A Fully Dynamic Multi-Mode CMOS Vision Sensor With Mixed-Signal Cooperative Motion Sensing and Object Segmentation for Adaptive Edge Computing

Xiaopeng Zhong, *Student Member, IEEE*, Man-Kay Law, *Senior Member, IEEE*,
Chi-Ying Tsui, *Senior Member, IEEE*, and Amine Bermak, *Fellow, IEEE*

***Abstract*—This article presents a low-power multi-mode CMOS vision sensor with mixed-signal in-sensor computation capabilities targeting the next-generation wireless sensing applications. To support the always-on and scene-adaptive edge computing scenarios with low power and low bandwidth, the sensor is reconfigurable for three operation modes, namely: 1) motion sensing (MS); 2) object segmentation (OS); and 3) full imaging (FIM). A mixed-signal cooperative scheme of frame differencing (FD) and background subtraction (BS) is proposed to achieve high-accuracy MS with varying object sizes and speeds. The mixed-signal BS-based OS can minimize both object localizing and imaging efforts for object analysis upon motion triggering, while FIM enables complete scene recording for the identified object of interest. The complete CMOS vision sensor is implemented through reconfigurable and fully dynamic mixed-signal processing at both pixel and column levels cooperatively to achieve low power and compact area. Fabricated in a 0.18-μm CMOS, the 256 × 216 chip prototype achieves the cooperative MS with only 2.36 μW at 15 frames/s, when composed of 14 FD frames (147 nJ/frame) and 1 BS frame (302 nJ/frame). The OS mode consumes 1.44∼2.04 μJ/frame at 0%∼100% object occupancy, linearly corresponding to 41%∼16% power saving when compared with the conventional digital OS. The FIM mode operates with only 1.41 μJ/frame for complete scene recording. The achieved energy efficiencies for all operation modes compare favorably with the state of the art.**

***Index Terms*—Background subtraction (BS), CMOS vision sensor, cooperative motion sensing (MS), edge computing, frame differencing (FD), fully dynamic, in-sensor computation, mixed-signal processing, object segmentation (OS), wireless sensor networks (WSNs).**

I. Introduction

Wireless sensor networks (WSNs) are the enabling technologies for Internet-of-Things (IoT) development. With the WSNs widely adopted in various applications including intelligent transportation, inhabitation, agriculture, and many more, CMOS vision sensors continue to play an important role in scene interpretation [1], [2]. They are paired up with the vision processors to form wireless vision sensor nodes (WVSNs). Due to the limited energy available from batteries and/or energy-harvesting sources [3], [4], a WVSN requires low power and low bandwidth so as to achieve an always-on and long-term autonomous operation at a minimum maintenance cost [1], [2].

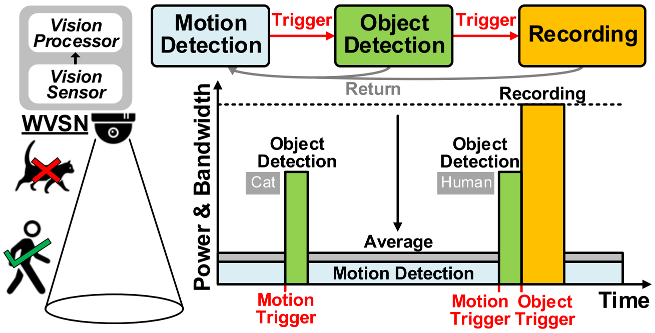


Fig. 1. Scene-adaptive edge computing paradigm of a WVSN for power and bandwidth optimization.

To fulfill the requirement of low power and low bandwidth, scene-adaptive edge computing is demanded in the WVSN, as illustrated in Fig. 1. Even using low-power sensors [5], [6] and in-sensor image compression [7]–[9], always-on image recording and streaming still consume a large bandwidth that can easily drain energy resources due to power-hungry wireless transmissions [10], [11]. Thus, object detection becomes popular to locally analyze the data before triggering recording and transmission for effective bandwidth and power reduction [1], [12], especially by using in-sensor computation architectures [12]–[15]. However, continuous object detection still incurs significant power and bandwidth burden for the long-term WVSN operation. Consequently, motion detection is widely employed as a trigger to minimize redundant object detections without missing the target objects [16]–[18].

Several low-power vision sensors with in-sensor motion sensing (MS) have been reported to work with a vision processor for the scene-adaptive WVSN. The mixed-signal designs [10], [16], [19]–[24] typically demonstrate superior power, speed, and area performances when compared with the

Manuscript received July 18, 2019; revised November 6, 2019; accepted December 11, 2019. Date of publication January 8, 2020; date of current version May 27, 2020. This article was approved by Associate Editor David Stoppa. This work was supported in part by the Hong Kong Innovation and Technology Fund (ITF), in part by Intel, and in part by Texas Instruments under Grant ITS/211/16FP. *(Corresponding author: Xiaopeng Zhong.)*

Xiaopeng Zhong and Chi-Ying Tsui are with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong (e-mail: xzhongab@ust.hk).

