An Energy Efficient Full-Frame Feature Extraction Accelerator With Shift-Latch FIFO in 28 nm CMOS

Dongsuk Jeon, *Student Member, IEEE*, Michael B. Henry, Yejoong Kim, Inhee Lee, *Student Member, IEEE*, Zhengya Zhang, *Member, IEEE*, David Blaauw, *Fellow, IEEE*, and Dennis Sylvester, *Fellow, IEEE*

***Abstract*—This paper presents an energy-efficient feature extraction accelerator design aimed at visual navigation. The hardware-oriented algorithmic modifications such as a circular-shaped sampling region and unified description are proposed to minimize area and energy consumption while maintaining feature extraction quality. A matched-throughput accelerator employs fully-unrolled filters and single-stream descriptor enabled by algorithm-architecture co-optimization, which requires lower clock frequency for the given throughput requirement and reduces hardware cost of description processing elements. Due to the large number of FIFO blocks, a robust low-power FIFO architecture for the ultra-low voltage (ULV) regime is also proposed. This approach leverages shift-latch delay elements and balanced-leakage readout technique to achieve 62% energy savings and 37% delay reduction. We apply these techniques to a feature extraction accelerator that can process 30 fps VGA video in real time and is fabricated in 28 nm LP CMOS technology. The design consumes 2.7 mW with a clock frequency of 27 MHz at $V_{dd} = 470$ mV, providing 3.5× better energy efficiency than previous state-of-the-art while extracting features from entire image.**

***Index Terms*—Energy efficient DSP, feature extraction, first-in first-out, near-threshold design, pipeline.**

I. Introduction

In the last decade, computer vision has been widely applied to many different fields. In medical imaging such as MRI or CT, images are analyzed using computer vision techniques to realize fully or partially automatic diagnosis [1], [2]. Recent advanced surveillance camera systems not only record video, but also provide functions including facial recognition and motion detection [3]. Automobile manufacturers incorporate various cameras on vehicles and analyze their external environment to improve driving safety or achieve self-driving functionality [4]. Although computer vision algorithms typically require substantial computing power to process multiple frames per second in real time, conventional applications such as those mentioned above have rather large power budgets and hence supporting these computational requirements has been feasible by using multi-core systems or GPUs that consume tens of watts [2], [4].

Manuscript received November 04, 2013; revised January 21, 2014; accepted February 26, 2014. Date of publication March 11, 2014; date of current version April 21, 2014. This paper was approved by Associate Editor Stefan Rusu.

The authors are with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: djeon@umich.edu).

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.2014.2309692

Recently, mobile battery-powered systems such as cellular phones, micro-robots and millimeter-sized sensor nodes have gained significant attention. Due to technology scaling and the development of new low-power techniques, these application areas continue to flourish, incorporating more functionality with time [5], [6]. Computer vision techniques can add significant value to these classes of systems, providing various useful features such as object recognition in phones or navigation and surveillance in micro-robots. However, the tight power constraints of these systems prevent practical implementations of computer vision algorithms. We therefore seek to significantly reduce hardware cost and power consumption associated with such algorithms.

In this paper, we propose a highly energy-efficient feature extraction accelerator design for visual navigation of micro-autonomous vehicles. The navigation algorithm must process 30 fps VGA video while consuming less than 30 mW power due to limited power budget of miniaturized system. We first propose a modified feature extraction algorithm that improves energy efficiency while maintaining feature extraction quality. We then apply architectural and circuit techniques including a robust low-power FIFO for subthreshold operation, further reducing power consumption. The resulting design achieves 2.7 mW power consumption at 470 mV supply voltage when extracting features from 640 × 480 VGA 30 fps video continuously at a low clock frequency of 27 MHz. The design realizes a 3.5× energy efficiency improvement over prior work.

II. Proposed Visual Feature Extraction Algorithm

A. Visual Feature Extraction

Visual feature extraction is a key step in many computer vision algorithms. Essentially it extracts useful information from a visual source such as an image, and this information can be used in a variety of applications including object recognition and pose estimation. Fig. 1 shows an example of the widely-used SIFT (scale-invariant feature transform) algorithm [7]. Feature extraction is performed on the original image (left), and small rectangles depict extracted features with different scales and orientations. These are then compared to features already stored in the database and finally some objects are recognized (right). Generally feature extraction should provide scale and rotation invariance for reliable extraction under different circumstances or viewpoints, as shown in Fig. 2.

SURF (Speeded-Up Robust Features) is a well-known variation of the SIFT algorithm. The authors of [8] claim it achieves identical or even superior extraction quality while

Fig. 1. An example of object recognition using extracted features [7].

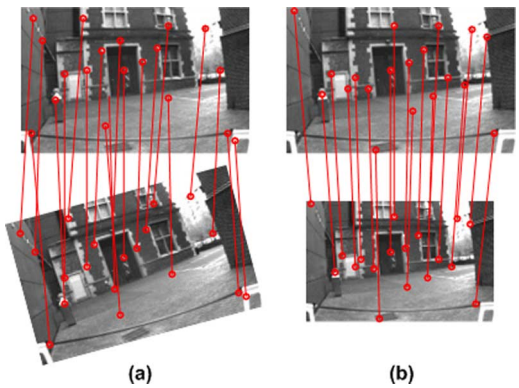


Fig. 2. Two key constraints of feature extraction algorithms: (a) rotation and (b) scale invariance.

reducing computational cost significantly, making it attractive for low-power applications. SURF consists of two distinct stages: detection and description. In the detection stage, an input image is first processed with multiple filters at different scales. Filter responses calculated simultaneously at different scales provide the scale invariance property. The algorithm then searches for interest points (local maxima) in 3-D scale-location pyramids. Although local maxima points can be extracted using simple digital comparators, the actual maxima point can reside somewhere between adjacent pixels and matrix-based equations are used to interpolate the maxima point in 3-D space. The description stage is responsible for describing each interest point and generating a corresponding final feature vector. For rotation invariance, the orientation of each interest point must be determined first. The description stage collects filter responses around it and searches for an angle which has largest filter responses using rotating sampling window as depicted in Fig. 3(a). After choosing orientation, a rectangular sampling region is rotated by that angle and filter responses are again collected in that region (Fig. 3(b)). This guarantees that collected responses around each interest point remain unchanged in images rotated by any angle. The sampling region is divided into small rectangles and a summation of sampling responses in each sub-region constitutes each dimension of the feature vector. Finally, this vector is normalized such that vectors extracted from different scale images have identical magnitude.

B. Proposed Hardware-Oriented Feature Extraction Algorithm

We apply the SURF algorithm to our design target, a MAV (micro air vehicle) with visual navigation shown in Fig. 4, where

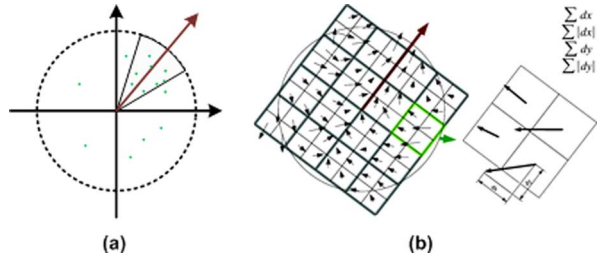


Fig. 3. Original feature vector generation process consisting of (a) orientation assignment and (b) feature vector generation [8].

feature extraction is a key function and a dominant power consumer. The MAV is designed to fly and navigate in indoor environments using various sensors to recognize obstacles and a camera for location search. Fig. 5 provides an overview of the visual navigation system [9]. First, an on-board camera captures 30 fps VGA video, which is fed into the proposed feature extraction accelerator. The feature extraction accelerator then extracts 64-dimensional SURF features that are compared to location database storing features from previously visited locations. If any match is found, it can be concluded that the test vehicle has returned to a location visited before and a loop closure is declared. Finally, this loop closure information is used in an algorithm called SLAM (Simultaneous Localization and Mapping [10]). SLAM continuously monitors the environment to determine current location and generate a map. Physical sensors such as gyroscopes, accelerometers, and lasers provide primary information on vehicle movements, but small errors accumulate over time and cause localization to fail at some point. Loop closure information from previous steps is used in this SLAM algorithm to compensate for these errors. In this class of system, feature extraction is one of the most computationally expensive steps, and our work therefore focuses on the design of a corresponding accelerator.

Since MAVs can move rapidly, they must perform both accurate and fast feature extraction. In addition, location monitoring should be done continuously, however a direct implementation on an X86 embedded processor consumes more than 1 W while processing only a few fps (frame per second) VGA video. Related work on custom-designed hardware for similar applications also report $>$50 mW power consumption [11]–[14] for processing partial images based on ROIs (Regions of Interest). However, this system has a tight power budget of $\sim$30 mW for digital processing due to a minimum required operation time without recharging. This power budget includes feature extraction as well as other functions such as feature mapping and navigation and arises based on the allocation of the vast majority of power consumption to actuation assuming a 3 W-hr 15 g Li-ion battery in a 100 g flier with 1 hour battery life.

