Low-Power, Real-Time Object-Recognition Processors for Mobile Vision Systems

A new low-power object-recognition processor achieves real-time robust recognition, satisfying modern mobile vision systems' requirements. The authors introduce an attention-based object-recognition algorithm for energy efficiency, a heterogeneous multicore architecture for data- and thread-level parallelism, and a network on a chip for high on-chip bandwidth. The fabricated chip achieves 30 frames/second throughput and an average 320 mW power consumption on test 720p video sequences, yielding 640 GOPS/W and 10.5 nJ/pixel energy efficiency.

Jinwook Oh
Gyeonghoon Kim
Injoon Hong
Junyoung Park
Seungjin Lee
Joo-Young Kim
Jeong-Ho Woo
Hoi-Jun Yoo
Korea Advanced Institute of Science and Technology

In recent years, object recognition has been widely adopted in various real-life applications. Microsoft's Kinect uses body-part recognition as a gaming interface, and automakers such as Toyota and BMW incorporate vehicle, pedestrian, and lane detection in their advanced driver-assistance systems. Smartphones that operate within a low power budget also use object recognition for booming applications such as augmented reality, face-recognition-based security, and gesture-recognition-based user interfaces.

In such applications, the Scale Invariant Feature Transform (SIFT) is the most popular candidate for how to extract some interest points out of the objects and describe them in a way that is robust against changes in translation, scaling, and rotation.[1] It has also proven to be one of the most robust among local invariant feature descriptors with respect to geometric changes, thanks to its invariant region detector and orientation-distribution-based descriptor. However, SIFT-based object recognition consumes a lot of power because of the heavy computations required in descriptor generation and matching.[2] In addition, today's high-resolution image sensors and tight power budgets make real-time SIFT implementation in mobile devices even harder; recent mobile cameras provide more than 720p resolution at 30 frames per second (fps), while the power consumption on mobile CPUs and GPUs ranges from roughly 0.1 W to 1 W. For this reason, until now, SIFT-based object-recognition applications in the mobile domain have only used Quarter Video Graphics Array (QVGA [320 × 240]) or Video Graphics Array

Published by the IEEE Computer Society

Table 1. BONE-V (Basic On-Chip Network—Vision) architectures.

Architecture	Processor	Die area, gate, SRAM	Target system, specifications	Performance, dissipated power	Attention model	Key technology
BONE-V1	0.18 μm 1P6M	38.5 mm^2, 0.8 M, 34 Kbytes	Robot vision QVGA, 16 frames per second (fps)	81.6 GOPS, 1.08 W	N/A	Memory-centric network on chip + 10 SIMD processing elements
BONE-V2	0.13 μm 1P6M	36 mm^2, 1.9 M, 228 Kbytes	Robot vision QVGA, 22 fps	125 GOPS, 583 mW	Single object	SIMD/MIMD reconfigurable processing elements
BONE-V3	0.13 μm 1P8M	49 mm^2, 3.73 M, 396 Kbytes	Car vision VGA, 60 fps	201.4 GOPS, 496 mW	Multiple-object perception	16-SIMD-chip multiprocessor
BONE-V4	0.13 μm 1P8M	50 mm^2, 2.92 M, 612 Kbytes	HMD VGA, 30 fps	228 GOPS, 345 mW	Unified visual attention model	Heterogeneous multicore
BONE-V5	0.13 μm 1P6M	32 mm^2, 2.4 M, 385 Kbytes	UAV 720p HD, 30 fps	342 GOPS, 320 mW	Context-aware visual attention model	SMT-enabled multicore

* GOPS: giga operation per second; HMD: head-mounted display; MIMD: multiple instruction, multiple data; QVGA: Quarter Video Graphics Array; SIMD: single instruction, multiple data; SMT: simultaneous multithreading; SRAM: static RAM; UAV: unmanned aerial vehicle; VGA: Video Graphics Array.

(VGA [640 × 480]) video at lower frame rates ranging from 2 to 10 fps, and they are limited to simple detection or instance tracking due to a lack of computing power and a low power budget.

To realize real-time SIFT-based object recognition that meets these high resolution and low power requirements, we propose an object-recognition processor using an attention-based recognition algorithm for energy efficiency, a heterogeneous multicore architecture for data and thread parallelism, and network-on-chip (NoC) communications for high bandwidth. The processor determines regions of interest (ROIs)—the parts of the image that likely contain target objects—which lets us perform the main recognition on only the selected regions, minimizing unnecessary computations. The heterogeneous multicore architecture provides several types of parallelism and so achieves high throughput and low power consumption for highly parallelizable recognition processing. The high-bandwidth NoC plays a role as the communications backbone for tens of processing cores while meeting the high-resolution video sequence's streaming demand of more than a few hundred megabytes per frame.

On the basis of these key three design ideas, we developed the BONE-V (Basic On-Chip Network—Vision) vision processors for different target applications (see Table 1). They have evolved to process higher-resolution video streams with better fidelity, throughput, and energy efficiency for more complex and noisy environments. From the early generation of simple, homogeneous single-instruction, multiple-data (SIMD) processors, the BONE-V architectures have exploited accurate attention models to increase the target systems' energy efficiency, and have employed different types of parallel processing elements for the main recognition pipeline. The NoC enabled the vision processor to implement flexible configurations of the several parallelization technologies: a chip multiprocessor (CMP), a combination of SIMD and multiple-instruction, multiple-data (MIMD) cores, and multithreading.

