

Smart Imager with Object Detection Exploiting Edge-Frame-Base Processing and Bounding Box Extraction for μW Power Purely-Harvested Sensor Nodes

Hayate Okuhara, Udari De Alwis, Liu Yue, Karim Ali Ahmed, Massimo Alioto
ECE Dept., National University of Singapore, Singapore, {hayate01, massimo.alioto}@nus.edu.sg

Abstract— Battery-less and cost-sensitive vision nodes are becoming essential in IoT-scale sensor networks, where in/near-sensor AI enables local recognition while minimizing data transmission. However, achieving multi-class object detection under available peak power budgets ($<10 \mu\text{W}$) and low-cost fabrication remains a major challenge. Existing smart imagers either lack on-chip intelligence or exceed such power budgets due to costly sensing and computing. This paper presents a fully-integrated smart imager performing multi-class object detection at $8.51 \mu\text{W}$ (equivalent to the power from a $7\text{mm} \times 6\text{mm}$ harvester at 300 lux) in standard 180nm CMOS. The system processes 1-bit edge-extracted frames, applies tile-level novelty detection for bounding-box ROI extraction, and computes CENTRIST features over cropped regions. A low-power approximate linear SVM classifies detected objects at 130 pW/pixel power. Unlike prior architectures, the proposed system maintains full image readout, supports flexible learning-based inference, and avoids custom optics and CIS processing. This makes it the first battery-less smart imager capable of flexible, multi-object detection in low-cost standard CMOS technology.

Keywords— battery-less, smart imager, vision, machine learning, edge image processing, always-on, sustainable semiconductors

I. INTRODUCTION

Image sensors with built-in vision processing, either on or near the sensor, are essential building blocks for advanced sensor networks, which enable sophisticated environment monitoring, smart manufacturing, and smart buildings, among the others. The rapid growth of connected devices (Internet of Things) calls for efficient computation distribution, favoring processing at the edge over centralized servers. In this context, some modern vision sensors implement data filters [1]-[10] or even AIs [11]-[17] locally to preprocess captured raw data and transfer only meaningful semantics to the cloud for more complex tasks.

Battery-less operation powered by energy harvesters is another key technology for sensor systems, thanks to the perpetual power production, leading to long-life node operations [17]-[19]. This shifts the primary design concern from energy efficiency to ensuring continuous operation under harvested power. However, edge imaging nodes face a tradeoff between more powerful local computing (integrated AI and traditional filters) and limited power availability. The available peak power is often limited to up to $10 \mu\text{W}$ generated by millimeter-scale low-cost harvesters under indoor lighting (hundreds of lux). At the same time, imagers must run at moderate frame rates, such

as a few frames per second, with rich AI vision capabilities. This corresponds to a few hundred pW per pixel for typical imaging resolutions such as QVGA (320×240).

Although broader low-power imaging sensors have been extensively studied in the last decades, few studies address both sub- $10 \mu\text{W}$ power and on-chip AI integration. Most prior works achieve the power targets by limiting on-chip processing, either with no integrated AI [1]-[3], [9] or with simple sensors only [20], [21]. On the other hand, imagers with inferences such as in-pixel 3-layer CNN (Convolutional Neural Networks) [14], or fixed application-specific models [11], [12] face huge power overheads reaching $10\times$ of acceptable power. A few smart imagers with the power budget compatible with a millimeter-scale harvester have been reported [13], [15]. However, they demonstrate a very limited model [13] or not a complete model [15] that can perform only primitive applications (e.g., face detection with limited resolution [13], digit detection [15]). [17] has an on-chip AI for image sensor and also meets the tight power requirement. However, it sacrifices the capability to read out raw imaging output, which significantly limits their range of applications. Hence, we must push the boundary of integrated intelligence while fully preserving the imaging capability within the tight power envelope.

Cost-effectiveness is another concern in edge nodes. Larger silicon areas and/or special design needs, such as CMOS image sensor process (CIS) [13] and a special optical layer [15], incur additional fabrication costs. Conventional smart imagers have a tradeoff between power and area efficiency [11]-[14]. Moreover, especially with integrated inferences, imagers with narrowly specialized vision tasks [11], [12] require extra system and design costs when their target application changes. Thus, it is preferable to keep ML models flexible through retraining in low-cost devices, as retraining models allows for solving different problems with the same hardware resources.

To address these challenges, this paper proposes a μW -class imaging to an AI smart imager for low-cost purely harvested systems. The contributions of this paper are listed as follows: 1. We propose a complete processing chain, from image capture to object detection, that enables a cost-effective μW -class system with ML capability. By using 1-bit edge extracted frames in which object contours are highlighted, we reduce both hardware and power costs by $4\times$ compared to grayscale image processing. The entire chain operates on-the-fly, eliminating the need for large frame buffers and reducing leakage power. 2. We integrate bounding-box extraction directly on-chip, which identifies

regions of interest (ROIs) based on significant interframe variations. Namely, it allows suppressing non-informative content in frames and reducing the data size processed in subsequent stages by $16\times$. 3. We design a low-power, information-rich feature extractor and pair it with a compact Support Vector Machine (SVM) classifier. This enables multi-class object detection in real-world environments with low energy cost, unlike [13] having a single face detection only. 4. The system is implemented in standard 180nm CMOS technology without costly fabrication steps like CIS [13] or optical stacks [15]. A test chip achieves $8.51 \mu\text{W}$ of total system power while retaining full imaging capability, enough to operate using a $7\text{mm} \times 6\text{mm}$ harvester under 300 lux indoor lighting.

