A 320 mW 342 GOPS Real-Time Dynamic Object Recognition Processor for HD 720p Video Streams

Jinwook Oh, *Student Member, IEEE*, Gyeonghoon Kim, *Student Member, IEEE*, Junyoung Park, *Student Member, IEEE*, Injoon Hong, *Student Member, IEEE*, Seungjin Lee, *Member, IEEE*, Joo-Young Kim, *Member, IEEE*, Jeong-Ho Woo, *Member, IEEE*, and Hoi-Jun Yoo, *Fellow, IEEE*

***Abstract*—A heterogeneous multi-core processor is proposed to achieve real-time dynamic object recognition on HD 720p video streams. The context-aware visual attention model is proposed to reduce the required computing power for HD object recognition based on enhanced attention accuracy. In order to realize real-time execution of the proposed algorithm, the processor adopts a 5-stage task-level pipeline that maximizes the utilization of its 31 heterogeneous cores, comprising four simultaneous multithreading feature extraction clusters, a cache-based feature matching processor and a machine learning engine. Dynamic resource management is applied to adaptively tune thread allocation and power management during execution based on the detected amount of tasks and hardware utilization to increase energy efficiency. As a result, the 32 mm^2 chip, fabricated in 0.13 μm CMOS technology, achieves 30 frame/sec with 342 8-bit GOPS peak performance and 320 mW average power dissipation, which are a 2.72 times performance improvement and 2.54 times per-pixel energy reduction compared to the previous state-of-the-art.**

***Index Terms*—Multi-core processor, object recognition, scale invariant feature transform, heterogeneous, low power processor, dynamic resource management, dynamic voltage and frequency scaling.**

Glossary of Abbreviations

GOPS	giga operations per second
SIFT	scale invariant feature transform
DLP	data-level parallelism
SIMD	single instruction multiple data
SIMT	single instruction multiple thread
GFLOPS	giga floating operations per second
NoC	network-on-chip
TLP	task-level parallelism
DRM	dynamic resource management
ROI	region-of-interests
CAVAM	context-aware visual attention model
SoC	system-on-a-chip
ILP	instruction-level parallelism
CMP	chip multiprocessor
DMA	direct memory access
MIMD	multiple instruction multiple data
SMT	simultaneous multithreading
IPC	instruction per cycle
SFEC	simultaneous multithreading feature extraction cluster
FMP	feature matching processor
MLE	machine learning engine
DRC	dynamic resource controller
DVPE	dual-threaded vector processing engine
SPE	scalar processing engine
TMU	task management unit
DVFS	dynamic voltage and frequency scaling
GF	Gaussian filtering
DoG	difference of Gaussian
LOC	localization
FD	feature description
FM	feature matching
SFU	special function unit
DMEM	data memory
IMEM	instruction memory
MAC	multiply-accumulate
GMACS	giga multiply-accumulate per second
ZLSH	zero-less locality sensitive hashing
CBM	cached-based matching

Manuscript received April 23, 2012; revised July 12, 2012; accepted September 13, 2012. Date of publication November 21, 2012; date of current version December 31, 2012. This paper was approved by Guest Editor Wim Dehaene. This work was supported by the Global Frontier R&D Program on Human-centered Interaction for Coexistence funded by the National Research Foundation of Korea grant funded by the Korean Government (MEST) (NRF-M1AXA003).

J. Oh, G. Kim, J. Park, I. Hong, S. Lee, and H.-J. Yoo are with the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 305-701, Korea (e-mail: jinwook.oh.0913@gmail.com).

J.-Y. Kim is with the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 305-701, Korea, and also with Microsoft Research, Redmond, WA 98052 USA.

J.-H. Woo is with the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 305-701, Korea, and also with Texas Instruments, Dallas, TX 75013 USA.

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/JSSC.2012.2220651

DBM	database-based matching
SAD	sum of absolute difference
LSH	locality sensitive hashing
PE	processing element
RPE	reconfigurable processing element
MT	multithreading
DTA	dynamic tile allocation
FIFO	first-in first-out
VFI	voltage-frequency island
UAV	unmanned aerial vehicle
IC	integrated circuit

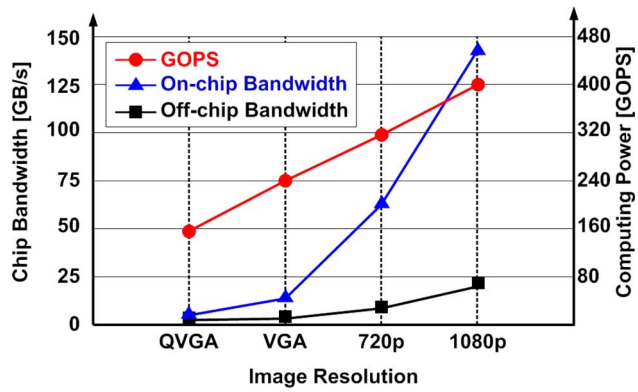


Fig. 1. On-chip memory bandwidth and computing power analysis of tile-based SIFT implementation.

I. Introduction

As THE resolution of the video applications are ever increasing, object recognition applications are becoming increasingly computationally intensive, requiring hundreds of giga operations per second (GOPS). Examples of such applications include augmented reality, image retrieval/reconstruction and scene analysis. Since these computationally complex applications are now being implemented in mobile vision platforms that are constrained by form factor, processing delay and power dissipation, a dedicated processor is necessary to obtain 30 frame/sec application throughput and sub-Watt power consumption with HD (720p or 1080p) video streams. However, the scale invariant feature transform (SIFT) [1]-based object recognition algorithm, which is popular for its invariance to scaling, rotation and illumination, is computationally complex due to its heavy workload required in local feature extraction and matching operation. As a result, conventional vision processors fail to achieve real-time performance while sustaining low power dissipation simultaneously.

Even the latest multi-threaded CPU [2] is only capable of achieving 4.48 frame/sec when performing SIFT-based recognition on 720p video, due to the limited computing power far below 100s of GOPS. Conventional single-threaded RISCs or VLIW DSPs [3] are even worse than this. In contrast, the extensive data-level parallelism (DLP) of GPU [4] or multi-core processors [5] which integrate multiple single instruction multiple data (SIMD) or single instruction multiple thread (SIMT) processing units, enable them to achieve high computing power of 100s of giga floating operation per second (GFLOPS), although at the cost of high power consumption approaching 200 W, which is far beyond power budgets of a mobile vision system. Thus, massively parallel DSPs, including IMAPCAR [6] and Strom-1 processor [7], are proposed to exploit high DLP with minimized power dissipation for specific applications, and achieve power efficiency of 50 GOPS/W and 24.4 GOPS/W respectively. However, their achievable computing power within limited power budget of mobile platforms is still insufficient for HD video-based object recognition, one of the most complex vision applications. Considering these problems of conventional vision processors, a new vision architecture should possess not only exceedingly high computing power but also high power efficiency for real-time SIFT implementation in mobile vision platforms.

Furthermore, since the tile-based SIFT implementation in a vision processor requires 60–150 GB/s on-chip bandwidth and 10–25 GB/s off-chip bandwidth for HD-based real-time object recognition due to its massively parallel architecture for high computing power, not just minimizing the number of off-chip accesses but also sustaining high utilization of on-chip bandwidth is important for a vision processor. Thus, high on-chip bandwidth of a highly parallel processor must be sustained based on high datapath utilization and network bandwidth in each core when performing compute-intensive operations such as convolution, cascaded feature extraction, and long-latency feature matching, so that the maximum possible throughput can be achieved.

With consideration of theses design constraints, in this paper, we introduce a real-time low-power object recognition processor which achieves 30 frame/sec throughput and sub-Watt power consumption for 720p video streams. A new visual attention-based object recognition algorithm is proposed that reduces more than 33% of the entire workload to relax the required computing power and on/off-chip bandwidth. It helps the vision processor overcome many of the above challenges in conventional SIFT implementation. In addition, the network-on-chip (NoC)-based heterogeneous multi-core architecture is proposed to obtain high computing power by utilizing different ILP, DLP and thread-level parallelism (TLP) of multiple processing cores. Lastly, we integrate a dynamic resource management (DRM) technique into the heterogeneous multi-core processor to minimize the power consumption of the processor as well as to increase utilization of on-chip bandwidth.

The rest of this paper is organized as follows. Section II describes the attention-based recognition algorithm and the highlights of the proposed architecture compared to the related architectures for vision applications. In Section III, the system architecture of the processor will be explained. The detailed core implementations will be covered at Section IV. Section V discusses the design and advantages of the DRM of the processor. The chip implementation and the system evaluation follow in Section VI. Finally, Section VII concludes this paper.

