18.4 A 345mW Heterogeneous Many-Core Processor with an Intelligent Inference Engine for Robust Object Recognition

Seungjin Lee, Jinwook Oh, Minsu Kim, Junyoung Park, Joonsoo Kwon, Hoi-Jun Yoo

KAIST, Daejeon, Korea

Fast and robust object recognition of cluttered scenes presents two main challenges: (1) the large number of features to process requires high computational power, and (2) false matches from background clutter can degrade recognition accuracy. Previously, saliency based bottom-up visual attention [1,2] increased recognition speed by confining the recognition processing only to the salient regions. But these schemes had an inherent problem: the accuracy of the attention itself. If attention is paid to the false region, which is common when saliency cannot distinguish between clutter and object, recognition accuracy is degraded. In order to improve the attention accuracy, we previously reported an algorithm, the Unified Visual Attention Model (UVAM) [3], which incorporates the familiarity map on top of the saliency map for the search of attentive points. It can cross-check the accuracy of attention deployment by combining top-down attention, searching for "meaningful objects", and bottom-up attention, just looking for conspicuous points. This paper presents a heterogeneous many-core (note: we use the term "many-core" instead of "multi-core" to emphasize the large number of cores) processor that realizes the UVAM algorithm to achieve fast and robust object recognition of cluttered video sequences.

In the attention-recognition loop of the UVAM, shown in Fig. 18.4.1, attention and recognition processes are iteratively optimized through the back-and-forth feedback between attention and recognition. During the feedback attention process, familiarity evaluation is performed by neuro-fuzzy inference using semantic features of objects such as size, orientation, and motion as clues. As a result, computationally expensive Scale Invariant Feature Transform (SIFT) [4] object recognition, consisting of feature detection, feature description and database matching, is performed only on regions-of-interest (ROI) selected by the attention feedback process.

Our processor exploits 3 key features to realize UVAM-based object recognition. First, the analog-digital mixed-mode intelligent inference engine (IIE) accurately distinguishes target objects from clutter using the adaptive neuro-fuzzy inference system (ANFIS) [5] to improve the accuracy of the attention feedback. Second, 4 feature extraction clusters (FEC) comprised of 4 SIMD vector-processing elements (VPE) and 32 MIMD scalar-processing elements (SPE) with hierarchical task management accelerate feature detection and generation stages. Third, per-frame power-mode control based on workload prediction by the IIE minimizes power consumption.

The overall block diagram of the heterogeneous many-core processor is shown in Fig. 18.4.2. A total of 51 IPs are connected by a hierarchical star NoC [6] and organized into 2 layers: the cognitive control layer (CCL), which performs global attention and power-management functions, and the parallel-processing layer (PPL), which performs feature extraction and matching. The CCL consists of the IIE, a RISC host processor, the power-mode controller (PMC), and several fixed-function units for accelerating feed-forward visual attention. The PPL consists of 4 FECs for feature detection and description, and 1 feature-matching processor (FMP) for database matching.

Each FEC consists of 1 SIMD VPE for exploiting data-level parallelism (DLP) of the feature-detection task, and 8 SPEs for exploiting task-level parallelism (TLP) of the feature-description task. The VPE is a 20-way 8b vector processor optimized for image windows between 32×32 pixels and 40×40 pixels. A 20B-wide, byte-addressable 40kB local memory, and 1kB of coefficient memory can be directly accessed by a register-programmed convolution controller to achieve 18.25MAC/cycle or over 91% utilization of the vector ALU during a Gaussian filter operation. The SPE is a 16b scalar processor for accelerating the control-intensive operations of the feature-description stage. Its 5-stage pipeline is capable of memory load and ALU execution in a single instruction and has hardware support for sine, co-sine, arc-tangent, square-root, division and modulo operations. As a result, 1 feature-detection task takes 180µs on the VPE, and 1 feature-description task takes 161µs on the SPE.

Figure 18.4.3 shows the task distribution between 4 VPEs and 32 SPEs. The global task-management unit (GTMU) in the CCL and 4 local task-management units (LTMU) in each FEC perform hierarchical task management of the VPEs and SPEs to achieve high utilization rate. The GTMU and LTMUs feature 16 entry command queues for buffering incoming task requests, which are processed at a rate of 1 per 20 cycles to support low-latency fine-grained task management. Moreover, the LTMUs enable SPE sharing between neighboring FECs to alleviate under-utilization or saturation of the SPEs. Thanks to SPE sharing, the average throughput of the 4 FECs is increased by 22% to 460 ROI per frame at 30fps.

