An Energy Efficient Real-Time Object Recognition Processor with Neuro-Fuzzy Controlled Workload-aware Task Pipelining

Joo-Young Kim, Minsu Kim, Seungjin Lee, Jinwook Oh, Kwanho Kim, Jeong-Ho Woo, 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: trample7@eeinfo.kaist.ac.kr

An energy efficient pipelined architecture is proposed for multi-core object recognition processor. The proposed neuro-fuzzy controller and intelligent estimation of the workload of input video stream enable seamless pipelined operation of the 3 object recognition tasks. The neuro-fuzzy controller extracts the fine-grained region-of-interest, and its task pipelining achieves 60.6fps, 5.8x higher performance. The 8.1mJ/frame energy efficiency is obtained by workload aware task scheduling and power management, and it is 6.9x higher efficiency compared to the previous multi-core architectures.
(Keywords: multi-core architecture, neuro-fuzzy, task pipelining, object recognition)

1. Introduction

Recently, object recognition has become the key technology for advanced video applications such as object-based video coding, scene analysis, and content based image retrieval [1-3]. Usually, the multi-core approach with single instruction multiple data (SIMD) processing elements (PEs) is adopted to process huge amount of video data in real-time [1]. The previous work [2] employs the visual attention mechanism mimicking the human vision system with a visual attention engine (VAE) to improve processing efficiency [4]. However, it just alleviated the calculation of the PEs by filtering useless features of the image by neural networks. In this paper, on top of the previous visual attention mechanism, a smart pipelining scheme is proposed to achieve both the workload reduction and the high speed operation resulting in high efficiency energy operation.

2. Energy Efficient Pipelined Architecture

Fig. 1 shows the comparison of the proposed Neuro-fuzzy (NF) controlled pipelining architecture with the conventional massively parallel PE architecture [1] and the VAE architecture [2]. In the massively parallel architecture, hundreds of small processing elements are linearly connected and run the same operation. This architecture has strength in low-level image processing which applies the same operations to each pixel of the large video data. However, it suffers from low efficiency in object oriented recognition processing because it processes the whole image area including meaningless background regions. In the VAE architecture [2], the VAE plays a role of pre-processing filter by eliminating the meaningless feature points due to global processing property of cellular neural networks. Since the workload of the VAE was not equal to that of the PEs, the VAE and PEs were hard to be pipelined. The post processor, adopted for database matching, was not pipelined either. In this architecture, the 3 object recognition tasks -visual perception (VP), main image processing (MIP), and post database matching (PDM)- are pipelined for high frame rate, high efficiency object recognition. The following 3 key features make the pipelining possible. First, the visual perception, an enhanced version of the previous visual attention, is separated from the PE processing. Second, the post database matching needs to process only the reduced number of feature vectors. Thirdly, the proposed neuro-fuzzy controller balances the workloads of the 3 stages well. It estimates the workload of the input stream by extracting the region-of-interests (ROIs) of the input image in 40 x 40 tile units, and the information on the number of ROIs, or workload information, can be used to balance the execution time of the following two stages.

Fig. 2 shows the details of how the NF controller controls the execution time. The workload aware task scheduling is proposed to adjust the execution time of the main image processing [Fig.2 (a)]. Since the NF controller estimates the workload of the following PEs by measuring the amount of the ROI tasks before scheduling, it can determine the number of operating PEs. That is, according to the workload levels, the attended number of ROIs versus the total number of ROIs, the number of PEs is linearly varied to keep the overall execution time constant. Initial workload thresholds for the control of the number of PEs can be modified to keep the execution time constant at the end of the pipeline with the feedback of the current execution time. The graph of the Fig. 2 (a) shows the results of the workload aware task scheduling with the 16 PEs and 4-workload levels. Fig. 2(b) shows control process to adjust the execution time of the database

matching stage by varying the applied database size before its processing. Based on the applied vector matching algorithm [5], the overall execution time of the post processor can be estimated by the following simple equation where α is a database transfer time, β is a unit cycle time to match a feature vector to the fixed database, F is the number of feature vectors, and the γ is the coverage rate of the database. The NF controller configures the database coverage rate of the post processor, so that the feature vectors from the PEs are matched by the post processor within the target pipeline time.

