J Real-Time Image Proc
DOI 10.1007/s11554-013-0357-y

ORIGINAL RESEARCH PAPER

# Exploiting task and data parallelism for advanced video coding on hybrid CPU + GPU platforms

Svetislav Momcilovic · Nuno Roma · Leonel Sousa

Received: 9 July 2012 / Accepted: 14 May 2013

**Abstract** Considering the prevalent usage of multimedia applications on commodity computers equipped with both CPU and GPU devices, the possibility of simultaneously exploiting all parallelization capabilities of such hybrid platforms for high performance video encoding has been highly quested for. Accordingly, a method to concurrently implement the H.264/ advanced video coding (AVC) inter-loop on hybrid GPU + CPU platforms is proposed in this manuscript. This method comprises dynamic dependency aware task distribution methods and real-time computational load balancing over both the CPU and the GPU, according to an efficient dynamic performance modeling. With such optimal balance, the set of rather optimized parallel algorithms that were conceived for video coding on both the CPU and the GPU are dynamically instantiated in any of the existing processing devices, to minimize the overall encoding time. The proposed model does not only provide an efficient task scheduling and load balancing for H.264/AVC inter-loop, but it also does not introduce any significant computational burden to the time-limited video coding application. Furthermore, according to the presented set of experimental results, the proposed scheme has proved to provide speedup values as high as 2.5 when compared with highly optimized GPU-only encoding solutions or even other state of the art algorithm. Moreover, by simply using the existing computational resources that usually equip most commodity computers the proposed scheme is able to achieve inter-loop encoding rates as high as 40 fps at a HD 1920 × 1080 resolution.

**Keywords** Video coding · GPU · Hybrid CPU + GPU Platforms · Load balancing · CUDA

## 1 Introduction

The increasing demand for high-quality video communication and the tremendous growth of video contents on Internet and local storages stimulated the development of highly efficient compression methods over the past decades. When compared with previous standards, the H.264/ MPEG-4 advanced video coding (AVC) [1] achieves compression gains of about 50 %, keeping the same quality of the reconstructed video [2]. However, such compression efficiency comes at the cost of a dramatic increase of the involved computational requisites, making real-time video coding hard to be achieved even on the most recent single-core central processing units (CPUs).

The latest generations of commodity computers, equipped with both multi-core CPUs and many-core graphics processing units (GPUs), already offer high computing performances to execute a broad set of signal processing algorithms. In particular, the GPU architectures consist of hundreds of cores especially adapted to exploit fine-grained parallelism, and as such are frequently applied to implement complex signal processing applications. On the other hand, on the multi-core CPUs the data-parallelism can be exploited either at a coarse grained level, by concurrently running multiple threads on different cores, or at a fine-grained level, by using vector instructions. Hence, the simultaneous exploitation of all these different

S. Momcilovic (✉) · N. Roma · L. Sousa
INESC-ID IST-TU Lisbon, Rua Alves Redol, 9,
1000-029 Lisbon, Portugal
e-mail: Svetislav.Momcilovic@inesc-id.pt

N. Roma
e-mail: Nuno.Roma@inesc-id.pt

L. Sousa
e-mail: Leonel.Sousa@inesc-id.pt

Published online: 02 June 2013