Attention-based object recognition

Unlike conventional SIFT-based recognition implementations, the BONE-V series

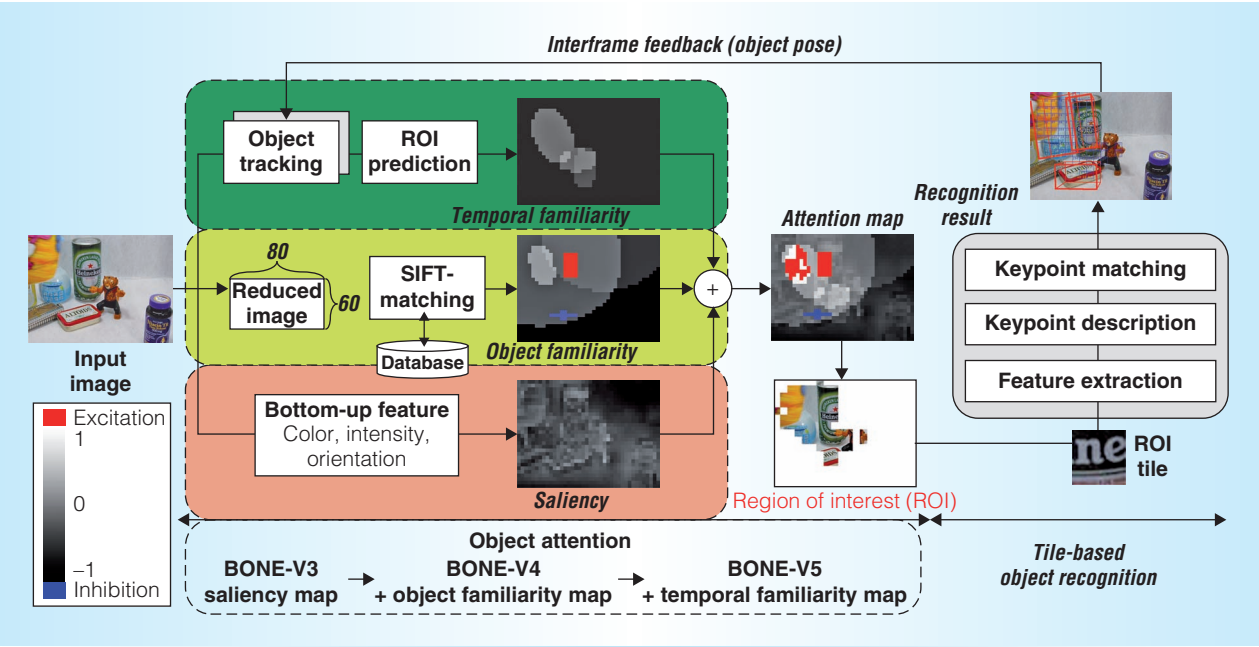

Figure 1. Flow diagram of the proposed attention-based object-recognition algorithm. The algorithm consists of two main stages, object attention and tile-based recognition. (SIFT: Scale Invariant Feature Transform.)

has adopted an attention-based recognition approach that extracts the ROI tiles from the input image to increase the processing throughput by executing multiple recognition threads on different tiles in parallel. It also reduces the required computation by eliminating background tiles that don't have any object features. The algorithm mainly consists of two stages, object attention and tile-based recognition, as Figure 1 shows. Because only the ROI tiles can be processed for the later stage, obtaining robust attention clues in the attention stage is critical to achieving high recognition accuracy. To do this, BONE-V3 uses bottom-up conspicuity information, such as color, intensity, and orientation, as attention clues to map the salient regions in the input image.[3] BONE-V4 additionally integrates object familiarity, measuring the similarity between the query object and target objects in a database by quick initial recognition with a reduced image size.[4]

As mobile vision processors extend their applications into dynamic environments with higher video resolution, however, they become more vulnerable to background clutter and distracting objects, and to dynamic noise such as occlusion, illumination, and motion blur. To mitigate these problems, BONE-V5 exploits the temporal familiarity,[5] which measures the temporal coherence of consecutive frames by tracking the recognized object in previous images to predict the next frame's ROI irrespective of dynamic noises. BONE-V5's proposed attention algorithm, called the *context-aware visual attention model* (CAVAM), integrates the saliency map, object familiarity, and temporal familiarity simultaneously; thus, BONE-V5 can obtain 50.4 percent attention accuracy, which is 1.44× higher than BONE-V4's attention model, and can reduce recognition complexity by 16 percent, on average, for dynamic object recognition with high-definition (HD) video streams.

The tile-based recognition consists of feature extraction, which includes feature detection and keypoint description, and feature matching. Feature detection performs pixel-level image processing with different types of image kernels for the input image and detects salient blobs as object keypoints. Because a chip bandwidth of approximately 100 Gbytes/second is required to process an entire 720p image, the system also requires a highly parallel data path in the feature-detection stage. Keypoint description

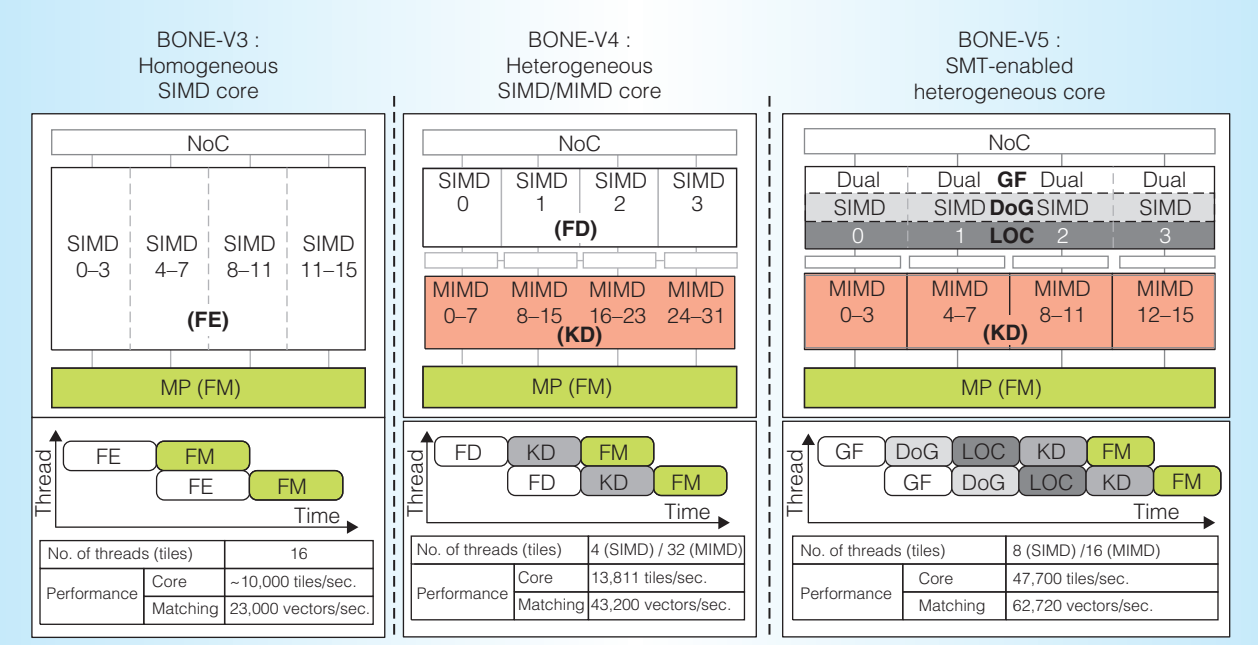

Figure 2. Evolution of the BONE-V multicore architecture. The instruction-level parallelism (ILP), data-level parallelism (DLP), and thread-level parallelism (TLP) have been increased to support accurate object recognition with higher video resolution. (DoG: difference of Gaussian; FD: feature detection; FE: feature extraction; FM: feature matching; GF: Gaussian filtering; KD: keypoint description; LOC: localization; MP: matching processor.)

generates a 128-dimensional keypoint vector that encrypts the detected keypoints' orientation and magnitude. Feature detection requires high bandwidth and high parallelism of the data path at the same time. Otherwise, keypoint description only requires a small bandwidth but still requires a high GOPS throughput.