One widely used technique to reduce power consumption in image processing is the extraction of ROIs. A low-cost pre-processing stage is inserted before the actual feature extraction step to search for small regions believed to have meaningful information or targeted objects. An input image is divided into many smaller tiles and only a subset of these is chosen for further processing. Although this can significantly reduce power consumption, the performance of the pre-processing algorithm dictates

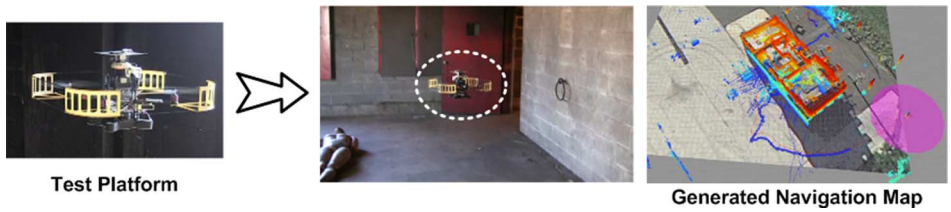


Fig. 4. A target application of MAV with visual navigation.

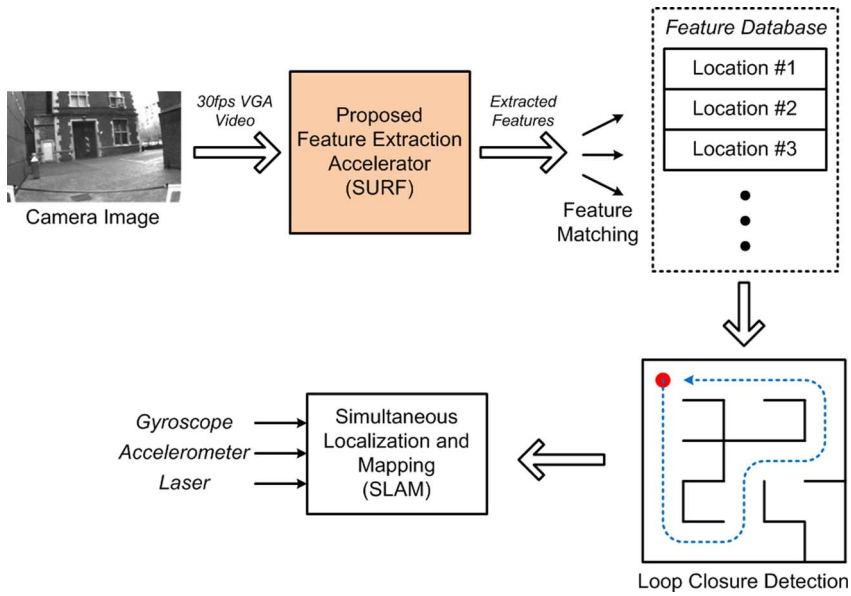


Fig. 5. An overview of the visual navigation algorithm flow.

the overall quality of feature extraction. ROI detection should also be trained in advance on the database containing specific classes of objects desired to be detected. However, to enable visual navigation in unknown environments, it is impossible to rely on specific objects to determine its current location. Therefore, our target application compares each captured image on a scenery basis (not individual objects), necessitating feature extraction from the entire frame.

To achieve low power while performing full-frame feature extraction, we optimize the original SURF algorithm with the goal of an energy-efficient hardware implementation without using an ROI-based approach. For the detector, first we use a single-octave scale space (Fig. 6(a)). In original SURF, the detection stage first builds scale pyramids to detect interest points in different scales [8]. Basically the filter size can be continuously increased until it reaches the entire image size for detection across all possible scales. However, to reduce computation we define a new pyramid after 4 filter size increases. In the new pyramid, both interest point search step and filter size increase are doubled for coarse searches in a larger scale. Since the target resolution is 640×480, only a small portion of features are extracted from larger objects or patterns and therefore reside in the second or higher scale pyramids. We choose the first (smallest) octave among them to extract dominant smaller features. Simulation results show significant amount of feature loss in this case, so we need to have (at least a part of) the 2nd octave to compensate for the loss. However, we need to add at least 3 new filters since the smallest and largest filters of each octave are only used for comparisons and do not have interest points. Instead, we extend the 1st octave and employ an additional filter (size 33) that lies between 1st and 2nd octaves to compensate for lost features. The resulting algorithm extracts more than 94% of originally extracted features while reducing filter power consumption by 38% compared to the original SURF algorithm with 5 octaves. After local maxima detection, the exact original location of the maxima is typically calculated using matrix-based arithmetic operations. Instead, we employ a fast localization technique for interpolation as described in Fig. 6(b).

In the description stage, a large and variable number of interest points marked by the detector must be processed. Previously a multi-core architecture has been proposed to deal with the variable throughput of this step [11]–[13]. As discussed in the previous section, for each interest point two separate filter response sampling steps are required for orientation assignment and actual description, respectively. In other words, the complete filter responses around each interest point should be transferred to a description core responsible for describing that point. These responses also have to be stored temporarily in data memory within each core for later steps. This necessitates a large buffer in each description core, which incurs a large area and power overhead.

We therefore propose a circular-shaped sampling region that unifies orientation assignment and description into one step as shown in Fig. 7. The authors in [15] compare polar grid sam-

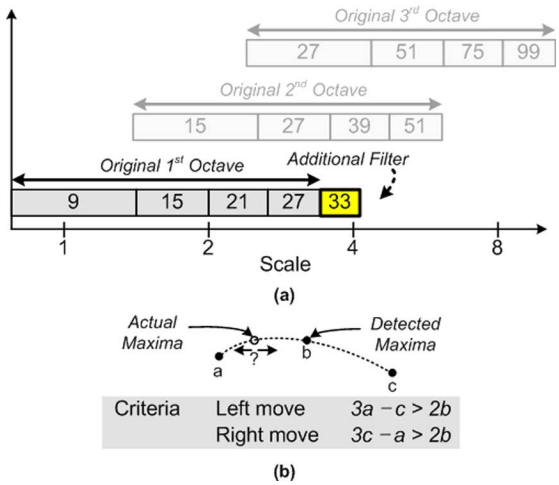


Fig. 6. Proposed (a) single-octave scale space; and (b) fast localization techniques for detector optimization.

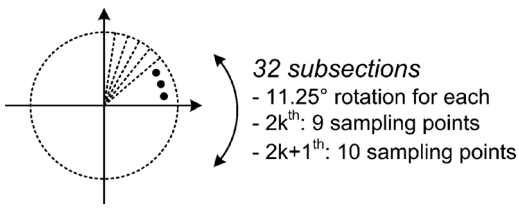


Fig. 7. Proposed circular-shaped sampling region approach.

plings and a rectangular grid, shedding light on the possibility of using a rotation invariant sampling region. However, to avoid two separate sampling methods and use all available sampling points within a circle, the proposed sampling region is still based on the original rectangular grid. Instead, it is divided into 32 subsections and a vector representing an interest point is generated based on the summation of filter responses in each subsection. Since the number of points in even- and odd-numbered subsections are different, the kth angle is composed of filter responses gathered in both kth and $(k+1)$th subsections such that all angles have the same number of sampling points. The interest point orientation can be easily determined by the subsection with the largest summation value.

Since the shape and coverage of a circular-shaped sampling region do not change when rotated by the assigned orientation, filter responses do not need to be re-collected for actual description. Furthermore, by restricting orientation angles to discrete values represented by each subsection, final feature vectors can be generated by simply re-ordering vector dimensions. Although this technique provides only discrete step rotation, the use of 32 subsections translates to a small rotation step of only 11.25° while providing the same feature dimensions as the original SURF algorithm. Due to the unified description, each description processing element does not have to store entire filter responses and instead just accumulates them into 32-dimensional vectors in real time, reducing memory requirements in each processing element by 89% and element area by 80%. This technique also enables other hardware design simplifications that are discussed in detail in the Section III.

Since our target application uses scenery-based matching, we tested the algorithm with actual videos captured by a robotic test vehicle [9] rather than conventional object-oriented testbenches. The modified SURF was tested on a database consisting of 100 frames extracted from these videos. Fig. 8 demonstrates the measured feature extraction quality metric, which is important in visual navigation and is defined by the ratio of the number of correctly matched features to the number of all matched features between original and re-scaled or rotated images. Fig. 8(a) and (b) confirm that the scale and rotation invariance performance of the proposed and original SURF algorithms are very similar. We observed that the proposed algorithm provides 30% fewer valid feature match count on average due to its use of limited filter scales. This is deemed an acceptable tradeoff for the targeted navigation application, for which feature match ratio was deemed more critical. To enable a larger scale invariance range for other applications, more Gaussian filters can be added to detect and describe larger features or the input image can be subsampled and processed through the accelerator repeatedly.

III. Energy-Efficient Hardware Architecture

A. Accelerator Architecture

Voltage scaling is a widely used and effective power-saving technique [16]–[18], but it incurs large performance penalties that are unacceptable in high throughput systems. Feature extraction algorithms are generally computationally expensive and SIFT/SURF algorithms require throughput on the order of GOPS or higher. In addition, the number of features in each frame varies widely and hence peak performance requirements can be much higher than typical performance. Therefore, a feature extraction accelerator must be designed carefully to effectively incorporate aggressive voltage scaling while also meeting high performance requirements.

Fig. 9 shows the overall architecture of the proposed accelerator design. To deal with the low clock frequencies associated with deep voltage scaling, the accelerator is uniquely designed to take only one pixel of input image per cycle at the low speed of 27 MHz. In addition, the entire accelerator operates at the same clock frequency, resulting in a matched-throughput system. A 640×480 8-bit grayscale input image is divided into 11 subsections, as shown in Fig. 10, and they are processed successively. Subsections are partially overlapped to allow the accelerator to extract features from the entire area including borders between subsections.