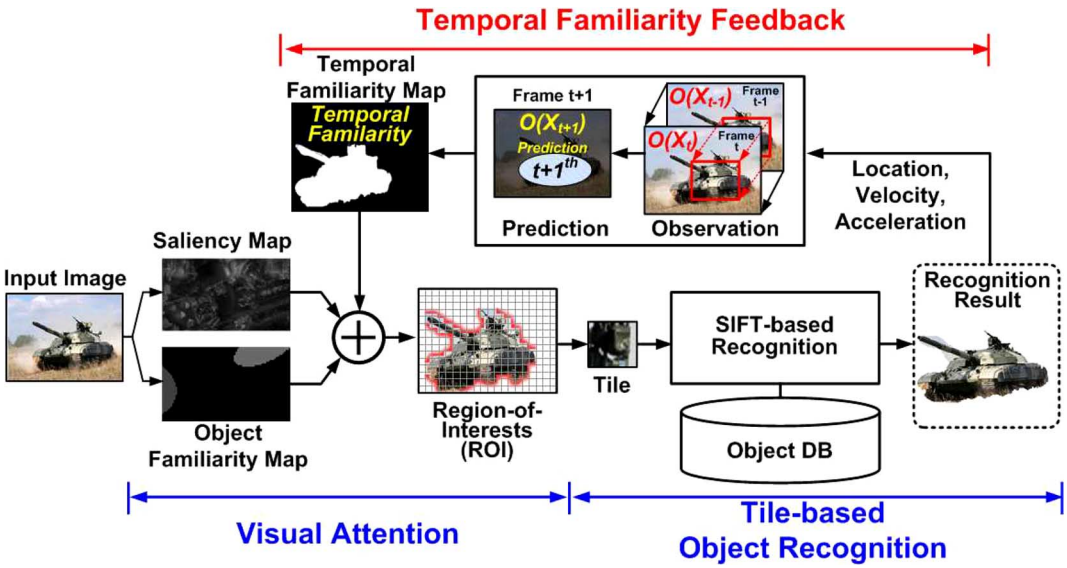


Fig. 2. Operation diagram of the context-aware visual attention model.

II. Background

A. Algorithm

The tile-based object recognition has been adopted to increase the system throughput by executing multiple recognition threads on different decomposed tiles in parallel. In addition, the attention operation was adopted to filter out the meaningless tiles with no object features and focus on the tiles containing object features as region-of-interests (ROIs). It helps the SIFT implementation increase its throughput and reduce power consumption by reducing the number of processing tiles dramatically. Previous attention model [8], which exploited not only the bottom-up conspicuity information but also the top-down object familiarity between the query and target objects in database (DB), obtained 35% background clutter tile rejection on average and 30% processing speed increase without any recognition accuracy degradation for 640×480 image-based object recognition.

However, when dynamic noises, such as motion blur, illumination and occlusion, fade SIFT features way from captured images, the previous model suffers from severe attention accuracy degradation and fails to accommodate HD object recognition due to its highly demanding computing power incurred from increased ROIs in an image. In order to solve this problem, we present more accurate attention algorithm to reduce the number of processing tiles further with increased processing speed.

Fig. 2 shows the proposed attention model for the HD object recognition in a mobile vision system, named context-aware visual attention model (CAVAM). The CAVAM integrates temporal familiarity which measures the temporal coherence of consecutive frames by tracking and prediction of the recognized objects in addition to the previous saliency and object familiarity. That is, the familiarity map reflects not only the spatial conspicuity but also temporal continuity so that the obtained ROI can track the target object movement accurately irrespective of dynamic noises. For example, the 77% loss of total SIFT features in an image by dynamic noises generates 3.3 times increase in the number of ROI tiles in the previous attention model, and requires additional $\sim$80 GOPS computing power and $\sim$30 GB/s on-chip bandwidth. However, in CAVAM, the required computing power and on-chip bandwidth are reduced by 16% and 46% respectively thanks to 1.44 times higher attention accuracy for dynamic object recognition with HD video streams on average. Thus, the 4.8% extra on-chip bandwidth for temporal familiarity generation is negligible compared to the performance gain.

B. Related Works

The proposed object recognition processor has some characteristics of application specific hardware accelerators such as IMAPCAR [6] and National Taiwan University's machine learning SoC [9]. In these systems, a highly parallel SIMD architecture and a high bandwidth dual memory architecture are adopted to accelerate in-vehicle image recognition and K-means clustering algorithm respectively, restricting redundant data computations and memory accesses for their targeted applications. However, unlike those application specific processors, the proposed processor is also optimized for general stream processing with application characteristics such as compute intensity, data parallelism, and produce-consumer locality [10] similar to CELL [11], GPUs [4] and the Strom-1 processor [7]. CELL includes data-parallel synergistic processing units, GPUs support many lightweight data-parallel threads, and the Strom-1 processor utilizes an optimized ALU and memory architecture for kernel and stream data processing with different ILP and DLP. In terms of its chip multiprocessor (CMP) or multi-core architecture, the proposed streaming architecture is more like Intel 80-Tile processor [12] and Toshiba's eight-core media processor [13] that process streams as threads in different cores. These multi-core processors exploit a packet-switched NoC and pipeline-based/thread-based parallel execution schemes for high computing power respectively. In comparison to those processors, while exploiting the high-performance technologies of the multi-core architectures, the proposed processor is much more power efficient due to its use of fixed-point ALUs instead of floating-point ALUs as

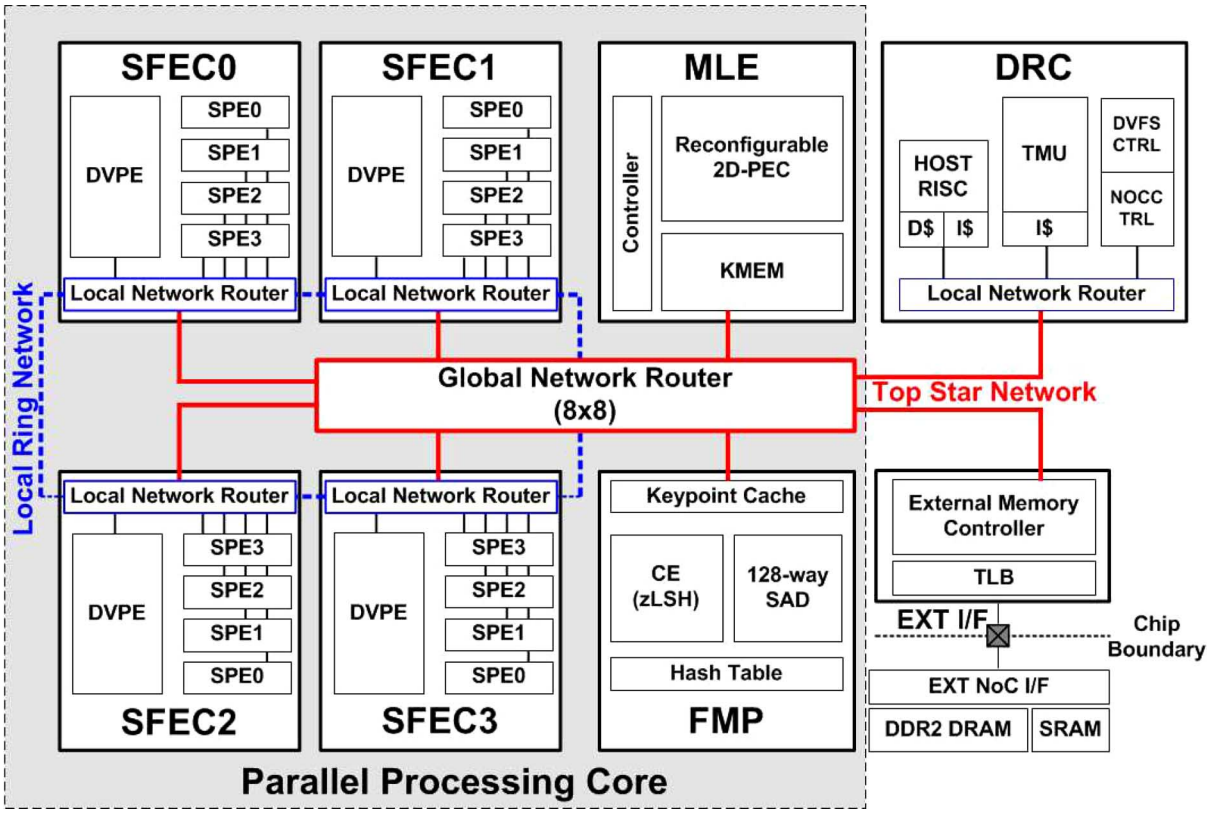


Fig. 3. Block diagram of the proposed multi-core processor for dynamic object recognition.

well as its use of 2-D direct memory access (DMA)-integrated NoC interfaces instead of a cache-based memory system that requires a power-hungry hierarchical memory architecture. Therefore, the processor can achieves higher computing power with lower power consumption compared to the multi-core processors [12], [13] and, also, SIMD-based parallel machines such as IMAPCAR and Xetal-II [14].

The previous generation of our vision processor [8] was designed to realize real-time object recognition for VGA images by employing a heterogeneous multi-core architecture containing SIMD and multiple instruction multiple data (MIMD) processing units in parallel. However, since its performance is not proportional to the number of processing cores due to its limited throughput and bandwidth of the top architecture, this processor is not capable of object recognition on HD 720p video streams. Thus, this chip adopts a dual-threaded SIMD/MIMD core cluster and latency/power optimized cores in addition to the NoC-based heterogeneous architecture by taking advantages of a new fine-grain object recognition pipeline. Coping with the machine learning-based DRM, the high level of ILP, DLP and TLP in the processor can realize object recognition for HD 720p video streams with enhanced throughput and power-efficiency for advanced mobile vision applications.

III. System Architecture

The proposed processor is based on a heterogeneous multi-core architecture for low-power real-time object recognition in a mobile vision system [15]. Achieving high computing power with low power consumption, the processor adopts several system-level technologies for high on-chip bandwidth with increased datapath utilization. It integrates a hierarchical NoC architecture with $\sim$100 GB/s aggregate on-chip bandwidth and heterogeneous processing cores with $\sim$350 GOPS computing power, which is measured by 8-bit integer operation, required for 720p video-based object recognition.