At the feature-matching stage, the system compares the generated keypoint vector to those for target objects in the database to find the nearest-neighbor keypoint vectors among the keypoint database of target objects to identify the query object in the image. This stage requires at least 5.4 Gbytes of external bandwidth to access the external database for matching.

In the original SIFT implementation, all three stages must be realized within 33.3 ms for a full-frame-rate video stream because of the mobile vision platform's limited image buffer size. However, thanks to the attention- and tile-based recognition operation, 3.2 ms of additional attention latency reduces 41.1 percent of processing tiles, and tile-based parallel processing increases the recognition throughput by tile- (thread-) level parallelism proportional to the number of parallel processing cores.

Object-recognition processors

To realize tile-based recognition with increased data-level parallelism (DLP) and thread-level parallelism (TLP), we proposed SIFT's task-level recognition pipeline for higher throughput and system utilization. We implemented 30 to 50 different parallel processing cores for each pipeline stage with optimized instruction-level parallelism (ILP) and DLP, and interconnected them through NoC, which features low latency and high throughput. We also modified each core's microarchitecture to support more threads and data in parallel with increased instruction per cycle (IPC) and utilization.

To accelerate each SIFT algorithm operation, we proposed using several multicore processors with optimized data parallelism and pipeline architecture (see Figure 2). BONE-V3's homogenous SIMD architecture used 16 very-long-instruction-word

(VLIW)-based SIMD processing elements for the feature-extraction stage. With the subsequent feature-matching stage executed in the matching processor, it comprised a two-stage task-level pipeline of SIFT-based recognition. In BONE-V3, because one SIMD processing element performs pixel-level parallel operations as well as cascade keypoint vector description with different configurations of its eight-lane SIMD, we adopted a multicasting NoC to modify the network bandwidth for each SIFT operation's traffic characteristics. However, this architecture only provides low SIMD throughput (about 10,000 tiles/second) owing to its long cascaded feature-detection and keypoint-description operations, and 23,000 vectors/second matching throughput owing to its simple architecture with no consideration of data compression and prefetch techniques.

We implemented the heterogeneous multicore BONE-V4, consisting of four SIMD cores and 32 MIMD processing elements, with a three-stage task-level pipeline for higher throughput. We optimized each SIMD and MIMD processing element to the feature-detection and keypoint-description stages, respectively; and, for the feature matching, we adopted the Huffman compression to reduce the external bandwidth requirement. Furthermore, the heterogeneous multicores are interconnected by a hierarchical NoC and perform tile-based recognition within a feature-extraction cluster containing one SIMD processing element and eight MIMD processing elements in parallel. The proposed three-stage task-level pipeline achieves approximately 1.5× the computing power of the previous architecture. However, it suffers from performance degradation due to comparatively long SIMD operation delay and low processing-element utilization, and its 13,811 tiles/second and 43,200 vectors/second throughput remains insufficient for object recognition on high-resolution video streams.

In BONE-V5, we propose a simultaneous multithreading (SMT)-enabled multicore architecture to increase not only system throughput but also high processing-element utilization. We propose the five-stage fine-grained pipeline to accelerate the overall processing speed of feature extraction and feature matching for a 16 × 16 pixel image ROI tile. To resolve BONE-V4's low SIMD data path utilization and throughput, which reduce feature-detection throughput, we divide the previous feature-detection stage into three more fine-grained stages—namely, Gaussian filtering (GF), difference of Gaussian (DoG), and localization (LOC)—which are realized with different special functional units (SFUs) of the SIMD processing elements. The MIMD processing element and feature-matching processor (FMP) perform keypoint description and feature matching, respectively, as in the previous architecture, but they increase throughput with utilization control and keypoint hashing and caching. In addition, the SMT technology is adopted to the SIMD processing element so that it performs GF, DoG, and LOC operations for two different ROIs simultaneously and doubles the SIMD processing element throughput, thereby not only obtaining approximately 80 Gbytes/second memory bandwidth but also increasing pipeline throughput by 3.45×. Coping with the five-stage task-level pipeline, we designed BONE-V5's NoC to control its router bandwidth in weighted round-robin fashion to obtain high bandwidth utilization for seamless SMT realization. Thanks to the fine-grained task-level pipeline and SMT-enabled multicore, we can obtain 47,700 tiles/second and 62,720 vectors/second SMT-enabled feature-extraction cluster (SFEC) cores and matching processor throughputs, respectively, for high-resolution image-based advanced object-recognition applications.

Although this architecture is targeted at SIFT-based recognition, it can also apply to different algorithms such as speeded-up robust feature (SURF) and binary robust invariant scalable keypoints (BRISK)-based recognition. The attention stage applies to any other algorithm as a preprocessing step for the ROI decision. In addition, because most of the local feature-point-based recognition consists of feature-detection, keypoint-description, and feature-matching stages in common, we can realize and

