A 170μW Image Signal Processor Enabling Hierarchical Image Recognition for Intelligence at the Edge

Hyochan An, Siddharth Venkatesan, Sam Schiferl, Tim Wesley, Qirui Zhang, Jingcheng Wang, Kyojin Choo, Shiyu Liu, Bowen Liu, Ziyun Li, Hengfei Zhong, Luyao Gong, David Blaauw, Ronald Dreslinski, Dennis Sylvester, and Hun Seok Kim
University of Michigan, Ann Arbor, USA. hyochan@umich.edu

Abstract: We propose an ultra-low power (ULP) Image Signal Processor (ISP) that performs on-the-fly in-processing frame (de)compression and hierarchical event recognition to exploit the temporal and spatial sparsity in an image sequence to achieve a 16× imaging system energy gain. The ISP is fabricated in 40nm CMOS and consumes only 170 μW at 5 fps for neural network-based intruder detection and 192× compressed image recording.

Introduction: Imaging is a highly desirable sensing modality in battery-operated IoT systems as it offers key contextual information about a sensor's environment. Prior IoT imaging systems have two key difficulties: 1) They process uncompressed frame images, resulting in a large frame buffer that increases chip size and leakage power. 2) Since they lack scene understanding, they cannot recognize useful information, and all frames must be transmitted fully, incurring considerable storage and radio power.

We address these challenges in the proposed ULP ISP, designed for size-constrained intelligent edge devices. First, we use macroblock(MCB)-based scene change detection using a new sparse census-transform encoding and JPEG compressed memory for input images, ensuring that full uncompressed images are never stored in their entirety on-chip. This reduces the required SRAM size for storing frames on-chip by 11.2× and leakage power by 26.9×. Second, we enable hierarchical event recognition through a programmable neural network (NN) engine that progressively prunes uninteresting areas or the entire image. Since relevant information typically occurs sparsely in time and space, image storage and transmission requirements can be reduced by >1000× (Fig. 1). In addition, the NN engine uses deep compression of all on-chip weights stored in a custom ultra-low leakage SRAM, further reducing system size and power. An H.264 engine compresses the final detected regions of interest, and the chip achieves a 192× total image size reduction ratio.

Architecture: Fig. 2 overviews the ISP design. All logic operates in a power-gated 0.6V domain. The imager interface block performs change detection (CD) on streamed-in images and stores the changed MCBs into the on-the-fly (during access) JPEG (de)compressed memory. The neural-engine (NE) processing element (PE) accelerates NN operations, controlled by a custom RISC processor, NCX. An H.264 Engine (H264E) performs intra-frame compression on an arbitrary (non-rectangular shaped) subset of MCBs and sends them off chip through the serial interface. An ARM Cortex-M0 orchestrates all blocks through the AHB bus.

Use Scenario: We demonstrate the proposed ISP in intruder detection and recording (Fig. 3) as follows: A companion imager chip with integrated motion detection [1] periodically inputs a sub-sampled image (32×20 pxls×1ch) into the ISP chip. The NE then performs NN-based person detection (consuming 14.4 μJ) on the sub-sampled image to determine if it contains a person. If it does, the ISP requests a full VGA frame from the imager (Bayer RGB format) and then performs on-the-fly MCB-based change detection against a reference frame, followed by JPEG compression (4.65μJ/frame with typical 12% change). The NE runs NN-based face detection, sweeping the region of changed MCBs on two scales (1 and 2× subsampling) with 16 pxl stride (255 μJ). If the changed region contains a face, NN-based facial recognition is performed (222 μJ). If this NN classifies the face as unregistered, the change-detected MCBs are H.264 compressed and stored in off-chip flash or radio transmitted. With an average of 12% change-detected MCBs and 23× H.264 compression ratio, the ISP achieves 192× overall size reduction for a VGA frame with 28.3dB PSNR and only transmits out those MCBs with unregistered face information. The three key techniques applied in the design are now described in more detail.