Even though CAVAM reduces 16% computing power and 36% on-chip bandwidth of HD object recognition, it is still difficult to satisfy the requirements at the same time in a vision processor. Thus, the proposed processor employs the 5-stage task-level pipeline and the simultaneous multithreading (SMT) [16] operations of ROI tile processing to increase the throughput of tile-based object recognition in CAVAM by increased ILP and TLP of the multi-core architecture. Furthermore, it integrates different types of parallel processing cores in a NoC architecture to reduce the processing delay of each pipeline stage as well as the visual attention stage in CAVAM, thereby achieving high computing power and on-chip bandwidth required in HD object recognition. For power efficiency of a mobile vision platform, resource management technique is applied to increase hardware utilization and throughput and to reduce power dissipation of idling cores. Using these technologies, the proposed mobile vision processor achieves 342 GOPS computing power and 640 GOPS/W power efficiency as well as 83.3 GB/s on-chip bandwidth required in CAVAM-based object recognition on 720p HD video streams.

A. SoC Architecture

Fig. 3 shows the overall block diagram of the proposed SoC. It contains of 4 simultaneous multithreading feature extraction cluster (SFEC) for SIFT feature extraction operation, a feature

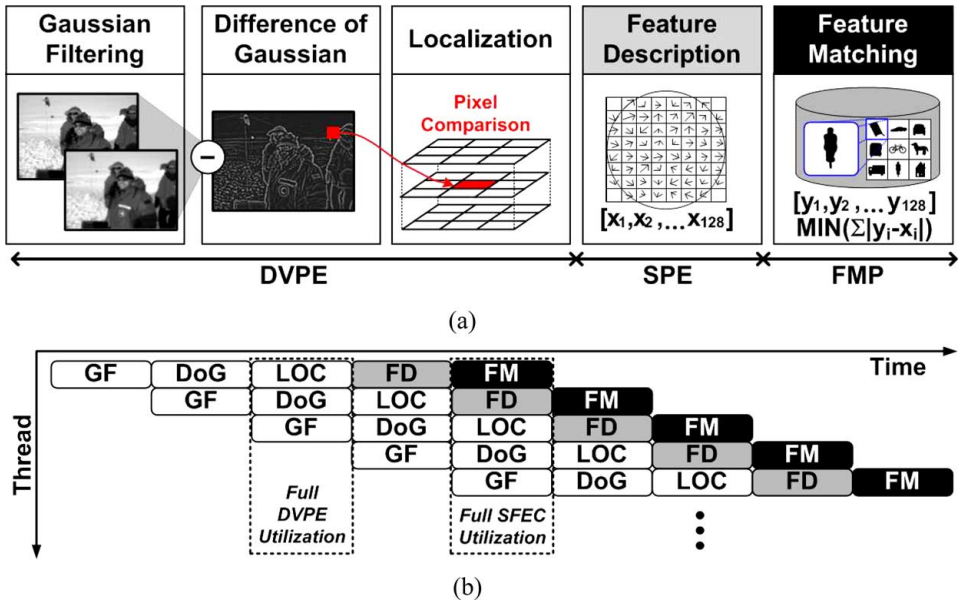


Fig. 4. (a) 5-stage fine-grain pipeline stages and (b) its pipeline operation with increased hardware utilization.

matching processor (FMP) for matching operation with SIFT descriptors in DB, a machine learning engine (MLE) for attention operation of CAVAM, a dynamic resource controller (DRC) for DRM implementation, and an external interface (EXT I/F) for NoC-based off-chip communication. Each core is connected by a hierarchical star-ring NoC [17] which 8 × 8 top router can obtain 640 MB/s/port theoretical bandwidth in respective directions. The local ring network that connects 4 SFEC cores with maximum 1-hop latency reduces top NoC occupancy of SFEC for congestion avoidance.

In the SFEC, one dual-threaded vector processing element (DVPE), integrating one 16-lane data-parallel SIMD processing element, and four scalar processing elements (SPEs) are contained together using a local NoC router. Once the 4 SFECs carry out feature extraction operation, then the FMP performs the following matching operation, which requires external memory access with minimized latency, for the extracted SIFT feature descriptors. The MLE is designed to realize several functions required in CAVAM by using its reconfigurable architecture with minimized power dissipation.

The DRC is the main controller which accounts for throughput and power consumption of the multi-core processor. It contains a task management unit (TMU) for software-managed workload allocation, a dynamic voltage and frequency scaling (DVFS) [18] controller for optimization of power dissipation in the SoC, a network-on-chip (NoC) controller for sustaining high NoC bandwidth, and a ARM10-based host RISC processor. With the help of the DVFS and NoC controllers, the TMU enables the maximum aggregate on-chip bandwidth for the object recognition processor reaches 83.3 GB/s.

B. 5-Stage Task-Level Pipeline

In order to realize the CAVAM for 720p HD video streams, the 5-stage task-level pipeline is proposed to accelerate overall processing speed of feature extraction and matching operation for 16 × 16 image tiles. The feature detection of previous SIFT algorithm [8] is divided into 3 fine stages, namely, Gaussian filtering (GF), difference of Gaussian (DoG), and localization (LOC), which are 3 main procedures of SIFT scale space generation and extrema detection. Along with conventional feature description (FD) and feature matching (FM) stages, the 5-stage object recognition is implemented to increase ILP/TLP with high system utilization as shown in Fig. 4(a). The DVPE performs GF, DoG and LOC with different special function units (SFUs) respectively, and the SPE and FMP perform FD and FM respectively. Fig. 4(b) shows the operation diagram of the proposed pipeline. Thanks to the increased datapath utilization of SFEC to 0.92 by this pipeline, the overall processing throughput of the SIFT pipeline is increased by 1.67 times compared to the conventional 3-stage pipeline.

C. SMT-Based SIFT Implementation

In SFEC, the SMT is adopted to process multiple ROI tiles at the same time for squeezing system throughput further out of the object recognition pipeline. Conventional singled-thread SIMD core suffers from low-datapath utilization of less than 0.3, since only small part of ALUs in each lane of SIMD core is activated for decoded instructions. Thus, to minimize the wastage of power and processing time, the SIMD unit is segmented into 3 different SFUs to support SMT operation, and the SFUs are corresponding to the GF, DoG and LOC respectively. Fortunately, the each ROI tile is totally independent to each other so that it is possible to exploit producer-consumer locality of tile-based SIFT processing.

Fig. 5 shows the proposed SMT-based SIFT implementation using SFECs. When the visual attention operation determines the ROIs from the image, the 128-entry TMU software queues accumulates the 32-bit pointers of image tiles and the TMU fetches image tiles from the external DRAM and to SFECs as a thread scheduler. The SFEC is designed to perform 3-stage feature detection operation for the maximum two threads simultaneously. The theoretical peak IPC of SFEC datapath is 1.92, and this can increase the throughput of overall architecture by 1.43 times compared to single-threaded SFEC. Thanks to the NoC

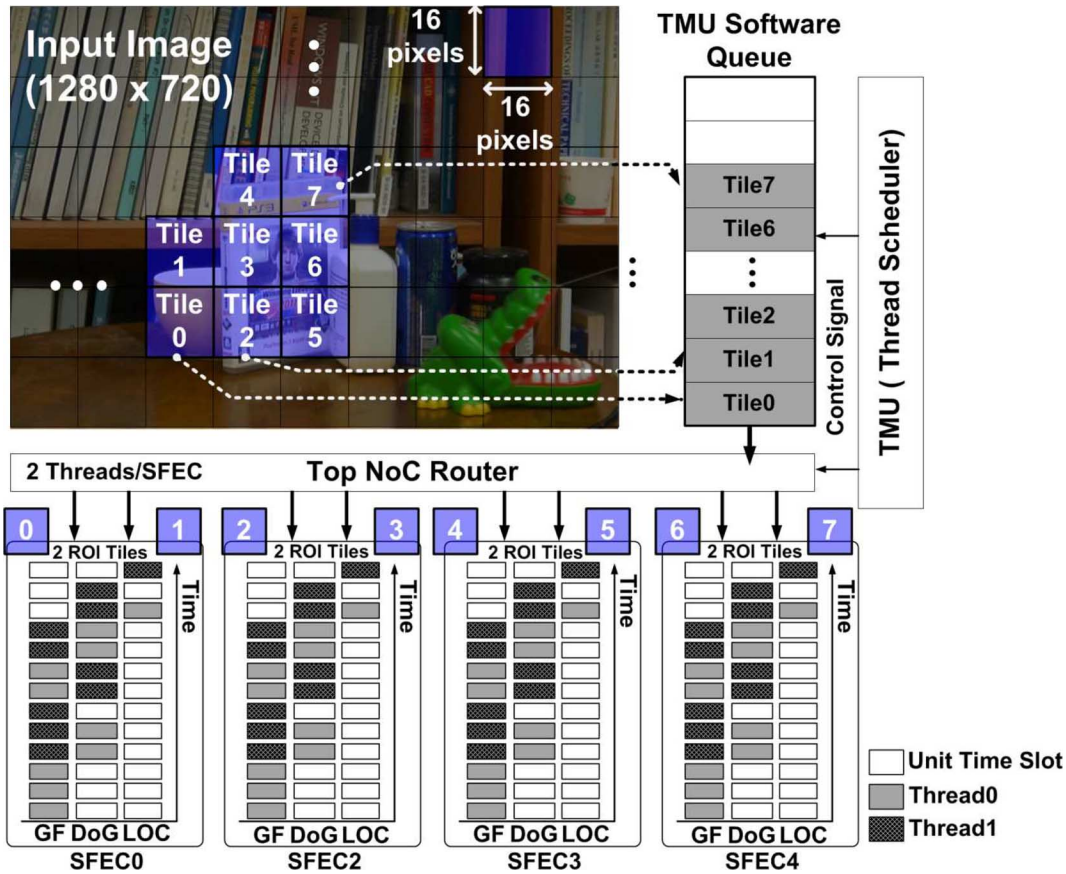


Fig. 5. Proposed SMT-based SIFT implementation using SFECs.

and SMT, the out-of-order execution of ROI-based SIFT operation can be easily implemented without complex scheduling and re-order buffer. With only 12% extra area overhead and 9.9 mW extra power consumption for a simple context switching controller, 16 general purpose registers, and 20 kB extra memory, the SFEC achieves at least 30% processing delay reduction for an ROI tile processing of 100–1000 instructions in an inner loop and 2.8 times processing speed increase for 10–100 ROI tile sequences based on hiding data fetch delay as well as increased pipeline stages. As a result, thanks to the dual-threaded processing of 8 ROI tiles and the fine-grain pipeline with increased throughput of 4 SFECs, the average performance of the proposed architecture reaches 47700 tiles/sec for feature extraction of SFEC and 62720 vectors/sec for the matching operation of FMP respectively.