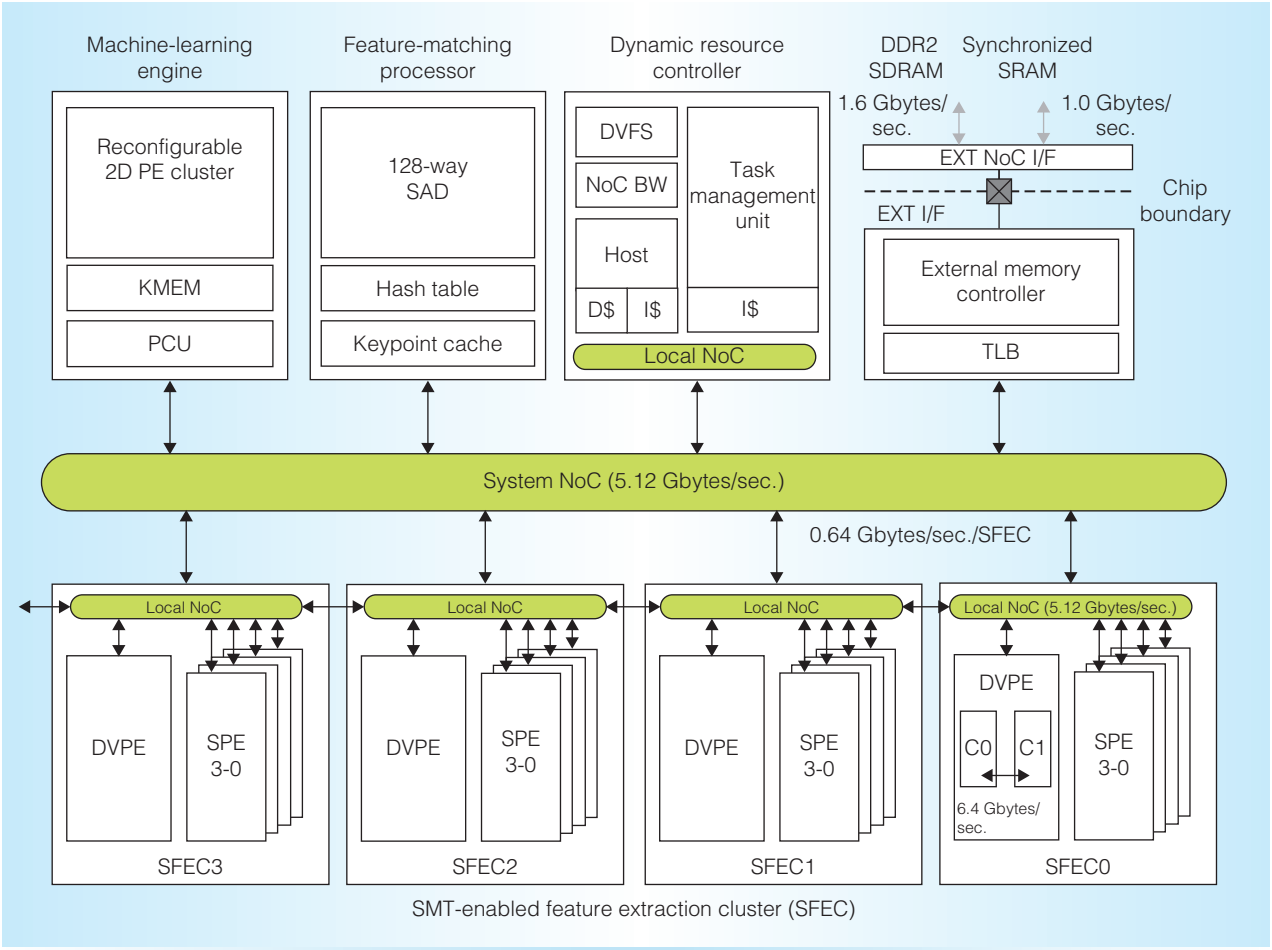

Figure 3. Top architecture of the SMT-enabled heterogeneous multicore processor. The BONE-V5 object-recognition processor uses a low-power heterogeneous multicore architecture. (BW: bandwidth; D$: data cache; DDR2: double-data rate; DVFS: dynamic voltage and frequency scaling; DVPE: dual-vector processing element; I$: instruction cache; I/F: interface; KMEM: kernel memory; NoC: network on a chip; PCU: processor control unit; PE: processing element; SAD: sum-of-absolute-difference unit; SDRAM: synchronous DRAM; SPE: scalar processing element; TLB: translation look-aside buffer.)

pipeline other recognition algorithms by programming SIMD, MIMD, and MP cores, respectively.

BONE-V5: SMT-enabled heterogeneous multicore processor

Figure 3 depicts the details of the BONE-V5 object-recognition processor, which uses a low-power heterogeneous multicore architecture with SMT operation. BONE-V5 consists of the main processing cores, a dynamic resource controller (DRC), and an external interface. The main processing cores include the throughput-optimized SFEC, the latency-optimized FMP, and the power-optimized machine learning engine (MLE) for different SIFT operations.

Once an ROI tile is allocated to an SFEC consisting of one dual-vector processing element (DVPE) and four scalar processing elements (SPEs), the SFEC performs the feature-detection and keypoint-description operations to generate the SIFT descriptors. The FMP compares the descriptors with those of the target objects in the database through the external interface. The MLE performs the attention operation by generating saliency and familiarity maps, and performing the machine-learning operation for dynamic resource management of DRC. The hardware-level resource control, based

on machine-learning techniques, changes the cores' thread allocation and operating voltage and frequency dynamically. A total of 31 cores are interconnected by the hierarchical star-ring network so that the processor's total aggregate bandwidth reaches 83.3 Gbytes/second to support real-time attention-based recognition on 720p HD video streams. The hierarchical star-ring network contains its 8 × 8 top NoC routers with 5.12 Gbytes/second bandwidth optimized for ROI tile and keypoints transaction between the cores and the external memories, the local NoC with 5.12 Gbytes/second throughput providing high bandwidth for feature-detection and keypoint-description operations among SFECs, and the 6.4 Gbytes/second internal communications channel of SFEC for high data-path utilization.

Throughput-optimized SFEC

The SFEC aims to maximize the throughput of SIFT feature extraction for an ROI tile. To this end, it achieves

- high energy efficiency by reducing as many instruction overheads as possible;
- memory locality with high bandwidth utilization; and
- increased processing throughput by ILP, DLP, and fine-grained TLP.

Each SFEC consists of a dual-vector processing element—the SMT-enabled 16-lane SIMD processing element—to exploit the SIMD's DLP and the SMT's TLP simultaneously for the feature-detection task, plus four SPEs—the MIMD processing element—to exploit the keypoint-description task's task-level parallelism (see Figure 4a). To support two ROI tiles per core and eight ROI tiles in four SFECs maximally, the SMT is integrated for the SIMD core with increased system utilization. The SMT only imposes a 12 percent higher hardware cost for register files and isolated memory architecture, and achieves at least a 30-percent processing delay reduction with dual-thread operation based on high memory locality.

Latency-optimized FMP

The FMP, depicted in Figure 4b, is a latency-optimized core that performs the feature-matching operation with good single-thread performance and a minimum off-chip bandwidth. Because more than 80 percent of the matching delay and 55 percent of the FMP power are consumed for the keypoint load/store of the external database, it is critical to minimize the keypoint access to optimize the core latency and extra power dissipation.

