A 36 Heterogeneous Core Architecture with Resource-Aware Fine-grained Task Scheduling for Feedback Attention based Object Recognition

Seungjin Lee, Jinwook Oh, Minsu Kim, Joonyoung Park, Joonsoo Kwon, Joo-Young Kim, and Hoi-Jun Yoo

School of EE&CS, Korea Advanced Institute of Science and Technology (KAIST)
373-1, Guseong-dong, Yuseong-gu, Daejeon, 305-701, Republic of Korea
E-mail: seungjin@eeinfo.kaist.ac.kr

Abstract: A heterogeneous SIMD/MIMD PE architecture is proposed to accelerate feedback attention based object recognition. Due to the difficulty of pipelining subsequent stages in an iterative recognition-attention feedback loop, parallelism within the recognition stage is exploited to improve performance. In addition, resource-aware fine-grained task scheduling is performed for high PE utilization, and voltage/frequency throttling based on neuro-fuzzy prediction enables low power real-time recognition on 30fps VGA video.
(Keywords: heterogeneous multi-core, computer architecture, task scheduling, object recognition)

1. Introduction

Robust object recognition is a computationally complex process that involves multiple cascaded transformations to account for variances between a training image and a novel input image [1]. Previous object recognition chips employed visual attention, inspired by the human vision system, that selectively processes regions of interest (ROI) containing objects to reduce computational complexity [2,3]. However, the feed-forward visual attention schemes in those works were fundamentally limited due to the lack of a feedback path from recognition to attention.

This work employs the unified visual attention model (UVAM) [4] which greatly improves attention precision, which is the number of selected tiles that actually contain target objects divided by the total number of selected tiles, by using feedback attention in addition to feed-forward attention, as shown in fig.1. High throughput of the attention-recognition feedback loop is achieved by 36 heterogeneous cores, comprised of both SIMD and multiple instruction multiple data (MIMD) PEs which exploit fine-grained parallelism within each recognition stage. High data reuse and high utilization are maintained by a hierarchical task scheduling scheme that schedules tasks based on local resource availability.

2. Heterogeneous SIMD/MIMD PE Architecture

Fig. 2 shows the feedback attention based UVAM compared to the feed-forward attention scheme of previous works [2,3]. In the UVAM, the attention and recognition stages constantly feedback information to one another, resulting in reduced calculations in the recognition stage, and improved accuracy in the attention stage. However, the feedback of the UVAM introduces interdependency between the attention and recognition stages, which makes parallelization difficult. Attention granularity, k, and attention delay, d, are introduced to address this issue, as shown in fig. 1(a),(b). Increasing k increases the possible parallelism of each recognition stage, while increasing d enables the recognition stage to be pipelined by deferring the interdependency between recognition and attention. However, a large value of k can degrade attention precision since background regions become more likely to be selected.

The value of k is kept relatively small by employing a small number of fast pipelines, instead of a large number of slow ones. For this, heterogeneous PEs exploit both data and task level parallelism to speed up recognition on each tile. SIFT [1] based object recognition consists of feature detection, feature description, and feature matching as shown in fig. 2. 4 SIMD Vector Processing Elements (VPE), each including a 20-way 8bit datapath, accelerate the pixel parallel feature detection tasks. Meanwhile, 32 MIMD Scalar Processing Elements (SPE), each including an area-efficient 16bit datapath, execute the control oriented feature description tasks in parallel. Feature matching is processed by a separate dedicated Feature Matching Processor (FMP). In order to exploit data locality, 1 VPE and 8 SPEs are grouped into a Feature Extraction Cluster (FEC), resulting in 4 FECs. Compared to the previous homogeneous SIMD PE based architecture [3,4], the 4 FECs provide approximately the same throughput as 16 SIMD PEs, but with 4x lower latency, as shown in fig. 3. Also, when k=4 and d=3, the FECs can in theory be fully utilized, as shown in fig. 4.

