# A Configurable Heterogeneous Multicore Architecture with Cellular Neural Network for Real-Time Object Recognition

Kwanho Kim, *Student Member, IEEE,* Seungjin Lee, *Student Member, IEEE,* Joo-Young Kim, *Student Member, IEEE,* Minsu Kim, *Student Member, IEEE,* and Hoi-Jun Yoo, *Fellow, IEEE*

***Abstract*—As object recognition requires huge computation power to deal with complex image processing tasks, it is very challenging to meet real-time processing demands under low-power constraints for embedded systems. In this paper, a configurable heterogeneous multicore architecture with a dual-mode linear processor array and a cellular neural network on the network-on-chip platform is presented for real-time object recognition. The bio-inspired attention-based object recognition algorithm is devised to reduce computational complexity of the object recognition. The cellular neural network is utilized to accelerate the visual attention algorithm for selecting salient image regions rapidly. The dual-mode parallel processor is configured into single instruction, multiple data (SIMD) or multiple-instruction-multiple-data modes to perform data-intensive image processing operations while exploiting pixel-level and feature-level parallelisms required for the attention-based object recognition. The algorithm's hybrid parallelization strategy on the proposed architecture is adopted to obtain maximum performance improvement. The performance analysis results, using a cycle-accurate architecture simulator, show that the proposed architecture achieves a speedup of 2.8 times for the target algorithm over conventional massively parallel SIMD architecture at low hardware cost overhead. A prototype chip of the proposed architecture, fabricated in 0.13 $\mu$m complementary metal-oxide-semiconductor technology, achieves 22 frames/s real-time object recognition with less than 600 mW power consumption.**

***Index Terms*—Cellular neural network, multicore, object recognition, parallelism, SIMD/MIMD.**

## I. Introduction

OBJECT RECOGNITION has been emerging as one of the most popular embedded applications and is widely used in various applications such as mobile robot vision systems, autonomous vehicle control, natural human–machine interfaces, and visual surveillance systems [1]–[3]. For example, object recognition plays an important role in autonomous navigation of intelligent mobile robots. The object recognition applications are characterized by complex and data-intensive computations with very large memory requirements. Especially, real-time performance and low power consumption are important factors for embedded systems. Programmability should be also considered to deal with a wide variety of recognition targets and algorithms.

Object recognition involves a series of complex image processing tasks ranging from low-level image processing to high-level image processing. In low-level processing (e.g., image filtering, feature extraction), simple arithmetic operations are regularly performed on a 2-D image array of pixels. On the contrary, high-level processing is irregular and performed on objects that are defined by groups of features extracted at the lower level. Since object recognition requires huge computation power for each stage, general-purpose architectures, such as microprocessor and digital signal processor cannot achieve a real-time processing due to its sequential processing. Although ASIC implementations like [5] can achieve real-time frame rates, there is a flexibility limitation to deal with complex and various recognition algorithms.

With the rise of multicore computing, parallelization and multiprocessor implementations have become increasingly important to increase the computing power. Because object recognition algorithm tends to exhibit huge amounts of computation and inherent parallelisms, how to maximally exploit its parallelism on the emerging multicore hardware is the key issue for achieving a real-time performance. Recently, scale invariant feature transform (SIFT) [4], the most popular object recognition algorithm, has been implemented on graphics processing unit [6] and multicore systems [7] while exploiting the parallelism. Though both works can achieve real-time performance with a large-size image, huge power consumption makes it difficult to be applicable for embedded systems. Several parallel processing architectures have been presented for vision applications. Massively parallel single-instruction-multiple-data (MP-SIMD) processors with linear processor array, such as Internet message access protocol [8] and Xetal [9] have been developed for low-level vision processing [10], [11]. However, these processors are not suitable for higher-level vision applications that exhibit more irregular and data-dependent behavior than low-level operations. Multiple-instruction-multiple-data (MIMD) multiprocessor architecture [12] has been realized to exploit task-level parallelism in

Manuscript received January 31, 2009; revised May 24, 2009. First version published September 1, 2009; current version published October 30, 2009. This paper was recommended by Associate Editor G. G. Lee.

The authors are with the Division of Electrical Engineering, Department of Electrical Engineering and Computer Science, Korea Advanced Institute of Science and Technology, Daemon 305-701, Korea (e-mail: kkh82@eeinfo.kaist.ac.kr; seungjin@eeinfo.kaist.ac.kr; trample7@eeinfo.kaist.ac.kr; beatin@eeinfo.kaist.ac.kr; hjyoo@ee.kaist.ac.kr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2009.2031516

vision applications. However, it is hard to achieve a real-time performance under low-power constraints due to its limited cost inefficiency and data synchronization overhead among the processors to access data in shared memory.

In this paper, a configurable heterogeneous multicore architecture is presented to achieve real-time object recognition performance for embedded systems. The main contributions of this paper are as follows. First, a VLSI architecture combining single instruction, multiple data (SIMD)/MIMD dual-mode parallel processor and cellular neural network on the network-on-chip (NoC) platform is proposed based on algorithmic considerations. Hybrid parallelization strategy of the algorithm on the proposed architecture obtains maximum parallel performance while exploiting pixel-level and feature-level parallelisms. Second, the proposed architecture is evaluated using a cycle-accurate architecture simulator in terms of the effect of hardware accelerator and dual-mode parallelism. The results show the proposed architecture achieves 2.8 times performance improvement over the conventional MP-SIMD architecture. Finally, the prototype chip proves the novelty of the proposed architecture and achieves 22 frames/s real-time object recognition with less than 600 mW power consumption.

This paper is organized as follows. In Section II, a bio-inspired attention-based object recognition algorithm is presented. In Section III, the target algorithm analysis including hardware acceleration and parallelism requirements is described. The proposed architecture and parallelization strategy of the algorithm are shown in Section IV. Performance analysis over the conventional MP-SIMD architecture is described in Section V. The prototype chip implementation is shown in Section VI. The conclusion of this paper will be made in Section VII.

## II. Bio-inspired Object Recognition Algorithm

Local feature-based object recognition algorithms, such as SIFT, have been widely used for robust and invariant object recognition. In the SIFT algorithm, scale invariant local key-points are extracted first by scanning the entire image over many scales. Then, key-point descriptors are generated from image gradients around the extracted key-points and matching for each key-point descriptor is performed by identifying its nearest neighbor on the database of key-points from training images. From the SIFT algorithm steps, it is easily found that the execution time of SIFT algorithm is proportional to the number of key-points extracted. Therefore, reducing the number of key-points extracted without sacrificing accuracy is a primary strategy to improve object recognition performance. To reduce the computational requirements of the object recognition, we introduce the additional step of finding salient image regions before key-points are extracted. In this paper, the concept of visual attention, inspired by human visual system, is integrated into the conventional object recognition algorithm.

### A. Saliency-Based Visual Attention

Humans focus on only visually-relevant parts of the available visual information. Visual attention is the ability of the human visual system to rapidly detect the salient image regions related to human interests. It is an essential role of the visual cortex in the human brain [13]. The visual attention mechanism controls the selection of the most informative parts of the image to process the huge amount of visual information gathered by two eyes. Many researchers have proven biologically and psychologically that such a visual attention strongly contributes to the high performance of the human vision system [14].

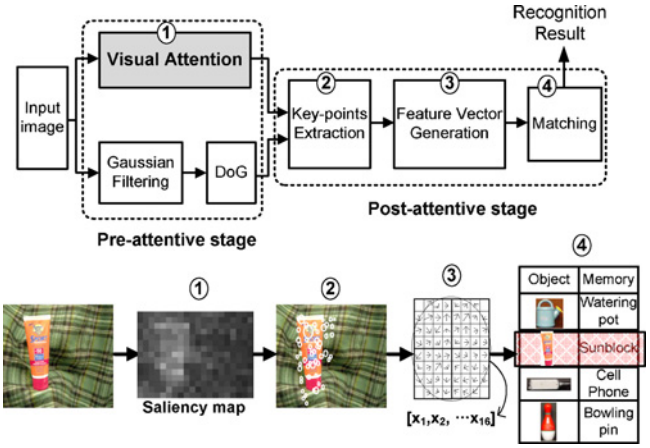



Fig. 1. Attention-based object recognition algorithm.