$$Execution\ time = (\alpha + F \times \beta) \times \gamma$$

3. Low Power SoC Architecture

Fig. 3 shows the block diagram of the overall processor [3] and the proposed NF controller. The processor consists of the NF controller, 16 coarse grained SIMD PEs which has 4 power domains, post processor, and 2 external interfaces. All IP blocks are inter-connected with network-on-chip (NoC) interface. The 4 power domains of the 16 PEs can be managed separately together with the workload aware task scheduling. The low power operation can be obtained by gating the unscheduled power domains. The 3 stages of the object recognition are directly mapped to the NF controller, 16 PEs, and the post matching processor, respectively. The NF controller is composed of neuro-fuzzy processing unit and ROI task management unit. The processing unit includes mixed mode circuits of cellular neural network and neuro-fuzzy classifier which are suitable to execute biologically inspired visual perception algorithm [3]. The cellular neural network is adopted to rapidly extract the static features such as intensity, color, and orientation from the image using its highly parallel and collective processing [4]. The bio-inspired neural and fuzzy processing is adopted to solve ambiguous classification problem in object detection and ROI extraction. The table based task management unit is responsible for scheduling ROI tasks toward the 16 PEs.

4. Performance Evaluation and Implementation Results

The 20 test images in VGA (640x480) size are evaluated for the 3 different architectures of Fig.1 to evaluate the effect of the NF controlled object recognition. The 3 architectures have the same 16 homogeneous PEs and post matching processor. We evaluate frame rate, average power, and energy per frame and Fig. 4 shows the experimental results. The VAE accelerates object recognition processing by 1.55x, but slightly increases power consumption. On the other hand, the proposed NF controlled task pipelining reduces the execution time by 5.83x with the fine-grained ROI extraction. Power consumption also goes down by 15% thanks to the workload aware task scheduling and power management. As a result, the proposed pipelined architecture improves energy efficiency for object recognition by 6.9x and 4.5x compared to the multi-core architecture without and with the VAE, respectively.

The proposed object recognition processor is fabricated in 0.13μm CMOS process. It contains 36.4M transistors including 3.73M gates and 396K byte on-chip SRAM in 49mm^2. The NF controlled task pipelining enables the 1.2V processor to achieve 60.6 fps object recognition with 496mW average power consumption at 200MHz. Fig. 5 shows the fabricated chip micrograph and summarizes its features.

5. Conclusion

3-stage pipelining is in the world-first proposed for energy efficient object recognition processor. The NF controller estimates the workloads of the input stream data by the number of fine-grained ROIs. The workload information is used to force the each execution time to be constant in the 3 pipeline stages, VP, MIP, and PDM. The proposed architecture can achieve 60.6 fps and 8.1mJ/frame energy efficiency, which is 5.8x and 6.9x improved result compared to the previous architectures. The object recognition SoC is fabricated in 0.13μm technology to operate 496mW at 200MHz. Its energy consumption per frame is the world lowest ever reported.

[1] A. Abbo, et al., "XETAL-II: A 107 GOPS, 600mW Massively-Parallel Processor for Video Scene Analysis," ISSCC Dig. Tech. Papers, pp. 270-271, Feb. 2007

[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 2009, session 8.3.

[4] S. Lee, et al., "The Brain Mimicking Visual Attention Engine: An 80x60 Digital Cellular Neural Network for Rapid Global Feature Extraction," IEEE Symp. on. VLSI circuits, pp. 26-27, 2008

[5] J. –Y. Kim et al., "A 66fps 38mW Nearest Neighbor Matching Processor with Hierarchical VQ Algorithm for Real-Time Object Recognition," IEEE A-SSCC, pp.177-180, 2008