To this end, we propose cache- and database-based matching for FMP. Cache matching uses the keypoints that were used to recognize the target object in the previous frame. Matching the cache's hit keypoints reduces external accesses by 98 percent. Otherwise, for the missed keypoints, the processor performs additional database matching on the basis of the proposed locality-sensitive hashing index to access the database's candidate keypoints. The system generates the new hash index from the locality-sensitive hash with a random permutation that removes redundant zero bits, thereby achieving up to 64.1 percent reduction in hash bin size compared to the conventional hashing algorithm. With the help of the new hash index, the database matching reduces access 86 percent over brute-force matching. With the proposed FMP, it only takes 5.53 ms to fetch all the keypoint vectors of the hash bins for one image, sufficient for real-time feature matching of 30 fps recognition.

Power-optimized MLE

BONE-V5's last core architecture is the power-optimized MLE, which generates saliency and object and temporal familiarity for attention as well as machine learning algorithms for DRC. Those algorithms require real-number representation to provide robust results. Otherwise, the MLE uses a fixed-point arithmetic system in the data path, which integrates simple integer arithmetic circuits to reduce dynamic power consumption. At the same time, to maintain the high fidelity of conventional recognition, it uses the 4 × 4 reconfigurable processing element (RPE) arrays with coarse granularity for different parallelism and bit resolution. The RPE contains four 8-bit-resolution processing elements, which includes a shifter, a multiplier, and an arithmetic logic unit

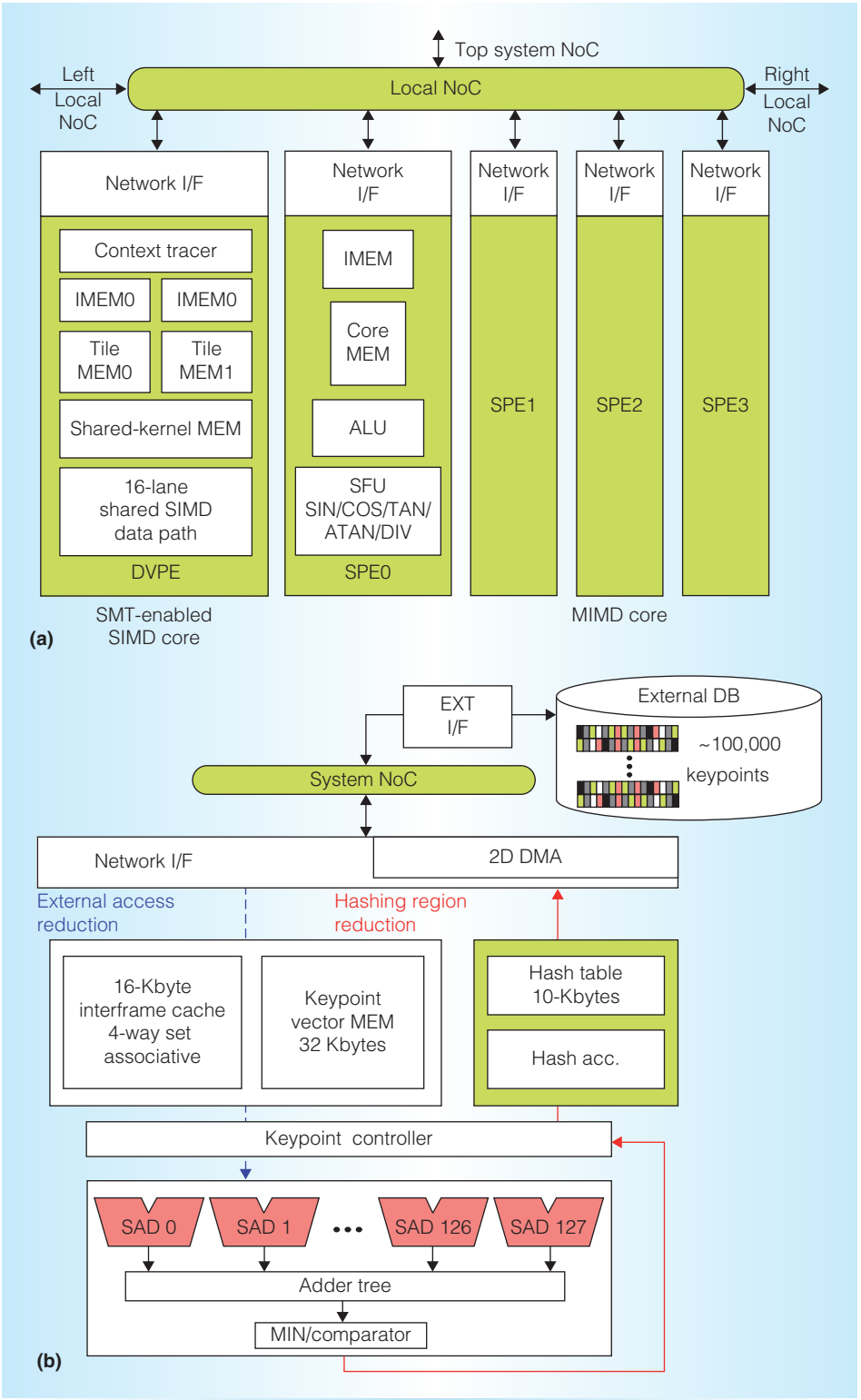

Figure 4. The proposed heterogeneous processing cores with different functionality and parallelism for the five-stage recognition algorithm. Throughput-optimized SMT-enabled feature-extraction cluster (SFEC) (a); latency-optimized matching processor (FMP) (b). (ALU: arithmetic logic unit; DMA: direct memory access; IMEM: instruction memory; MEM: memory; SFU: special functional units.)