For computational modeling of the human-like visual attention mechanism, our attention system is based on a saliency-based model of the visual attention presented by Itti [15]. It consists of four main steps. First, the model starts with extracting a set of features like color, intensity, and orientation in parallel. Then, a conspicuity map is generated for each feature by using multiscale center-surround filters, which highlights the parts of the image that strongly differ from their surroundings. After that, a set of conspicuity maps is merged into a single map of attention called saliency map. Finally, given the saliency map, the most salient locations are selected by means of a winner-take-all mechanism. Because the saliency-based model is based on simple and bottom-up feature extraction methods suitable for massively parallel operation, a real-time performance can be achieved on dedicated parallel hardware.

### B. Attention-Based Object Recognition

In this paper, the saliency-based visual attention is incorporated into the local feature-based object recognition such as SIFT to reduce the computational complexity of the object recognition. The proposed attention-based object recognition algorithm flow is shown in Fig. 1. The algorithm is largely divided into two stages according to the visual attention: pre-attentive and post-attentive stages. During the pre-attentive stage, in contrast to the conventional object recognition, visual attention is performed in advance. Visual attention can be regarded as a pre-processing step which allows a rapid selection of the sensory information. To select the image regions of interest from the saliency map, we use the attention threshold whose value is parameterizable. Some image pre-processing operations like difference of Gaussian (DoG) are also performed in parallel. In the post-attentive stage, key-points are extracted as local maximum or minimum of the

DoG images across scales on the pre-selected salient image regions provided by the visual attention mechanism. Key-point descriptor vectors are then generated using local image gradients in the region around each key-point, providing robustness to scale, rotation, and intensity changes. Finally, we can recognize the object by matching individual key-point descriptor vectors to a database of features from training images using a nearest neighbor search algorithm [4].

The main purpose of visual attention in object recognition is cueing subsequent visual processing stages to improve performance by reducing the computation cost of higher-level vision tasks. Thus, instead of trying to interpret the entire scene, the recognition module focuses on the scene parts previously provided by the visual attention module. By the visual attention mechanism, the number of key-points extracted is reduced as shown in Fig. 1, avoiding background clutter. The key-points only in the attended image region need to be matched to the object database, making it faster and easier to recognize the object. As a result, computational requirements of the object recognition are reduced by 36% on average when we test 50 objects with a background image. The visual attention improves the recognition performance in the presence of large amounts of clutter by up to an order of magnitude. Moreover, a number of computer vision applications such as object tracking and image segmentation can benefit from the visual attention mechanism as well.

## III. Target Algorithm Analysis

### A. Dedicated Hardware Acceleration

The first step of algorithm analysis is to identify the most performance critical parts, which need to be accelerated through a dedicated hardware or architectural optimizations. The execution time profile of the attention-based object recognition algorithm on 2.4 GHz CPU is shown in Fig. 2. According to the profile results, we found that visual attention and matching are the most computationally demanding tasks. The execution time overhead of the visual attention should be minimized to reduce performance penalty by the additional step. Also, matching is an essential step and occupies larger portion of total execution time when the size of database or descriptor vector dimension becomes larger. Therefore, we decided the visual attention and nearest neighbor matching are accelerated by a dedicated hardware for better cost efficiency. On the contrary, other parts of the algorithm such as feature extraction and feature vector generation are variable depending on a variety of recognition algorithms and target applications. It is clear that a programmable multiprocessor is more suitable to perform such data-intensive and various image processing tasks. In this paper, it is unique that the cable news network (CNN) is adopted to accelerate saliency-based visual attention operation. The reason why the CNN is beneficial in the visual attention will be described in the next subsection.

### B. Cellular Neural Network to Accelerate Visual Attention

The CNN is a 2-D array of locally connected cells as an alternative to fully connected neural networks [16]. The CNN

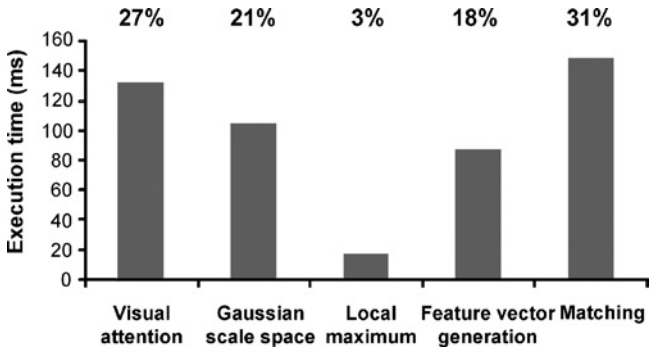



Fig. 2. Execution time profile of the attention-based object recognition algorithm on a 2.4 GHz CPU.

operation is characterized by a set of template parameters. Due to its enormous computational capabilities, the CNN enables the realization of real-time complex image processing applications, such as pattern recognition [17] and video processing [18]. The uniform local connections of the CNN make it suitable for VLSI implementation. Although the cells of the CNN are connected only to its neighbor cells, the propagation effects of the dynamics of the network allow global feature extraction in images.

In this paper, a CNN mechanism is utilized to accelerate visual attention operation as shown in Fig. 3. The most time-consuming step of the saliency-based visual attention is to calculate various center-surround filters such as Gaussian filter for intensity feature and Gabor filter for orientation feature [15]. The CNN can complete the center-surround filters in real-time because of its inherent cell-level parallel computation model compared to digital processing systems. We use some well-known CNN templates for implementation of CNN-based Gaussian filter and Gabor filter [19]. The templates can also be obtained by a learning process based on genetic algorithms [20]. As a result, the saliency-based visual attention can be accelerated on the CNN-based dedicated hardware to minimize the overhead of the visual attention.

### C. Required Parallelism on Target Algorithm

The attention-based object recognition has different operation characteristics based on before-and-after visual attention, which requires two different types of parallelisms as shown in Fig. 4. In the pre-attentive stage, all pixels on the entire image perform the same operation for image pre-processing. Massively parallel SIMD processing is appropriate in such regular and data-intensive operations to exploit pixel-level parallel processing. Various image filtering used in scale-space image generation belong to this type of operation. On the contrary, in the post-attentive stage, image processing restricts to only interesting image regions (i.e., image objects) around extracted key-points selected by the visual attention module. Image objects may be of irregular shape and dynamically changed according to the algorithmic parameters or the location of the selected key-points. It makes pure SIMD processing less effective and cannot fully utilize the computing capability of the hardware. MIMD supports independent processing of overlapping image segments to exploit feature-level parallel

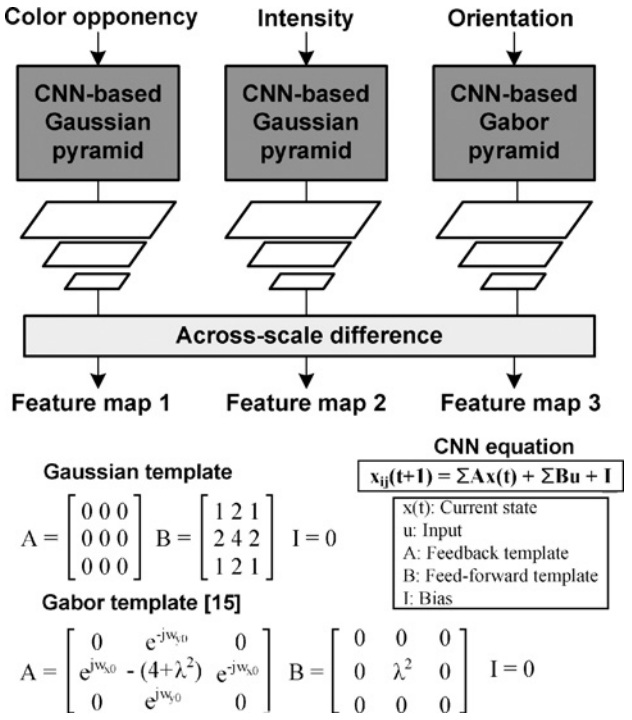



Fig. 3. CNN-based center-surround mechanism based on [15].