Each subsection has 640×124 pixels. In each cycle, only one pixel of the input image is fed into the proposed accelerator. The input image flow is first integrated in 2-D and Gaussian box filters with different scales are applied. Filter responses form a 3-D scale-location space and a local maxima detector searches for the interest points in this space. Finally, an interpolator determines the exact location of maxima using the proposed simplified maxima detection technique. While the detector is searching for interest points, the input image must be delayed temporarily. Therefore, we delay the input image using a 7067-entry FIFO at the input stage of descriptor shown

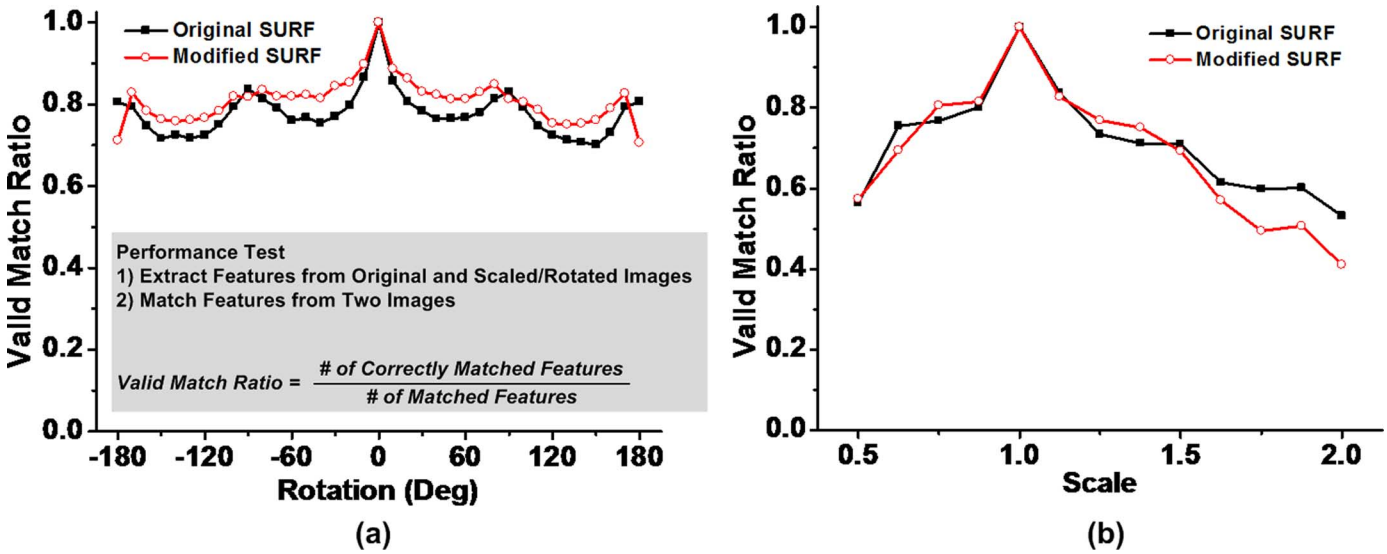


Fig. 8. (a) Rotation; and (b) scale invariance performance comparisons.

in Fig. 9. Since the largest radius of sampling regions in the detector and descriptor are 16 and 40 pixels, respectively, the buffer must store 56 lines while the detector searches for local maxima. Additional margins from pipeline and control signals translate to a 7067-entry FIFO. Then it is integrated in 2-D in the same way as in the detection stage. Although the input image is integrated identically in the detector and descriptor, the use of separate integrators actually reduces silicon area by minimizing FIFO size. Since the original 8-bit input image becomes 18-bit after integration due to larger dynamic range, FIFO area is reduced by 95,000 μm^2 (56%) while overhead from the additional integrator is only 9,700 μm^2.

The integrated image goes through Haar wavelet filters in different scales, which provides the necessary filter responses for feature description. While the interest point information from the detector is passed to descriptor processing elements in real time, one of the idle processing elements is assigned to each interest point. Each processing element captures the Haar wavelet filter responses around each point and generates feature vectors. The proposed design uses 40 processing elements in total, and they are power-gated when not in use. The number of processing elements is chosen to provide more than $2\times$ margin compared to the maximum number of features being extracted simultaneously at one location in actual test images. Finally, a post processor reorders, normalizes, and rotates generated feature vectors and produces the final output. Additional hardware techniques are applied to further optimize each component, and these will be described in the following sections.

[19] presents an early effort to adopt a similar dataflow and architecture. However, it is not fully matched-throughput system and remains partially based on the use of reconfigurable cores, which requires $> 3\times$ faster clock frequency for the same video throughput (increasing power). In addition, a large buffer memory of 2.8 Mb (compared to 56 kb FIFO for delaying the image in the proposed design) is required before the descriptor due to multi-stage description, and peak performance is limited to 890 interest points per frame.

B. Parallelized Filters and Arithmetic Blocks

Two different types of filters are used in the detector and descriptor, but their operation is very similar and is based on simple arithmetic operations on the integrated image. Both Gaussian box filters and Haar wavelet filters are based on the summation of an image, which can be easily achieved by 2-D integrated image and simple arithmetic operations such as addition and subtraction as shown in Fig. 11. In conventional multi-core architectures, this can be calculated using a single arithmetic unit and processing one (or a few using a SIMD architecture) set of data in each cycle. However, the entire image must be stored in a large memory and power overhead is incurred in accessing this large memory every cycle. In addition, multiple operations are required to obtain filter responses at one point and, therefore, the system must operate at a much higher clock frequency, limiting aggressive voltage scaling. Although each summation over a rectangular region requires only 4 data read and 3 arithmetic operations, the current approaches still consume significant power when applied over an entire frame.

To mitigate this, we apply a fully unrolled and parallelized architecture to Gaussian box filters and Haar wavelet filters. First, the input image is delayed by differing numbers of cycles using different size FIFOs. As the input image continues to be processed, images with varying delays appear at the FIFO outputs and they are used for filter response calculation at this point. Once all FIFOs are completely filled with data, three arithmetic operations can be performed simultaneously using a $3\times$ lower clock frequency. In the final design, due to deeply parallelized filter architectures one pixel of the input image is fed into the accelerator at a fixed speed; this allows all processing including detection and description to be done at the same low clock frequency. Therefore, this architecture allows for a single clock domain of 27 MHz over the entire accelerator and provides greater headroom for voltage scaling compared with an X86 single core implementation that requires 1 GHz clock frequency for a few fps throughput. In addition, each cycle data is generated by relatively small FIFOs instead of a large memory, which reduces energy consumed in data readout as well. Different size filters

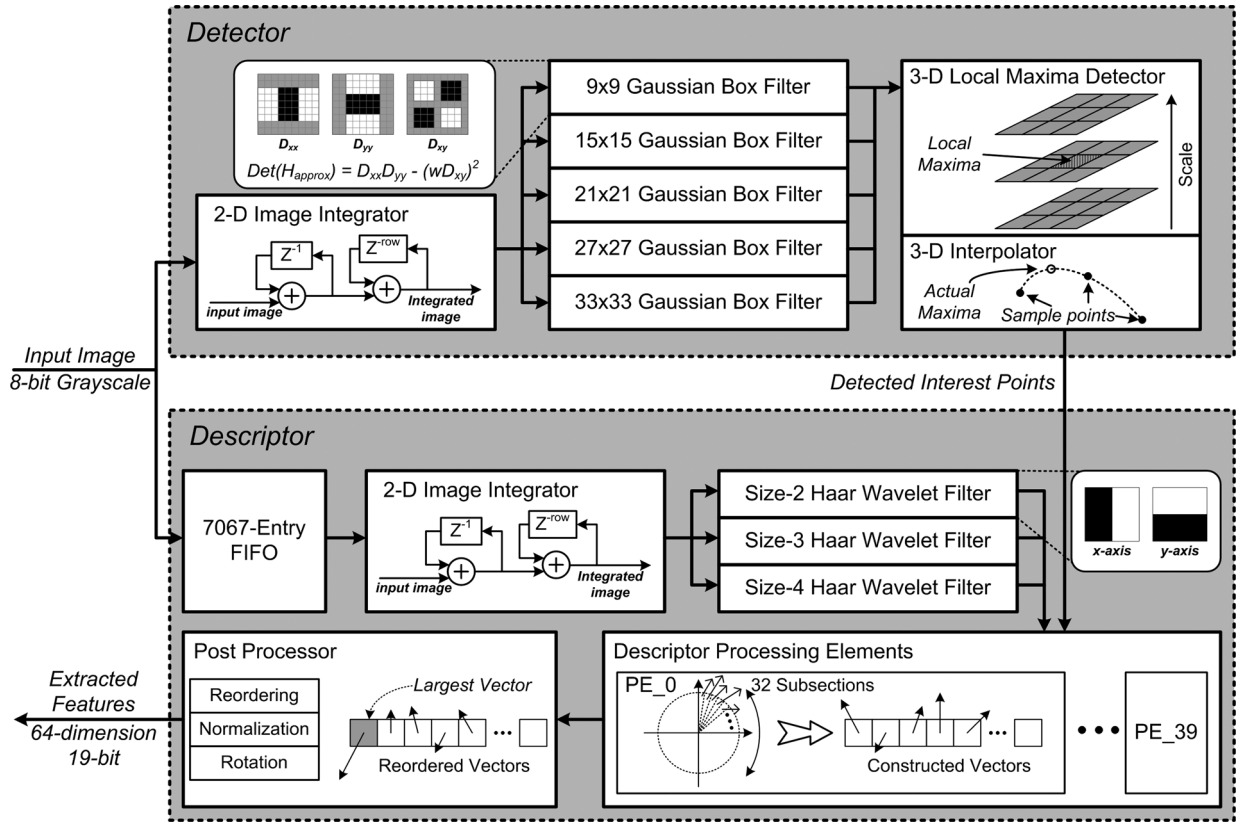


Fig. 9. Proposed feature extraction processor architecture.