IV. Core Architecture

A. Dual-Threaded Vector Processing Element

The detailed architecture of DVPE is depicted in Fig. 6, consisting of 3 SFUs for a 16-lane data-parallel SIMD unit. For 2 different threads, the SIMD datapath utilization can be increased to 0.92 on average for SIFT feature detection. Since each thread or an image tile in tile-based recognition is totally independent, data consistency and race condition of two threads can be eliminated by isolating each memory space. Therefore, the SFEC contains 2 16 kB DMEM and 4 kB IMEM for different threads

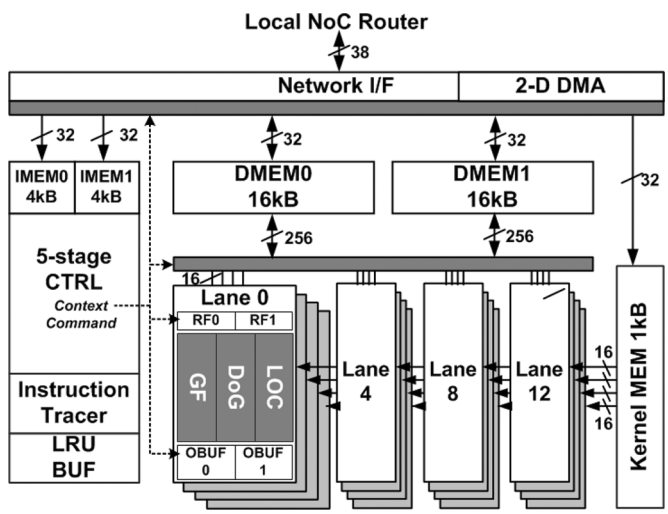


Fig. 6. Details of DVPE architecture of SFEC.

which are executed with unraveled dependency between consecutive pipeline stages. The GF unit performs scale space generation for 2-D Gaussian filtering with different size of kernels, realizing 5 scale spaces of each octave with a 16×16 tile. For memory access reduction, the lane size of SIMD unit is optimized for 1-D convolution size, which substitutes for redundant access pattern of 2-D Gaussian filtering by line-wise operation. And also, the one-cycle-latency 3-operand MAC unit is employed for 1-D convolution operation which incurs 3.1 times

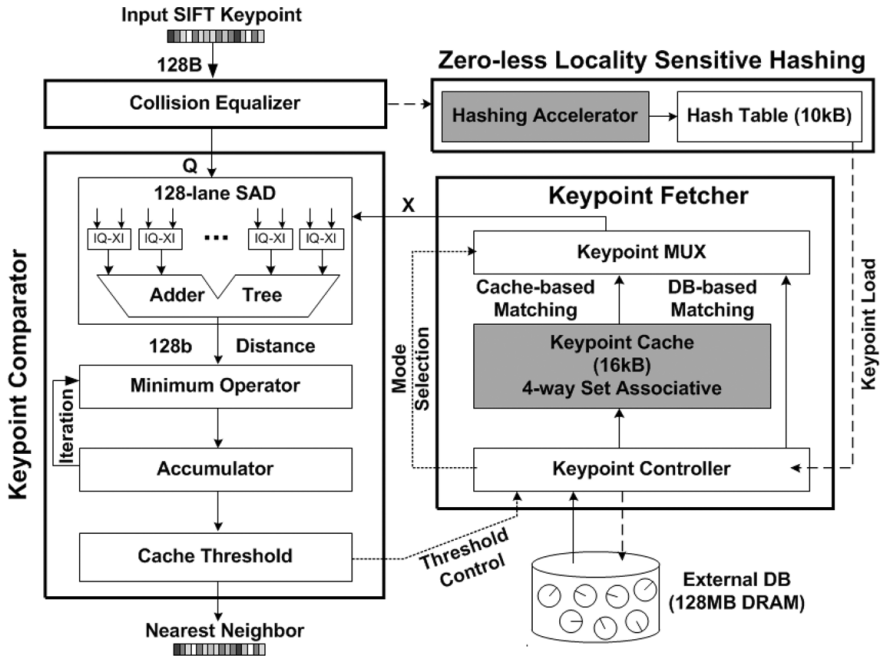


Fig. 7. Hardware diagram of proposed feature matching operation.

speed-up for Gaussian filtering and 15% throughput increase in SFEC pipeline. With the help of proposed technique, the DVPE can obtain 1.46 IPC and 0.82 utilization, which are 1.87 times and 2.7 times improvement respectively. As a result, the proposed DVPE can perform feature extraction operation for an ROI tile within 15000 cycles, 3.45 times improvement from the previous SIMD core [8].

B. Feature Matching Processor

Fig. 7 shows the hardware diagram of the cache-based FMP for keypoint matching with the zero-less locality sensitive hashing (ZLSH). Since more than 80% of matching delay is consumed at external DB access for feature searching, minimization of keypoint searching regions in DB is essential to increase matching throughput as well as the overall recognition speed. To this end, in FMP, there are two keypoint matching mechanism; the primary cache-based matching (CBM) and the secondary DB-based matching (DBM). The CBM uses the keypoints which are previously used at the last matching and stored in the keypoint cache for searching the nearest neighbor. If the keypoint is matched, the matching ends with 98% reduction of external accesses. If the keypoint is not matched, the additional DBM is performed with the ZLSH index to access the candidate keypoints of DB, which still results in 86% access reduction than the brute-force matching.

The 16 kB inter-frame cache is implemented with 4-way set associative structure since each way is corresponding for 32 keypoint vectors and the size of a cache line is 128-byte corresponding to the size of a keypoint vector. The 32 kB vector memory contains the detected keypoint vectors for the ROIs that are compared with cached vectors through the 128-way SAD array for CDM. The 1024-bit wide 1 kB configuration SRAM is used as operand register files to reduce redundant memory access for higher throughput. The hashing accelerator and a 10 kB hash table are implemented to realize ZLSH to minimize the degree of uneven binning of hashing for maximum reduction of external memory access, and achieve 64% reduction in the largest bin size compared with previous locality sensitive hashing [19]. As a result, it only consumes less than 3.2 μs on matching a keypoint with DB for the proposed task-level object recognition pipeline.

C. Machine Learning Engine

The architecture of MLE is as shown in Fig. 8(a) and it is designed for accelerating CAVAM's Kaman filter operation and DRC's reinforcement learning algorithm with different processing granularity to optimize power consumption. The architecture of the MLE is very similar to application specific reconfigurable processors, such as ADRES [20], Montium Tile processor [21] and eXtreme Processing Platform (XPP) [22], which adopts coarse-grained architecture for multimedia and communication applications. They exhibit high level of ILP and/or DLP but less control flow. The MLE also utilizes data-parallel, compute-intensive operations by adopting SIMD computation model. The ALU of SIMD processing core is optimized for sub-word-level operations, including ADD, SUB, MUL and SHIFT, supporting one shared instruction set architecture.