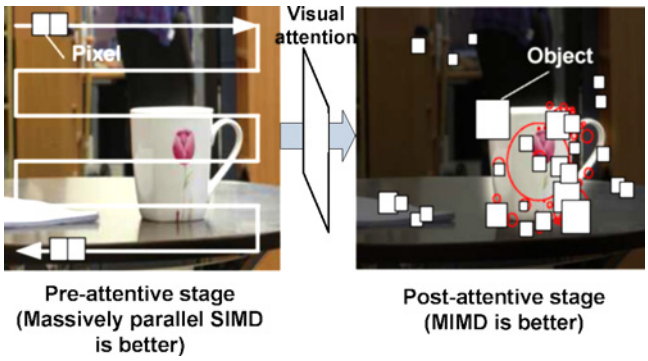



Fig. 4. Different parallelism on the target algorithm: pixel-level parallelism at the pre-attentive stage and feature-level parallelism at the post-attentive stage.

processing. Therefore, the system architecture needs to support both MP-SIMD and MIMD parallelisms in order to exploit pixel-level and feature-level parallel processing, respectively.

## IV. Proposed Multicore Architecture for Real-Time Object Recognition

### A. Overall Architecture

Fig. 5 shows the proposed NoC-based heterogeneous multicore architecture for real-time object recognition, which consists of a main processor, a visual attention engine (VAE), a matching accelerator (MA), linear array of programmable edge clusters (PECs) and an external interface. Different from a conventional massively parallel SIMD architecture [8]–[11], the linear array of $N$ simple PEs with nearest neighbor connections is equally divided into $M$ PECs, each of which

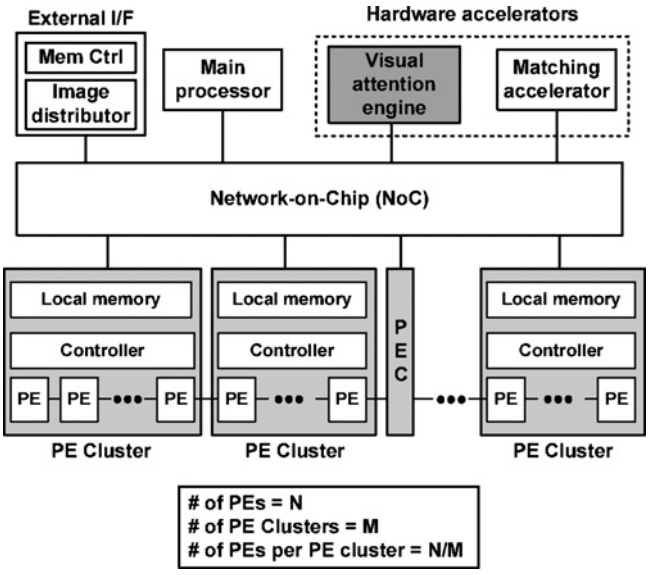



Fig. 5. Block diagram of the proposed heterogeneous multicore architecture.

contains a linear array of $N/M$ PEs and a controller to allow independent processing of each PEC. The ARM10-compatible 32-bit main processor controls the overall system operations. The VAE [21], an 80 × 60 digital cellular neural network, rapidly detects the salient image regions on the subsampled image (80 × 60 pixels) by neural network algorithms like contour and saliency map extraction. The $M$ linearly connected PECs perform data-intensive image processing applications such as image gradients and histogram calculations for further analysis of the salient image parts (i.e., the objects) provided by the VAE. The MA boosts nearest neighbor search to obtain a final recognition result in real-time [22]. The direct memory access (DMA)-like external interface controlled by the main processor distributes automatically the corresponding image data from external memory to each PEC as well as the hardware accelerators such as the VAE and the MA in order to reduce system overhead. Initially, the 2-D image plane is equally divided into $M$ PECs according to the image size specified by the main processor. A larger image area with neighboring pixels across the PEC boundary is initially assigned to each PEC for filtering operation. The NoC is used as on-chip interconnection to provide a good scalability and achieves a low-latency packet transfer between IPs by employing a dual-channel crossbar switch and adaptive circuit and packet switching [23]. Each core is connected to the NoC via a network interface.

### B. PEC Design

The PEC is an in-order, $N/M$-way SIMD processor designed to accelerate image processing tasks. Fig. 6(a) shows the block diagram of the PEC. It contains $N/M$ linearly-connected PEs controlled by a cluster controller, a cluster processing unit (CLPU), local shared memory (LSM), a LSM controller, and a PE load/store unit. The $N/M$ PEs operate in a SIMD fashion and perform image processing operations in a column-parallel (or row-parallel) manner. Each PE utilizes a 4-way very long instruction word (VLIW) architecture to execute up to four

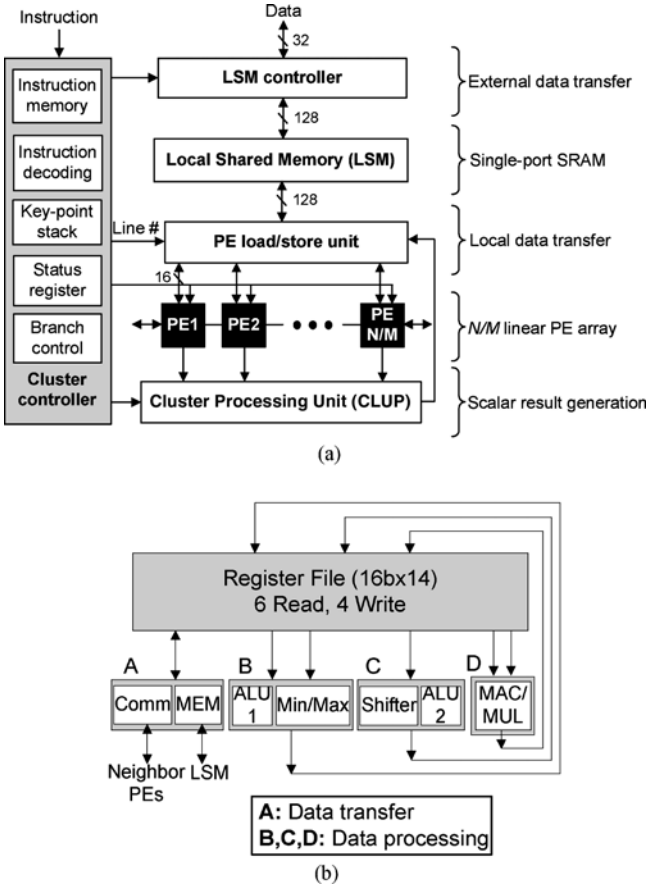



Fig. 6. Block diagram of: (a) PE cluster and (b) 4-way VLIW PE.

instructions in a single cycle as shown in Fig. 6(b). It consists of a few 16-bit data path units and a 10-port register file. The VLIW PE enables to exploit instruction-level parallelism by executing data transfer and processing instructions concurrently because memory access patterns are well predictable for many low-level image processing tasks. The CLPU, which consists of an accumulator and a comparator, generates a single scalar result from the parallel output processed by the PE array. The LSM is used as an on-chip frame memory or local memory for each PEC to store the input or processed image data and objects. The LSM controller is responsible for data transfer between external memory or other PECs and the LSM while the PE load/store unit can access the LSM only for local data transfer.