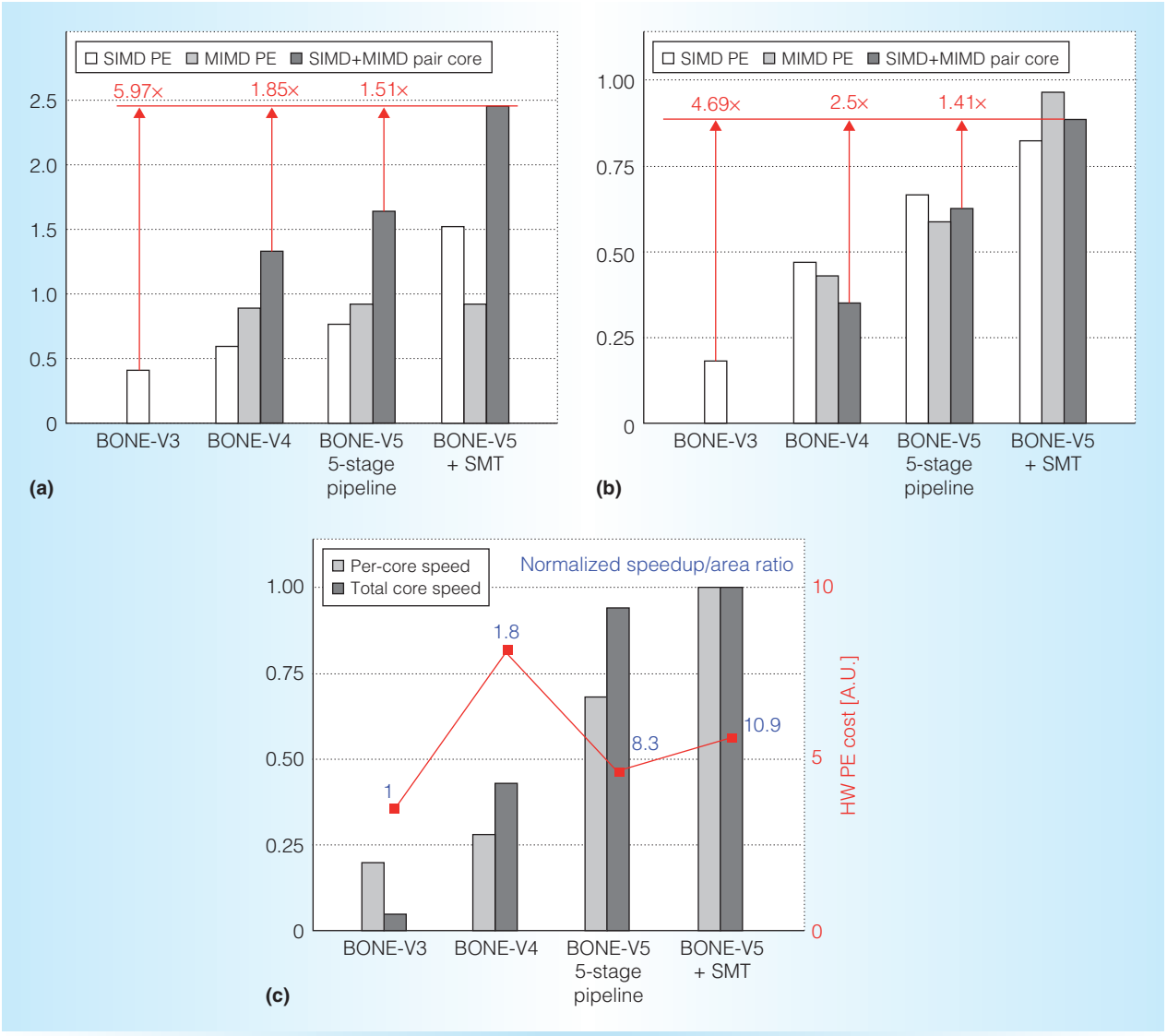

Figure 5. Performance comparison of SMT-enabled BONE-V5 with previous BONE-V architecture. Instruction per cycle (IPC) (a), utilization (b), and processing speed over hardware cost (c) when performing tile-based SIFT feature extraction.

(ALU); and the four processing elements also can be reconfigured from 8-bit-resolution pixel-level operations to 32-bit-resolution complex numerical operations. As a result, different 8-, 16-, 24-, and 32-bit-resolution processing elements are employed with different MLE parallelism levels, such as 16 to 64 ALUs, for CAVAM and DRC. It drastically eliminates up to 71 percent of the unnecessary power dissipation when performing only the reinforcement learning algorithm with 38.1 mW.

Parallel processing cores

Figure 5 summarizes the IPC, utilization, and cost-normalized average speedup of different types of cores in BONE-V generations when performing SIFT's feature-detection, keypoint-description, and feature-extraction operations, respectively. In BONE-V5, the IPC and utilization of the 16-lane SIMD processing element are increased by utilizing not only the fine-grained pipeline but also the SMT operation based on the pipelined SIMD data path. In addition, to minimize

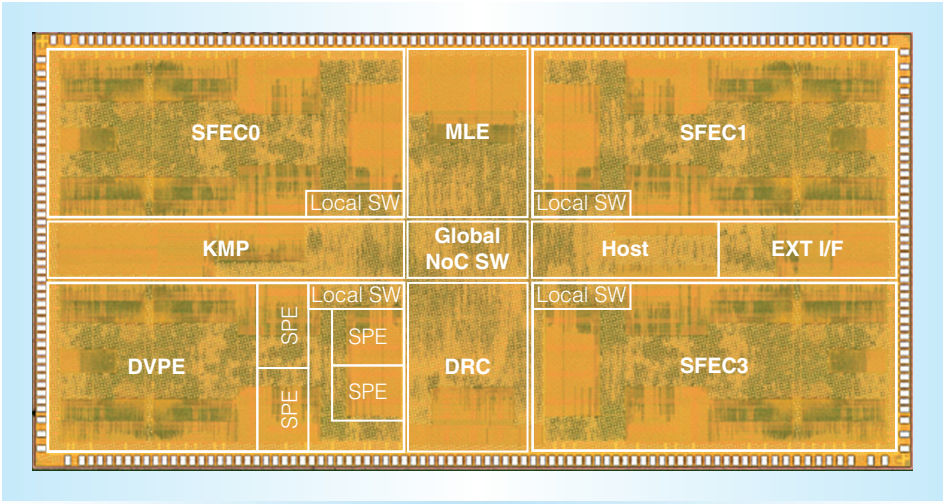

Figure 6. BONE-V5 chip photograph.

the pipeline stalling of the five-stage pipeline controller, the shared SIMD ALUs are segmented into three SFUs for two instruction decoders, resolving each stage's dependency on the basis of the SIFT algorithm's deterministic instruction sequence. Thanks to these technologies, the proposed SFEC can increase IPC 1.85× compared to BONE-V4. In particular, because the five-stage pipeline and SMT also increase the average IPC while achieving higher peak IPC, overall utilization of the proposed feature-extraction cluster achieves 2.5× higher performance than the previous architecture. Therefore, the proposed architecture can obtain 2.32× higher throughput than BONE-V4 for HD-based object recognition.