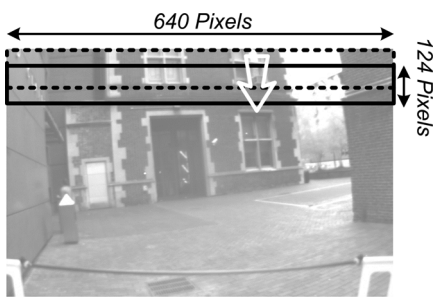


Fig. 10. Overlapped image subsections are processed successively to allow for proper feature extraction at boundaries.

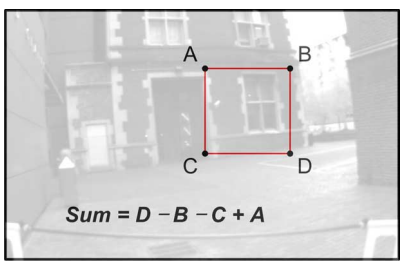


Fig. 11. Image summation in a rectangular region implemented with 3 arithmetic operations on a 2-D integrated image.

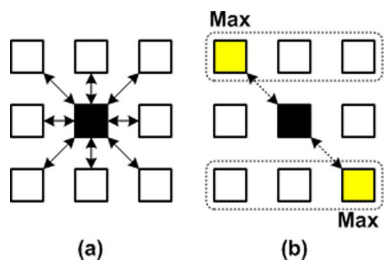


Fig. 12. (a) Original and (b) proposed local maxima detection schemes. In (b), maximum point of each row is already stored and only one comparison per row is required.

can be easily implemented using the same architecture with adjusted delays.

Similarly, the 2-D image integrator can be implemented using only two adders and one 124-entry FIFO, which produces one pixel of the integrated image per cycle in real-time. A 3-D local maxima detector applied after the Gaussian box filters searches for local maxima in the $3 \times 3 \times 3$ location-scale space. A total of 26 subtractions must be calculated in each cycle to determine if a given point is larger than all neighboring points. However, the amount of computation can be reduced significantly by reusing previous results. In each cycle, the lower 3 pixels of each scale are processed and the location of the maximum value among them is attached to the lower middle pixel as an additional 2 bits. Each target point can then be compared to only 8 pixels (maxima of each row) rather than 26 (Fig. 12), reducing the number of comparisons by 69%.

The proposed accelerator processes one pixel to determine if it is a local maxima in 3-D space and it requires 5 filter responses calculations per cycle. Since each filter performs 32 operations for D_{xx}, D_{xy}, and D_{yy} calculations, the proposed architecture

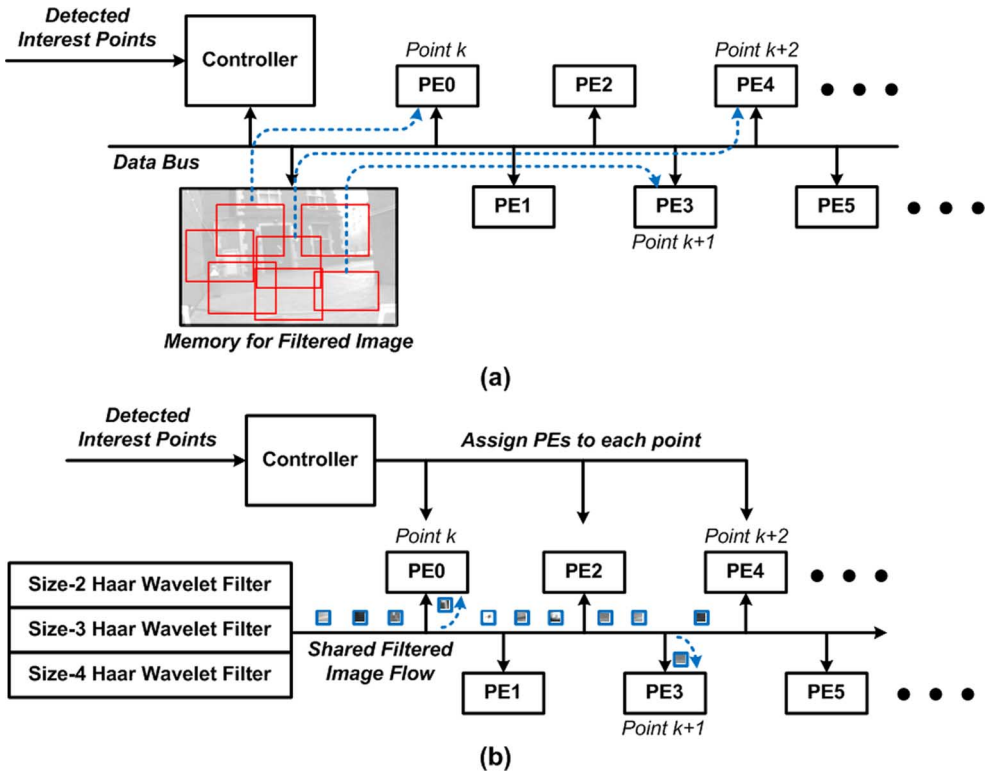


Fig. 13. (a) Conventional multi-core architecture where each core communicates through a shared data bus independently. (b) Proposed architecture where a single response flows continuously through shared data bus and each core reads in only its required blocks.

has roughly 160 arithmetic blocks in parallel, providing a large degree of parallelization.

C. Single Stream Descriptor

Interest points extracted by the detector are continuously passed to the descriptor with each point assigned to an idle processing element (PE). Based on responses of Haar wavelet filters, the set of PEs must simultaneously process a large number of interest points depending on the input image. Therefore, the descriptor must offer high peak performance while maintaining low power consumption. This is handled through the use of many PEs, however this incurs high hardware cost, particularly for data memory used to temporarily store filter responses around an interest point. A conventional design uses a multi-core architecture as shown in Fig. 13(a). An independent controller manages filter responses stored in a large central data memory, and the entire sampling region around an interest point should be passed to a PE once the controller makes a PE assignment. When the number of interest points is high, significant data is transferred through a shared data bus, which requires a high-speed data bus operating at a high clock frequency [13]. Furthermore, overlapping regions sent to multiple PEs incur further overhead. After each PE receives sampling responses and stores them in local memory, it calculates feature vectors through orientation assignment and the actual description step.

However, the proposed circular-shaped sampling region discussed in Section II-B unifies these two steps while removing the need for storing responses in local memory. Based on this algorithm-architecture co-optimization, we propose the single stream descriptor described in Fig. 13(b). In this architecture, filter responses continuously flow through a shared data channel at a fixed low speed such that all processing elements see the same data stream. Filter responses leave the filter bank and reach all of processing elements in the same cycle. At a low operating voltage, wire delay is negligible compared to logic delay and filter responses are simply repeated using inverters. Since interest points are assigned in advance, PEs can easily identify the proper filter responses and capture data from the channel at the appropriate time interval. Since entire filter responses are transferred through a shared data channel (regardless of the number of interest points), this channel can be realized with a matched-throughput low speed data bus. This point removes the need for bus synchronizer and makes it possible to run the bus in the same low voltage domain, which removes overheads from an additional voltage regulator. In addition, this removes redundant data transmission for overlapped sampling regions, eliminating unnecessary switching.

IV. Latch-Based Low-Power and Robust FIFO Design

The proposed accelerator architecture requires a large number of storage elements (FIFOs) across all sub-blocks. In particular, the 7067-entry FIFO at the input stage of the descriptor can consume appreciable leakage and switching power, and both the Gaussian box filters and Haar wavelet filters have many smaller FIFO blocks. It is therefore critical to choose a low-power FIFO block that also offers robust behavior at near- or sub-threshold regime to facilitate aggressive voltage scaling. This last requirement is challenging as there are several known problems in low-voltage memory design.

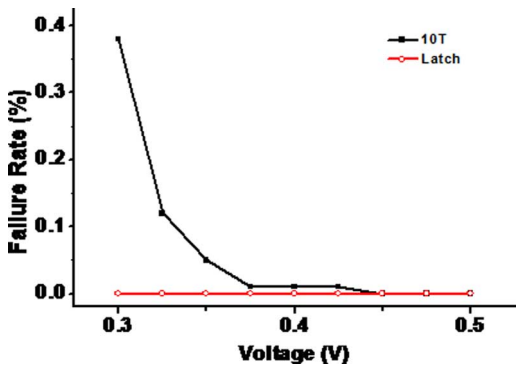


Fig. 14. 10T SRAM and latch write operation failure rates from 10 k Monte-Carlo SPICE simulations.