II. RELATED WORK

A. Low-power image sensor with integrated image filter

With the rise of edge computing demands, several low-power imagers integrate vision tasks such as motion detection, edge detection, and Local Binary Pattern (LBP) filtering within sub-mW power budgets [1]-[10]. Kim et al. proposed a smart imager with motion detection, which identifies frame-to-frame pixel changes, under a limited dynamic range (38.5 dB) [1]. However, its system peak power reaches nearly 1000 pW per pixel, which vastly exceeds the target power envelope. Choi et al. proposed in-sensor motion detection and Histogram of Oriented Gradients (HOG) feature extraction for collecting only objects of interest at a power of tens μW [2]. Motion detection is widely used to reduce redundant frame captures and avoid wasteful energy consumption [2], [4]-[7]. Similarly, saliency and novelty detections, which highlight attention-worthy content, have been implemented in some vision sensors for more effective redundancy suppression than motion detection [7], [9], [10]. In [8], a programmable in-sensor kernel is proposed to achieve various kernel capabilities.

B. Near/in-sensor AI tasks

Recent advances in near/in-sensor vision tasks have enabled on-chip inference, even for sub-mW class low-power imagers, as reported in [11]-[17]. They have been implemented by analog [12], [16], [17], mixed-mode [11], [14], [15], or purely digital processing [13]. In [11] and [12], the multiple face detection task is implemented with specialized classifiers (Viola-Jones 2-stage classifier [11] and a classifier with 25 fixed features [12]). Some imagers include small neural networks such as CNNs [14], [17], and binary neural networks, BNNs (3-layer [15], 1st layer [16]). With digital processing, [13] implements a single face detection with 5b-SVM at a power budget of $5.6 \mu\text{W}$ at 5 FPS.

Despite recent advances, achieving a μW -class low-cost smart imager with decent machine learning capabilities is still challenging for sensor nodes operating under millimeter-scale harvesters. Imagers with analog/mixed-mode scene processing [11], [12], [14] have achieved state-of-the-art energy efficiency that is essential for battery-operated systems (33.8 pJ per pixel energy as in [14]). Nevertheless, they suffer from unacceptable minimum power consumption reaching hundreds of μW (thousands of pW/pixel), which is not suitable for purely-harvested systems. With near QVGA resolution, hundreds of pW/pixel of power efficiency is necessary. The imager in [15] achieved a μW order of the entire system with 143pW/pixel of

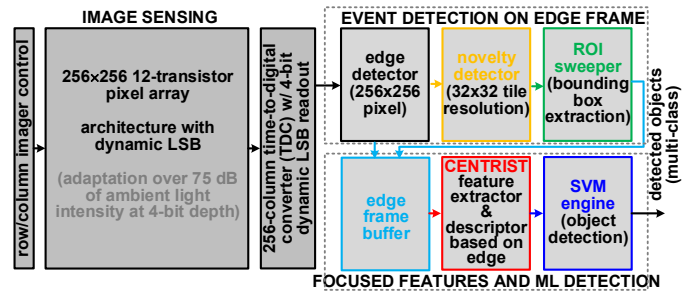


Fig. 1. Block diagram of the proposed smart imager SoC

efficiency, but the 3-layer BNN omits the fully connected (FC) layer and does not complete the inference on a chip. Moreover, its model capability is limited for simple tasks (digit recognition and eye orientation tracking demonstrated) even with the assistance of off-chip computations. Additionally, model flexibility is a critical consideration for low-cost systems. Highly specialized machine learning models, such as those in [11] and [12], lack flexibility. When the application changes, these models cannot be reused, requiring redesigns of the system. Similarly, analog-based processing lacks design portability among process technologies, unlike purely digital designs.

Digital-based scene processing is promising for μW -class power consumption while maintaining moderate frame rates (300 pW/pixel at 5 FPS [13]). Unlike analog and mixed-mode approaches, digital processing avoids a steady on-current, which significantly reduces power consumption at moderate frame rates. Additionally, it offers better design portability across process technologies than analog processing. The drawback is the relatively large fabrication area, as many computations must be managed. Indeed, considering the active area per pixel counts normalized by F^2 (F = process feature size), [13] is 28000, while the one with mixed-mode processing [14] is 4970. Even with the expensive cost, the imager in [13] implements a 5-bit SVM in 32×24 resolution for limited detection capability (single-face detection). Moreover, system fabrication should be simple for low-cost devices. Employing CIS processes like [13] incurs extra integration/fabrication costs. Similarly, mounting the optical layer [15] increases the system cost and causes less flexibility for vision tasks.

III. PROPOSED SMART IMAGER

A. Entire architecture

Fig. 1 shows a block diagram of the proposed imager SoC. It is composed of the dynamic least significant bit (LSB) imager, edge detector, novelty detector adopted from the architectures in [9], a frame buffer, ROI sweeper, the CENTRIST (CENSus TRansform hISTogram) feature extractor [22], and an SVM engine, covering from imaging to object detection. The edge-based processing contributes to shrinking the computational units for the consecutive vision processing. Also, we employed a classical linear SVM classifier, unlike other imagers with CNNs/BNNs, to reduce SRAM capacity requirements, leading to less leakage power and a suitable design for moderate FPS operations. The imager generates 4-bit grayscale images in 256×256 resolution and automatically adjusts the LSB based on environmental light intensity (over 75dB of the range [9]). Then,