A single-port static random access memory (SRAM) is used for the LSM to reduce area overhead. The LSM provides a single-cycle access and is shared between the PE load/store unit, the LSM controller, and the CLPU. Arbitration for the LSM is performed on a cycle-by-cycle basis to improve the LSM utilization. The LSM controller is first in priority and accesses the LSM 128-bit in a single cycle, providing lots of bandwidth with little interference to the PE load and store. The LSM controller, which is an independent processing unit optimized for data transfer like the DMA engine, enables the data transfers in parallel with PE execution to hide excessive external memory latency. Normally, large intermediate data of object recognition algorithm, such as the SIFT should be stored in external memory due to limited on-chip memory capacity. Therefore, concurrent PE computation and data transfer of the PEC improve performance by minimizing external memory access penalty.

### C. Parallelization Strategy on Proposed Architecture

In this section, we describe a hybrid parallelization strategy of the attention-based object recognition algorithm on the proposed architecture (see Fig. 7). For parallelization at the pre-attentive stage, a column-wise mapping of the image to each PE is initially assumed. All $N$ linear array PEs perform the same operation on the entire image by broadcasting instructions and iterating it the number of times equal to the number of image lines, while exploiting pixel-level parallelism. The processing result is stored in a collection of all PEC's local memories. By fully utilizing the hardware resource in the MP-SIMD mode, image pre-processing operations, such as Gaussian scale-space generation, can obtain maximum performance.

Parallelization at the post-attentive stage is achieved in the MIMD mode by maintaining a key-point stack in each PEC's local memory. Given salient points by the VAE, key-points are first extracted within a certain region around the salient point and each PEC pushes extracted key-points location into a key-point stack. Then, feature-level processing is independently executed on each PEC via two steps: 1) collecting object image data around the key-point location popped from the stack, and 2) processing the corresponding image object. By repeating steps 1 and 2 until all key-point stacks are empty, feature-level recognition tasks can be performed on each PEC in parallel by sweeping through all salient image regions (i.e., objects) within the 2-D image. Each PEC also exploits $N/M$-way pixel-wise data-level parallelism to process the objects. In addition, the PEC performs the object processing and the next object data pre-fetching simultaneously, which leads to the increased amount of parallelism by improving utilization of the PEC.

As an example, we mapped the SIFT algorithm on the proposed configurable architecture. Various image processing tasks for the SIFT algorithm are divided into two parts based on the operation characteristics. Gaussian scale space and DoG image pyramid generation are performed at all pixels (i.e., PEs) in the MP-SIMD mode while exploiting $N$-way pixel-level parallelism. Then, key-point localization is performed on each PEC in MIMD mode and the final key-points are pushed in the corresponding PEC's key-point stack. Orientation assignment and SIFT descriptor computation around the key-points are also performed on each PEC while exploiting $M$-way feature-level parallelism. As a result, a hybrid parallelization strategy that combines pixel-level and feature-level parallelisms can achieve optimal parallel performance for the SIFT algorithm as well as the attention-based object recognition.

### D. SIMD/MIMD Dual-Mode Configuration

The proposed architecture supports both MP-SIMD and MIMD parallelisms on a single hardware platform with low hardware cost by adaptively selecting a switching mode of the

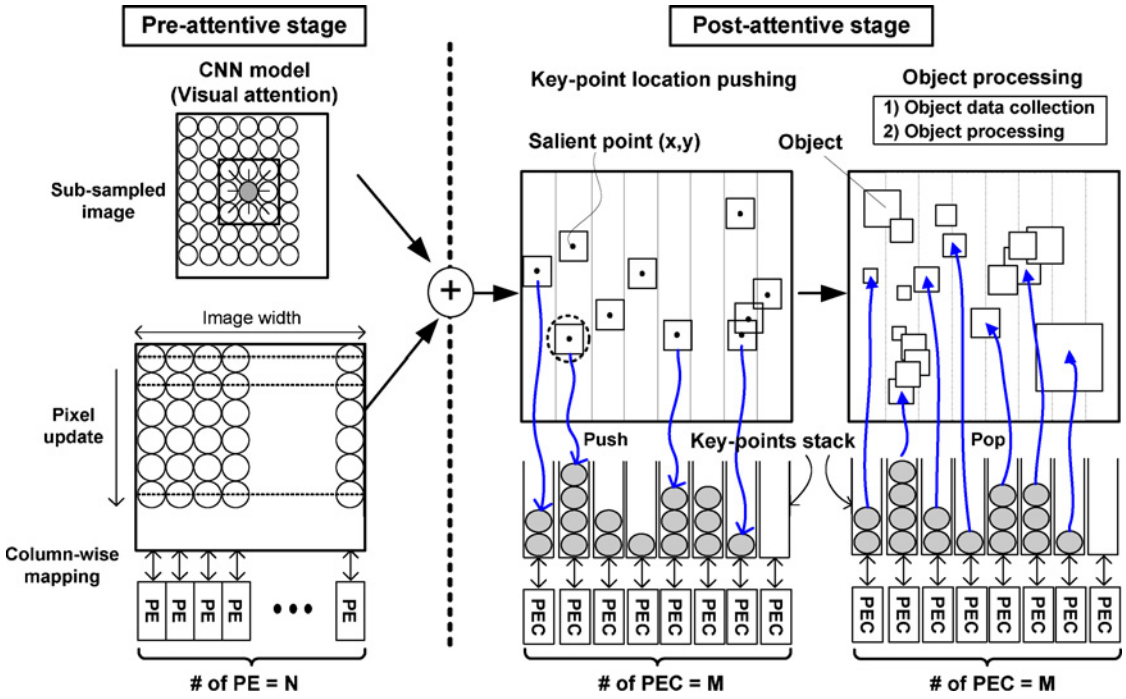



Fig. 7. Parallelization strategy of the target algorithm on the proposed architecture.

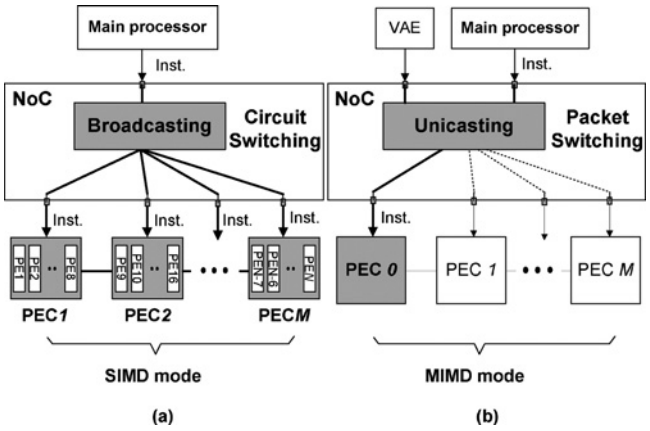



Fig. 8. Dual-mode configuration: (a) SIMD mode and (b) MIMD mode.

NoC as shown in Fig. 8. In a circuit switching NoC, the main processor broadcasts instruction and data to all PE array. In this mode, the system exploits $N$-way massively parallel SIMD operation for image pre-processing to fully take advantage of multicore processing capabilities. On the contrary, in a packet switching NoC, each PEC is responsible for the objects, each of which contains image data around the extracted key-points. In the MIMD mode, the $M$ PECs operate independently in parallel for object-parallel processing.