3. Resource Aware Fine-grained Task Scheduling

Resource aware fine-grained task scheduling, shown in fig. 5, minimizes external memory accesses and maximizes utilization of the FECs to achieve high performance. The global task management unit (GTMU)

maintains a Tile Memory Allocation Table (TMAT) that tracks the input image tiles loaded within each VPE. When the GTMU schedules new tile tasks for the VPEs, the TMAT is used to minimize the number of new tiles that must be loaded. As a result, external memory access is reduced by 53% compared to when input image tiles are not reused, and by 32% compared to sequential task scheduling.

SPE sharing among FECs through the collaboration of the Local Task Management Units (LTMU) enable high utilization of the SPEs. A major challenge in achieving high utilization in our heterogeneous multi-core pipeline is that the number of features detected, N_F, varies for each tile as shown in fig. 2. Therefore, if N_F is less than 8, then some of the 8 SPEs will be unutilized. If N_F is greater than 8, then the recognition pipeline will have to stall to accommodate additional iterations of the SPEs, thus seriously degrading the recognition throughput. However, with SPE sharing, unutilized SPEs are made available to neighboring FECs with N_F greater than 8, thus effectively averaging out the variation of N_F among the 4 FEC. As a result, pipeline stall occurrences are reduced by 82%, and the average tile throughput is increased by 22% to 13811 tiles per second. This translates to a worst case performance of 33.8 fps for a 640x480 pixel input when all tiles in the frame are selected.

4. Neuro-fuzzy Prediction based Voltage/Frequency Throttling

The neuro-fuzzy Intelligent Inference Engine (IIE) is used to throttle the voltage and frequency of the FECs and FMP. There is high potential for power savings through voltage/frequency throttling since the workload of the FECs and FMP varies widely depending on the number of selected tiles. Since frames are input at a fixed interval in a real-time application (33ms for a 30fps video stream), power consumption can be minimized by using the lowest possible voltage/frequency level that provides enough performance to process each frame within that time period. However, the required performance is not known for certain until processing on a frame is finished, due to the iterative tile selection of feedback attention. The IIE uses previous workload history and the result of the feed-forward attention algorithm to predict the total "attention budget" required to successfully recognize the objects in a frame. The actual voltage and frequency throttling is performed in 8 steps from 1.2V/200MHz to 0.65V/50MHz. As a result, 48% power reduction is achieved with less than 1% failed recognitions due to an attention budget deficit.

5. Implementation Results

The proposed chip [5], shown in fig. 6, is implemented in a 0.13um CMOS process and contains 2.93M gates and 612kB of SRAM. The recognition performance is tested on a 60s 30fps video sequence of a scene containing high amounts of background clutter, as shown in fig. 7. The UVAM achieves high average attention precision of 76%, and on average only 14% of the tiles in each frame are selected as ROI. Thanks to the high throughput of the recognition pipeline and voltage/frequency throttling, power consumption is reduced to 345mW, with 96% recognition rate and less than 1% false positive rate.

6. Conclusion

For the first time, an object recognition chip architecture optimized for feedback attention is proposed. Its heterogeneous SIMD/MIMD PEs reduce latency of the tile recognition task by 4x compared to homogeneous SIMD PEs, thus enabling high attention precision of 76% through a low attention granularity of 4. The 4 FECs, which each process 1 tile during an attention-recognition cycle, maintain high throughput by resource-aware fine-grained task scheduling, which minimizes external memory accesses, and maximizes SPE utilization. Meanwhile, power consumption is minimized by attention budgeting performed by neuro-fuzzy workload prediction of the IIE and subsequent voltage/frequency throttling of the FECs and FMP. As a result, this work achieves 96% recognition rate while consuming just 8.5mJ per frame.

[1] D.G. Lowe, "Distinctive image feature from scale-invariant keypoints," International Journal of Computer Vision, vol.60, no.20, pp. 91-110, 2004.