First, very low on-off current ratios significantly degrade read and write margins, impeding robust operation. Second, the impact of process variation at low voltage is magnified, causing problems for large memory arrays where any single storage element could fail. Conventionally, FIFOs are implemented with shift registers or 6T SRAM and a cyclic address generator [20], [21]. SRAM is an attractive solution in the super-threshold regime due to its small area and low power consumption. However, under aggressive voltage scaling its operating margins nearly disappear with common failures below some $V_{cc,min}$ value. Furthermore, SRAM bitcell suffers from large variability due to small device sizes and read/write tradeoff and their relatively slow access time can become a bottleneck at the system level in throughput-constrained applications. Robustness issues can be overcome by adding more transistors (e.g., 8T or 10T), at the cost of area and power, while slow access speeds remain [22], [23]. On the other hand, shift registers are both very fast and robust even at very low operating voltages. However, the density is several times worse than SRAM since each storage cell consists of two latches. Master and slave latches switch every cycle and, therefore, a shift register approach also consumes much higher switching power, exacerbated by the need to propagate data in every cycle. Fig. 14 shows write operation reliability comparisons between a 10T bitcell (known to be among the best low voltage SRAM bitcells) and a latch. In 10 k Monte-Carlo simulations at 27 °C the 10T cell starts to fail at 425 mV while the latch operates without failure down to 300 mV. In this case the high variability of the 28 nm technology limits voltage scalability of 10T SRAM, and although voltage boosting techniques could be applied to the SRAM, this adds design complexity that is not needed in the proposed design using latches. Furthermore, the proposed design requires >100 FIFO blocks with 20 different sizes; using latch-based memory will reduce design cost significantly.

To overcome these issues in conventional FIFO designs, we propose a new FIFO architecture based on latches. The approach starts with a conventional shift register and replaces all storage cells with latches; hence this approach is called *shift-latch*. It is impossible to move all data simultaneously since latches are level-sensitive such that enabling all latches would lead to the entire path becoming transparent. However, data can be propagated using a one-hot encoded enable signal that moves in the opposite direction each cycle, as depicted in Fig. 15. Initially only the 4th latch is enabled and the value from the previous latch is written into this latch. As a result, both the 3rd and 4th latches now have identical values with the 3rd latch becoming a redundant cell, which we call a *bubble*. In the next cycle, the enable signal is asserted at a location one stage earlier, i.e., the 3rd latch is enabled in Fig. 15. This latch then accepts data from the 2nd latch, which then becomes the bubble. In the following cycle, enable signal is staggered again and the 2nd latch is enabled. As a result, data moves forward and the bubble moves backward again. Finally, the 1st latch is enabled and input data

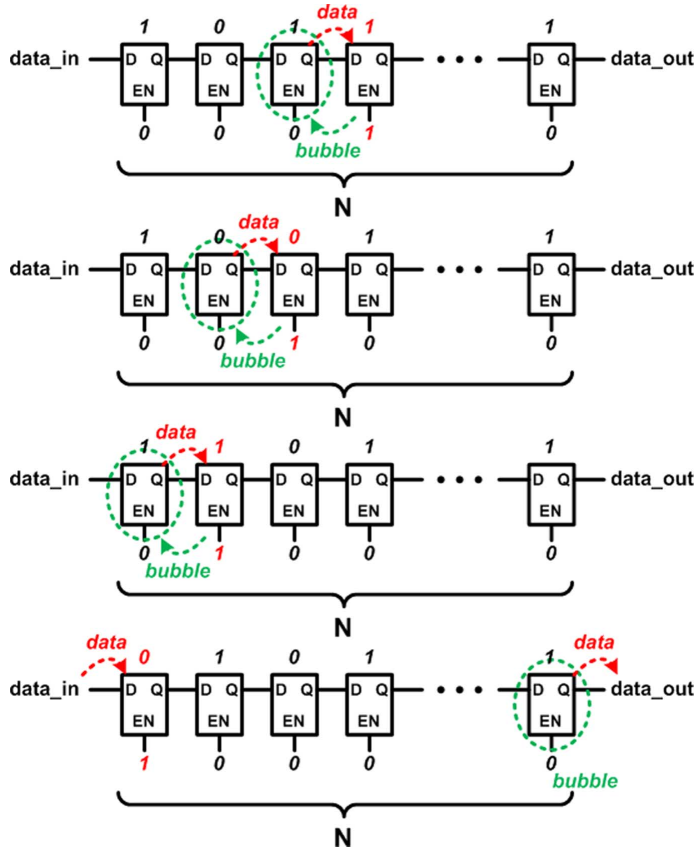


Fig. 15. Proposed single-lane shift latch propagating data and a bubble in opposite directions at each cycle.

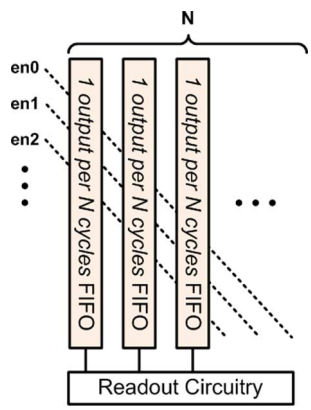


Fig. 16. A one-output-per-cycle FIFO consisting of N lanes and shared readout circuitry.

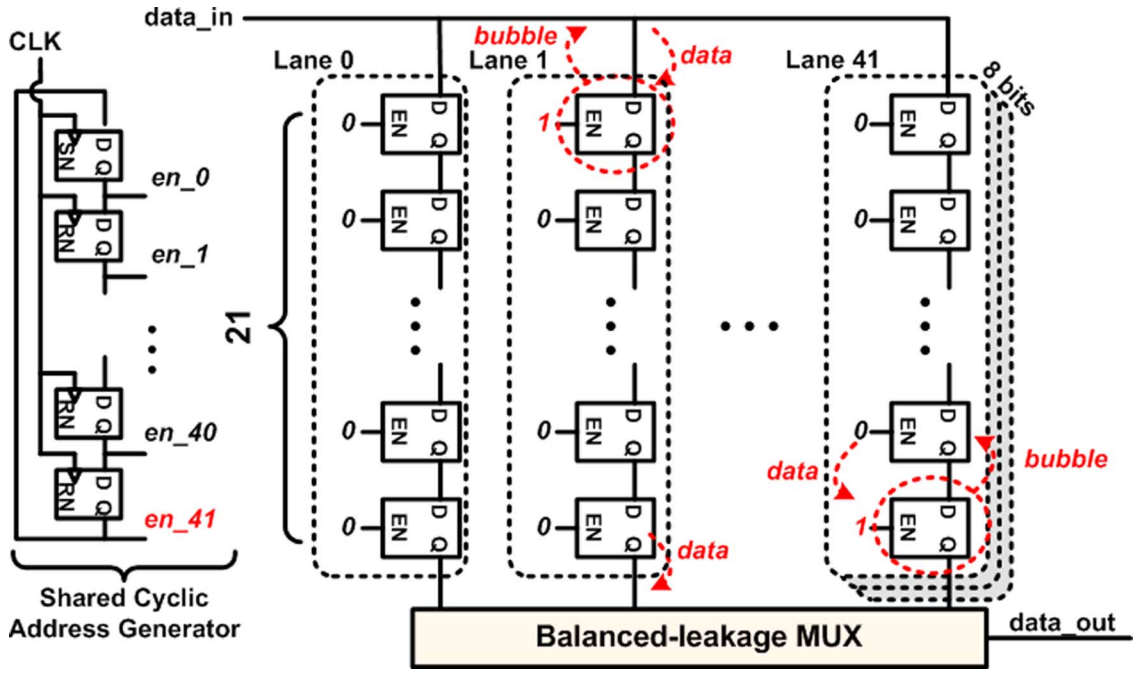


Fig. 17. An 8-bit 840-entry FIFO based on the proposed shift-latch architecture.

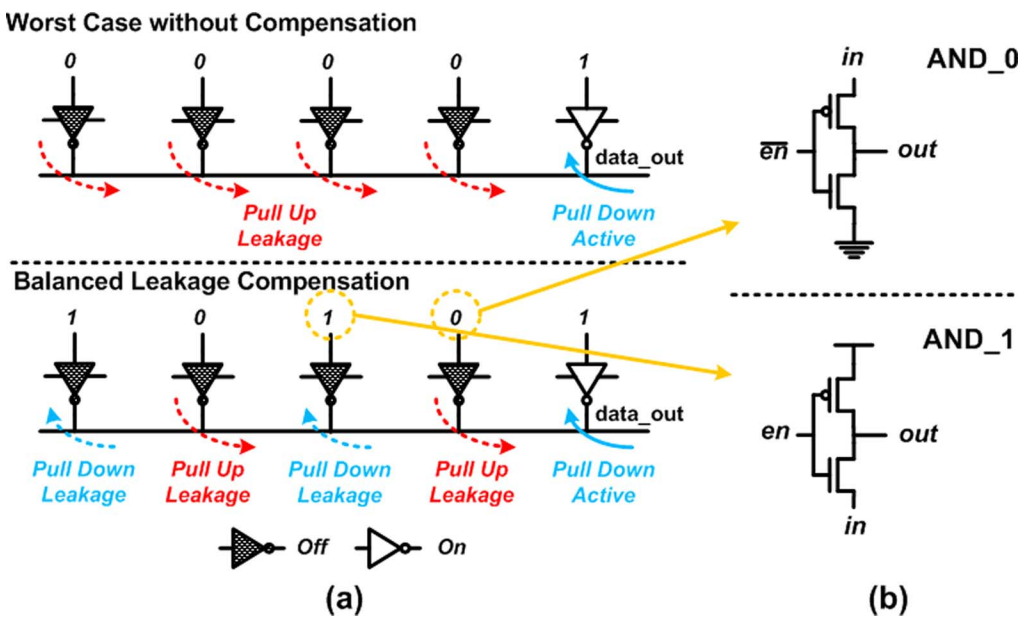


Fig. 18. (a) Worst-case scenario for leakage current affecting bitline pull-down with and without leakage compensation technique. (b) Proposed 2-transistor AND gates.

is written to it. At the same time, data stored in the last latch is read out through the output port and it becomes the bubble.

