出了渐进的运动矢量精度算法。该算法首先引入了 1/8 像素运动矢量精度，然后将 1/4 像素和 1/8 像素运动矢量精度限制在围绕 MVP 的方形区域内。同时，基于运动矢量区域划分，提出了一种新颖的分段MVD 压缩算法，在不损失 MVD 信息量的情况下减小了 MVD 的幅值，进而节省了MVD 编码码率。实验结果表明在 HEVC 参考软件平台上，该算法能降低编码端的复杂度，同时获得非常显著的编码性能提升。

其次，本文提出了基于 CPU-GPU 并行的 HEVC 快速帧间预测。该算法主要针对插值和运动估计进行了 GPU 并行优化，并提出基于 GPU 返回结果的快速模式决策。具体来说，帧间的并行算法包括以下几个方面：1) 以灵活的 CTU 窗口单位的 CPU-GPU 流水线。这种方式的优点是使得 GPU 并行计算量可配置，便于移植。2) 基于时域相关性的 CTU 级 MVP 推导；3) 低复杂度的 GPU 端线程分配和快速可变块 SAD 合并。4) 基于 GPU 返回结果的快速帧间模式决策。实验结果表明，该算法在可接受的性能损失前提下，显著降低了 HEVC 编码复杂度。

最后，本文提出了基于 CPU-GPU 并行的 HEVC 快速帧内预测。该算法主要包括以下两个方面：1) GPU 并行帧内预测；2) 基于并行梯度计算的快速 CU 划分决策。对于 1)，帧内预测基于原始帧进行，去除了块依赖性。为避免帧内参考点获取的分支，该算法构建了参考点图像偏移表，以支持快速的参考点获取。此外，针对不同的块大小定制了低延迟的线程分配策略，最大限度降低了 GPU 计算延时。对于 2)，在编码一帧之前，利用 GPU 并行地评估整帧的纹理复杂度，并提出基于 GPU 返回纹理梯度的快速 CU 划分决策。实验结果表明本文提出的 CPU-GPU 帧内并行算法能显著降低 HEVC 帧内编码复杂度。

关键词：视频编码，HEVC，GPU，异构平台，运动矢量

Research on CPU-GPU Heterogeneous Platform Based HEVC Coding Optimization

Juncheng Ma (Computer Applied Technology)
Directed by Siwei Ma

ABSTRACT

With the fast development of video industry, the video content is growing rapidly and the video resolution is getting higher and higher. Video compression is always an important research topic due to the limited storage capacity and bandwidth. Since the eighties, many international video coding standard (including the widely used MPEG-2 and H.264/AVC) have been published, constantly refreshing the upper bound of coding performance. In year 2013, High Efficiency Video Coding (HEVC) standard was published as the new generation of video coding standard. It nearly double upgraded the coding performance compared to H.264/AVC, because of the new coding techniques, e.g. large block partition and more coding modes. However, in the mobile Internet era, real-time video coding is another main factor which affects the user experience except compression performance. For HEVC, the high coding performance is achieved with the expense of coding complexity increment. Therefore, there is still a long way to meet the market demand of real-time HEVC coding.

In recent years, with the improvement of general purpose GPU computing power, GPU gradually transformed into a coprocessor, assisting CPU to complete high-performance parallel computing. With low-cost, low-power and lightweight processor array, GPU is superior to CPU in parallel computing capability. Therefore, CPU-GPU based parallel computing framework is widely used in data-intensive applications such as machine learning, numerical analysis, video encoding and so on.

Based on the current bottlenecks of HEVC development, and GPU development background, this paper researches HEVC video encoding optimization from two aspects: coding complexity optimization and parallel encoder acceleration. In coding optimization, this paper analyses the space for HEVC encoding complexity optimization, and proposes an efficient HEVC performance optimization algorithm. In parallel encoder acceleration, based on the CPU-GPU heterogeneous platforms, an efficient inter and intra parallel HEVC coding framework is proposed.

At first, considering coding complexity optimization, we start from HEVC motion vector coding. Based on the correlation of HEVC motion vector complexity and rate-distortion performance, we proposed a progressive motion vector resolution algorithm. Firstly, 1/8 pixel motion vector accuracy is introduced in the algorithm. Then 1/4 pixel and 1/8 pixel pixel motion vector accuracy are limited in a square area around the MVP, respectively. Moreover, based on limited motion vector area, a piecewise MVD derivation method is applied to code the MVD efficiently. Without losing the amount of information, this method reduces the MVD amplitude, thus saving the MVD coding bit rate. Experimental results show that the proposed algorithm can reduce the encoding complexity, while achieving significant coding performance compared to the HEVC reference software.

Secondly, a CPU-GPU parallel based fast inter prediction algorithm for HEVC is proposed. The motion estimation and interpolation module of HEVC is accelerated using GPU, and the returned results are used for fast mode decision on CPU side. Specifically, the inter-parallel algorithms include the following aspects: 1) flexible CTU window based CPU-GPU pipeline. The advantage of this approach is to make the GPU parallel computing capacity configurable and easy to transplant. 2) temporal correlation based CTU-level MVP derivation; 3) low complexity GPU-side thread allocation and fast SAD generation for variable blocks. 4) GPU returned information assisted mode decision. Experimental results show that the algorithm can reduce HEVC coding complexity significantly with acceptable coding performance loss.

Finally, a CPU-GPU parallel based fast intra prediction algorithm for HEVC is proposed. The algorithm includes the following two aspects: 1) parallel GPU intra prediction; 2) fast CU partition decision based on parallel gradient calculation. For 1), intra prediction is processed on the original frame to get rid of block dependency. To avoid branch when fetching intra reference sample, a reference-sample offset table is constructed to support fast reference acquisition. In addition, for different block sizes, customized low latency thread allocation strategy are designed to minimize GPU computing delay. For 2), before encoding a frame, a parallel GPU texture complexity assess procedure is launched for the entire frame. Base on the GPU-returned texture gradients, we propose a fast CU partition decision method. Experimental results show that the proposed intra CPU-GPU parallel algorithms can significantly reduce HEVC intra coding complexity.

KEY WORDS: Video coding, HEVC, GPU, Heterogeneous platform, Motion vector