In terms of reconfigurable processing core architecture, the MLE is analogous to the MorphoSys [23] which comprise the reconfigurable cell arrays, an RISC control processor, context memory, frame buffer and DMA controller. Similarly, the MLE adopts 4×4 reconfigurable processing element (PE) arrays, consisting of 16 reconfigurable processing elements (RPEs), and 32 kB kernel memory for parameter operands, a light-weight control RISC processor for reconfiguration control and instruction fetch/decode. A RPE is composed of 4 processing elements to modify the parallelism of MLE which

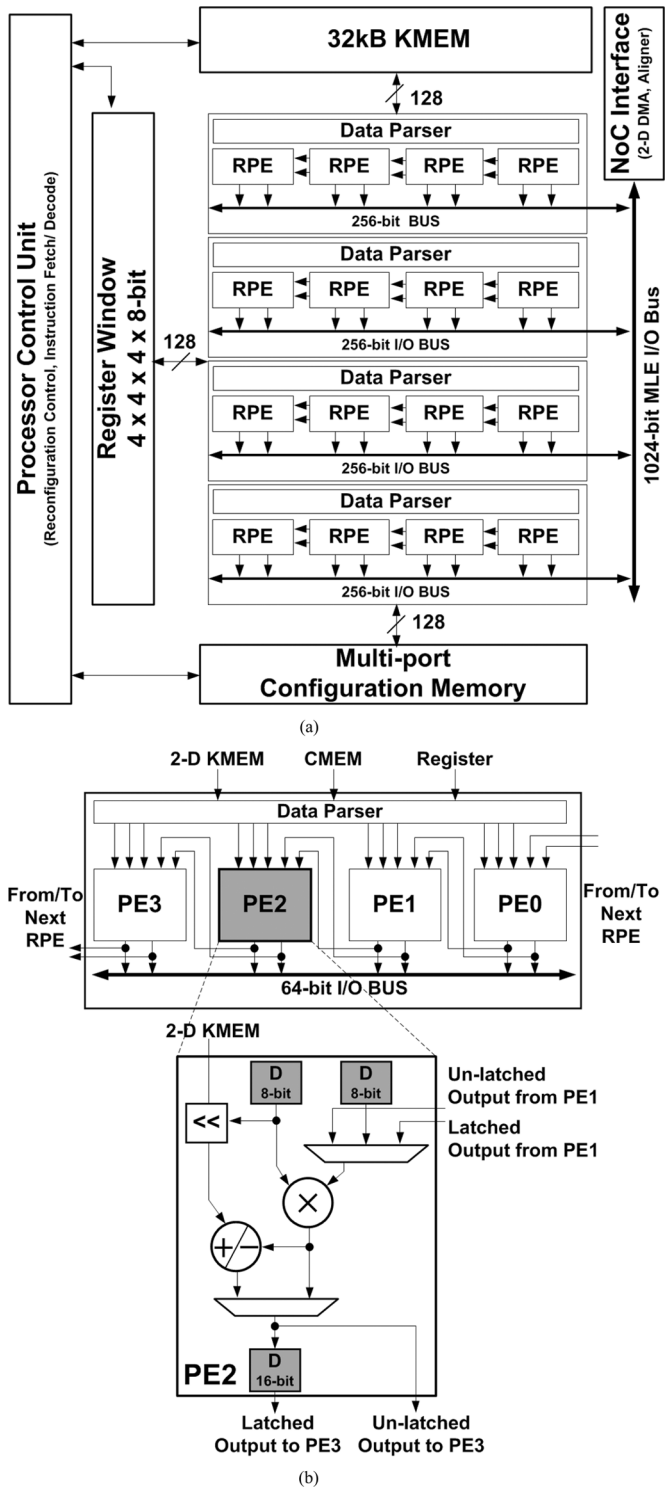


Fig. 8. (a) The architecture of machine learning engine and (b) block diagram of reconfigurable processing element.

possesses 8 bit-resolution granularities for different operations. The block diagram of RPE is depicted in Fig. 8(b). The 4 processing elements can be reconfigured from 8-bit resolution pixel-level operators to 32-bit resolution complex sequential learning operators. Each processing element can propagate result data to the next processing element as new operand.

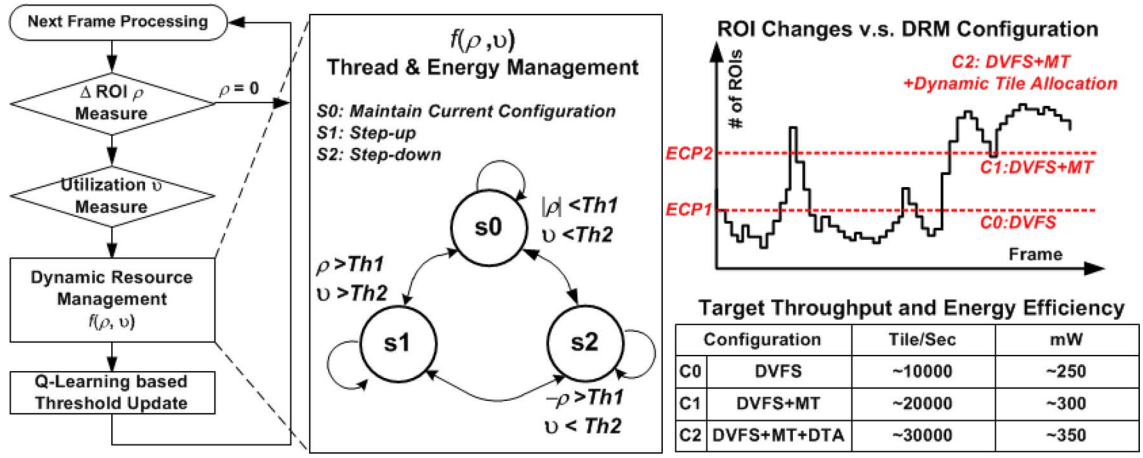
As a result, different bit-resolution of 8/16/24/32-bit processing element can be applied for different types of target algorithms with minimized power consumption. For example, in two extreme cases, such as 16 32-bit configuration for high precision learning algorithm and 64 8-bit configuration for pixel-level operation of saliency map generation, the power consumption varies from 33 mW to 123 mW, while reducing at most 71% of unnecessary power dissipation on un-used registers and ALUs compared to the SIMD-core based implementation. In addition, it only takes 4.4 ms based on 12.5 16-bit giga multiply-accumulates per second (GMACS) for running reinforcement learning algorithm of DRM.

V. Dynamic Resource Management

DRM [24], which is famous technology adaptively tuning hardware resource of multi-core processors or data centers, is employed to handle workload allocation and voltage and frequency configuration of the proposed architecture. The DRC operates the DRM operation with its sub IPs as hardware resource controller. The TMU performs ROI tiles allocation for DVPE and keypoints allocation for SPE and FMP to sustain maximum throughput by keeping the core from going idle frequently. Coping with the TMU's workload allocation, the DVFS controller and NOC controller is configured according to the performance margin for the 30 frame/sec real-time requirement. Since the implemented processor is designed to satisfy the maximum workload scenario of actual use cases, a large power saving through aggressive voltage and frequency scaling can be obtained by DRM.

A. DRM Implementation

Because the configuration is carried out based on one thread which is a 16×16 ROI tile-based SIFT operation, the DRM of the object recognition processor needs less than a few μs response time, or 10–500 cycles. Therefore, we adopt hardware-implemented DRC with software programmability with about 100 times speed-up compared to the middleware-based approach [25]. While using the on-line learning ability of MLE, the DRC can change the throughput and power characteristics of multi-core system with precise workload prediction for better energy efficiency as shown in Fig. 9. The DRC adjusts the power management configuration of SFEC based on the amount of ROIs, ρ, and utilization per frame, υ. Since the dynamic resource management can estimate the optimized state transition point by adopting reconfigurable thresholds, th_1 and th_2, of ROI and utilization the optimum energy and throughput management can be selected by one of three different configurations; C0: DVFS, C1: DVFS + multithreading (MT), C2: DVFS+MT+ dynamic tile allocation (DTA). The DTA will be discussed on the following sub-section for utilization control. Based on the configurations, the multi-core architecture can change its throughput and power efficiency to sustain 30 frame/sec with lowest energy consumption. The control parameters, which are 2 state transition thresholds and 2 energy configuration ROI points, are updated by with Q-learning-based on-line learning operation [26] to minimize DRC miss prediction rate less than 2.2% for prohibiting severe performance degradation. As a result, 9.6 mJ/frame or 10.5 nJ/pixel energy efficiency can be obtained with 320 mW average power consumption.



Configuration		Tile/Sec	mW
C0	DVFS	~10000	~250
C1	DVFS+MT	~20000	~300
C2	DVFS+MT+DTA	~30000	~350

Fig. 9. Operation diagram of dynamic resource management.

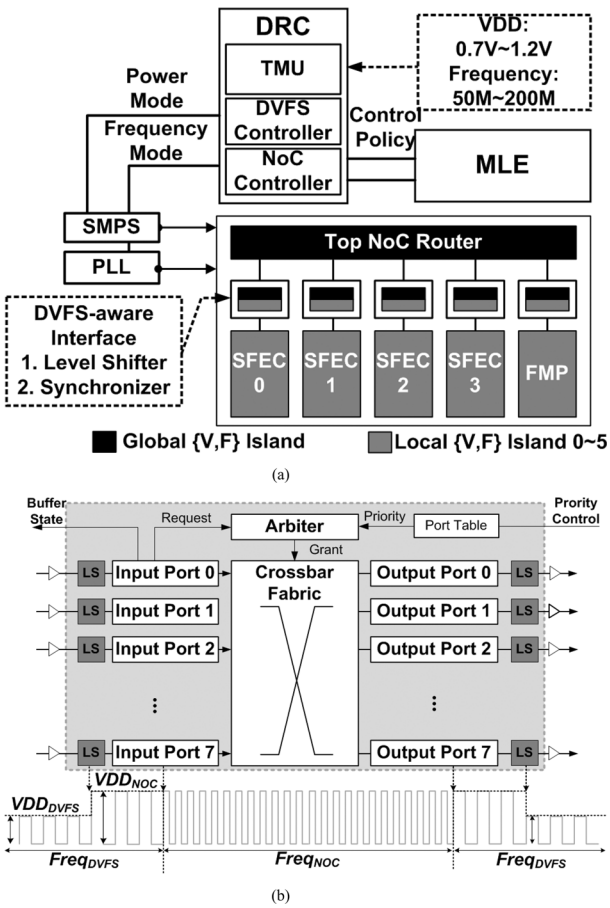


Fig. 10. (a) System diagram of NoC-based DVFS implementation and (b) DVFS-aware NoC router architecture.