It takes about a few tens of cycles to change the NoC configuration depending on the network traffic status due to circuit establishment and release time overhead for the circuit switching NoC. For object recognition applications, however, the operation mode conversion occurs only twice during the recognition period of 1-frame image: SIMD to MIMD conversion after the pre-processing stage and MIMD to SIMD conversion after completing the recognition. Therefore, such a dual-mode architecture is suitable for object recognition with negligible impact on the overall system performance.

### *E. VAE Design*

To deal with complex algorithms like visual attention, analog CNN processing even achieving 6 to 7 bit accuracy is a challenge with reasonable size transistors and it is not suitable to be integrated into SoC due to analog-to-digital or digital-to-analog conversion overhead. The VAE is an $80 \times 60$ digital CNN optimized for small area and energy efficiency. Fig. 9 shows the block diagram of the VAE, which is composed of four arrays of $20 \times 60$ cells, 120 visual PEs (VPEs) shared by the cell arrays, and a controller with 2 kB instruction memory. Previous digital CNN [24] can integrate only a small number of cells due to the large size of digital arithmetic blocks. On the contrary, the VAE integrates $80 \times 60$ cells that each correspond to a pixel in an $80 \times 60$ resolution image. This is possible because the cells of the VAE only perform storage and inter-cell data transfer to minimize area while a smaller number of shared VPEs are responsible for processing the cells data. Each cell consists of two elements: four 6 T SRAM cell-based register file for data storage and 4-directional shift register for data transfer between neighboring cells. An $80 \times 60$ shift register array, distributed among the cells, eliminates data communication overhead in convolution operations of arbitrary kernel size and shape, which is the most frequently used operation in the CNN. The VAE controller generates the control signals for sequencing the operation of the cells and the VPEs.

The CNN operation on the VAE is performed by a spiraling shift sequence because the VAE has only connections to four neighbor cells in contrast to the conventional CNN hardware.

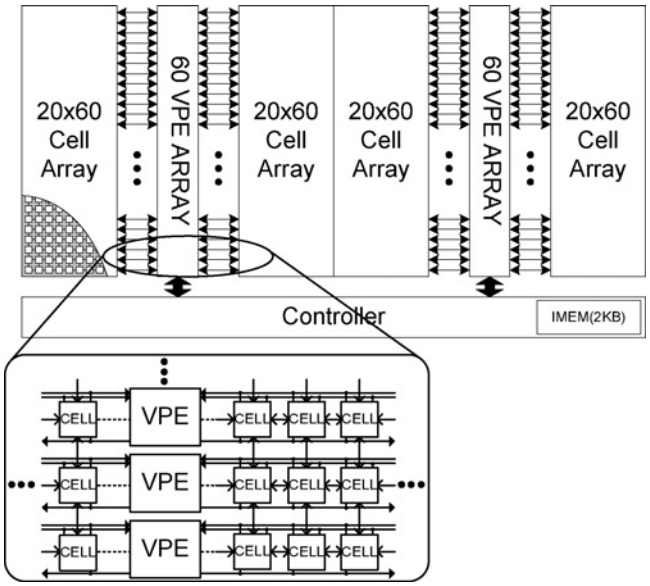



Fig. 9. Block diagram of the VAE.

Thanks to 1-cycle shift operation, it takes only 4.3 $\mu$s to complete a 3 $\times$ 3 CNN template. As a result, the VAE takes only 2.4 ms to complete a saliency map extraction of an 80 $\times$ 60 image. More details of the VAE are described in [21].

## V. Performance Analysis

### A. Simulation Setup

An architecture analysis is performed to show superiority of the proposed architecture for the attention-based object recognition algorithm. We built a cycle-accurate architecture simulator, including the PEC's instruction-set simulator and the parameterized NoC simulator, which enables co-simulation of the NoC and computation units. MP-SIMD architecture is modeled as a reference multicore architecture for the target algorithm and compared to the proposed architecture in terms of execution time. MP-SIMD paradigm has been widely adopted for vision processing because of the good match with pixel-level parallelism present in most low-level vision processing algorithms. The MP-SIMD processor provides high computation efficiency with reduced control circuit overheads. For performance analysis, we use 50 test images of 640 $\times$ 480 pixels and a database of 100 objects with about 20 000 total key-points, which are selected considering a real office environment. The computing power of the PE is assumed to be the same as that in Fig. 6(b) for both architectures. The number of PEs per PEC (*N/M*) is assumed to be 8. Total execution time is measured by the sum of the number of clock cycles at pre-attentive and post-attentive stages. The execution time of matching is not counted as the number of cycles at the post-attentive stage because it is assumed that the matching is performed on a hardware accelerator for both architectures.

### B. Performance Analysis

We analyzed the performance of the proposed architecture in terms of three key factors: the VAE, dual-mode parallelism, and workload balance.

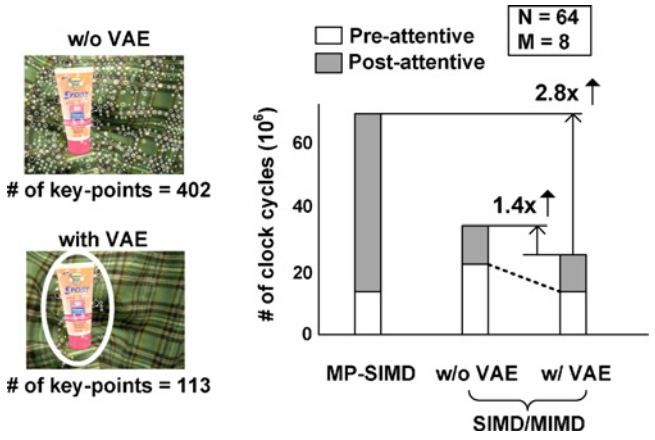



Fig. 10. Effect of the VAE on performance.

1) *Effect of the VAE:* With the help of the VAE, the number of key-points extracted is reduced as shown in Fig. 10; therefore, computation cost of higher-level tasks at the post-attentive stage is also reduced. The VAE takes only 2.4 ms to complete visual attention operation for an 80 $\times$ 60 subsampled image, which means that the overhead of the VAE is very little compared to total execution time. Although the low-resolution image is mapped on the VAE, it does not cause any loss of recognition accuracy because the role of the VAE is just to make a rough selection of the salient image regions before the detailed processing. As a result, the simulation results in case of $N = 64$ and $M = 8$ show that the proposed heterogeneous architecture, in which the VAE selects the interesting parts of the scene rapidly, achieves a 2.8 times improvement over a conventional MP-SIMD architecture, as shown in Fig. 10.

The visual attention model [15] can be implemented on a linear array of PEs in MP-SIMD mode instead of the VAE. For an 80 $\times$ 60 image, the VAE achieves about four times better performance over the programmable linear PE array because the CNN-based VAE can easily accelerate the center-surround mechanism. Fig. 10 shows the VAE contributes to 1.4 times performance improvement over the proposed architecture without the VAE by reducing the overhead of performing the visual attention on the programmable linear PE array. As a result, we can clearly show the justification of integrating the VAE on the proposed architecture.