(A) Compression of Memory-Intensive Data: To reduce SRAM size and hence leakage, we extensively employed data compression (Fig. 4). Specifically, the input image data, NN weights, and output image data are all stored or transmitted in compressed format. For the input image, on-the-fly JPEG compression is performed on the streamed-in image, achieving 11.2× reduction in required memory size (7.4 Mb to 0.66 Mb) to store two VGA frames (reference and current). The JPEG codec is customized to remove interdependencies between MCBs and allow MCB-wise (de)compression. For NN weights, we use pruning, non-uniform quantization, Huffman encoding for convolution layers, and index-based encoding for sparse fully connected layers [2]. These combined techniques enable convolution layer compression of 2.3b per weight on average with <1% accuracy degradation. Compressed weights for all three NNs used in the intruder detection scenario (680 kb, 850 kb, and 1.9 Mb) are stored on chip. The H.264 algorithm was customized to reduce the number of MCBs required from the JPEG compressed memory for the H.264 intra-mode prediction by interpolating the upper-left corner pixel and skipping Diagonal Down Left and Vertical Left prediction modes. This reduces the number of required MCBs by 2.6× with negligible loss (<0.1 dB PSNR) (Fig. 5). Combined, these techniques reduce on-chip SRAM size by 5× (45 Mb to 9 Mb) and total leakage power by 12×, which includes 2.4× leakage power reduction via a custom-designed 0.3 V bitcell/0.6 V peripheral SRAM array with 8σ hold margin.

(B) Spatial Pruning using Change Detection Engine (CDE): The CDE performs spatial pruning at the MCB level (Fig. 6). First, the CDE encodes each 16×16 pixel MCB (3072b) of a reference image to a 64b pattern vector. Each element of a pattern vector is the ternary comparison result of two pixel intensities at predefined positions of the MCB. This new sparse census transform encoding is tolerant to uniform illumination change. For every newly streamed-in image, a 64b pattern vector is prepared and compared to that of the reference image. MCB change is flagged when the hamming distance between two vectors exceeds a tunable threshold. To improve coverage, flagged MCBs are also dilated (neighboring MCBs are flagged in a tunable manner). At the same time, only flagged MCBs are JPEG compressed using a pointer-based data structure to accommodate the variable length of compressed MCBs. Processors can access arbitrary MCBs in raw uncompressed format with natural (fixed-length) block addressing as the decompression happens on-the-fly. These combined techniques reduce on-chip VGA image size by 110× from 460 kB to 4.2 kB (with typical 12% change) while achieving 95% coverage and 5% false positive rate for CD.

(C) Event Recognition using NE: The highly programmable NE accelerates multiple NNs through its PE designed for efficient memory accesses (Fig. 7). With NN-specialized instructions, NCX controls the PE to support heterogeneous NNs. The PE is a computational core with a 512 8b MAC array and buffers to enable high MAC utilization. For a convolutional layer, a set of weights is decompressed once and swept across the entire input. To save memory for intermediate outputs, the convolved values after ReLu are shifted back to memory as 8b values. For large sparse fully-connected layers, the combination of the outer product-based matrix-vector multiplication and index-based encoded weight maximize the activity rate of MACs by only computing non-zero weights. The NE achieves 1.5 TOPS/W (op = 8b mult/add) at 0.58 V while operating at 153 kHz (allowing 5 fps person detection).

Measurements: The ISP is fabricated in 40nm LP CMOS (Fig. 8) and operates at 153 kHz. Latency (system energy) of person detection, face detection, and face recognition processing is 0.19s (31.9 μJ), 3.22s (541 μJ), and 2.85s (478 μJ), respectively. The LFW dataset is used for NN training, yielding accuracy results given in Fig. 9. Continually executing each step in the intruder detection and recording scenario (Fig. 3) consumes 170 μW on average. The energy consumption of the full data flow to produce a 192× compressed output image (12% MCB change) is 1.5 mJ per frame, broken down into Cortex M0 (16.6 μJ), CD & JPEG (10.7 μJ), H.264 compression (317 μJ), and clock tree (151 μJ). Fig. 10 compares the ISP with prior works.

978-1-7281-9942-9/20/$31.00 ©2020 IEEE

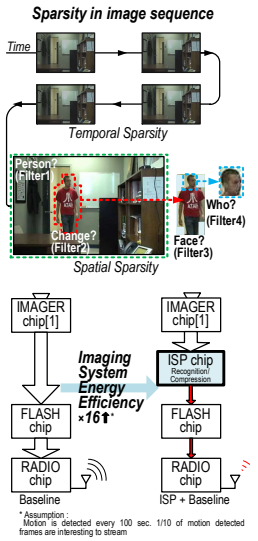


Fig. 1. Motivation of hierarchical image recognition.