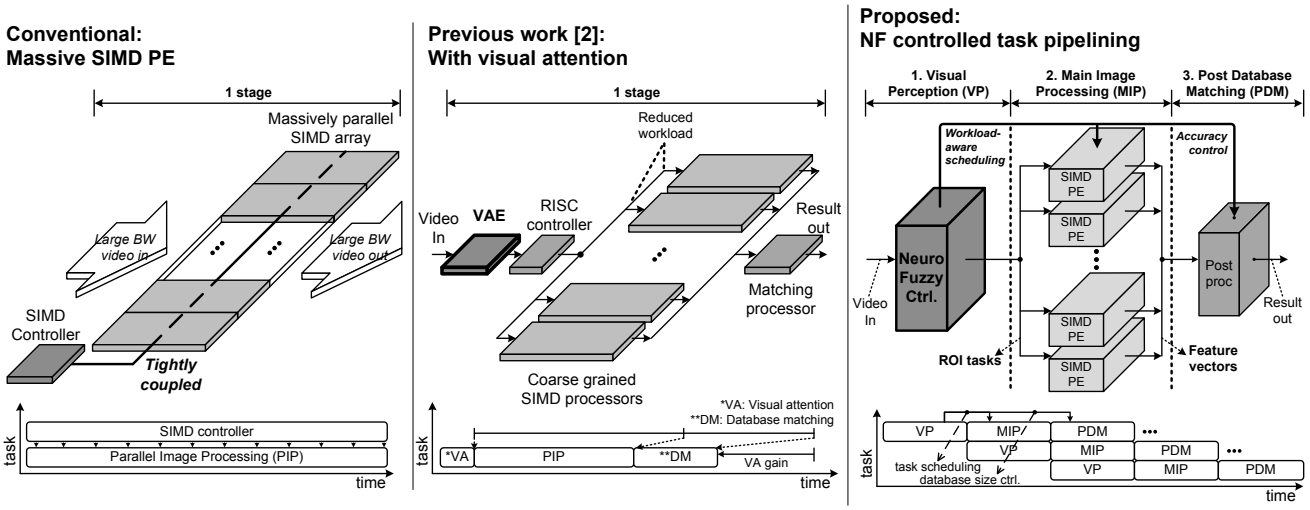
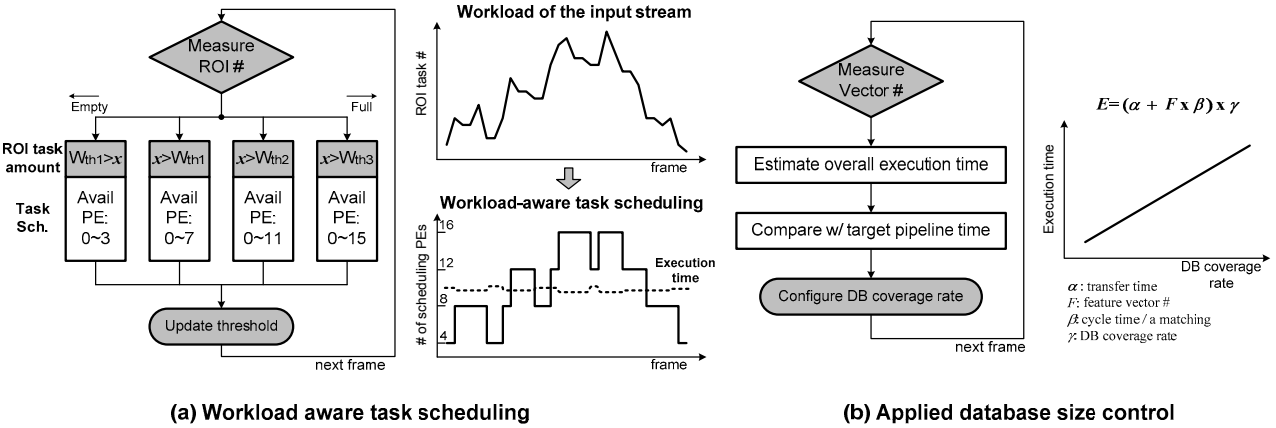