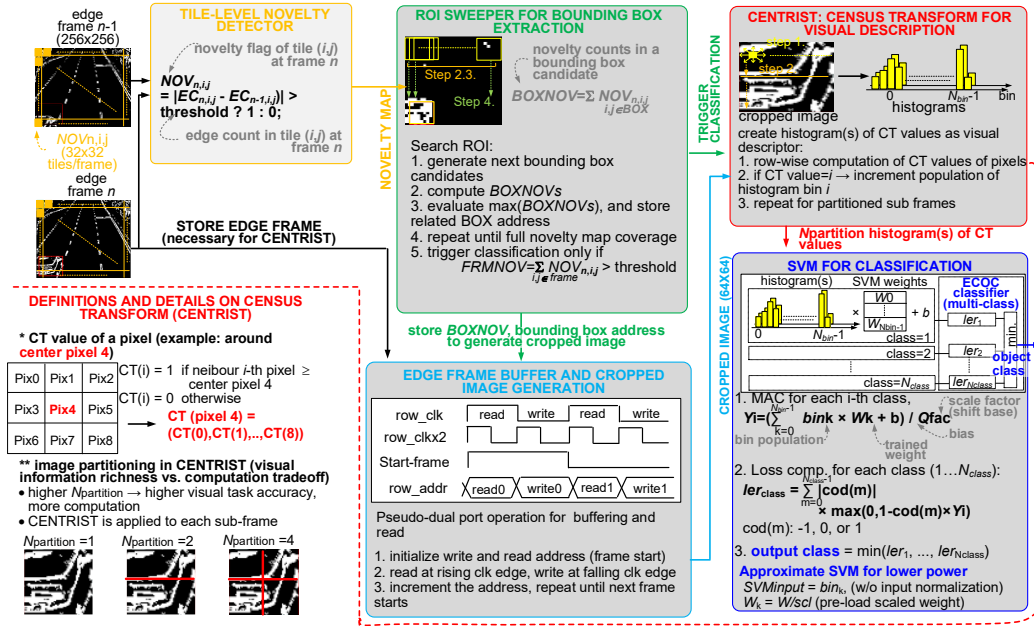


Fig. 2. Proposed processing algorithm from edge detection to object detection

its readout frames are converted into 1-bit edge images that highlight contours of objects in frames. The frame buffer stores a whole-edged image equivalent to 8kB (1/4 of the grayscale data size). The novelty detector checks if a new object appears in the current frame by evaluating the difference between the current and previous frame with a 32×32 tile map, and the ROI sweeper extracts it as a bounding box (64×64 in an edge frame). CENTRIST, which generates descriptive histograms from edge-based input, receives only the images cropped by the bounding box ($16 \times$ smaller), further reducing the processing energy. Finally, the SVM outputs the classification results accordingly.

B. Proposed processing algorithm

Fig. 2 depicts details of the proposed processing chain. After the edge extraction, the novelty detector evaluates the number of edge counts ($EC_{n,i,j}$) in the novelty tile located at (i, j) in frame n , dividing the frame into a 32×32 resolution. The flag noted as $NOV_{n,i,j}$ is asserted if the edge difference from the previous to the current one is more than a predefined threshold. The ROI sweeper rasters the generated novelty map with the bounding box size to find the location with the maximum novelty count ($BOXNOV$). The ROI sweeper also monitors if the current frame has sufficient novelty content to activate the classification by evaluating the summation of the entire novelty map ($FRMNOV$).

The bounding box extraction takes one frame cycle to search the entire pane and determine the location of the bounding box. Thus, the edge frame under the evaluation requires a buffer. However, if the capacity is identical to one frame, accessing a single port memory for reading the previous frame and writing the current one conflicts with each other. This issue is addressed in our algorithm by allowing the buffer to operate as a pseudo dual-port memory with twice the speed of the edge detector (assigning the read operation to the clock phase of "1" and the write operation to "0"). Once the bounding box extraction is completed, the cropped image (ROI) is seamlessly read out before the previous frame is overwritten.

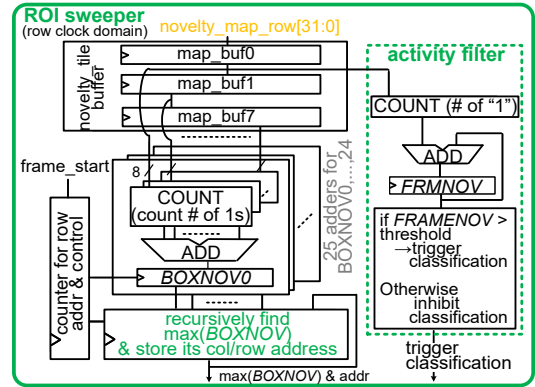


Fig. 3. ROI sweeper with parallel BOXNOV computations

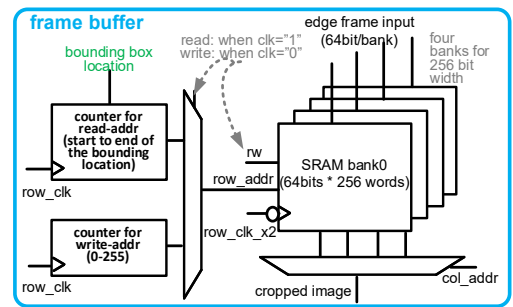


Fig. 4. Circuit diagram of the frame buffer to store a full frame of 1-bit edges

The feature extractor is based on the CENTRIST algorithm [22] that highlights object contour features and is well suited to edge-based processing. It first converts the pixel information to the CT values, as shown in Fig. 2. Then, it counts the population of the CT values as bins and generates histograms. The image features can be enhanced by partitioning the input image ($N_{\text{partition}}$) and having multiple histograms corresponding to each sub-image at the cost of extra computations. Then, the linear SVM receives generated feature maps for object detection.