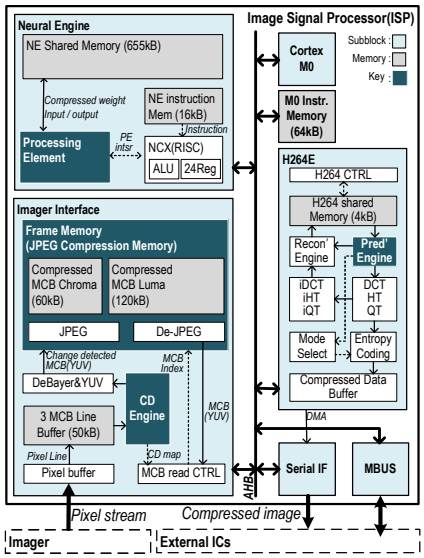


Fig. 2. Top-level architecture of the Image Signal Processor.

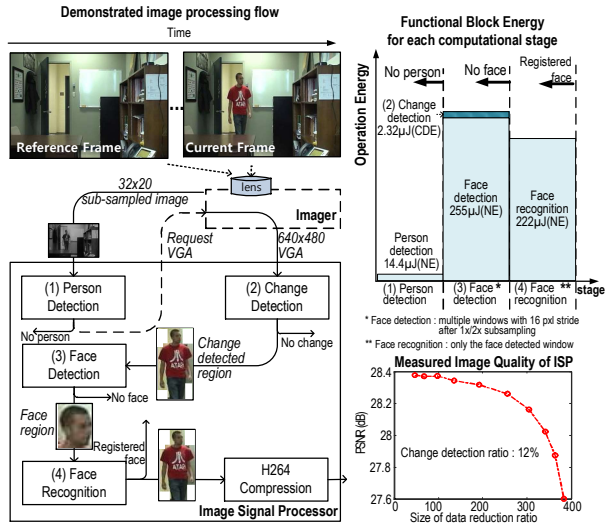


Fig. 3. Demonstration of the Image Signal Processor (ISP): (left) Demonstrated image processing flow; (top right) Functional block energy consumption for each computational stage; (bottom right) Quality of output image of ISP.

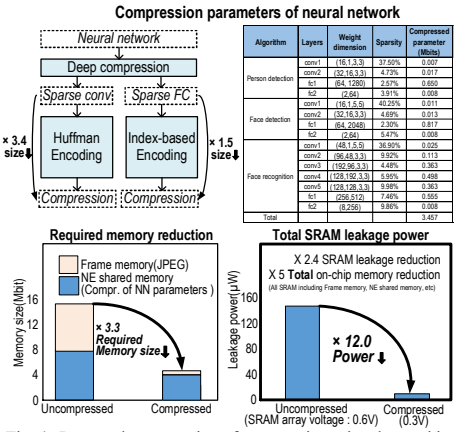


Fig. 4. Proposed compression of memory intensive data entities: (top) Compression of parameters of neural networks; (bottom) Leakage power reduction from compression of memory size.

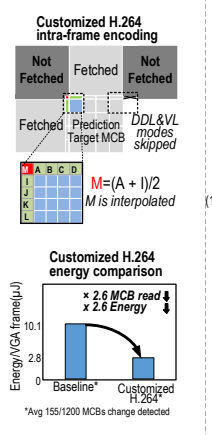


Fig. 5. Customized H.264 intra frame encoding.

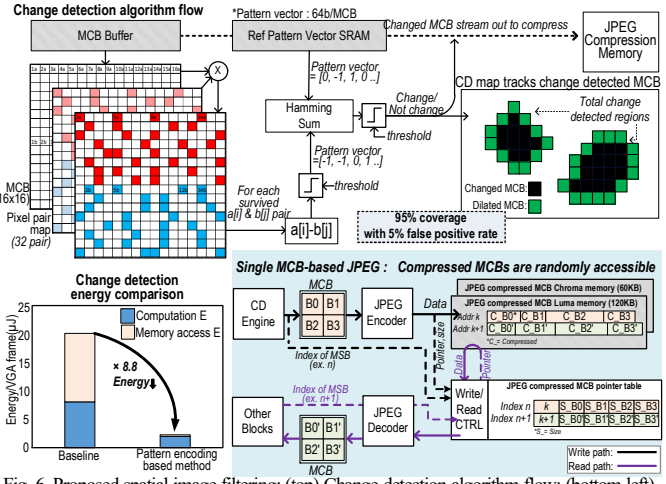


Fig. 6. Proposed spatial image filtering: (top) Change detection algorithm flow; (bottom left) change detection energy comparison; (bottom right) Architecture of MCB-based JPEG compression memory.