Fig. 1 Proposed neuro-fuzzy controlled multi-core architecture



(a) Workload aware task scheduling (b) Applied database size control

Fig. 2 Pipeline time balancing schemes

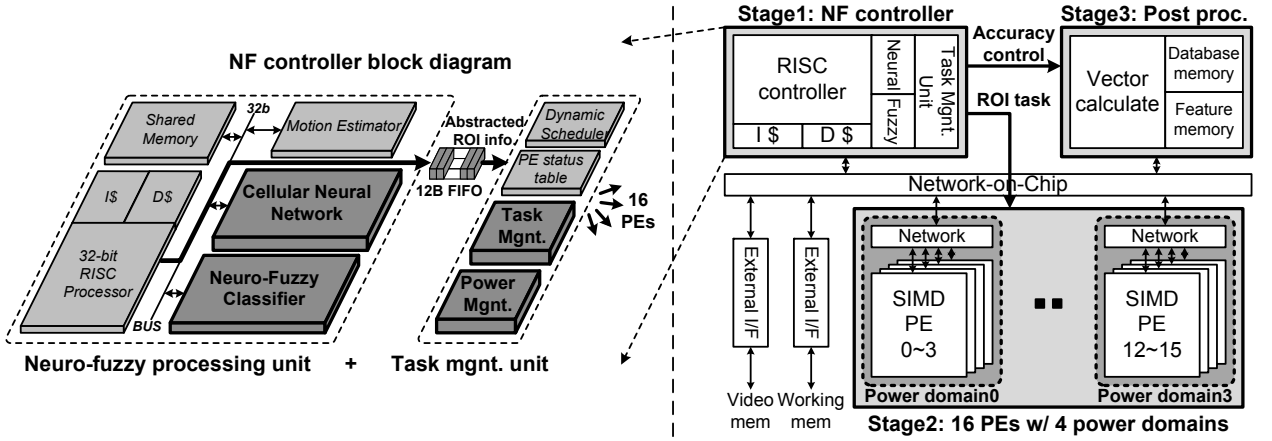
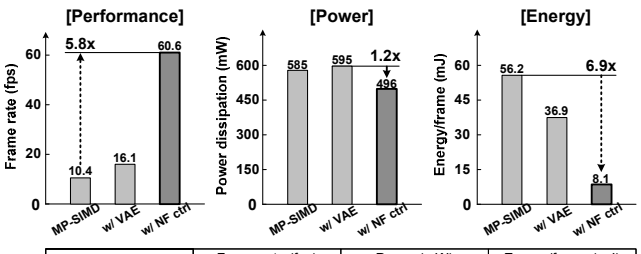
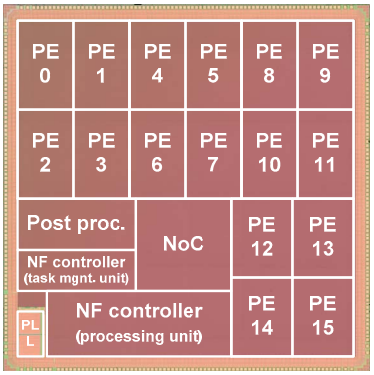


Fig. 3 Block diagram of the proposed NF controller and overall processor



	Frame rate (fps)	Power (mW)	Energy/frame (mJ)
MP-SIMD (1)	10.4	585	56.2
/w VAE (2)	16.1	595	36.9
/w NF control (3)	**60.6**	**496**	**8.1**
Gain (1-3 / 2-3)	5.83x / 3.76x	1.18x / 1.20x	6.93x / 4.56x

Fig. 4 Experimental results



Technology	0.13mm 1P8M CMOS	
Die Size	7mm x 7mm	
Power supply	1.2V core, 2.5 I/O	
Operating freq.	400MHz(45FO4) for NoC / 200MHz (90FO4) for IPs	
# of transistors (gates, memory)	36.4M transistors 3.73M gates / 396KB SRAM	
Peak Performance	16 SIMD PEs	128 GOPS
	NF controller	54 GOPS
	Post proc.	19.4 GOPS
	Total	**201.4 GOPS**
Power consumption	**496mW** (full application running)	
Power efficiency	**290 GOPS/W**	
Frame rate	**60.6 fps**	
Energy/frame	**8.1 mJ**	
Video size	VGA(640x480 pixels)	

Fig. 5 Chip micrograph and feature summary