Significance-Driven Adaptive Approximate Computing for Energy-Efficient Image Processing Applications

Special Session Paper: Extended Abstract

Dave Burke[†], Dainius Jenkus[†], Issa Qiqieh[†], Rishad Shafik[†], Shidhartha Das[‡] & Alex Yakovlev[†]
[†]Electrical and Electronic Engineering, Newcastle University, NE1 7RU, UK
[‡]ARM, 110 Fulbourn Rd, Cambridge CB1 9NJ, UK
D.Burke2@ncl.ac.uk, D.Jenkus1@ncl.ac.uk, I.Qiqieh1@ncl.ac.uk, Rishad.Shafik@ncl.ac.uk, Shidharhta.Das@arm.com, Alex.Yakovlev.ncl.ac.uk

1 OVERVIEW

With increasing resolutions the volume of data generated by image processing applications is escalating dramatically. When coupled with real-time performance requirements, reducing energy consumption for such a large volume of data is proving challenging.

In this paper, we propose a novel approach for image processing applications using significance-driven approximate computing. Core to our approach is the fundamental tenet that image data should be processed intelligently based on their informational value, i.e. significance. Using quantified definition of significance, for the first time, we show how the complexity of data processing tasks can be drastically reduced when computing decisions are synergistically adapted to significance learning principles. A variable-kernel convolution filter case study running on an Odroid XU-4 platform is demonstrated to evaluate the effectiveness of our approach, with up to 45% energy reduction for a given performance requirement.

2 INTRODUCTION & RATIONALE

Image processing applications, which include acquisition, processing and analysis of real-world digital images have two major challenges posed by their conflicting requirements of performance and energy efficiency. With continued advancement of camera and sensing technologies, there is a persistent demand for higher resolution of the captured frames (i.e. images) that require decoding at real-time [1]. As such, the volume of data to be processed over a given time is increasing rapidly.

Approximate computing has recently emerged as a promising approach, which leverages the intrinsic resilience of these applications to imprecision [2]. Existing approximate computing practices commonly fall into two categories: the first being application-specific hardware (HW) or software (SW) [3], which take advantage of low-complexity algorithms and/or HW tailored to the application needs, and the second being the design of application-independent HW systems [4, 5], which process data using generic computing resources at low complexity, both in favour of improved performance and reduced energy consumption. These are often coupled with system-level controls, such as dynamic voltage/frequency scaling (DVFS) [6, 7], HW/SW co-design [2] and/or mapping [3]. These techniques have no knowledge of the significance or the informational value of the data being processed.

Images typically consist of areas where the contrast between colors define the artefacts and features of the image more than those without any contrast [8]. We postulate that these informational values, i.e. significance, can be used to modulate the computation efforts with the aim of achieving energy minimization under quality and performance constraints. Based on this tenet, we make the following **contributions**:

1. For the first time we present a quantifiable definition of significance in the context of image processing applications,
2. We propose a parallel HW/SW resource (approximate or precise) allocation approach adapted to significance of the image blocks for optimized performance, energy and quality (PEQ) trade-offs, and
3. A GPU-based variable-kernel parallel convolution filter is used as a case study to validate the proposed approach.

Section 3 defines significance in image processing, which is then used for adaptive approximate computing and a validation case study in Sections 4 and 5.

Figure 1: Significance of image with different threshold levels.

3 SIGNIFICANCE IN IMAGE PROCESSING

We define significance in the context of an image, as areas where the deviation is significantly different to a local mean, such that, it exposes information features arising from changes in visual effects and perception. The research originated by investigating, by means of a software demonstrator, if significance in still images can be estimated through parallel inference of mean and standard deviation per image block. The calculation of mean and standard deviation was based on the the work of [9] and initially used integral images. Three methods were used to generate the image masks. Method 1 generates Standard deviation using Integral Images with `sum` and `square sum` matrices on 32x32 clusters, chosen to constrain the Integral mean computations to 16-bit integers, these are further sub-divided into smaller 4x4 blocks. Method 2 generates deviation by utilising the absolute difference between sample and mean, avoiding the use of the Integral images `square sum` matrix and subsequent square roots. Method 3 generates an Approximate Absolute deviation by direct computation of a single value from each of 4 adjacent 4x4 blocks. Figure 1 shows four images generated by the said demonstrator. The original image, Figure 1(a), was clustered in smaller 4x4 blocks with thresholded deviation mask applied to a gray scale of the original image (< threshold is black and non-significant, >= threshold is grey and significant). Variable number

CODES/ISSS '17 Companion, Seoul, Republic of Korea

of clusters per image can also be applied with PEQ trade-offs (not reported in this paper).

Figure 1(b) and (c) demonstrate that at low threshold levels, 3 and 20, the deviation figures for each block using the three methods don't show immediately discernible differences in the image masks. Figure 1(c) shows a zoomed-in red area of image (a). The top image shows little difference between Methods 1 and 2, the lower image shows the sparser image results of Method 3. Method 1 utilises compute-intensive operations, leading to up to 180 ms latency per 20Mpixel image. Method 2 uses the less intensive `abs()` function to generate absolute variance. This reduced the latency to ≈160 ms. Method 3 in Figure 1(d) used only four samples from each 4x4 block to compute absolute standard deviation using simplified summation, with only ≈ 6ms latency per image.

Varying these thresholds can generate optimistic (too few significant blocks) or pessimistic (too many significant blocks) outcomes. This will be used as a control knob for meeting specified quality requirements in our proposed approach (Section 4).