[2] K. Kim, et al., "A 125GOPS 583mW Network-on-Chip Based Parallel Processor with Bio-inspired Visual Attention Engine," ISSCC Dig. Tech. Papers, pp.308-309, Feb. 2008.

[3] J.-Y. Kim et al., "A 201.4 GOPS 496mW Real-Time Multi-Object Recognition Processor with Bio-inspired Neural Perception Engine," ISSCC Dig. Tech. Papers, pp.150-151, Feb. 2009.

[4] S. Lee, et al., "Familiarity based unified visual attention model for fast and robust object recognition," Pattern Recognition, vol.43, pp.1116-1128, Mar. 2010.

[5] S. Lee, et al., "A 345mW Heterogeneous Many-Core Processor with an Intelligent Inference Engine for Robust Object Recognition," ISSCC 2010, session 18.4.

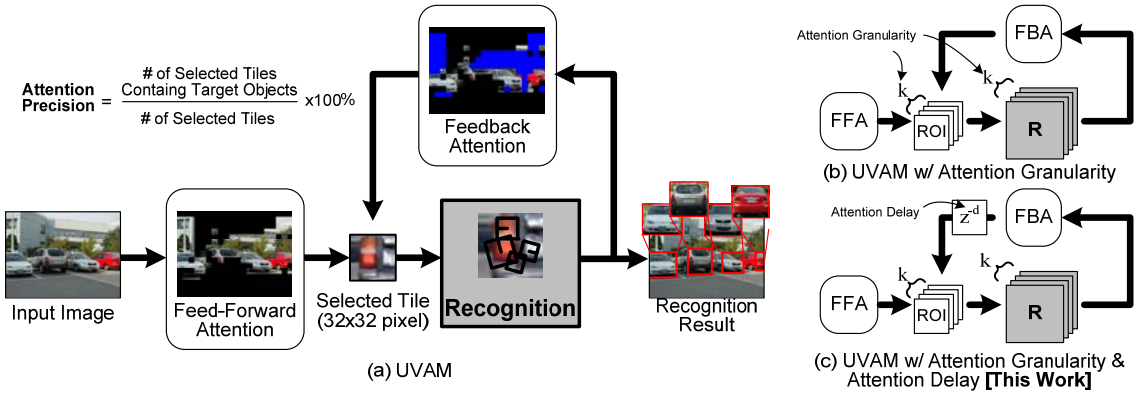


Figure 1. The Recognition Attention Feedback Loop of the Unified Visual Attention Model