2) *Effect of SIMD/MIMD Dual-Mode Parallelism:* Fig. 11(a) shows the execution time comparison of MP-SIMD and the proposed architecture when the number of PEs is increased. To evaluate only the effect of dual-mode parallelism on performance, the number of key-points affected by the VAE is assumed to be the same for both architectures. The MP-SIMD architecture does not perform well on the post-attentive stage because it is difficult to achieve feature-level parallelism required for the post-attentive tasks. After the pre-attentive stage, image regions of interest around the key-points may be scattered and overlapped over the entire image as shown in Fig. 4, which leads to a sequential examination of each object due to high cost of image data redistribution

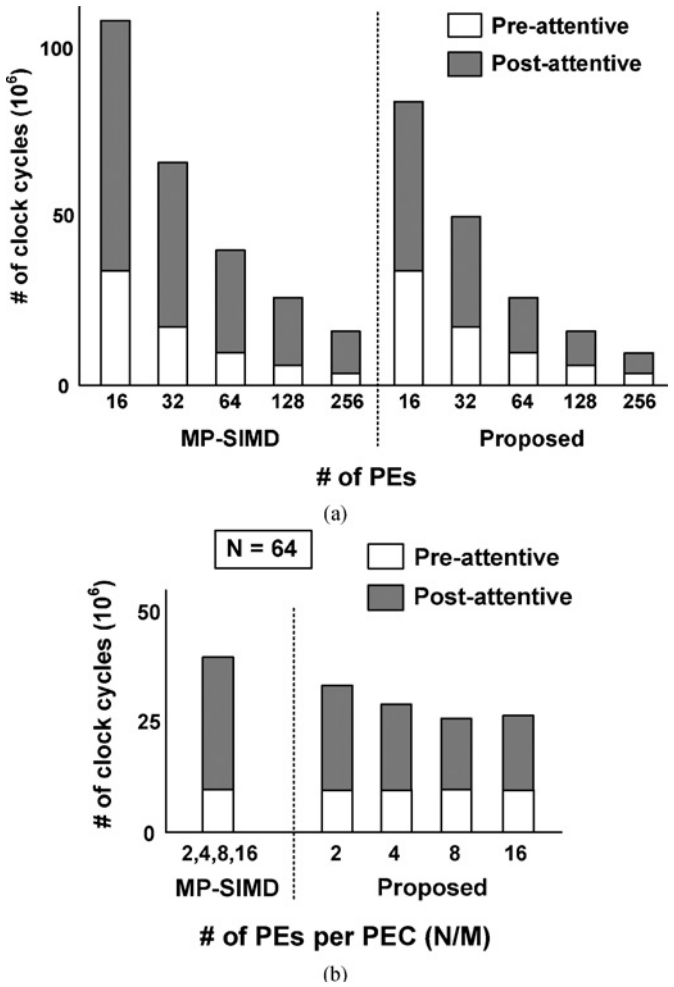



Fig. 11. Performance comparison of MP-SIMD and the proposed architecture.

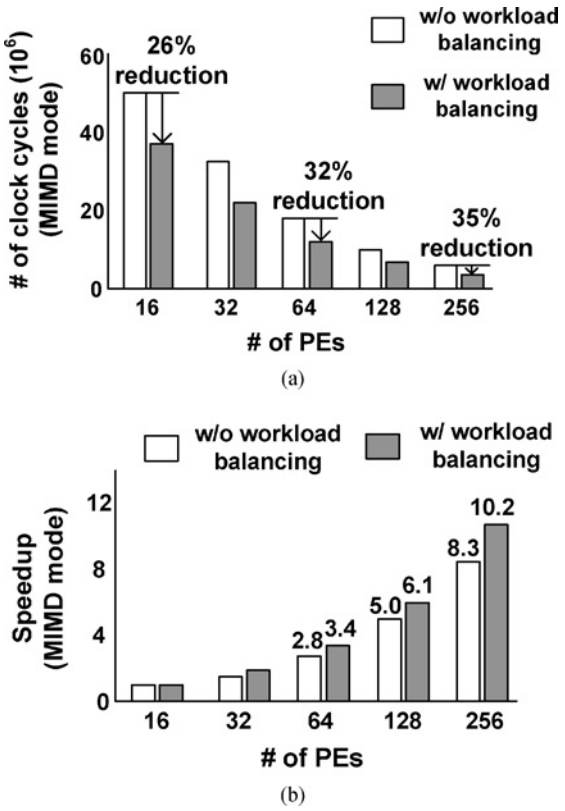



Fig. 12. (a) Performance improvement and (b) parallel speedup by the workload balancing mechanism.

on the MP-SIMD architecture. Moreover, the objects to be processed are dynamic and variable depending on the number of key-points extracted and their location on the image. Such irregular and dynamic characteristics of the algorithms make the pure SIMD architecture less effective and cannot fully utilize the computing capability of the hardware. As a result, the proposed architecture supporting dual-mode parallelism provides better performance over the conventional MP-SIMD architecture due to the increased amount of parallelism on the post-attentive stage.

To study the effect of the PEC configuration, Fig. 11(b) shows the performance comparison according to the number of PEs per PEC (i.e., N/M) when the number of PEs is assumed to be 64. The processing time on the post-attentive stage increases as the N/M is lower. This is because each PEC's performance is not large enough to deal with feature-level tasks, such as orientation assignment and descriptor generation. On the contrary, a larger number of PEs per PEC ($N/M = 16$) do not show better performance compared to the case of $N/M = 8$ due to lower available feature-level parallelism by a smaller number of PECs. As a result, it is crucial to select a suitable number of PEs per PEC on the proposed architecture considering trade-off between a single PEC's performance and the amount of feature-level parallelism.

3) *Effect of Workload Balance:* Fig. 11 shows the post-attentive tasks do not scale well on the proposed architecture while the pre-attentive tasks scale linearly. Workload imbalance is one of the most primary limitations for performance scalability on a multicore processor. According to the parallelization strategy of the algorithm (see Fig. 7), the number of extracted key-points is considered as the number of tasks to be processed on the post-attentive stage (i.e., MIMD mode). For the attention-based object recognition algorithm, the location of extracted key-points tends to be concentrated on attended image regions, which results in workload imbalance among the PECs in MIMD mode. To improve the performance in MIMD mode, workload balancing mechanism is applied to increase the utilization of the PECs. In this experiment, the main processor keeps track of the operating status of all PECs as a central controller and makes all PECs keep busy by transferring a key-point (i.e., task) from the most heavily loaded PEC to the idle PEC. After the key-point is transferred, image data redistribution around the key-point from the source PEC to the idle PEC is required. Such data redistribution can be efficiently performed by fully utilizing bandwidth provided by the NoC. As a result, Fig. 12(a) shows the post-attentive tasks achieve 32% additional performance improvement on average when applying the workload balancing mechanism. Fig. 12(b) shows the parallel speedup of the proposed architecture. The workload balance offers better scalability performance.

TABLE I
CHIP SPECIFICATIONS

| | | |
|---|---|---|
| Process | 0.13 um 1p 8M CMOS technology | |
| Die Size | 6 mm × 6 mm | |
| Power Supply | 1.2-V for core, 2.5-V for I/O | |
| Operating Frequency | 400 MHZ for Noc, 200 MHZ for IPs | |
| # of TRs (gates, memory) | 1.9 M gates, 228 KB SRAM | |
| Power Consumption | < 583 mW (for all applications) | |
| Peak Performance | 8 PE clusters | 96 GOPS |
| | VAE | 24 GOPS |
| | MA | 4.8 GOPS |
| | Main processor | 0.2 GOPS |
| | Total | 125 GOPS |
| Object Recognition Speed | 22 frame/sec @ 320 × 240 image | |

TABLE II
POWER BREAKDOWN

| Modules | Power (mW) | Power Percentage |
|---|---|---|
| 8 PECs | 392 | 67.2% |
| VAE | 84 | 14.4% |
| NoC | 46 | 7.9% |
| RISC | 14 | 2.4% |
| MA | 45 | 7.7% |
| EXT I/F | 2 | 0.4% |
| Total | 583 | 100% |

TABLE III
PROCESSING TIME EVALUATION AT 320 × 240 IMAGE

| Tasks | Processing Time | Ratio |
|---|---|---|
| Gaussian scale space + VAE | 14.1 ms | 31.1% |
| Key-point localization | 3.5 ms | 7.7% |
| Orientation assignment | 5.2 ms | 11.4% |
| Descriptor generation | 8.4 ms | 18.5% |
| Matching | 14.2 ms | 31.3% |
| Total | 45.4 ms | 100% |