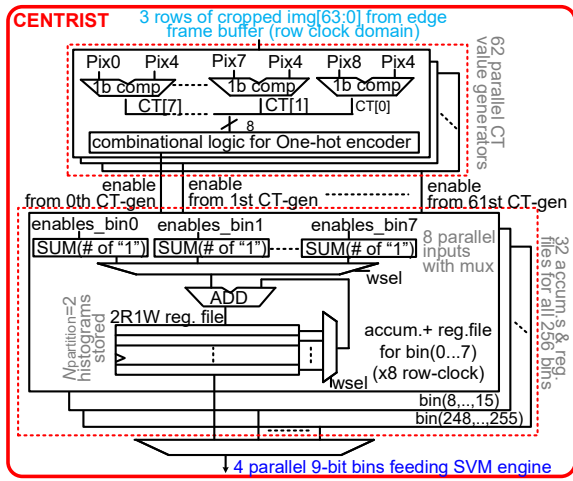


Fig. 5. The implemented CENTRIST utilizing parallel data processing

While the conventional SVM requires some divisions for the scaling factor and normalization, hardware dividers are expensive in terms of power and area. Hence, we introduce approximations in the SVM: 1. The scaling factor (Q_{fac}) is applied to weight values pre-stored in weight memory modules. 2. Normalization is omitted at the cost of a few % accuracy loss compared to an ideal implementation with software. Thanks to the approximation, the power consumption is a few hundred nanowatts in the target frequency range, as shown later.

C. Circuit architecture

The details of the image sensor, edge, and novelty detectors' architectures are depicted in [9]. The edge detector generates an edge frame row every imaging row cycle by checking gradients of neighboring pixels. And the novelty detector outputs a row of the novelty map at every eight rows of imaging cycles as it accumulates the edge counts in the 8×8 window and compares them with the ones from the last frame.

The proposed ROI sweeper implementation is shown in Fig. 3. The input buffer stores eight lines of the input novelty map, corresponding to the bounding box height (= 64 pixels). The input is updated every eight row cycles with an enable signal from the novelty detector. Each adder tree assigned to a bounding box candidate evaluates $BOXNOV$ based on the map stored in the line buffer. Then, the maximum $BOXNOV$ among the 25 trees and its location is stored. These computations are repeated until the entire novelty map is swept. Finally, the location that has the highest novelty value among the whole map is obtained and sent to the frame buffer. The activity filter evaluates $FRMNOV$ with a simple accumulator. The classification trigger is generated by a digital comparator.

The frame buffer in Fig. 4 consists of four 2kB SRAM banks with a 64-bit data width single-port, available in [23]. The buffer operates at $2 \times$ faster row clock so that both read/write operations can be treated in one row cycle and behave as a pseudo dual-port SRAM. The counter receives the start address of the bounding box and keeps incrementing until the whole cropped image read out. Also, another counter counts from address 0 to 255 for the write address. The read operation is assigned earlier than write operation in a row cycle so that the previous frame information

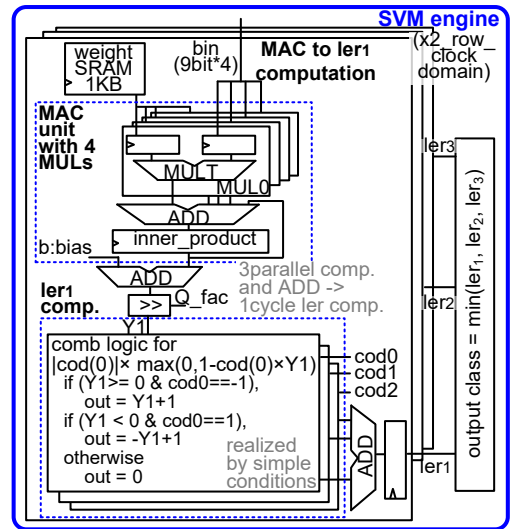


Fig. 6. Implemented SVM engine with parallel processing elements.

is not overwritten before the cropped image read out.

The proposed CENTRIST implementation is shown in Fig. 5. The 62 parallel CT value generators convert input pixel information into CT values (see Fig. 2) from 0 to 255, which are then encoded into the one-hot codes. These codes represent the population of each bin; thus, consecutive logic trees count the number of asserted flags. The accumulator fused with the register file computes the histogram based on the inputs and is shared with eight bins by a multiplexer that switches the input at an $8 \times$ faster cycle. The CENTRIST module has 32 accumulators in total. Thus, all 256 bins (= eight bins per accumulator \times 32) laid in the whole input row are considered in one row cycle. As shown later, we decided that $N_{partition}$ is set to two after design search exploration, which means the CENTRIST module generates two histograms (i.e., 512 features).

The ECOC-based SVM engine in Fig. 6 is designed for solving three-class problems with 512 input features. It does not need buffers for temporal results unlike BNNs/CNNs, enabling hundreds nW power. Three parallel processing elements conduct the computation from the inner product to the loss computation for each class. Then, the final $min.$ logic finds the detected class. Inside each processing element, four multipliers (MULs) share one adder and register, leading to lower clocking power while accelerating inner product operation. Accordingly, the SVM engine reads four bins in parallel from the CENTRIST module. The loss computation is conducted with simple combinational logic. Thus, by operating with a $2 \times$ faster row frequency, the classification is completed within 64 (=512/4) plus a few after the inner product cycles in total.