B. NoC-Based DVFS Implementation

Fig. 10(a) shows the NoC-based DVFS implementation of the proposed processor. It has 6 different voltage-frequency islands (VFIs), including one global island for the top NoC router, the MLE and the DRC and 5 different local islands for SFECs and a FMP core. When there is packet transition between two different cores, the level shifting and the synchronization have to be performed for different power and clock frequency domains. To simplify the implementation complexity and increase system robustness, the monolithic design of NoC router is proposed by merging a level shifter and synchronizing dual-clock FIFO at the TX and RX ports of NoC. The DRC configures each local VFI from 0.7 V–1.2 V VDD range, and 50 MHz–200 MHz operating frequency range by using the external switched-mode power supply IC and the PLL. For higher throughput of the NoC, the proposed switch can configure the priority ports of the weighted round robin arbiter based on the DRM configuration to reduce the packet conflicts.

The proposed NoC router architecture is shown in Fig. 10(b). An arbiter controls the cross fabric to connect input ports to output ports and each port contains 38-bit wide and 8-words deep queues. The network interface supports 640 MB/s/port bandwidth at 200 MHz to all the switches in the respective directions. As Philips' Æthereal NoC [27], the proposed compact NoC router is designed with simple packet switching but obtains high processing speed only with 0.31 mm^2 for the 8×8 top switch. Thanks to the monolithic NoC routers, the hierarchical star-ring network can be easily implemented without extra IP or back-end support, supporting multi-core DVFS system.

C. SFEC Utilization Control

To achieve highest SFEC throughput for SIFT feature extraction along with the 5-stage task-level pipeline, the utilization of each processing core should be sustained as high as possible. In order to increase utilization of total 4 DVPEs and 16 SPEs of 4 SFECs, the processor adopts the DTA based on the distance between processing ROIs.

Since the neighboring 8 tiles of the ROI tile are additionally required to perform the scale space generation, at most 6 of 9 processing ROIs tiles can be shared with the next ROI processing. Based on the ROI distance which indicates the number of tiles that can be shared between two different ROI processing, the TMU performs the DTA of ROI tiles to the SFECs by utilizing different sustained bandwidth between SFEC cores, such as 3.2 GB/s internal data channel of SFEC, 512 MB/s/port local ring NoC, and 426 MB/s/port top star NoC. Through the internal bus and local NoC, the shared tiles can be rapidly transferred between two different ROI threads,

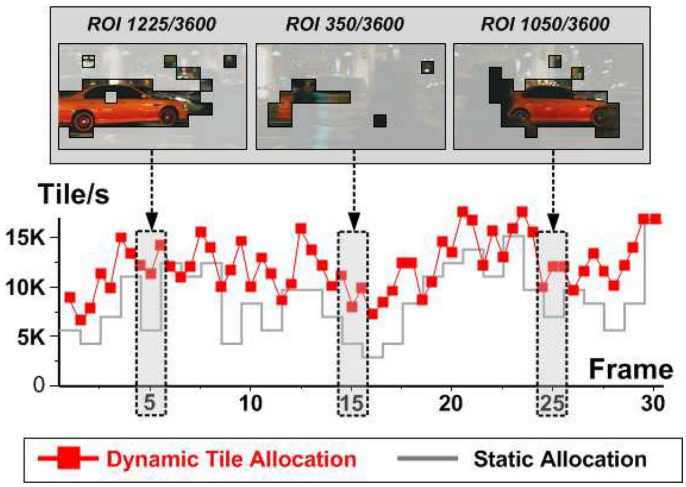


Fig. 11. Measurement results of throughput enhancement by MLE operation.

as a result, the DTA can increase 17% processing speed by reducing top channel occupancy compared to the conventional sequential thread allocation.

D. On-Line Learning-Based Energy Control

With the help of MLE, the DRC can perform on-line learning based dynamical resource control to minimize the power consumption for the different amount of tasks in each frame. After deciding the number of ROIs in the image by attention process, the DRC can measure the required timing margin to process all ROIs so that it configures the thread allocation and DVFS strategy based on the DRM policy. The MLE performs on-line learning operation for varying patterns of the current task and the hardware utilization of processing cores to estimate optimized hardware resource and power margin for the current frame.

The measurement result of the DRC operation for 30 continuous HD video frames is shown in Fig. 11. The number of ROIs for frames is fluctuating as 50–1800 that possibly incurs performance degradation or power wastage. Therefore, the MLE updates the new energy configuration point based on the optimized throughput and power consumption for its task. For the given test frames, it updates two thresholds, th_1 and th_2, and two ECPs of DRM policy as control parameters by on-line learning of ROI variation as a monitoring parameter. Then, the power and throughput are updated to provide optimized system performance based on four parameters of DRM. For the test video scene, the updated control parameters of $\{47, 0.44, 224, 1288\}$ enable the processor achieve 279 mW average power consumption and 14341 tiles/sec throughput. As a result, the DRC obtains 1.3 times higher frame rate and 66% energy reduction compared to the static allocation.

VI. Implementation and Evaluation

A. Chip Summary

The proposed chip in Fig. 12 is implemented with 0.13 μm CMOS process and occupies 32 mm^2 with 2.4 M NAND2 gate count of and 382 kB on-chip SRAM. Table I summarizes the chip specification. A total 31-IP multi-core processor consumes

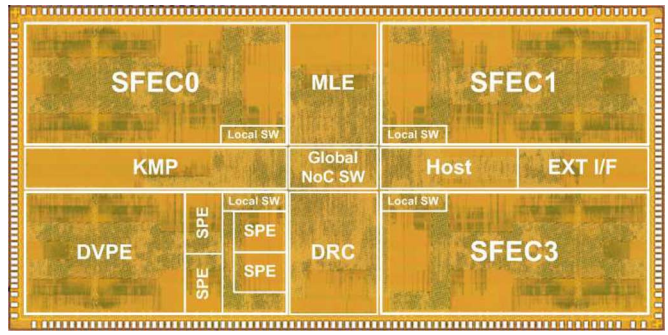


Fig. 12. Chip microphotograph.

TABLE I
Chip Specification

Number of Cores	31 IPs
Process	0.13μm 1P6M Logic CMOS
Chip Size	4.0 x 8.0 mm^2
Nominal Frequency	200MHz
Nominal Voltage	1.2V
Gate / SRAM	2.4M / 382kB
Power Dissipation	534mW (Peak) / 320mW (Ave.)
Peak Performance	342GOPS
Area Efficiency	10.69GOPS/mm^2
Power Efficiency	640GOPS/W
Per-frame Energy	9.6mJ/Frame
Per-pixel Energy	10.5nJ/Pixel

534 mW peak power and 320 mW average power with 0.7–1.2 V and 50 MHz–200 MHz DVFS configuration. For 8-bit 342 GOPS peak performance, this chip achieves 10.69 GOPS/mm^2 area efficiency and 640 GOPS/W power efficiency and, for application, obtain 9.6 mJ per-frame energy efficiency and 10.5 nJ per-pixel efficiency.

Table II lists the comparison of four vision processors which have similar vision applications with this work. As compared with four architectures, namely, CMOS sensor integrated camera chip [28], a massively parallel image processor [29] and our previous arts [30], [8], this work reduces at least 51.5%, 14.8%, 54.6% and 49.3% power efficiency (GOPS/W) respectively. Thanks to the 5-stage fine-grain pipeline and SMT-enabled multi-core architecture, this chip obtains 1.5 times higher GOPS, which is 342 GOPS, even with 18% reduced gate counts compared to our latest work [8]. In addition, the DRM-based DVFS enhances energy efficiency of the multi-core processor and enable the chip to obtain 640 GOPS/W consequently. As a result, for object recognition applications, the obtained 10.5 nJ/pixel energy dissipation is the lowest ever and 2.54 times less than the state-of-the-art object recognition processors.

Table III shows the computing power (GOPS) and power consumption breakdown of the proposed processor with gate count. The total peak performance amounts 342 GOPS when 534 mW is dissipated. The 4 SFECs, the largest component of the processor with 1.88 M gate, accounts for the highest 112 GOPS and consume 236 mW, 44% of total power consumption. The FMP performs 98 GOPS with 86 mW power consumption

TABLE II
PERFORMANCE COMPARISON WITH RELATED WORKS

	Ref.[28]	Ref.[29]	Ref.[30]	Ref.[8]	This Work
Process	65nm	65nm	130nm	130nm	130nm
Gate	20.4M	N/A	3.73M	2.93M	2.4M**
On-chip SRAM	2.175MB	288kB	396kB	626kB	382kB
Power Supply(V)	1.0	1.0	1.2	0.65~1.2	0.65~1.2
Frequency(MHz)	500	200~560	200	50~200	50~200
Peak Power(mW)	783 @250MHz	330 @200MHz	695 @200MHz	704 @200MHz	534 @200MHz
GOPS	512	154	201.4	228	342
GOPS/W*	311	425	290	324	640
Per-frame Energy	N/A	N/A	8.2 mJ	11.4 mJ	9.6 mJ
Per-pixel Energy	N/A	N/A	26.7nJ	37.1 nJ	10.5 nJ

*Scaled to 0.13μm Process ** The reported gate count 1.4M in [15] should be corrected to 2.4M

TABLE III
BREAKDOWN OF COMPUTING POWER, POWER DISSIPATION AND GATE COUNTS

Core	GOPS	Power (mW)	Gates (K)
4 SFEC	112	236	1888
FMP	98	86	123
MLE	90	123	159
DRC	42	49	140
NoC	-	30	57
ETC	-	10	48
Total	342	534	2415

and achieves very high power efficiency thanks to the dense functional units of comparison and hashing units. The DRC integrated with MLE can obtain 132 GOPS of high computing power for different algorithm realizations while accounting for 32.2% of total power consumption only with 12.3% of total gate count. The top NoC router and the miscellaneous control circuits consume 40 mW in total with 105 K gates for global switch communication and IP control operation.