The proposed processor can also reduce the implementation cost, such as power and area consumption. We cut off the scratchpad's size by reducing the ROI tile size from 32 × 32 to 16 × 16 while increasing TLP and ILP by using the recognition pipeline and SMT. When we evaluate each core's normalized speedup over hardware cost, the proposed architecture achieves more than 10× higher cost efficiency than BONE-V3 and 6× higher cost efficiency than BONE-V4.

Implementation result

Figure 6 shows the floorplan of the BONE-V5 prototype chip, and Table 2 shows the chip summary. It is fabricated with a 0.13 µm CMOS process, occupying 32 mm^2 with a NAND2 equivalent gate count of 2.4 M and 382 Kbytes of on-chip static RAM (SRAM). A total of 31 processing and control cores operate with 200 MHz as the nominal operating frequency and 1.2 V as the nominal V_{DD}. Peak power dissipation is 534 mW, and the average power dissipation is 320 mW with the help of the dynamic voltage and frequency scaling (DVFS). For 342 GOPS peak performance, this chip achieves 10.69 GOPS/mm^2 area efficiency and 640 GOPS/W power efficiency and, for the application, obtains 9.6 mJ per-frame energy efficiency and 10.5 nJ per-pixel efficiency while processing the proposed attention-based SIFT recognition with 720p HD video streams.

Table 2. BONE-V5 chip summary.

Parameter	Value
Process	0.13 µm 1P6M logic CMOS
Chip size	4.0 × 8.0 mm^2
Gate, SRAM	2.4 M, 382 Kbytes
V_{DD}, frequency (DVFS)	200 to 50 MHz, 1.2 to 0.7 V
Power dissipation	534 mW (peak), 320 mW (average)
Peak performance	342 GOPS
Power efficiency	640 GOPS/W
Energy efficiency	9.6 mJ/frame, 10.5 nJ/pixel

Table 3 compares the BONE-V5 with the GPGPU (general-purpose computing

Table 3. Performance comparison.

Processor	Category	Size	Frequency, V_{DD}	Power
Nvidia 8800 GTX[2]	Desktop graphic processor	480 mm^2 @ 90 nm CMOS	1.35 GHz, N/A	~185 W
CELL-BE IBM	Game console	235 mm^2 @ 90 nm CMOS	3.2 GHz, 0.9~1.3 V	~50 W
OMAP 4430 TI[6]	Mobile multimedia	69.7 mm^2 @ 45 nm CMOS	1 GHz, 1.2 V	~550 mW
Blackfin analog device	Automotive	24 mm^2	600 MHz, 1.2 V	280 mW (average)
Image recognition SoC Sony[7]	Automotive	44.54 mm^2 @ 40 nm CMOS	266 MHz, 1.1 V	748 mW
EFFEX University of Michigan[8]	Feature extraction CV	16.5 mm^2	1 GHz, N/A	~7 W
BONE-V	General object recognition	32 mm^2 @ 130 nm CMOS	200 MHz, 0.7 to 1.2 V	320 mW (average)

* PVP: Pipelined Video Processor.

on GPUs) and mobile vision processors for SIFT feature extraction or object tracking and recognition. Although Nvidia's GPGPU has 518 Gflops computing power with ~185 W power consumption,[2] its processing speed is limited to only 13 fps owing to its inefficient hardware architecture for SIFT operation, as IBM's Cell Broadband Engine (CELL-BE) is. On the other hand, low-power vision processors, such as Texas Instruments' OMAP4430 for mobile multimedia,[6] Analog Device's Blackfin, and Sony's image recognition SoC for automotive applications,[7] increase their energy efficiency by adopting vision-optimized architectures, achieving high energy efficiency (as much as 24.6 mJ/frame). Including the similar heterogeneous multicore processor EFFEX[8] and our previous work,[4] the previous mobile vision processors failed to achieve high computing power and high energy efficiency simultaneously. In contrast, the BONE-V5 achieves 9.6 mJ/frame and 10.5 nJ/pixel energy efficiency with 342 GOPS computing power, which are at least 2.56× and 7.34× higher than the state-of-the-art architectures, respectively.

However, BONE-V5's energy-efficient architecture could sacrifice recognition accuracy. Processing elements' fixed-point ALUs incur degraded recognition accuracy owing to the lack of precision, especially for keypoint description's numerical operations. Fortunately, because of SIFT's distinctive feature selection, the precision errors in keypoint description incur less than 1 percent accuracy degradation at the final decision stage. The attention accuracy also affects energy efficiency and recognition accuracy. The unvisited ROI tiles containing salient object keypoints could reduce recognition accuracy; on the other hand, the misguided ROI tiles containing noise keypoints could degrade energy efficiency. Thus, high accuracy in the attention-based recognition stage is critical to achieving high energy efficiency as well as recognition accuracy.

We designed BONE-V5 to support 30 fps 720p video streams, but it can support 1080p video streams as well by configuring recognition accuracy and throughput. When sustaining recognition accuracy, the throughput decreases to 16.7 fps along with the average 1.8× increase in ROI tiles, while attention helps to alleviate the performance variation. When sustaining throughput, we reduce the accuracy to maintain the number of processing keypoints in the image by configuring the kernel or ROI tile size. Reducing the amount of Gaussian space in SIFT by half degrades accuracy by approximately 18.9 percent accuracy on the test sequences. Or, because the NoC-based multicore architecture is highly scalable, the multiple-chip integration could easily support higher resolutions such as 1980 × 1080 or 2560 × 1440 with increased numbers of SFECs, FMPs, and MLEs.

The fabricated chip is integrated with an application multimedia board in the unmanned aerial vehicle (UAV) system. Because the CAVAM's high fidelity is critical for realizing 30 fps throughput with high energy efficiency, we believe that the UAV's dynamic video sequences can be used to evaluate the proposed algorithm and chip exhaustively. The UAV's target object includes 22 different toy tanks and cars and building miniatures. Even with severe

Performance	Test algorithm	Video	Throughput	Energy
518.43 GFLOPS	SIFT feature extraction	1,024 × 768	13 fps	14.2 J/frame, 18 μJ/pixel
256 GFLOPS	SIFT feature extraction	UVGA (1,600 × 1,200)	0.48 fps	104 J/frame, 54 μJ/pixel
1.3 GFLOPS	SIFT feature extraction	VGA (640 × 480)	0.1 fps	5.5 J/frame, 17.9 μJ/pixel
~1 GMAC, 25 GOPS (PVP*)	SIFT feature extraction	CIF (352 × 288)	0.88 fps	318 mJ/frame, 3.1 μJ/pixel
464 GOPS	Object tracking	WVGA (800 × 400)	30 fps	24.6 mJ/frame, 77.1 nJ/pixel
N/A	SIFT feature extraction	1,024 × 768	25 fps	280 mJ/frame, 356 nJ/pixel
342 GOPS	Object recognition	720p HD (1,280 × 720)	30 fps	9.6 mJ/frame, 10.5 nJ/pixel

occlusion, illumination, and motion blurs, the CAVAM successfully performs recognition with high accuracy, achieving a true positive rate of approximately 98.2 percent and a false positive rate of less than 1.1 percent. Thanks to the CAVAM-based recognition, the proposed chip is an average of 3.7× (up to 5.6×) faster than the state-of-the art processor when running at 200 MHz with a 1.2 V supply voltage without DVFS.