The IIE, shown in Fig. 18.4.4, consists of a 5-stage current-mode analog datapath for neuro-fuzzy inference, and a digital controller for loading inputs and parameters. The high and low boundaries of the parameterized Gaussian membership function's transfer curve can be controlled by V_{ref1} and V_{ref2} as shown in the waveforms. The slope of the Gaussian function is controlled by varying the g_m of M1 through M4. A 4kB internal cache reduces memory access overhead by 86%, thereby improving inference throughput by 21%. As a result, the IIE achieves 1M fuzzy logic inferences per second (FLIPS), and area and power consumption of the analog datapath are 0.176mm^2 and 1.2mW, or 54% and 15%, respectively, compared to an equivalent digital implementation.

Figure 18.4.5 shows the perturbation-learning [7] scheme employed by the IIE to achieve real-time adaptation. The evaluation result, $E(w_{ij}^{n})$ of the outer large circle, and the perturbed results, $E(w_{ij}^{n}+\delta)$ of the inner 9 iterative calculation paths, are used to calculate the antecedent parameters w_{ij}^{n+1} for the next epoch, where w_{ij}^{n} are the current antecedent parameters, x_t is the input, $y(x_t, w_{ij}^{n})$ is the familiarity output, y_d is the desired output, and δ denotes perturbation. One iteration, or epoch, of perturbation learning takes 3.5µs and learning with <5% error is achieved in just 20 epochs, or 70µs.

Figure 18.4.6 outlines the power-management scheme and its measurement results. The PMC performs per-frame power-mode control on the voltage and frequency island (VFI) containing the PPL, which consumes 78% of the chip's peak power. Workload prediction with an average error rate of 6% is performed by the IIE using the workload of the previous frame and the bottom-up saliency map as inputs. As a result, the average power consumption of the chip is reduced by 48%, compared to when only saliency-based attention is used without PMC.

The chip (Fig. 18.4.7) occupies 50mm^2 in a 0.13µm 8 metal CMOS process and contains 2.93M equivalent gates and 626kB of SRAM. Peak performance is 228GOPS while peak power efficiency is 545 GOPS/W with the PPL throttled down to 50MHz/0.65V. 96% recognition accuracy and 8.5mJ/frame energy efficiency are achieved on a 30fps VGA video stream of a cluttered scene, showing that the UVAM based heterogeneous many-core chip enables fast and robust object recognition in real-life scenarios.

References:

[1] 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.

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

[3] S. Lee, et al., "Familiarity based unified visual attention model for fast and robust object recognition," Pattern Recognition, doi:10.1016/j.patcog.2009.07.014, 2009.

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

[5] J.-S.R. Jang, "ANFIS: Adaptive-network-based fuzzy inference system," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 23, no. 3, pp. 65-685, May 1993.

[6] H.-J. Yoo, et al., "Low-Power NoC for High-Performance SoC Design," CRC Press, 2008.

[7] M. Jabri, "Weight Perturbation: An Optimal Architecture and Learning Technique for Analog VLSI Feedforward and Recurrent Multilayer Networks," Neural Computation, vol.3, no.4, pp. 546-565, 1991.

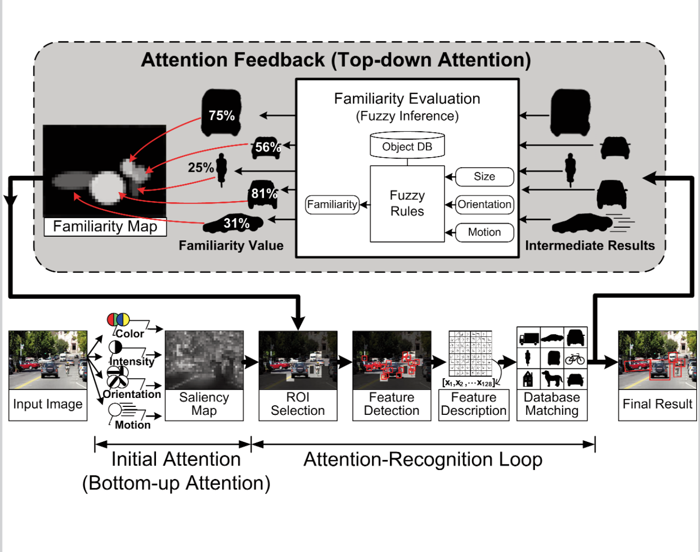


Figure 18.4.1: Attention-recognition loop of the unified visual attention map.