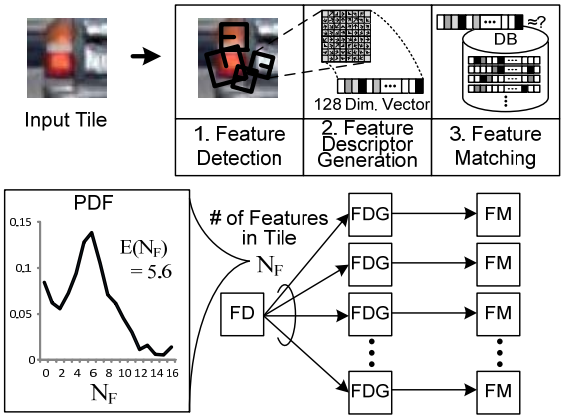


Figure 2. Steps of the tile recognition task

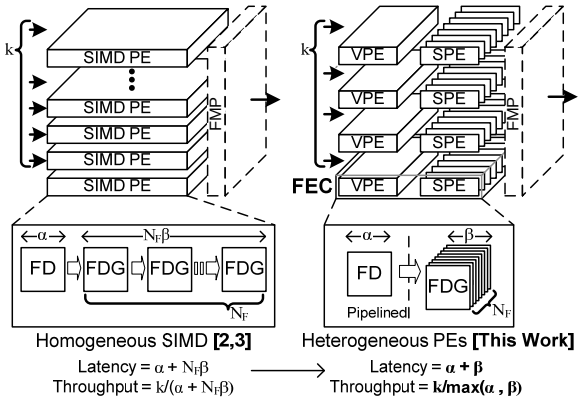


Figure 3. Heterogeneous SIMD/MIMD PE architecture

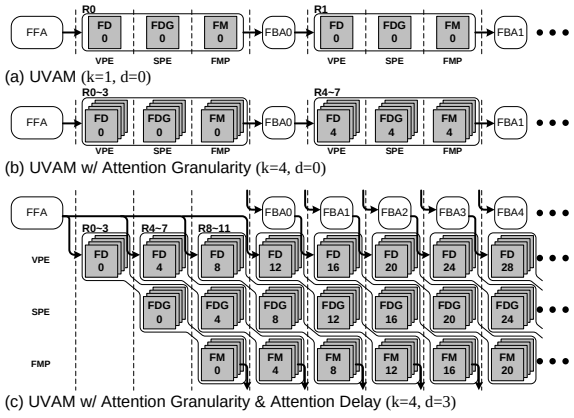


Figure 4. Parallelization and pipelining of the UVAM

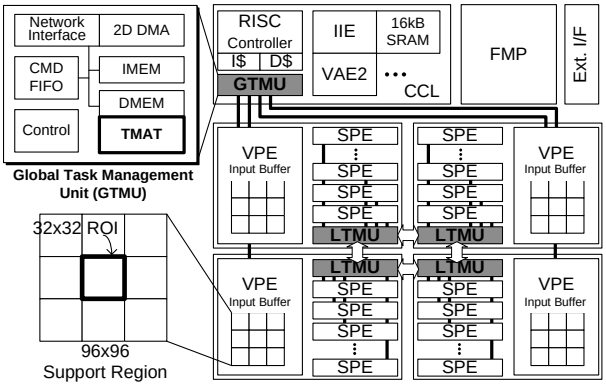
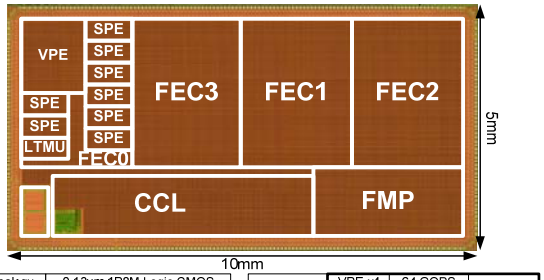


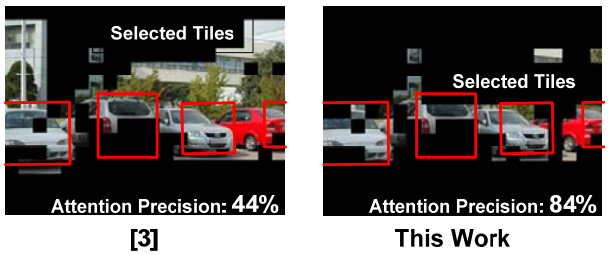
Figure 5. Top block diagram and resource-aware fine-grained task scheduling



Technology	0.13um 1P8M Logic CMOS	
Die Size	10.0mm x 5.0mm	
Gates / SRAM	2.92M Gates / 612 kB	
Power Supply	CCL	1.2 V
	FEC/FMP	0.65 ~ 1.2 V
Operating Frequency	CCL	200MHz
	FEC/FMP	50 ~ 200MHz

Peak Performance	VPE x4	64 GOPS	83.2 GOPS
	SPE x32	19.2 GOPS	
	CCL	66 GOPS	
	FMP	76.8 GOPS	
	Total	228 GOPS	
Power Consumption	Peak	704 mW	
	Average	345 mW	

Figure 6. Chip photograph and summary



Average results (30fps 1min video sequence):

	[3]	This Work
Attention Type	Feed-forward	**UVAM**
Attention Precision	58%	**76%**
Power Consumption (30fps)	496mW	**345mW**
False Positive Rate	8%	**1%**

Figure 7. Evaluation results