B. System Evaluation

The fabricated chip is integrated with an application multimedia board and applied to the unmanned aerial vehicle (UAV) system as shown in Fig. 13. The chip is evaluated in a real demonstration system for 30 frame/sec 720 p video streams of the UAV. The proposed object recognition processor, named BONE-V5, is integrated with the Texas Instrument's OMAP4430-based multimedia board by using the FPGA extension board. The video processing, such as capturing and displaying image is carried out Linux operating system in OMAP at the main board. The software system of BONE-V5 includes the custom compilers for converting ANSI C-based SIMD and MIMD core programs into separate assembly codes and the linker for merging them with custom assembly code for kernel functions of vision applications. The generated assembly of each core is managed by the control program of TMU which also bases the same compiler. Otherwise, the host RISC program uses the ARM compiler separately due to its different architecture. The communication between two processors is conducted through the NoC interface in FPGA, and the processor can access the 128 MB DDR2 SDRAM and the 4 MB SRAM in the extension board by the FPGA.

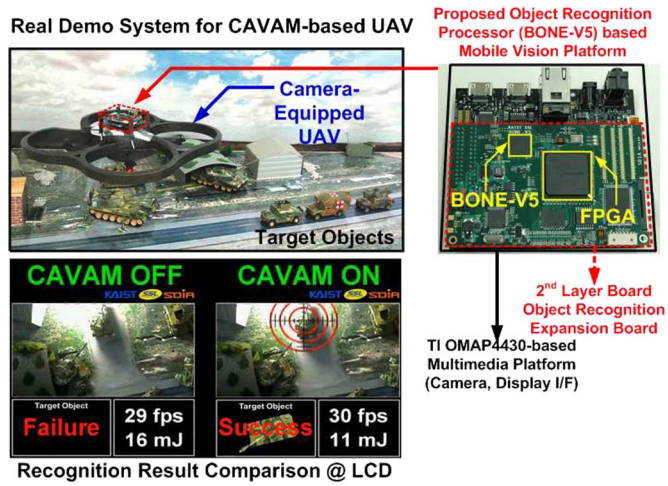


Fig. 13. The mobile vision platform and the UAV-based real demonstration system.

The recognition accuracy measured in terms of the true positive rate is approximately 98.2% with a false positive rate of less than 1.1% for 22 different target objects such as toy tanks, cars and building miniatures. The recognition accuracy of the CAVAM is even to the SIFT implementation without attention operation that obtains theoretical maximum accuracy, while reducing processing time more than 40%. With the help of CAVAM, this processor can obtain 73% attention accuracy, which is 1.44 times higher compared to the previous object recognition processor [8]. As a result, the CAVAM-based processor can provide high-quality recognition performance for 720p HD video applications with low-power consumption.

VII. CONCLUSION

In this paper, we present a real-time object recognition processor for HD 720p video streams in mobile vision system. The context-aware visual attention model is proposed to reduce the on-chip bandwidth of HD video-based object recognition at least 46%. Along with the proposed 5-stage task-level pipeline of SIFT-based object recognition, the heterogeneous multi-core processor employs the simultaneous multithreading clusters for feature extraction and the latency-optimized matching processor for feature matching, and achieves 47700 tiles/sec

and 62720 vectors/sec throughput respectively. With the help of machine learning engine, the dynamic resource controller increases the system utilization and power efficiency at the same time. As a result, the fabricated SoC achieves 30 frame/sec dynamic object recognition for UAV with 720p video streams while dissipating 320 mW at 200 MHz on average, achieving 2.54 times higher energy efficiency with 10.5 nJ/pixel compared to the state-of-the-art vision processors.

REFERENCES

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[2] M. Martinez, A. Collet, and S. S. Srinivasa, "MOPED: A scalable and low latency object recognition and pose estimation system," in *Proc. IEEE Int. Conf. Robotics and Automation*, May 2010, pp. 2043–2049.

[3] S. Agarwala, T. Anderson, A. Hill, M. D. Ales, R. Damodaran, P. Wiley, S. Mullinnix, J. Leach, A. Lell, M. Gill, A. Rajagopal, A. Chachad, M. Agarwala, J. Apostol, M. Krishnan, D. Bui, Q. An, N. S. Nagaraj, T. Wolf, and T. T. Elappuparackal, "A 600-MHz VLIW DSP," *IEEE J. Solid-State Circuits*, vol. 37, no. 11, pp. 1532–1544, Nov. 2002.

[4] C. M. Wittenbrink, E. Kilgariff, and A. Prabhu, "FERMI GF100 GPU architecture," *IEEE Micro*, vol. 31, no. 2, pp. 50–59, 2011.

[5] A. Dehnhardt, M. B. Kulaczewski, L. Friebe, S. Moch, P. Pirsch, H.-J. Stolberg, and C. Reuter, "A multi-core SoC for advanced image and video compression," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Mar. 2005, vol. 5, pp. 665–668.

[6] S. Kyo, S. Okazaki, T. Koga, and F. Hidano, "100 GOPS in-vehicle vision processor for pre-crash safety systems," in *Proc. IEEE Int. Symp. VLSI Circuits*, June 2008, pp. 28–29.

[7] B. K. Khailany, T. Williams, J. Lin, E. P. Long, M. Rygh, D. W. Tovey, and W. J. Dally, "A programmable 512 GOPS stream processor for signal, omage, and video processing," *IEEE J. Solid-State Circuits*, vol. 43, no. 1, pp. 202–213, Jan. 2008.

[8] S. Lee, J. Oh, J. Park, M. Kim, and H.-J. Yoo, "A 345 mW heterogeneous many-core processor with an intelligent inference engine for robust object recognition," *IEEE J. Solid-State Circuits*, vol. 46, no. 1, pp. 42–51, Jan. 2011.

[9] T.-W. Chen, C.-S. Tang, S.-F. Tsai, C.-H. Tsai, S.-Y. Chien, and L.-G. Chen, "Tera-scale performance machine learning SoC (MLSoC) with dual stream processor architecture for multimedia content analysis," *IEEE J. Solid-State Circuits*, vol. 45, no. 11, pp. 2321–2329, Nov. 2010.

[10] S. Rixner, W. J. Dally, U. J. Kapasi, B. Khailany, A. Lopez-Lagunas, P. Mattson, and J. D. Owens, "A bandwidth-efficient architecture for media processing," in *Proc. ACM/IEEE Int. Symp. Microarchitecture*, Nov. 1998, pp. 3–13.

[11] B. Flachs, S. Asango, S. H. Dhong, H. P. Hofstee, G. Gervais, R. Kim, T. Le, P. Liu, J. Leenstra, J. Liberty, B. Michael, H.-J. Oh, S. M. Mueller, O. Takahashi, A. Hatakeyama, Y. Watanabe, N. Yano, D. A. Brokenshire, M. Peyravian, V. To, and E. Iwata, "The microarchitecture of the synergistic processor for a cell processor," *IEEE J. Solid-State Circuits*, vol. 41, no. 1, pp. 63–70, Jan. 2006.

[12] S. R. Vangal, J. Howard, G. Ruhl, S. Dighe, H. Wilson, J. Tschanz, D. Finan, A. Singh, T. Jacob, S. Jain, V. Erraguntla, C. Roberts, Y. Hoskote, N. Borkar, and S. Borkar, "An 80-tile sub-100-W TeraFLOPS processor in 65-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 43, no. 1, pp. 29–41, Jan. 2008.

[13] T. Mori, Y. Ueda, N. Nonogaki, T. Terazawa, M. Sroka, T. Fujita, T. Kodaka, T. Mori, K. Morita, H. Arakida, T. Miura, Y. Okuda, T. Kizu, and Y. Tsuboi, "A power, performance scalable eight-cores media processor for mobile multimedia applications," *IEEE J. Solid-State Circuits*, vol. 44, no. 11, pp. 2957–2965, Nov. 2009.

[14] A. A. Abbo, R. P. Kleihorst, V. Choudhary, L. Sevat, P. Wielage, S. Mouy, B. Vermeulen, and M. Heijligers, "Xetal-II: A low-power massively-parallel processor for video scene analysis," *IEEE J. Solid-State Circuits*, vol. 43, no. 1, pp. 192–201, Jan. 2008.

[15] J. Oh, G. Kim, J. Park, I. Hong, S. Lee, and H.-J. Yoo, "A 320 mW 342 GOPS real-time moving object recognition processor for HD 720p video streams," in *IEEE ISSCC 2012 Dig. Tech. Papers*, 2012, pp. 220–221.

[16] D. M. Tullsen, S. J. Eggers, and H. M. Levy, "Simultaneous multithreading: Maxmizing on-chip parallellism," in *Proc. IEEE Conf. Computer Architecture*, June 1995, pp. 392–403.

[17] J.-Y. Kim, J. Park, S. Lee, M. Kim, J. Oh, and H.-J. Yoo, "A 118.4 GB/s multi-casting network-on-chip with hierarchical star-ring combined yopology for real-time object recognition," *IEEE J. Solid-State Circuits*, vol. 45, no. 7, pp. 1399–1409, July 2010.