D. Entire system operation

Fig. 7 shows a detailed timing chart of the entire processing chain. Every row cycle generates a row of edge frames, and the novelty detector accumulates eight of them to generate one row of the novelty map (see the leftmost box). After filling up the line buffer of the ROI sweeper with the first eight rows, parallel evaluations for $BOXNOV$'s start. Once the whole novelty map is rastered, the ROI sweeper sends the bounding box address to the frame buffer.

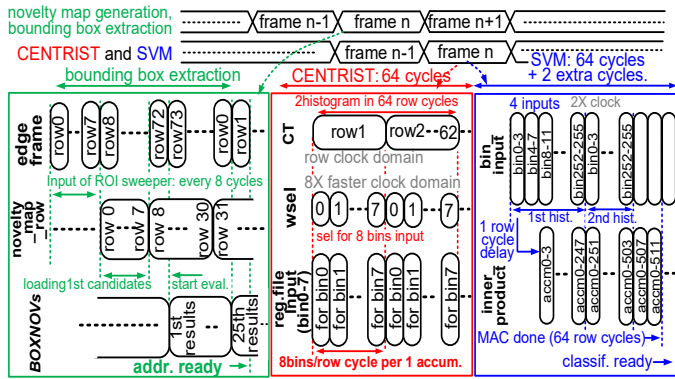


Fig. 7. Timing chart from edge frame to SVM

The CENTRIST module receives one row of cropped images (64×64) from the frame buffer every row cycle, and combinational logic trees generate the enable signals for bins population at the same cycle (see the middle box). Then, those enables are accumulated into the register file with the $8 \times$ row clock by multiplexing eight parallel inputs, which allows one histogram to be generated in 32 cycles ($= 64$ cycles for two histograms). The SVM module conducts the inner product with 512 features (rightmost box). As discussed above, thanks to four parallel MULTs, they are completed in 64 cycles by operating with a $2 \times$ row clock frequency. The classification result is ready just half a row cycle later than the inner product computations. Since CENTRIST and SVM computations take less than 256 row cycles (one frame cycle), they can form a vision pipeline with novelty detection and bounding box extraction.

As shown in Fig. 7, the processing chain computes the classification and the feature extraction for the previous frame while it performs the bounding box extraction of the current frame. This on-the-fly operation enables all frames generated by the imaging sensor to be considered for object detection.

IV. CHIP DEMONSTRATION AND EXPERIMENTAL SETUP

The proposed imager is fabricated in standard TSMC 180-nm technology (Fig. 9, chip size of 7×5 mm²). It also shows the area breakdown of the implemented system. Most of the area (68%) is occupied by the imaging sensor and its periphery. Then, the edge detector is the second largest module (9%), which manages raw 4-bit grayscale inputs. Nevertheless, the rest of the modules have small area occupations, thanks to the edge-based processing. The size of the frame buffer is especially 7%, which will be $4 \times$ if the raw grayscale image must be stored. CENTRIST is larger than other modules, even after the edge detection, because of the register files for two histograms. We trained the SVM model with the STL-10 [24] and INRIA person

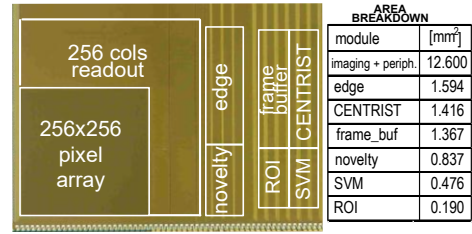


Fig. 9. Chip photo and area breakdown

[25] datasets so that it detects car, animal, and person. The obtained model is utilized for accuracy measurements.

V. MEASUREMENT/EVALUATION RESULTS

Fig. 8 (a) reveals the measured power consumption of the entire system at various frame rates. The proposed imager consumes $8.51 \mu\text{W}$ in total at 1FPS and 0.8V (0.9V) of V_{DD} for digital (analog). The minimum V_{DD} voltage is limited by SRAM. Nevertheless, the power can be further reduced to $5.07 \mu\text{W}$ at 0.5 FPS with the same V_{DD} . Fig. 8 (b) shows the power breakdown among the imaging periphery, image sensing, and the entire processing chain. The first one (see [9] for details) dominates the system power, followed by the third. They determine the entire system's power behavior as having a digital nature and leading to lower minimum power than analog-based processing. The proposed processing, from edge detection to classification, occupies 26% of the entire system power ($2.19 \mu\text{W}$), as further detailed in Fig. 8 (c). The frame buffer and edge detector consume the majority of power (58%) among them; hence, minimizing the memory capacity with edge-based processing contributes a significant reduction in the system power. Also, thanks to the reduced data size for the classification, the total power consumption with SVM and CENTRIST is 544 nW .