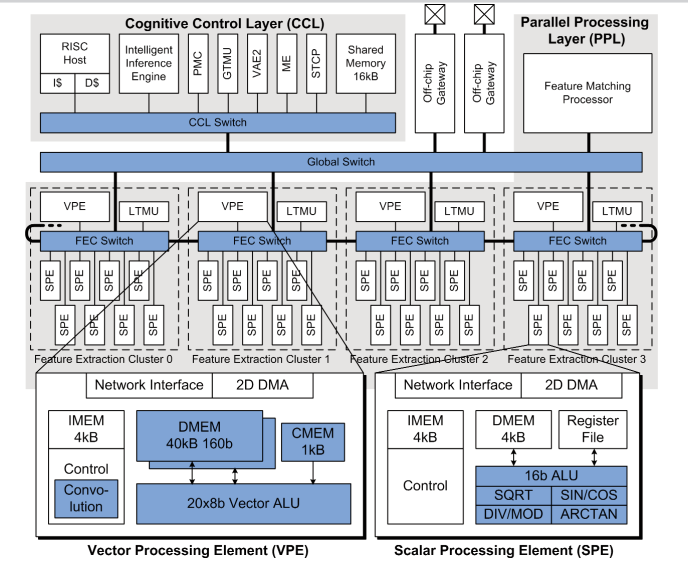


Figure 18.4.2: Block diagram of the heterogeneous multi-core processor.

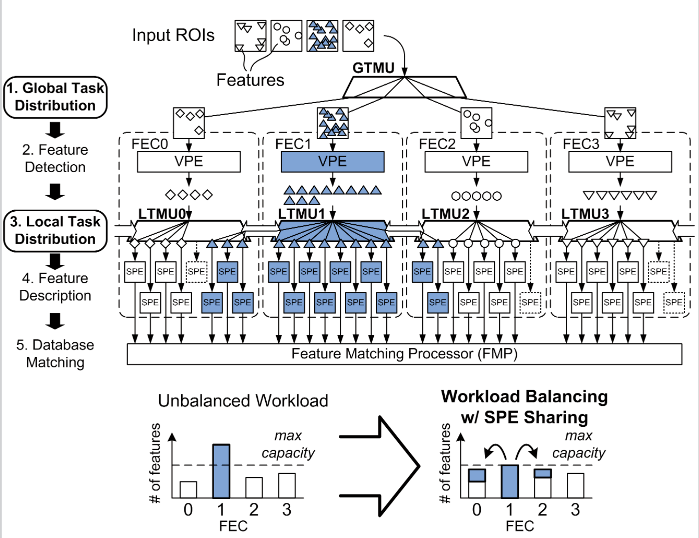


Figure 18.4.3: Hierarchical task management of the VPEs and SPEs.

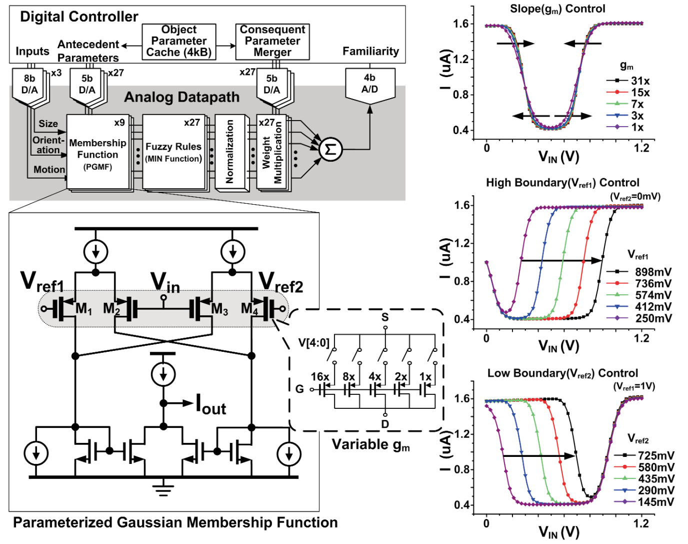
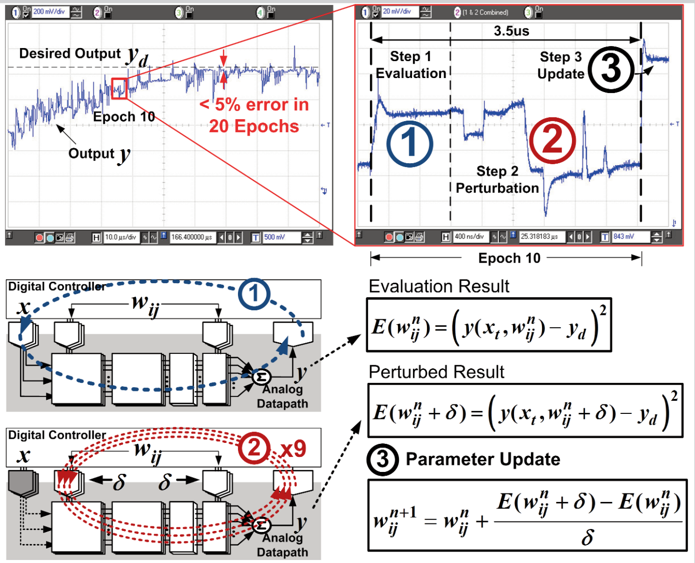


Figure 18.4.4: Mixed-mode intelligent inference engine and parameterized Gaussian membership function.