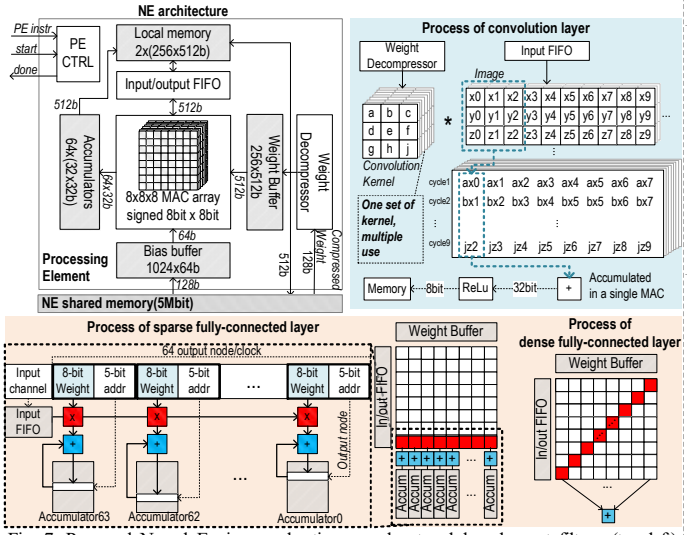
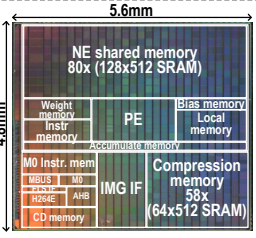


Fig. 7. Proposed Neural Engine accelerating neural network-based event filters: (top left) Architecture of the Processing Element; (top right) Process of convolution layer; (bottom left) Process of sparse fullyconnected layer; (bottom right) Process of dense fully-connected layer.



Process	40nm
Chip size	4.8mm x 5.6mm
On-chip SRAM	9Mbit
Supply voltage	0.58V~0.7V(Core) 0.29V~0.35V(Memory) 1.2V(IO)
Imge size& throughput*	VGA 0.16fps ~ 5fps
Frequency	153KHz
Power	170μW

* Various depending on contents of picture

Fig. 8. Chip specification & die photo.

	This work	ISSCC '17[2]	VLSI '18[3]
Techonology	40nm	65nm	65nm
Die area	4.8mm x 5.6mm	4mm x 4mm	1.784mm x 1.784mm
Event detecion/recognition algorithm	**NN-based Person detection Face detection face recognition**	Harr-like cascade classifier NN-based face recognition	Harr-like cascade classifier Face alignment NN-based Face recognition
Image processing algorithm	**Debayer YUV Change detection**	NA	NA
On-chip memory	**9Mb**	1.3Mb	0.056Mb
# of NNs and on-chip weights	**3**	1	1
Neural network weight compression	**Y**	N	N
External-memory access	**N**	N	Y
NN multiplication bit precision	8	16	1
Peak energy efficiency[TOPS/W]	1.5	NA	13.3
Max Resolution	**VGA**	QVGA	NA
Power / FPS	**170μW(0.16fps~5fps)***	620μW(1fps)	200μW(1fps)

*Frame rate can vary depending on scenario

Fig. 10. Comparison table.

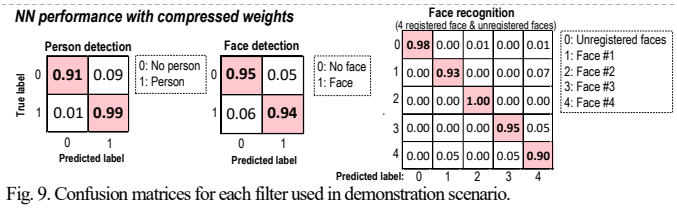


Fig. 9. Confusion matrices for each filter used in demonstration scenario.

Acknowledgement

We thank Sony Semiconductor Solutions Corp./Sony Electronics Inc. for supporting this work.

Reference

[1] K. D. Choo *et al.*, *ISSCC*, pp. 96-97, Feb. 2019.
[2] S. Han *et al.*, *arXiv preprint arXiv:1510.00149*, 2015.
[3] K. Bong *et al.*, *ISSCC*, pp. 248-249, Feb. 2017.
[4] S. Kang *et al.*, *Symp. On VLSI Circuits*, pp. 137-138, 2018.