To observe the effect of CENTRIST, Fig. 10 presents simulation results showing the accuracy improvement of CENTRIST to sweeps various $N_{\text{partition}}$ s. In this evaluation, three classes from STL-10 (car, deer, boat) are used. As can be seen, CENTRIST improves 5.3% of the classification accuracy without frame divisions. $N_{\text{partition}} = 2$ provides a further 3.5% of the improvement. However, we can see the diminishing accuracy returns at $N_{\text{partition}} = 4$ (only 0.4% of improvement) while it requires $2 \times$ further computations. Hence, the optimal value that we employed in our design is two. We further evaluate the accuracy of the implemented SVM using ideal image inputs from the STL-10 and INRIA datasets, focusing on the inherent SVM accuracy without any noise from imaging. The results are summarized in a confusion matrix in Figs. 11 and Fig. 12 for

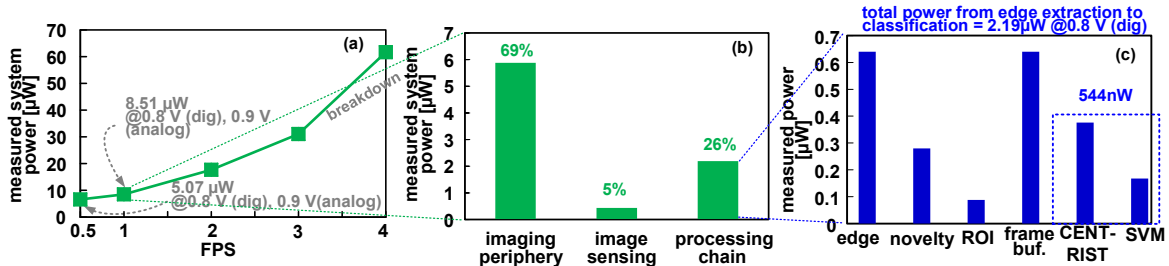


Fig. 8 (a) entire system power (b) power breakdown with image periphery, image sensing and processing chain (c) power break down inside the processing chain.

$N_{\text{partition}} \rightarrow$	0 (no CENTRIST)	1	2	4
CENTRIST iterations	0	1X	2X	4X
accuracy loss	-9.2%	-3.9%	-0.4%	0%

CENTRIST improves accuracy by 5.3% on avg
diminishing accuracy returns

Fig. 10. CENTRIST accuracy improvement with $N_{\text{partition}}$ (3 class). See Fig. 2 for how the frame is partitioned.

		predicted classes			
actual classes	class	car	living thing	TPR	FNR
	car	475	20	0.960	0.040
	living thing	25	980	0.975	0.025
	avg. accuracy	0.97			

living thing = person/animal

Fig. 11. Confusion matrix of implemented SVM with two classes. (car: STL-10, living thing with person: INRIA and STL-10)

		predicted classes				
actual classes		car	animal	person	TPR	FNR
	car	475	16	4	0.960	0.040
	animal	10	341	33	0.888	0.112
	person	15	143	463	0.746	0.254
	avg. accuracy	85.2		SW implementation (no approximations): 88.7%		

Fig. 12. Confusion matrix of implemented SVM with three classes.

the two-class and three-class problems, respectively. Regardless of the approximation, the SVM achieved an average accuracy of 97% for simple two-class object detection. As expected, the accuracy decreases for three-class object detection. Nevertheless, it still achieves an average accuracy of 85.2%. The accuracy loss compared to an ideal linear SVM software implementation was 3.5% compared to the three-class result.

To evaluate the inherent accuracy of the entire processing chain, from edge detection to SVM, Fig. 13 presents the object detection accuracy with the classifier and frame skip rate from the ROI sweeper, using ideal indoor/outdoor video sequence inputs from the CDnet 2014 [26] and CAMEL [27] datasets. Cars do not appear in indoor scenes, so we suppressed the class for such videos with two-class detection (i.e., person and animal are used), while the outdoor scenes are evaluated with living things and cars. Classification accuracy is defined as the ratio of correct predictions among all triggered classification events by the ROI sweeper. The average accuracy with three(two)-class detection is 88 % (90 %). Also, the blue bar shows that the ROI sweeper efficiently skips 60 % of the redundant frames (i.e., frames having negligible novelties) on average.

VI. COMPARISON WITH STATE OF THE ART

Table I compares other image sensor SoCs that integrate imaging to detection processing. Our system uniquely achieves multi-class object detection, which requires more powerful processing than face/digit detection within the target power budget (state-of-the-art of 130 pW/pixel). The novelty detector and ROI sweeper enable 60 % frame redundancy suppression and spotting the bounding box for flexible object targets, unlike others ([11] and [12] detect face only). Although [13] achieves similar power efficiency, it is limited to single-face detection, and our per-pixel power remains 2.3 \times lower. [15] achieves comparable power efficiency but sacrifices ML flexibility and cost due to its optical layer. In contrast, our imager achieves 10% better power efficiency using a standard CMOS process, which leads to lower fabrication costs as well. As another cost metric, the active area per pixel count represents the fabrication cost efficiency. Some imagers are more area-efficient, but their power