## VI. PROTOTYPE CHIP IMPLEMENTATION

To demonstrate the performance of the proposed architecture, the prototype chip [25] with 64 PEs and 8 PECs is fabricated in the 0.13 $\mu$m 1P8M complementary metal–oxide–semiconductor (CMOS) process. The chip micrograph is shown in Fig. 13 and the chip specifications are shown in Table I. The chip die size is $6 \times 6$ mm$^2$, including 1.9 M gate count and 228 kB on-chip SRAM. Operating frequency of the chip is 200 MHz for the IPs and 400 MHz for the NoC. The peak performance is 125 GOPS at 200 MHz in the case of 8-bit fixed point operations in SIMD mode and a sustained performance of 42 GOPS is achieved on the target application. The power consumption is less than 600 mW at 1.2-V power supply while object recognition application is running at 22 frames/s on the QVGA (320 × 240) image. The power breakdown of the chip is shown in Table II and the processing time at each stage is shown in Table III. The tree-based topology NoC with three crossbar switches provides 76.8 GB/s aggregated bandwidth. The NoC uses the source synchronous scheme [26] in which a strobe signal is transmitted along with the packet for a timing reference at a receiver end, thus, some delay variation among the PECs in broadcasting packets is tolerable under 400 MHz operating frequency. The NoC consumes 9% of the die area and 8% of the power consumption, which means that the NoC cost is amortized over the processing units. In addition, due to the simple control circuit in the SIMD architecture, the control part of the PEC (i.e., the cluster controller including 2 kB instruction memory) occupies only 6% of the total PEC area, which results in high computation efficiency. Table IV shows the power efficiency comparison with the previous recognition processors. GOPS/W and energy/pixel are adopted as a performance index for the normalization. As a result, the prototype chip achieves up to ten times higher GOPS/W in case of 8-bit fixed-point operation and the lowest energy per pixel for object recognition task with the help of the VAE and dual-mode parallelism.

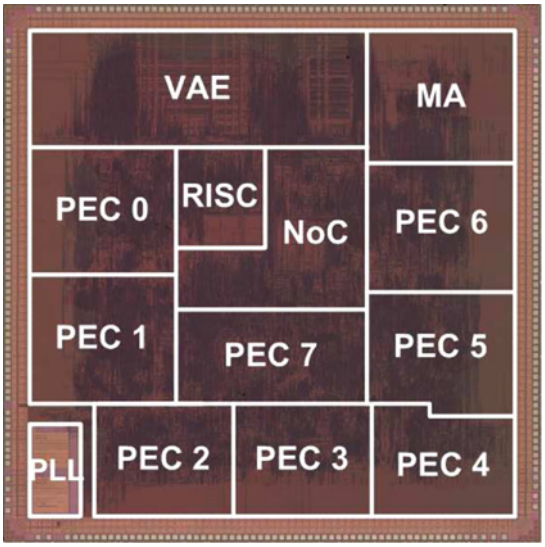



Fig. 13. Chip micrograph.

TABLE IV
POWER EFFICIENCY COMPARISON WITH THE PREVIOUS RECOGNITION PROCESSORS

| | NEC[6] | KAIST[8] | **This work** | |
|---|---|---|---|---|
| Peak performance(GOPS) | 51.2 | 81.6 | 125 | |
| Power(mW) | 2500 | 800 | 583 | |
| GOPS/W** | 20.5 | 58.3 | 214 | |
| Image resolution | 256 × 240 | 320 × 240 | 320 × 240 | 640 × 480 |
| Recognition frame rate (frames/s) | 30* | 9.3 | 22 | 8.7 |
| Energy per pixel (nJ/pixel)*** | 1356 | 1120 | 345 | 218 |

* Applied recognition algorithm is simpler than this paper.

** 1 OP = 8-bit fixed-point operation.

*** Energy per pixel = $\frac{\text{Power}}{\text{Image resolution} \times \text{Frame rate}}$.

## VII. CONCLUSION

We presented the attention-based object recognition algorithm and its novel hardware architecture based on the target algorithm analysis. The proposed configurable heterogeneous multicore architecture combines MP-SIMD/MIMD dual-mode parallel processor and cellular neural network on the NoC platform for real-time object recognition. The cellular neural network is utilized to accelerate the visual attention algorithm for selecting salient image regions rapidly. The dual-mode parallel processor performs data-intensive object recognition tasks while exploiting pixel-level and feature-level parallelisms. Hybrid parallelization strategy of the target algorithm on the proposed architecture is adopted to obtain maximum parallel performance. The performance analysis results showed the proposed architecture achieves 2.8 times speed-up over the conventional MP-SIMD architecture for the target algorithm. The prototype chip implementation demonstrates 22 frames/s real-time object recognition while dissipating less than 600 mW. The proposed architecture is targeted for the object recognition accelerator chip and can be effectively used in various embedded systems such as a mobile robot and a mobile phone for real-time object recognition.

## REFERENCES


[1] S. Se, D. Lowe, and J. Little, "Vision-based mobile robot localization and mapping using scale-invariant features," in *Proc. Int. Conf. Robotics Automation*, 2001, pp. 2051–2058.

[2] S. Ahn, M. Choi, J. Choi, and W.-K. Chung, "Data association using visual object recognition for EKF-SLAM in home environment," in *Proc. IEEE Int. Conf. Intell. Robots Syst.*, 2006, pp. 2760–2765.

[3] Z. Sun, G. Bebis, and R. Miller, "On-road vehicle detection: Review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 694–711, May 2006.

[4] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[5] V. Bonato, E. Marques, and G. A. Constantinides, "A parallel hardware architecture for scale and rotation invariant feature detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 12, pp. 1703–1712, Dec. 2008.

[6] N. Sinha, J.-M. Frahm, M. Pollefeys, and Y. Genc, "Feature tracking and matching in video using programmable graphics hardware," *Mach. Vision Appl.*, pp. 1–11, Mar. 2007.

[7] H. Feng, E. Li, Y. Chen, and Y. Zhang, "Parallelization and characterization of SIFT on multicore systems," in *Proc. IEEE Symp. Workload Characterization*, Sep. 2008, pp. 14–23.

[8] S. Kyo, T. Koga, S. Okazaki, and I. Kuroda, "A 51.2-GOPS scalable video recognition processor for intelligent cruise control based on a linear array of 128 four-way VLIW processing elements," *IEEE J. Solid-State Circuits*, vol. 38, no. 11, pp. 1992–2000, Nov. 2003.

[9] A. Abbo, R. Kleihorst, V. Choudhary, L. Sevat, P. Wielage, S. Mouy, and M. Heijligers, "XETAL-II: A 107 GOPS, 600 mW massively-parallel processor for video scene analysis," in *Proc. IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, 2007, pp. 270–271.

[10] H. Noda, M. Nakajima, K. Dosaka, K. Nakata, M. Higashida, O. Yamamoto, K. Mizumoto, T. Tanizaki, T. Gyohten, Y. Okuno, H. Kondo, Y. Shimazu, K. Arimoto, K. Saito, and T. Shimizu, "The design and implementation of the massively parallel processor based on the matrix architecture," *IEEE J. Solid-State Circuits*, vol. 42, no. 1, pp. 183–192, Jan. 2007.

[11] S. Arakawa, Y. Yamaguchi, S. Akui, Y. Fukuda, H. Sumi, H. Hayashi, M. Igarashi, K. Ito, H. Nagano, M. Imai, and N. Asari, "A 512GOPS fully-programmable digital image processor with full HD 1080p processing capabilities," in *Proc. IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2008, pp. 312–313.

[12] D. Kim, K. Kim, J.-Y. Kim, S. Lee, and H.-J. Yoo, "An 81.6 GOPS object recognition processor based on NoC and visual image processing memory," in *Proc. Custom Integr. Circuits Conf.*, 2007, pp. 443–446.