[18] G. Semeraro, G. Magklis, R. Balasubramonian, D. H. Albonesi, S. Dwarkadas, and M. L. Scott, "Energy-efficient processor design using multiplie clock domains with dynamic voltage and frequency scaling," in *Proc. IEEE Int. Symp. High-Performance Computer Architecture*, Feb. 2002, pp. 29–40.

[19] S. Lee, J. Kwon, J. Oh, J. Park, and H.-J. Yoo, "A 92 mW 76.8 GOPS vector matching processor with parallel Huffman decoder and query re-ordering buffer for real-time object recognition," in *Proc. IEEE Asian Solid-State Circuits Conference*, Nov. 2010, pp. 1–4.

[20] B. Mei, S. Vernalde, D. Verkest, H. D. Man, and R. Lauwereins, "ADRES: An architecture with tightly coupled VLIW processor and coarse-grained reconfigurable matrix," in *Proc. Int. Conf. Field-Programmable Logic Application*, 2003, pp. 61–70.

[21] P. M. Heysters, G. J. M. Smit, and E. Molenkamp, "Energy-efficiency of MONTIUM reconfigurable tile processor," in *Proc Int. Conf. Engineering of Reconfigurable Systems and Algorithm*, 2004, pp. 38–44.

[22] V. Baumgarte, G. Ehlers, F. May, A. Nückel, M. Vorbach, and M. Weinhardt, "PACT XPP—A self-reconfigurable data processing architecture," *Int. J. Supercomputing*, vol. 26, no. 2, pp. 167–184, 2003.

[23] M.-H. Lee, H. Singh, G. Lu, N. Bagherzadeh, and F. J. Kuradahi, "Design and implementation of the MorphoSys reconfigurable computing processor," *J. VLSI Signal Processing Systems*, vol. 24, pp. 147–164, 2000.

[24] A. Motakis, G. Kornaros, and M. Coppola, "Dynamic resource management in modern multicore SoCs by exposing NoC services," in *Proc. IEEE Int. Workshop on Reconfigurable Communication-Centric Systems-on-Chip*, June 2011, pp. 1–7.

[25] W. Kim *et al.*, "System level analysis of fast, per-core DVFS using on-chip switching regulators," in *Proc. IEEE Int. Symp. High Performance Computer Architecture*, 2008, pp. 123–134.

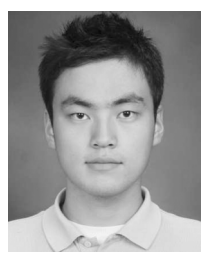
[26] E. Ipek, O. Mutlu, J. F. Martıınez, and R. Caruana, "Self-optimizing memory controllers: A reinforcement learning approach," in *IEEE Int. Symp. Computer Architecture*, June 2008, pp. 39–50.

[27] K. Goossens, J. Dielissen, and A. Radulescu, "Æthereal network on chip: Concepts, architecture, and implementations," *IEEE Design & Test of Computers*, pp. 414–421, Sept.–Oct. 2005.

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

[29] T. Kurafuji, M. Haraguchi, M. Nakajima, T. Gyoten, T. Nishijima, H. Yamasaki, Y. Imai, M. Ishizaki, T. Kumaki, Y. Okuno, T. Koide, H. J. Mattausch, and K. Arimoto, "A scalable massively parallel processor for real-time image processing," in *IEEE ISSCC Dig. Tech. Papers*, 2010, pp. 334–335.

[30] J.-Y. Kim, M. Kim, S. Lee, J. Oh, K. Kim, and H.-J. Yoo, "A 201.4 GOPS 496 mW real-time multi-object recognition processor with bio-inspired neural perception engine," *IEEE J. Solid-State Circuits*, vol. 45, no. 1, pp. 32–45, Jan. 2010.



Jinwook Oh (S'08) received the B.S. degree in electrical engineering and computer science from Seoul National University, Seoul, Korea in 2008 and the M.S. degree in electrical engineering and computer science from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2010. He is currently working toward the Ph.D. degree in electrical engineering and computer science from KAIST.

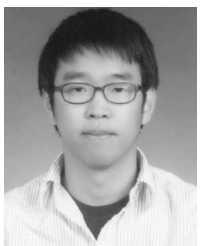
His research interests include digital signal processors for computer vision. Recently, he is involved with multi-core processor design with dynamic resource management.

Gyeonghoon Kim (S'10) received the B.S. and M.S. degrees from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2009 and 2011, respectively, where he is currently pursuing the Ph.D. degree in electrical engineering.

His research interests include low-power digital processors with dynamic resource management for computer vision and Network-on-Chip (NoC) based SoC design.

Junyoung Park (S'09) received the B.S. and M.S. degrees in electrical engineering and computer science from KAIST, Daejeon, Korea, in 2009 and 2011, respectively, where he is currently working toward the Ph.D. degree in the same department.

His previous and recent research interests include development of the parallel processors for computer vision and many-core architecture and VLSI implementation for bio-inspired vision processor.



Injoon Hong (S'11) received the B.S. degrees in electrical engineering and computer science from KAIST, Daejeon, Korea, in 2011, where he is currently working toward the M.S. degree in the same department.

His previous and recent research interests include development of the digital accelerator and micro-architecture for computer vision, and VLSI implementation for machine learning algorithms.

Seungjin Lee (S'06–M'12) is a researcher at the Department of Electrical Engineering in the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, where he received the B.S., M.S., and Ph.D. degrees in 2006, 2008, and 2011, respectively.

His previous research interests include low power digital signal processors for digital hearing aids and body area communication, and novel architectures for accelerating neural networks on a system-on-chip. Currently, he is investigating system-on-chip architectures for computer vision efficient heterogeneous processing.

Joo-Young Kim (S'05–M'12) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2005, 2007, and 2010, respectively. His Ph.D. research was focused on system-on-a-chip (SoC) architectures and circuit innovations for low energy object recognition processing. Three parallel SoCs architected and designed by him won two ISSCC presentations and two ISSCC/DAC design contests from 2008 through 2010.

Since 2010, he has been with Microsoft Research, Redmond, WA, where he is currently working as a research hardware design engineer. He is engaged in the research and development of client and cloud applications which target disruptive end-to-end consumer experiences enabled by innovative natural user interfaces and cloud-backed services. More specifically, he is playing a key role in inventing customized architectures and circuits to enable computationally intensive workloads such as 3D computer vision, motion extraction, and data compression for mobile devices.

Dr. Kim has authored and coauthored more than 35 journal publications and conference presentations in the solid-state circuits area including IEEE International Solid-State Circuits Conference (ISSCC), Symposium on VLSI Circuits (VLSIC), and the IEEE JOURNAL OF SOLID-STATE CIRCUITS. He has been invited to talk about his research on vision processors at industrial and research organizations in Intel, nVidia, Texas Instruments, and IMEC.

Jeong-Ho Woo (A'12) received the B.S., M.S., and Ph.D. degrees in electrical engineering from KAIST, Korea, in 2002, 2004, and 2008, respectively.

During the period in KAIST, he developed mobile unified shader for low-power fully programmable 3D graphics and low-power multimedia SoC capable of image, video, and 3D graphics. His research interests include low-power high-performance digital circuits and multimedia system design with specific interest in 3D computer graphics and multimedia processing architecture. Currently, he is with Texas Instruments, Inc., Dallas, TX, designing multimedia system architecture for mobile devices.

Hoi-Jun Yoo (M'95–SM'04–F'08) graduated from the Electronic Department of 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, in 1985 and 1988, respectively.

Since 1998, he has been on the faculty of the Department of Electrical Engineering at KAIST and now is a full Professor. From 2001 to 2005, he was the Director of Korean System Integration and IP Authoring Research Center (SIPAC). From 2003 to 2005, he was the full-time Advisor to Minister of Korea Ministry of Information and Communication and National Project Manager for SoC and Computer. In 2007, he founded SDIA (System Design Innovation & Application Research Center) at KAIST. Since 2010, he has served the general chair of Korean Institute of Next Generation Computing. His current interests are computer vision SoC, Body Area Networks, biomedical devices and circuits. He is a co-author of *DRAM Design* (Korea: Hongleung, 1996), *High Performance DRAM* (Korea: Sigma, 1999), *Networks on Chips* (Morgan Kaufmann, 2006), *Low-Power NoC for High-Performance SoC Design* (CRC Press, 2008), *Circuits at the Nanoscale* (CRC Press, 2009), *Embedded Memories for Nano-Scale VLSIs* (Springer, 2009), *Mobile 3D Graphics SoC from Algorithm to Chip* (Wiley, 2010), and *Bio-Medical CMOS ICs* (Springer, 2011).

Dr. Yoo received the Electronic Industrial Association of Korea Award for his contribution to DRAM technology in 1994, Hynix Development Award in 1995, the Korea Semiconductor Industry Association Award in 2002, Best Research of KAIST Award in 2007, Scientist/Engineer of this month Award from Ministry of Education, Science and Technology of Korea in 2010, Best Scholarship Awards of KAIST in 2011, and Order of Service Merit from Ministry of Public Administration and Security of Korea in 2011 and has been co-recipient of ASP-DAC Design Award 2001, Outstanding Design Awards of 2005, 2006, 2007, 2010, 2011 A-SSCC, Student Design Context Award of 2007, 2008, 2010, 2011 DAC/ISSCC. He has served as a member of the executive committee of ISSCC, Symposium on VLSI, and A-SSCC, and the TPC chair of the A-SSCC 2008 and ISWC 2010, IEEE Fellow, IEEE Distinguished Lecturer ('10-'11), and Far East Chair of ISSCC ('10-'11).