After N cycles all data values have propagated forward by one entry and one output is produced from the last latch, completing one *period*. After $N-1$ periods, the value initially stored in the first latch is shifted to the last latch and can be passed to a readout circuit. Therefore, this can be viewed as a single FIFO *lane* with $N(N-1)$ total FIFO depth and throughput of *one output per N cycles*. Hence, a conventional *one output per cycle* FIFO is built by arranging N identical lanes in parallel and connecting their enable signals diagonally (Fig. 16). In each cycle, exactly one output is generated from different FIFOs and we can obtain a conventional 1 output per cycle throughput by adding additional readout circuitry to choose the appropriate output among N FIFO lanes. If we implement a same depth (delay) FIFO using shift register, we would need $N(N-1)$ flip-flops for the same delay. Fig. 17 shows an example of an 840-entry FIFO based on the proposed shift-latch FIFO architecture. A cyclic address generator automatically generates the one-hot encoded enable signals shared across all lanes. In the final design, each lane is activated only every other cycle to avoid overlap in enable signals of adjacent cycles and enhance robustness (i.e.,

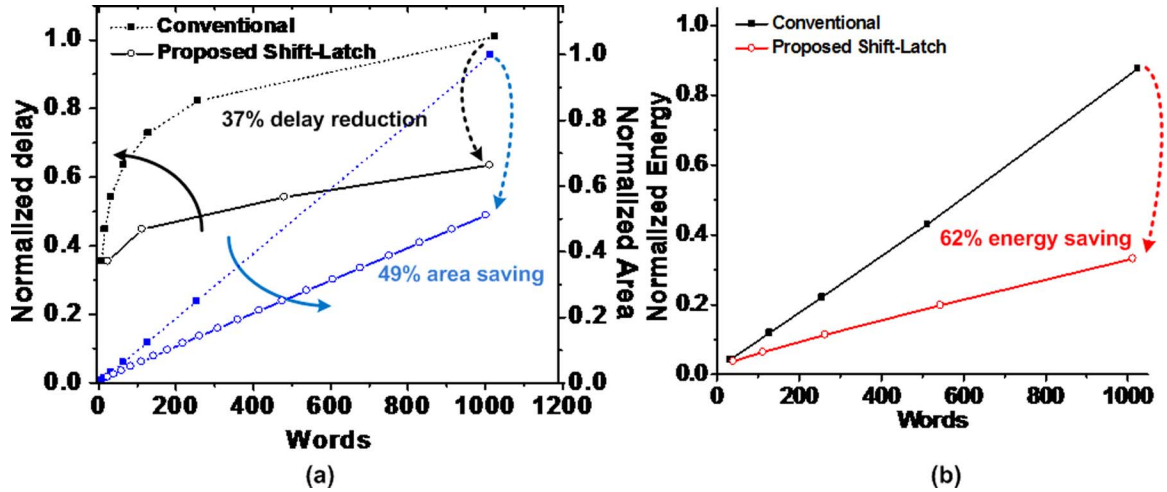


Fig. 19. Simulated (a) delay, area, and (b) energy consumption of baseline and proposed FIFO designs as a function of FIFO size.

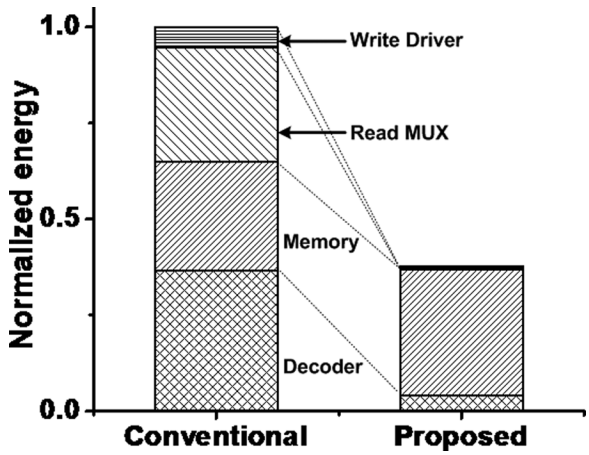
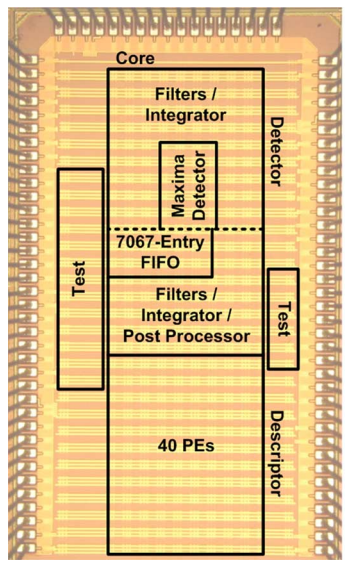


Fig. 20. Simulated energy savings in each component of a 1 k-entry FIFO.



Technology	28nm LP CMOS
Vdd	470mV
Clock Freq.	27MHz (102FO4)
Core Area	0.85x2.61mm^2
Input Video	640x480 30fps
Power	2.7mW
Performance	149.3GOPS
Efficiency	55.3TOPS/W

Fig. 21. A microphotograph of the fabricated feature extraction accelerator and summary table.

guaranteeing non-overlapping enable signals in the low voltage operating regime is very challenging). Even-numbered enable signals are connected to even-numbered lanes and the same applies to odd-numbered enable signals. This FIFO has 21 latches in each lane and 42 lanes in total and they are connected to a shared MUX readout circuitry.

In the near- and sub-threshold regimes, significantly lower MOSFET on-off current ratio degrades read operation reliability and limits the number of storage cells that can be tied to a single bitline [24]. Fig. 18(a) (upper) shows the worst-case scenario where an activated driver tries to pull a bitline down while all other disabled drivers exhibit pull-up leakage currents. To mitigate this, we propose a leakage compensation technique that minimizes the effect of leakage current, as shown in Fig. 18(a) (lower). Inactive cells are preset to have an equal number of ones and zeros at the input, resulting in roughly balanced pull-up and pull-down leakage currents on the bitline. This can be implemented by adding additional AND gates before access transistors to force values feeding into the readout driver to pre-determined values. To minimize this overhead, we employ two distinct 2-transistor AND gates (Fig. 18(b)), which is enabled by guaranteed pre-charge and pre-discharge of output nodes arising from the sequential readout property of FIFOs. This technique suppresses the impact of PVT variations and improves readout delay variation (σ) by 34% with 4% speedup despite the added AND gate delay. Although a shift register would not require a wide MUX for readout, the proposed readout circuitry does not contribute appreciably to total power consumption (less than 5% of total power), allowing the proposed FIFO to be more energy efficient than a standard shift register.

The proposed FIFO design was simulated and compared against prior work in low power queues. The baseline design is a latch-based memory with a logic-based readout [25], representing one of the most energy efficient and robust designs at low voltages. It uses a cyclic address generator and each storage cell is accessed through a logic-based readout path for fast and robust readout. Fig. 19 provides simulation results that show the proposed shift-latch FIFO improves readout delay and energy efficiency with smaller area compared to the baseline. For a 1 k-entry FIFO, the proposed design is 37% faster, 49%

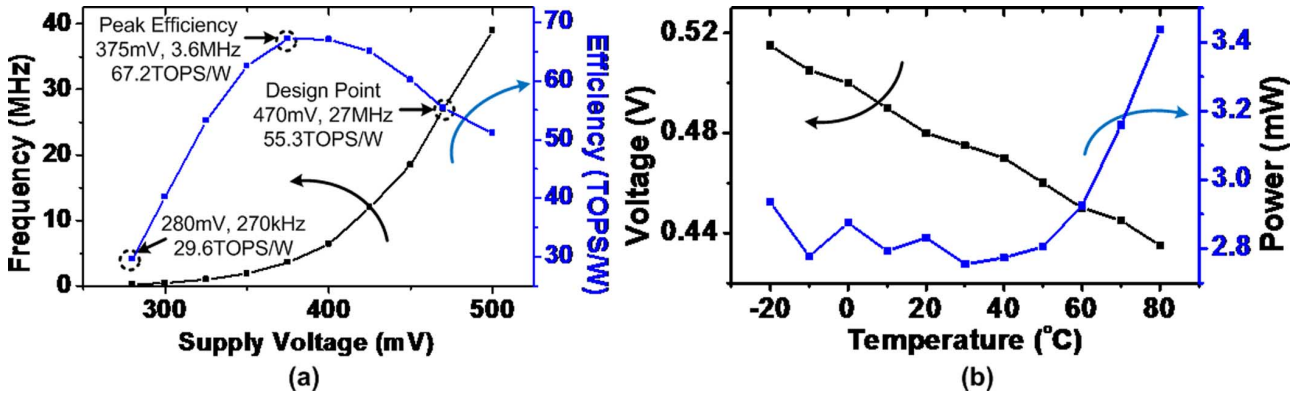


Fig. 22. Measurement results across different operating (a) voltages and (b) temperatures.