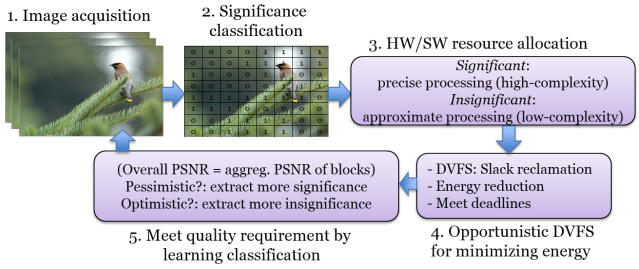

Figure 2: Proposed adaptive approximate computing approach

4 ADAPTIVE APPROXIMATE COMPUTING

Underpinning the low-cost evaluation of significance (Method 3, Figure 1), a significance-driven approximate computing approach is developed. The aim is to reduce energy consumption with given soft real-time and expected quality requirements. Figure 3 shows the proposed approach consisting of five key steps. After capturing the real-time images, they are clustered in predefined number of blocks, and significance is estimated using Method 3 (Section 3) in the second step using the default classification thresholds. In the third step, the significant blocks are then processed using precise algorithms or hardware blocks, while non-significant blocks are processed using approximate and low-complexity ones. Reduced complexity of algorithms or hardware processing generates opportunities for DVFS for slack reclamation in the fourth step based on the soft real-time deadlines. The impact of such adaptive processing is then estimated using objective image quality, such as peak signal-to-noise ratio (PSNR). Since the impact of approximation can be pre-characterized using worst-case PSNRs, the overall quality can be estimated as an aggregation of the block PSNRs. The removes the need for the developer to write extra software routines, which can be marked as a major advantage of our approach.

If the evaluated quality is well below or well above the requirement, current classification threshold is marked as pessimistic or optimistic. In either cases, suitable classification threshold is learnt through an iterative reinforcement learning (RL) algorithm to find relationship between the classification threshold applied and the expected quality requirement.

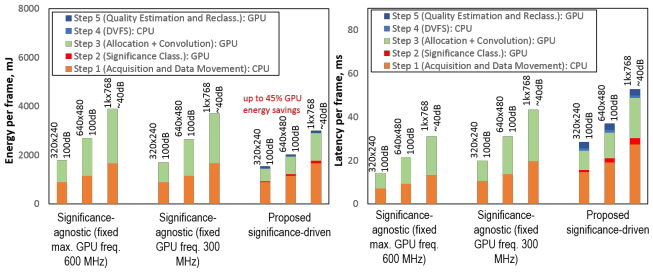

Figure 3: Comparative PEQ trade-offs

5 CASE STUDY & DISCUSSIONS

To evaluate the effectiveness of our proposed approach a real-time convolution filter is used, running on a heterogeneous Odroid XU-4 platform (where the CPU cores offload the convolution task to GPUs). The aim was to replace the existing significance-agnostic 5x5 kernel convolution filter kernel by a variable kernel filter that can use 3x3 kernel (approximate) for non-significant blocks and 5x5 kernel (precise) otherwise. The inference of significance and the allocation of kernels were adapted based on 20 frames per second (fps) with 40dB PSNR minimum quality requirement. The use of low-complexity kernel allowed for opportunistic DVFS to minimize energy through slack reclamation with a given real-time soft deadline.

Figure 3 shows the PEQ trade-offs for the given case study. Two key observations can be made. Firstly, as can be seen, the existing parallel convolution filter approaches (with max. GPU frequency or lower) cannot leverage dynamic allocation of approximate or precise resources based on significance. As such, it performs poorly in terms of GPU energy when compared with our proposed significance-driven approach. Our approach adopts reduced-complexity convolution filters in image areas that are of low significance (Section 3), and saves energy by up to 45%, while maintaining the said performance (20 fps) and quality (40 dB PSNR) requirements. Secondly, as expected, with increasing image resolutions our approach continues to benefit from dynamic resource allocations, coupled with DVFS, leading to further energy savings in the GPU-based convolution filter.

Further development will explore significance inference and image processing onto FPGA, to optimise PEQ trade-offs with built-in OpenCL runtime kernels.

REFERENCES

[1] Beckett, J. P. Apparatus and method for digital camera and recorder having a high resolution color composite image output, 1998.

[2] Venkataramani *et al.* Scalable-effort classifiers for energy-efficient machine learning. *Proc. 52nd Annu. Des. Autom. Conf. - DAC '15* (2015), 1–6.

[3] Alioto, M. Energy-Quality Scalable Adaptive VLSI Circuits and Systems beyond Approximate Computing. In *DATE* (2017), IEEE, p. 127–132.

[4] Qiqieh *et al.* Energy-efficient approximate multiplier design using bit significance-driven logic compression. In *DATE* (2017), pp. 7–12, Switzerland.

[5] Chippa *et al.* Approximate computing: An integrated hardware approach. In *Conf. Rec. - Asilomar Conf. Signals, Syst. Comput.* (2013).

[6] Sampson *et al.* EnerJ: Approximate data types for safe and general low-power computation. In *ACM SIGPLAN Not.* (2011), vol. 46, ACM, pp. 164–174.

[7] Sampson *et al.* Significance Driven Computation: A Voltage-scalable, Variation-aware, Quality-tuning Motion Estimator. In *ISLPED* (2009), pp. 195–200, San Fancisco, CA, USA.

[8] Preston, K. The need for standards in image processing. *Nature 333*, 6174 (1988), 611–612.

[9] Viola, P., and Jones, M. Robust real-time face detection. *Int. J. Comput. Vis. 57*, 2 (2004), 137–154.