ALGORITHM ACCURACY AND NOVELTY-INDUCED FRAME SKIP RATE FOR VARIOUS DATASETS

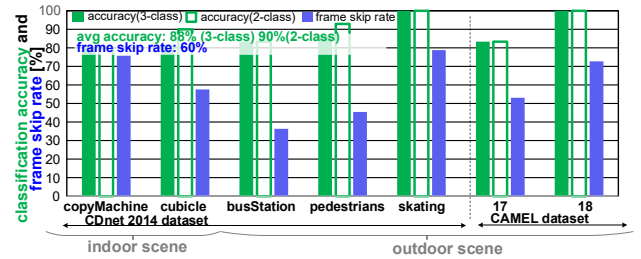


Fig. 13. System accuracy with video datasets: CDnet 2014 and CAMEL

TABLE I. COMPARISON WITH PRIOR ART

	This work	[15]	[14]	[11]	[12]	[13]
Tech.	180 nm CMOS	180 nm CMOS	180 nm CMOS	65 nm CMOS	180 nm CIS	90 nm CIS
resolution [pixels]	256 \times 256	128 \times 128	126 \times 126	160 \times 128	240 \times 240	160 \times 120
area*/pix [F ² /pixel]	8700 (edge frame buf.)	36100 (no buf.)	4,970 (no buf.)	33,000 (no buf.)	7,140 (analog buf.)	28,000 (no buf.)
frame rate [FPS]	0.5 - 4	1-32	250	24 - 268	120	1 - 5
image bit	4	10	8	8	10	4 / 6
power/pixel [pW/pixel]	130 @ 1 FPS	143 @ 1FPS	8,470	2,050 - 10,000	5,035 @ 120 fps	300 @ 5 FPS
energy / (pixel x frame)** [pJ]	130	143	33.8	2.5 - 104 (7.2-303 ***)	438.4	59 (120****)
integrated vision processing	edge +novelty + bbox**** + CENTRIST + SVM	Optical / normal conv.+ ReLU / MP	Conv.+ ReLU / MP / FC	Conv.+ Cortex-M0 +bbox (face only)	analog log Haar-like filters + bbox (face only)	motion + edge + SVM
ML model / classifier	9b SVM 64 \times 64	3-layer BNN	3-layer CNN 126 \times 126	Viola-Jones 64 \times 64	classifier (25 features)	5b SVM 32 \times 26
ML model demo.	multi-class object classif.	digit/eye orientat. classif.	(single) face det. only	(multiple) face det. only	(multiple) face det. only	(single) face det. only

*extrapolated from chip-photo and normalized by process feature size **irrelevant metric for purely harvested system *** scaled to 180 nm by 0.7X/gen. (average factors announced by foundry) **** bounding box

per pixel is an order of magnitude worse than ours, which achieves better FoMs in both area and power.

VII. CONCLUSION

In this paper, we proposed a smart imager with multi-class object detection fabricated in 180-nm standard CMOS technology. An edge-frame-based processing chain and the bounding box extraction with novelty detection effectively reduce the on-chip processing amount and required buffer size. Moreover, the proposed algorithm operates on-the-fly by utilizing parallel processing to avoid lags and huge frame buffers. As a result, the imager achieved 8.51 μ W of power consumption, which is 130pW/pixel and equivalent to the power generated by a 7mm \times 6mm harvester at 300lux. We demonstrated that the proposed system works with the prevalent video datasets (CDnet2014 and CAMEL), achieving a detection accuracy of 90% for two classes, which demonstrates the proposed imager's suitability for real-world applications.

ACKNOWLEDGMENT

This work was supported by the Singapore Ministry of Education (T2EP50223-0040) and TSMC for chip fabrication.