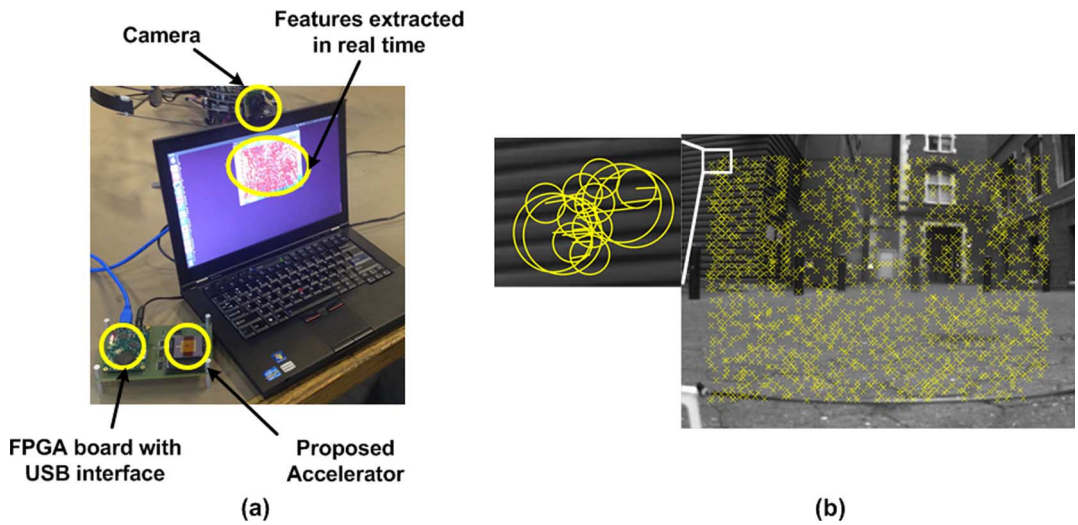


Fig. 23. (a) A real-time feature extraction test setup and (b) a sample image marked with 1421 extracted features from measurements.

smaller, and consumes 62% less energy due to energy savings from shared address generator and readout circuitry. Fig. 20 shows detailed energy savings in each component. Although more energy is consumed in storage cells because of shifting data and higher switching activity, energy savings from read and write circuitry dominate due to the slow logarithmic increase of interface size for the proposed shift-latch FIFO.

V. Measurement Results

A feature extraction accelerator based on the proposed hardware and algorithm techniques is fabricated in 28 nm LP CMOS technology. Fig. 21 shows a microphotograph of the fabricated design along with a summary table. It operates at the design point of 470 mV with a clock speed of 27 MHz to process 30 fps VGA video input. While continuously processing input video, the accelerator consumes only 2.7 mW with 149.3 GOPS performance, yielding a 55.3 TOPS/W energy efficiency. Fig. 22(a) shows measurement results over a range of operating voltages. This design can operate down to 280 mV, which represents the deep sub-threshold regime in this process, largely due to robust FIFO design and careful standard cell selections (selecting only cells with stack heights less than 3). As voltage scales down, energy efficiency starts to decrease at some point due to dominating leakage power and increasing cycle time [17]. A peak efficiency of 67.2 TOPS/W is obtained at 375 mV. The accelerator design can process 4 fps at this operating point. Fig. 22(b) shows required operating voltage for the given throughput (30 fps) at different temperatures, which confirms less than 20% power consumption variation from -20°C to 80°C.

Fig. 23(a) shows a test setup where the accelerator processes video from a camera and returns extracted features in real time. The accelerator is incorporated into the full navigation processing flow in simulation using FPGA board, while the remainder of the navigation algorithm is performed by a host X86 host CPU. Fig. 23(b) presents a sample image from a camera on the robotic test vehicle, along with 1421 features extracted using the fabricated accelerator. Detected points near frame edges have sampling regions overlapped with image borders and they are ignored for reliable extraction in this case. A part of the image in the red box has clear parallel patterns and extracted features in this region have very similar orientations, confirming proper feature extraction operation.

Table I provides comparisons between our work and recent prior works. The proposed accelerator is targeted solely for feature extraction and extracts features from the entire frame in contrast to other ROI-based designs. Although it was designed for VGA input video, the proposed accelerator architecture does not vary with video size and it can be adjusted

TABLE I
COMPARISON OF PRIOR WORK AND THE PROPOSED DESIGN

	Proposed	[12]	[13]	[14]
Technology	**28nm**	65nm	130nm	40nm
Design Target	**Feature Extraction**	Object Recognition	Object Recognition	Object Recognition
Base Algorithm	**SURF**	SIFT	SIFT	Haar-like
Extraction Scope	**Entire Frame**	ROI only	ROI only	ROI only
Input Video	**640×480**	1920×1080	1280×720	1280×960
Core Voltage	**470mV**	1V	0.7~1.2V	0.9V
Power	**2.8 mW**	52.5mW	320mW	69.3mW
Scaled Efficiency	**55.3TOPS/W***	12.4TOPS/W**	6.6TOPS/W**	15.7TOPS/W**

Scaled Efficiency = Reported Efficiency * (Technology/28nm) * (Voltage/470mV)^2
***Average efficiency with equivalent number of operations **Peak efficiency**

TABLE II
SUMMARY OF POWER SAVINGS FROM PROPOSED TECHNIQUES

Approach	Energy savings
Parallelized filters	38%
Single stream descriptor & circular-shaped description	48%
Shift-latch FIFO	22%
Optimized detector	27%

to process 1280 × 720 HD video with 81 MHz clock frequency and 12 mW power consumption at 600 mV. For comparison, energy efficiency was scaled with respect to operating voltage and technology, and OPS/W was used for comparison against other works with different functionalities. The proposed design achieves 3.5× better energy efficiency over prior work. Any size of video can be divided into multiple 640 × 124-pixel subsections and they can be processed independently in the accelerator, where clock frequency and operating voltage are selected differently to accommodate performance requirement. The proposed algorithm and architecture optimization techniques are also applicable to other feature extraction algorithms that share similar properties, such as rotation-invariant description and multi-scale pyramid construction.

VI. CONCLUSIONS

This paper proposes various hardware and algorithm techniques to realize a highly energy-efficient feature extraction accelerator. Hardware-oriented algorithm optimizations reduce hardware cost (e.g., area and power consumption) significantly while maintaining extraction quality. The proposed accelerator architecture is focused on maximizing the benefits of deep voltage scaling while meeting high throughput requirements. A new shift-latch FIFO architecture provides a practical and efficient solution in the near- and sub-threshold regimes. A feature extraction accelerator using these techniques is fabricated in 28 nm LP CMOS technology and measurement results confirm that it processes 30 fps VGA video at supply voltages as low as 470 mV at a low clock speed of 27 MHz. Overall the design provides 3.5× higher energy efficiency than prior state-of-art and offers full-frame feature extraction. Table II summarizes power savings obtained by the proposed algorithm, architecture and circuit techniques.

REFERENCES

[1] J. Zheng, D. Kuai, Z. Liu, Y. Teng, and T. Zhang, "Salient feature volume and its application in brain MRI image registration," in *Proc. Int. Conf. Biomedical Engineering and Informatics*, Oct. 2011, pp. 477–481.

[2] M. P. Heinrich, M. Jenkinson, M. Brady, and J. A. Schnabel, "MRF-based deformable registration and ventilation estimation of lung CT," *IEEE Trans. Medical Imaging*, vol. 32, no. 7, pp. 1239–1248, Jul. 2013.

[3] P. Kumar, A. Mittal, and P. Kumar, "A multimodal audio visible and infrared surveillance system (MAVISS)," in *Proc. Int. Conf. Intelligent Sensing and Information Processing*, Dec. 2005, pp. 151–156.

[4] S. Segvic, A. Remazeilles, A. Diosi, and F. Chaumette, "Large scale vision-based navigation without an accurate global reconstruction," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Jun. 2007, pp. 1–8.

[5] Y. Lee, G. Kim, S. Bang, Y. Kim, I. Lee, P. Dutta, D. Sylvester, and D. Blaauw, "A modular 1 mm^3 die-stacked sensing platform with optical communication and multi-modal energy harvesting," in *2012 IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2012, pp. 402–404.

[6] Y. Zhang, F. Zhang, Y. Shakhsheer, J. D. Silver, A. Klinefelter, M. Nagaraju, J. Boley, J. Pandey, A. Shrivastava, E. J. Carlson, A. Wood, B. H. Calhoun, and B. P. Otis, "A batteryless 19 μW MICS/ISM-band energy harvesting body sensor node SoC for ExG applications," *IEEE J. Solid-State Circuits*, vol. 48, no. 1, pp. 199–213, Jan. 2013.

[7] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE Int. Conf. Computer Vision*, Sep. 1999, pp. 1150–1157.

[8] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[9] S. Shen, N. Michael, and V. Kumar, "Autonomous multi-floor indoor navigation with a computationally constrained MAV," in *Proc. IEEE Conf. Robotics and Automation*, May 2011, pp. 20–25.

[10] G. Grisetti, C. Stachniss, and W. Burgard, "Improving grid-based SLAM with Rao-Blackwellized particle filters by adaptive proposals and selective resampling," in *Proc. IEEE Conf. Robotics and Automation*, Apr. 2005, pp. 2432–2437.