[13] M. I. Posner and S. E. Petersen, "The attention system in human brain," *Annu. Rev. Neurosci.*, vol. 13, pp. 25–42, Mar. 1990.

[14] D. Heinke, G. W. Humphreys, "Computational models of visual selective attention: A review," in *Connectionist Models in Psychology*, G. Houghton, Ed. East Sussex, U.K.: Psychology Press, 2005, pp. 273–312.

[15] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[16] L. O. Chua and L. Yang, "Cellular neural network: Theory and applications," *IEEE Trans. Circuit Syst.*, vol. 35, no. 10, pp. 1257–1290, 1988.

[17] G. Costantini, D. Casali, and M. Carota, "CNN-based unsupervised pattern classification for linearly and non linearly separable data sets," *Trans. Circuits Syst.*, vol. 4, no. 5, pp. 448–452, May 2005.

[18] A. Kananen, A. Paasio, M. Laiho, and K. Halonen, "CNN applications from the hardware point of view: Video sequence segmentation," *Int. J. Circuit Theory Appl.*, vol. 30, nos. 2–3, pp. 117–137, Mar. 2002.

[19] B. E. Shi, "Gabor-type filtering in space and time with cellular neural networks," *IEEE Trans. Circuits Syst. I: Fundamental Theory Appl.*, vol. 45, no. 2, pp. 121–132, Feb. 1998.

[20] J. A. Nossek, "Design and learning with cellular neural networks," *Int. J. Circuit Theory Appl.*, vol. 24, no. 1, pp. 15–24, Jan. 1996.

[21] S. Lee, K. Kim, M. Kim, J.-Y. Kim, and H.-J. Yoo, "The brain mimicking visual attention engine: An 80 × 60 digital cellular neural network for rapid global feature extraction," in *Proc. IEEE Symp. Very-Large-Scale Integr.*, 2008, pp. 26–27.

[22] J.-Y. Kim, K. Kim, S. Lee, M. Kim, and H.-J. Yoo, "A 66 fps 38 mW nearest neighbor matching processor with hierarchical VQ algorithm for real-time object recognition," in *Proc. IEEE Asian Solid-State Circuits Conf.*, 2008, pp. 177–180.

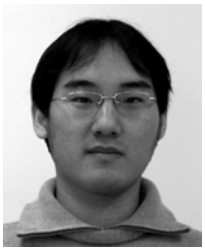
[23] K. Kim, J.-Y. Kim, S. Lee, M. Kim, and H.-J. Yoo, "A 76.8 GB/s 46 mW low-latency network-on-chip for real-time object recognition processor," in *Proc. IEEE Asian Solid-State Circuits Conf.*, 2008, pp. 189–192.

[24] P. Keresztes, A. Zarándy, T. Roska, P. Szolgay, T. Bezák, T. Hidvégi, P. Jónás, and A. Katona, "An emulated digital CNN implementation," *J. VLSI Signal Process. Syst.*, vol. 23, nos. 2–3, pp. 291–303, Nov.–Dec. 1999.

[25] K. Kim, S. Lee, J.-Y. Kim, M. Kim, D. Kim, J.-H. Woo, and H.-J. Yoo, "A 125GOPS 583 mW network-on-chip-based parallel processor with

bio-inspired visual attention engine," in *Proc. IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, 2008, pp. 523–524.

[26] K. Lee, S.-J. Lee, and H.-J. Yoo, "Low-power networks-on-chip for high-performance soc design," *IEEE Trans. VLSI Syst.*, vol. 14, no. 2, pp. 148–160, Feb. 2006.

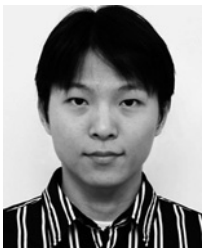


**Kwanho Kim** (S'04) received the B.S. and M.S. degrees in electrical engineering and computer science from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2004 and 2006, respectively. He is currently working toward the Ph.D. degree in electrical engineering and computer science from the Division of Electrical Engineering, Department of Electrical Engineering and Computer Science, KAIST.

In 2004, he joined the Semiconductor System Laboratory, KAIST, as a Research Assistant. His research interests include VLSI design for object recognition, architecture, and implementation of NoC-based SoC.

**Seungjin Lee** (S'06) received the B.S. and M.S. degrees in electrical engineering and computer science from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2006 and 2008, respectively. He is currently working toward the Ph.D. degree in electrical engineering and computer science from the Division of Electrical Engineering, Department of Electrical Engineering and Computer Science, KAIST.

His previous research interests included low power digital signal processors for digital hearing aids and body area communication. Currently, he is investigating parallel architectures for computer vision processing.



**Joo-Young Kim** (S'05) received the B.S. and M.S. degrees in electrical engineering and computer science from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2005 and 2007, respectively. He is currently working toward the Ph.D. degree in electrical engineering and computer science from the Division of Electrical Engineering, Department of Electrical Engineering and Computer Science, KAIST.

He has been with the Division of Electrical Engineering, Department of Electrical Engineering and Computer Science, KAIST. Since 2006, he has been involved with the development of parallel processors for computer vision. Currently, his research interests include parallel architecture and sub-block design for computer vision systems.

**Minsu Kim** (S'07) received the B.S. degree in electrical engineering and computer science from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2007. He is currently working toward the M.S. degree in electrical engineering and computer science from the Division of Electrical Engineering, Department of Electrical Engineering and Computer Science, KAIST.

His research interests include network-on-chip-based SoC design and VLSI architecture for computer vision processing.

**Hoi-Jun Yoo** (M'95–SM'04–F'08) received the B.S. degree from the Department of Electronics, Seoul National University, Seoul, Korea, in 1983, and received the M.S. and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 1985 and 1988, respectively. His Ph.D. work concerned the fabrication process for gallium arsenide vertical optoelectronic integrated circuits.

From 1988 to 1990, he was with Bell Communications Research, Red Bank, NJ, where he invented the 2-D phase-locked vertical cavity surface emitting laser array, the front-surface-emitting laser, and the high-speed lateral heterojunction bipolar transistor. In 1991, he became the Manager of the DRAM Design Group, Hyundai Electronics, and designed a family from fast-1 M DRAMs and 256 M synchronous DRAMs. In 1998, he joined the Faculty of the Department of Electrical Engineering, KAIST, where he is now a Full Professor. From 2001 to 2005, he was the Director of System Integration and IP Authoring Research Center, funded by the Korean Government to promote worldwide IP authoring and its SoC application. From 2003 to 2005, he was the Full Time Advisor to the Minister of the Korea Ministry of Information and Communication, and National Project Manager for SoC and Computer. In 2007, he founded the System Design Innovation and Application Research Center, KAIST, to research and develop SoCs for intelligent robots, wearable computers, and bio systems. He is the author of the books *DRAM Design* (Seoul, Korea: Hongleung, 1996; in Korean), *High Performance DRAM* (Seoul, Korea: Sigma, 1999; in Korean), and chapters of *Networks on Chips* (New York, Morgan Kaufmann, 2006). His current interests include high-speed and low-power networks on chips, 3-D graphics, body area networks, biomedical devices and circuits, and memory circuits and systems.

Dr. Yoo received the Electronic Industrial Association of Korea Award for his contribution to DRAM technology in 1994, the Hynix Development Award in 1995, the Korea Semiconductor Industry Association Award in 2002, the Best Research of KAIST Award in 2007, the Design Award of 2001 Asia and South Pacific Design Automation Conference, and the Outstanding Design Awards for the 2005, 2006, and 2007 Asian Solid-State Circuits Conference (A-SSCC). He is a Member of the Executive Committee of International Solid-State Circuits Conference, the Symposium on VLSI, and A-SSCC. He is the Transaction Processing Performance Council Chair of the A-SSCC 2008.