Evaluation Result

$$E(w_{ij}^n) = \left(y(x_t, w_{ij}^n) - y_d\right)^2$$

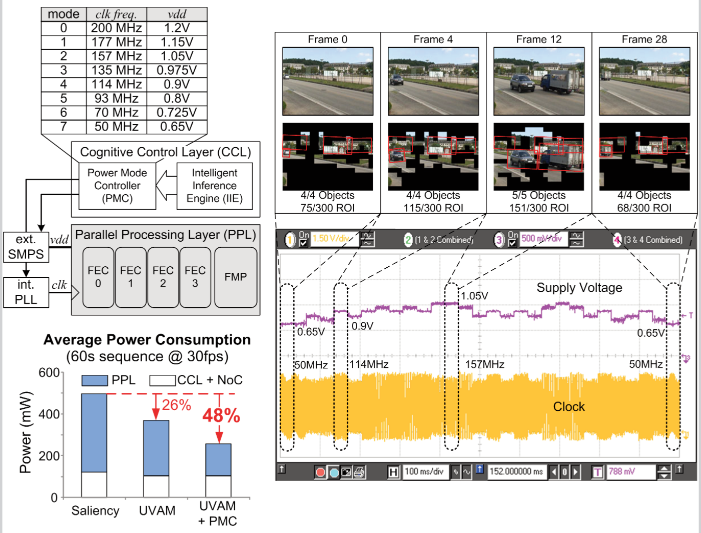
Perturbed Result

$$E(w_{ij}^n + \delta) = \left(y(x_t, w_{ij}^n + \delta) - y_d\right)^2$$

3 Parameter Update

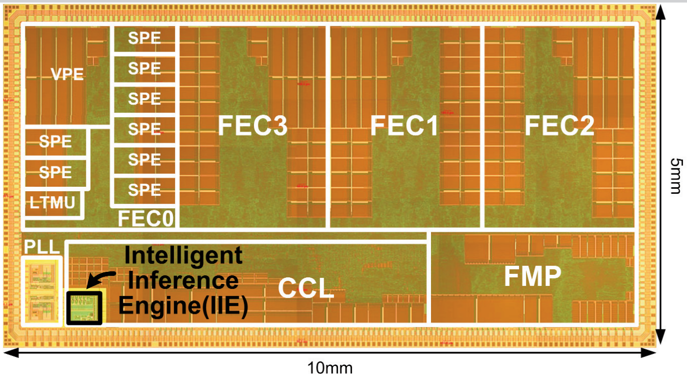
$$w_{ij}^{n+1} = w_{ij}^n + \frac{E(w_{ij}^n + \delta) - E(w_{ij}^n)}{\delta}$$

Figure 18.4.5: Perturbation learning in the intelligent inference engine.



mode	clk freq.	vdd
0	200 MHz	1.2V
1	177 MHz	1.15V
2	157 MHz	1.05V
3	135 MHz	0.975V
4	114 MHz	0.9V
5	93 MHz	0.8V
6	70 MHz	0.725V
7	50 MHz	0.65V

Figure 18.4.6: Per-frame power-mode management based on intelligent workload prediction.



Technology	0.13um 1P8M Logic CMOS	
Die Size	10.0mm x 5.0mm	
Gates / SRAM	2.92M Gates / 612 kB	
NoC IPs	51	
Power Supply	CCL & NoC	1.2 V
	PPL	0.65 ~ 1.2 V
Operating Frequency	Global NoC	400MHz (45FO4)
	CCL	200MHz (90FO4)
	PPL	50 ~ 200MHz (45FO4)

Peak Performance	CCL	66 GOPS
	VPEs	64 GOPS
	SPEs	19.2 GOPS
	FMP	76.8 GOPS
	Total	228 GOPS
IIE Performance	Learning	11.35 MCUPS/mm^2
	Inference	1 MFLIPS (27 rule)
Power Consumption	Peak	704 mW
	Average	345 mW

Figure 18.4.7: Chip micrograph and summary.