REFERENCES

- [1] G. Kim et al., "A 467nW CMOS visual motion sensor with temporal averaging and pixel aggregation," 2013 IEEE International Solid-State Circuits Conference Digest of Technical Papers, San Francisco, CA, USA, 2013, pp. 480-481.
- [2] J. Choi, S. Park, J. Cho and E. Yoon, "A 3.4- μ W Object-Adaptive CMOS Image Sensor With Embedded Feature Extraction Algorithm for Motion-Triggered Object-of-Interest Imaging," in *IEEE Journal of Solid-State Circuits*, vol. 49, no. 1, pp. 289-300, Jan. 2014.
- [3] X. Zhong, Q. Yu, A. Bermak, C. -Y. Tsui and M. -K. Law, "A 2PJ/Pixel/Direction MIMO Processing Based CMOS Image Sensor for Omnidirectional Local Binary Pattern Extraction and Edge Detection," in Proc. *IEEE Symp. VLSI Circuits*, Honolulu, HI, USA, 2018, pp. 247-248.
- [4] K. D. Choo et al., "5.2 Energy-Efficient Low-Noise CMOS Image Sensor with Capacitor Array-Assisted Charge-Injection SAR ADC for Motion-Triggered Low-Power IoT Applications," in *IEEE ISSCC Dig. Tech. Papers*, San Francisco (CA), Feb. 2019, pp. 96-98.
- [5] Y. Zou, M. Gottardi, M. Lecca and M. Perenzoni, "A Low-Power VGA Vision Sensor with Embedded Event Detection for Outdoor Edge Applications," in *IEEE Journal of Solid-State Circuits*, vol. 55, no. 11, pp. 3112-3121, Nov. 2020.
- [6] M. -Y. Chiu et al., "A Multimode Vision Sensor With Temporal Contrast Pixel and Column-Parallel Local Binary Pattern Extraction for Dynamic Depth Sensing Using Stereo Vision," in *IEEE Journal of Solid-State Circuits*, vol. 58, no. 10, pp. 2767-2777, Oct. 2023.
- [7] T. -H. Hsu et al., "A 0.8 V Multimode Vision Sensor for Motion and Saliency Detection With Ping-Pong PWM Pixel," in *IEEE Journal of Solid-State Circuits*, vol. 56, no. 8, pp. 2516-2524, Aug. 2021.
- [8] T. -H. Hsu et al., "A 0.5-V Real-Time Computational CMOS Image Sensor With Programmable Kernel for Feature Extraction," in *IEEE Journal of Solid-State Circuits*, vol. 56, no. 5, pp. 1588-1596, May 2021.
- [9] K. A. Ahmed, H. Okuhara and M. Alioto, "55pW/pixel Peak Power Imager with Near-Sensor Novelty/Edge Detection and DC-DC Converter-Less MPPT for Purely Harvested Sensor Nodes," 2023 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2023, pp. 102-104.
- [10] J. Vohra, A. Gupta, M. Alioto, "Imager with In-Sensor Event Detection and Morphological Transformations with 2.9 pJ/pixel \times frame Object Segmentation FOM for Always-On Surveillance in 40 nm," in *IEEE ISSCC Dig. Tech. Papers*, San Francisco (CA), Feb. 2024, pp. 104-105.
- [11] M. Lefebvre, L. Moreau, R. Dekimpe and D. Bol, "7.7 A 0.2-to-3.6TOPS/W Programmable Convolutional Imager SoC with In-Sensor Current-Domain Ternary-Weighted MAC Operations for Feature Extraction and Region-of-Interest Detection," in *IEEE ISSCC Dig. Tech. Papers*, San Francisco (CA), Feb. 2021, pp. 118-120.
- [12] H. Song, S. Oh, J. Salinas, S. -Y. Park and E. Yoon, "A 5.1ms Low-Latency Face Detection Imager with In-Memory Charge-Domain Computing of Machine-Learning Classifiers," in Proc. *IEEE Symp. VLSI Circuits*, Kyoto, Japan, Jun. 2021.
- [13] A. Verdant et al., "A 3.0 μ W@5fps QVGA Self-Controlled Wake-Up Imager with On-Chip Motion Detection, Auto-Exposure and Object Recognition," in Proc. *IEEE Symp. VLSI Circuits*, Honolulu, HI, USA, Jun. 2020, pp. 1-2.
- [14] Hsu, Tzu-Hsiang et al., "A 0.8V Intelligent Vision Sensor with Tiny Convolutional Neural Network and Programmable Weights Using Mixed Mode Processing-in-Sensor Technique for Image Classification," in *IEEE ISSCC Dig. Tech. Papers*, San Francisco (CA), Feb 2022, pp.1-3.
- [15] Xuecheng Wang, Zheng Huang, Tianyi Liu, Wanxin Shi, Hongwei Chen, Milin Zhang, "A 0.35V 0.367TOPS/W Image Sensor with 3-Layer Optical-Electronic Hybrid Convolutional Neural," in *IEEE ISSCC Dig. Tech. Papers*, San Francisco (CA), Feb. 2024, pp. 116-117.
- [16] H. Xu et al., "Senputing: An Ultra-Low-Power Always-On Vision Perception Chip Featuring the Deep Fusion of Sensing and Computing," in *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 69, no. 1, pp. 232-243, Jan. 2022.
- [17] M. Nazhamaiti et al., "Selfputing: A 0.57 μ W @ 15 fps Vision Chip with Self-powered In-Pixel Computing and In-Memory Computing for Visual Perception on the Edge," 2024 *IEEE European Solid-State Electronics Research Conference (ESSERC)*, Bruges, Belgium, 2024, pp. 585-588.
- [18] A. Y. -C. Chiou and C. -C. Hsieh, "A 137 dB Dynamic Range and 0.32 V Self-Powered CMOS Imager With Energy Harvesting Pixels," in *IEEE Journal of Solid-State Circuits*, vol. 51, no. 11, pp. 2769-2776, Nov. 2016.
- [19] S. Cho, S. Choi, J. Woo, A. Kim and B. -G. Nam, "A self-powered always-on vision-based wake-up detector for wearable gesture user interfaces," 2017 *IEEE Asian Solid-State Circuits Conference (A-SSCC)*, Seoul, Korea (South), 2017.
- [20] S. Hanson, Z. Foo, D. Blaauw and D. Sylvester, "A 0.5 V Sub-Microwatt CMOS Image Sensor With Pulse-Width Modulation Read-Out," in *IEEE Journal of Solid-State Circuits*, vol. 45, no. 4, pp. 759-767, April 2010.
- [21] K. A. Ahmed, L. Lin, P. S. Salamani and M. Alioto, "Imager with Dynamic LSB Adaptation and Ratiometric Readout for Low-Bit Depth 5- μ W Peak Power in Purely-Harvested Systems," 2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits), Honolulu, HI, USA, 2022, pp. 50-51.
- [22] J. Wu and J. M. Rehg, "CENTRIST: A Visual Descriptor for Scene Categorization," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1489-1501, Aug. 2011.
- [23] Dolphin Design, "Single Port SRAM compiler - Memory optimized for ultra low power and high density - Dual Voltage - compiler range up to 512 k," URL: https://my.dolphin-design.fr/memory/spram-pluton-elc-hd-rr-dv_tsmc_180nm_g_generator/ (last access:17 March 2024).
- [24] A. Coates, A. Ng, H. Lee, "An Analysis of Single-Layer Networks in Unsupervised Feature Learning". In Proc. The International Conference on Artificial Intelligence and Statistics, 15:215-223, 2011.
- [25] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," In Proc. IEEE CVPR, San Diego, CA, USA, 2005, pp. 886-893.
- [26] Y. Wang, P. -M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth and P. Ishwar, "CDnet 2014: An Expanded Change Detection Benchmark Dataset," in *IEEE CVPR Workshops*, Columbus, OH, USA, 2014, pp. 393-400.
- [27] E. Gebhardt and M. Wolf, "CAMEL Dataset for Visual and Thermal Infrared Multiple Object Detection and Tracking," In Proc. IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 2018, pp. 1-6.