Our BONE-V multicore processors have evolved to realize real-time object recognition in low-power mobile vision platforms with higher-resolution images. Exploiting an attention-based object-recognition algorithm and the NoC-based multicore architecture, the latest BONE-V5 architecture achieves outstanding SIFT performance—better than the other commercial state-of-the-art processors—with higher energy efficiency and low power consumption. This technology will impact mobile application processors' vision architecture in the near future, merging conventional MPU with the proposed dedicated object-recognition processor for real-time recognition-based advanced vision applications with high accuracy and fidelity. Future BONE-V processors will further lower the power consumption while achieving high computing power and high on-chip bandwidth with more accurate attention-based recognition. MICRO

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant no. 2012008937 funded by the Korea government (MEST).

References

1. D.G. Lowe, ''Distinctive Image Features from Scale-Invariant Keypoints,'' *ACM Int'l J. Computer Vision,* Jan. 2004, pp. 91-110.
2. S.N. Sinha et al., ''GPU-Based Video Feature Tracking and Matching,'' *Workshop on Edge Computing Using New Commodity Architectures,* 2006.
3. J.-Y. Kim et al., ''A 201.4GOPS 496mW Real-Time Multi-Object Recognition Processor with Bio-Inspired Neural Perception Engine,'' *Proc. IEEE Int'l Solid-State Circuits Conf.,* IEEE CS, 2008, pp. 150-151.
4. S. Lee et al., ''A 345mW Heterogeneous Many-Core Processor with an Intelligent Inference Engine for Robust Object Recognition,'' *Proc. IEEE J. Solid-State Circuits,* Jan. 2011, pp. 42-51.
5. J. Oh et al., ''A 320mW 342GOPS Real-Time Moving Object Recognition Processor for HD 720p Video Streams,'' *Proc. IEEE Int'l Solid-State Circuits Conf.,* IEEE CS, 2012, pp. 220-221.
6. J. Clemons et al., ''MEVBench: A Mobile Computer Vision Benchmarking Suite,'' *Proc. IEEE Int'l Conf. Workload Characterization,* IEEE CS, 2011, pp. 91-102.
7. Y. Tanabe et al., ''A 464GOPS 620GOPS/W Heterogeneous Multi-Core SoC for Image-Recognition Applications,'' *IEEE Int'l Solid-State Circuits Conf.,* IEEE CS, 2012, pp. 222-223.
8. J. Clemons et al., ''EFFEX: An Embedded Processor for Computer Vision Based Feature Extraction,'' *Proc. IEEE Design Automation Conf.,* IEEE CS, 2011, pp. 1020-1025.

Jinwook Oh is a PhD student in electrical engineering at the Korea Advanced Institute

of Science and Technology. His research interests include low-power digital signal processors for computer vision and, more recently, very-large-scale integration (VLSI) implementation of network-on-chip (NoC)-based heterogeneous multicore architectures. Oh has an MS in electrical engineering from the Korea Advanced Institute of Science and Technology. He's a student member of IEEE.

Gyeonghoon Kim is a PhD student in electrical engineering at the Korea Advanced Institute of Science and Technology. His research interests include low-power digital processors with dynamic resource management for computer vision and NoC-based system-on-chip (SoC) design. Kim has an MS in electrical engineering from the Korea Advanced Institute of Science and Technology. He's a student member of IEEE.

Injoon Hong is an MS student in electrical engineering at the Korea Advanced Institute of Science and Technology. His research interests include development of the digital accelerator and microarchitecture for computer vision, and VLSI implementation for machine learning algorithms. Hong has a BS in electrical engineering from the Korea Advanced Institute of Science and Technology. He's a student member of IEEE.

Junyoung Park is a PhD student in electrical engineering at the Korea Advanced Institute of Science and Technology. His research interests include development of parallel processors for computer vision and many-core architecture and VLSI implementation for bioinspired vision processors. Park has an MS in electrical engineering from the Korea Advanced Institute of Science and Technology. He's a student member of IEEE.

Seungjin Lee is a researcher in the Department of Electrical Engineering at the Korea Advanced Institute of Science and Technology. His research focuses on efficient heterogeneous system-on-chip architectures for computer vision processing. Lee has a PhD in electrical engineering from the Korea Advanced Institute of Science and Technology. He's a member of IEEE.

Joo-Young Kim is a researcher and hardware design engineer at the eXtreme Computing Group (XCG) at Microsoft Research. His research interests include multicore architecture, parallel programming, field-programmable gate array (FPGA) acceleration, and SoC implementation. Kim has a PhD in electrical engineering from the Korea Advanced Institute of Science and Technology, where he performed the work for this article. He's a member of IEEE.

Jeong-Ho Woo is an OMAP multimedia architect at Texas Instruments. His research interests include low-power, high-performance digital circuits and multimedia system design, particularly 3D computer graphics and multimedia processing architecture. Woo has a PhD in electrical engineering from the Korea Advanced Institute of Science and Technology, where he performed the work for this article. He's a member of IEEE.

Hoi-Jun Yoo is a full professor in the Department of Electrical Engineering at the Korea Advanced Institute of Science and Technology. His research interests include networks on chips, 3D graphics, bioinspired vision processors, body area networks, biomedical devices and circuits, and memory circuits and systems. Yoo has a PhD in electrical engineering from the Korea Advanced Institute of Science and Technology. He's a fellow of IEEE.

Direct questions and comments about this article to Jinwook Oh, Dept. of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), 373-1, Guseong-dong, Yuseong-gu, Daejeon, 305-701, Republic of Korea; jinwook.oh.0913@gmail.com.