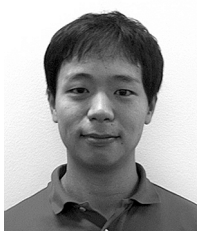
[11] S. Lee, J. Oh, M. Kim, J. Park, J. Kwon, and H.-J. Yoo, "A 345 mW heterogeneous many-core processor with an intelligent inference engine for robust object recognition," in *2010 IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2010, pp. 332–333.

[12] Y.-C. Su, K.-Y. Huang, T.-W. Chen, Y.-M. Tsai, S.-Y. Chen, and L.-G. Chen, "A 52 mW full HD 160-degree object viewpoint recognition SoC with visual vocabulary processor for wearable vision applications," in *Proc. IEEE Symp. VLSI Circuits*, Jun. 2011, pp. 258–259.

[13] J. Oh, G. Kim, J. Park, I. Hong, S. Lee, and H.-J. Yoo, "A 320 mW 342GOPS real-time moving object recognition processor for HD 720 p video streams," in *2012 IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2012, pp. 220–222.

[14] Y.-M. Tsai, T.-J. Yang, C.-C. Tsai, K.-Y. Huang, and L.-G. Chen, "A 69 mW 140-meter/60 fps and 60-meter/300 fps intelligent vision SoC for versatile automotive applications," in *Proc. IEEE Symp. VLSI Circuits*, Jun. 2012, pp. 152–153.

[15] S. A. J. Winder and M. Brown, "Learning local image descriptors," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Jun. 2007, pp. 1–8.
[16] A. Chandrakasan and R. Brodersen, *Low-Power CMOS Design*. New York, NY, USA: Wiley-IEEE Press, 1998.
[17] B. Zhai, D. Blaauw, D. Sylvester, and K. Flautner, "Theoretical and practical limits of dynamic voltage scaling," in *Proc. Design Automation Conf.*, May 2005, pp. 868–873.
[18] A. Wang and A. Chandrakasan, "A 180-mV subthreshold FFT processor using a minimum energy design methodology," *IEEE J. Solid-State Circuits*, vol. 40, no. 1, pp. 310–319, Jan. 2005.
[19] F.-C. Huang, S.-Y. Huang, J.-W. Ker, and Y.-C. Chen, "High-performance SIFT hardware accelerator for real-time image feature extraction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 3, pp. 340–351, Mar. 2012.
[20] S. Yoshizawa, K. Nishi, and Y. Miyanaga, "Reconfigurable two-dimensional pipeline FFT processor in OFDM cognitive radio systems," in *Proc. IEEE Int. Symp. Circuits and Systems (ISCAS)*, May 2008, pp. 1248–1251.
[21] C.-C. Wang, J.-M. Huang, and H.-C. Cheng, "A 2 K/8 K mode small area FFT processor for OFDM demodulation of DVB-T receivers," *IEEE Trans. Consumer Electronics*, vol. 51, no. 1, pp. 28–32, Feb. 2005.
[22] L. Chang, R. K. Montoye, Y. Nakamura, K. A. Batson, R. J. Eickemeyer, R. H. Dennard, W. Haensch, and D. Jamsek, "An 8T-SRAM for variability tolerance and low-voltage operation in high-performance caches," *IEEE J. Solid-State Circuits*, vol. 43, no. 4, pp. 956–963, Apr. 2008.
[23] C.-H. Lo and S.-Y. Huang, "P-P-N based 10T SRAM cell for low-leakage and resilient subthreshold operation," *IEEE J. Solid-State Circuits*, vol. 46, no. 3, pp. 695–704, Mar. 2011.
[24] B. Zhai, S. Hanson, D. Blaauw, and D. Sylvester, "Analysis and mitigation of variability in subthreshold design," in *Proc. Int. Symp. Low Power Electronics and Design (ISLPED)*, Aug. 2005, pp. 20–25.
[25] M. Seok, D. Jeon, C. Chakrabarti, D. Blaauw, and D. Sylvester, "A 0.27 V 30 MHz 17.7 nJ/transform 1024-pt complex FFT core with super-pipelining," in *2011 IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2011, pp. 342–344.

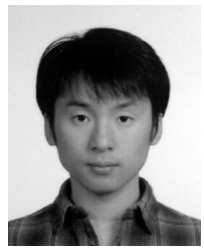


Dongsuk Jeon (S'10) received the B.S degree in electrical engineering from Seoul National University, Seoul, Korea, in 2009. He is currently pursuing the Ph.D. degree in electrical engineering at the University of Michigan, Ann Arbor, MI, USA.

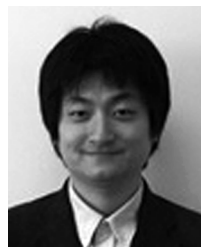
His research interests include energy-efficient signal processing, subthreshold circuit design, and error-resilient systems.

Mr. Jeon is the recipient of the Samsung Scholarship for graduate student.

Michael B. Henry is the owner of Isocline Engineering LLC, an engineering research firm in Ann Arbor, MI, USA. He is also a visiting scholar at the University of Michigan.



Yejoong Kim received the B.S. degree in electrical engineering from Yonsei University, Seoul, Korea, in 2008, and the M.S. degree in electrical engineering and computer science from the University of Michigan, Ann Arbor, MI, USA, in 2012. He is currently working toward the Ph.D. degree at the University of Michigan, Ann Arbor. His research interests include subthreshold circuit designs, ultra low-power SRAM, and the design of millimeter-scale computing systems and sensor platforms.



Inhee Lee (S'07) received the B.S. and M.S. degrees in electrical and electronic engineering from Yonsei University, Seoul, Korea, in 2006 and 2008, respectively. He is pursuing the Ph.D. degree in University of Michigan, Ann Arbor, MI, USA.

His current research interest includes a low-power $\Delta\Sigma$ ADC, a low-power capacitive energy harvester and power management circuit, a low-power battery monitoring circuit, and a micro-scale wireless sensor node.

Zhengya Zhang (S'02–M'09) received the B.A.Sc. degree in computer engineering from the University of Waterloo, Ontario, Canada, in 2003, and the M.S. and Ph.D. degrees in electrical engineering from the University of California, Berkeley, CA, USA, in 2005 and 2009, respectively.

Since 2009, he has been on the faculty of the University of Michigan, Ann Arbor, MI, USA, as an Assistant Professor in the Department of Electrical Engineering and Computer Science. His current research interests include low-power and high-performance VLSI circuits and systems for computing, communications and signal processing.

Dr. Zhang was a recipient of the National Science Foundation CAREER Award in 2011, the Intel Early Career Faculty Honor Program Award in 2013, the David J. Sakrison Memorial Prize for outstanding doctoral research in EECS at UC Berkeley, and the Best Student Paper Award at the Symposium on VLSI Circuits. He is an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS I: REGULAR PAPERS.

David Blaauw (M'94–SM'07–F'12) received the B.S. degree in physics and computer science from Duke University, Durham, NC, USA, in 1986, and the Ph.D. degree in computer science from the University of Illinois, Urbana, IL, USA, in 1991.

Until August 2001, he worked for Motorola, Inc., Austin, TX, USA, where he was the manager of the High Performance Design Technology group. Since August 2001, he has been on the faculty at the University of Michigan, Ann Arbor, MI, USA, where he is a Professor. He has published over 400 papers and holds 40 patents. His work has focused on VLSI design with particular emphasis on ultra-low-power and high-performance design.

Dr. Blaauw was the Technical Program Chair and General Chair for the International Symposium on Low Power Electronic and Design. He was also the Technical Program Co-Chair of the ACM/IEEE Design Automation Conference and a member of the ISSCC Technical Program Committee.

Dennis Sylvester (S'95–M'00–SM'04–F'11) received the Ph.D. degree in electrical engineering from the University of California, Berkeley, CA, USA, where his dissertation was recognized with the David J. Sakrison Memorial Prize as the most outstanding research in the UC-Berkeley EECS department.

He is a Professor of Electrical Engineering and Computer Science at the University of Michigan, Ann Arbor, MI, USA, and Director of the Michigan Integrated Circuits Laboratory (MICL), a group of ten faculty and 70+ graduate students. He has held research staff positions in the Advanced Technology Group of Synopsys, Mountain View, CA, USA, Hewlett-Packard Laboratories, Palo Alto, CA, USA, and visiting professorships at the National University of Singapore and Nanyang Technological University, Singapore. He has published over 350 articles along with one book and several book chapters. His research interests include the design of millimeter-scale computing systems and energy efficient near-threshold computing. He holds 19 U.S. patents. He also serves as a consultant and technical advisory board member for electronic design automation and semiconductor firms in these areas. He co-founded Ambiq Micro, a fabless semiconductor company developing ultra-low power mixed-signal solutions for compact wireless devices.

Dr. Sylvester has received an NSF CAREER award, the Beatrice Winner Award at ISSCC, an IBM Faculty Award, an SRC Inventor Recognition Award, and eight best paper awards and nominations. He is the recipient of the ACM SIGDA Outstanding New Faculty Award and the University of Michigan Henry Russel Award for distinguished scholarship. He has served on the technical program committee of major design automation and circuit design conferences, the executive committee of the ACM/IEEE Design Automation Conference, and the steering committee of the ACM/IEEE International Symposium on Physical Design. He has served as Associate Editor for IEEE TRANSACTIONS ON CAD and IEEE TRANSACTIONS ON VLSI SYSTEMS, and Guest Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II.