Man-Kay Law is with the State Key Laboratory of Analog and Mixed-Signal VLSI, University of Macau, Macau, China.

Amine Bermak is with the Information and Computing Technology Division, College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSSC.2019.2961848

digital counterparts [17], [18], [25]. However, existing mixed-signal MS vision sensors still exhibit several limitations. First, they only support one single MS technique (i.e., either frame differencing (FD) [10], [16], [19]–[22] or background subtraction (BS) [23], [24]), resulting in limited sensitivity to various object speeds and sizes. Even though combining FD and BS is viable for improving the MS robustness [17], [26], the corresponding mixed-signal in-sensor implementation is yet to be demonstrated. Second, prior vision sensors lack on-chip object segmentation (OS) capability to differentiate objects from the background (BG) upon motion triggering, which is necessary to reduce data movement bandwidth, avoid tedious object window searching, as well as improve detection accuracy [1], [27]. Although digital post processing is feasible for OS, the induced imaging and processing overhead can be significant [20], [28]. Finally, due to the incompatibility among the processing modules for multi-mode operations and the associated static power consumption, optimizing the energy efficiencies for all the operation modes still remains a challenge.

In this article, we present a low-power CMOS vision sensor, embedding multiple mixed-signal operation modes for the scene-adaptive WVSN under static/rarely changed BGs. The MS mode is applied for motion triggering [17] instead of activity monitoring/tracking [26]. To the best of our knowledge, it is the first attempt to combine FD and BS in a compatible mixed-signal architecture and operate them cooperatively to improve the MS robustness against varying object sizes and speeds. Based on the BS results, mixed-signal OS is directly realized during the column-level analog-to-digital conversion to achieve high segmentation accuracy while reducing the processing power. Full imaging (FIM) mode is also seamlessly implemented through the reconfigurable processing elements (RPEs) to support full-scene recording. Featuring high reconfigurability and fully dynamic operations with mixed-signal pixel and column processing, this article demonstrates both high energy efficiency and compact area for all the operation modes.

The remainder of this article is organized as follows. The MS techniques and the proposed multi-mode vision sensor are introduced in Section II. Section III describes the sensor architecture and the pixel implementation. The reconfigurable processing architecture is elaborated in Section IV. Experimental results are discussed in Section V, followed by the conclusions in Section VI.

II. Multi-Mode Vision Sensor

This section first outlines the two existing MS techniques (i.e., FD and BS) and their respective limitations, and then discusses the implementation and operation principle of the proposed mixed-signal multi-mode CMOS vision sensor.

A. FD

FD has been widely used for low-power MS due to its simplicity. Its operating principle is expressed as (1) and illustrated in Fig. 2(a). F_{n-1} is the previous frame and F_n is the current frame. $M_{\mathrm{FD}n}$ is a 1-bit motion image and TH is a threshold for comparison. When the absolute difference between $F_n(i,j)$ and $F_{n-1}(i,j)$ is larger than TH, $M_{\mathrm{FD}n}(i,j)$ is set to "1" to indicate a motion event. Thresholding is used to filter false detections caused by noise, so TH is usually set right above the noise level. Then, TH determines the minimum detectable brightness contrast, which is harder to achieve by objects in the dark regions.

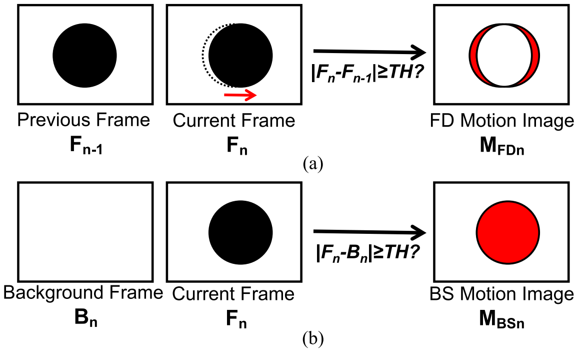


Fig. 2. MS techniques. (a) FD. (b) BS.

$$M_{\mathrm{FD}n}(i,j)=\begin{cases}1, & |F_n(i,j)-F_{n-1}(i,j)|>TH\\ 0, & |F_n(i,j)-F_{n-1}(i,j)|\leq TH\end{cases} \tag{1}$$

The total number of motion events (N_{FD}) is often used as a criterion for triggering [17], [18], [22], which is proportional to the object size (S) and displacement (D) between the two frames

$$N_{\mathrm{FD}} \propto (S, D) \tag{2}$$

where $D = V/\text{frame rate}$ (V is the object speed). As high frame rate is required to maintain a prompt response to fast motion, the minimum detectable S and V are limited in FD. Although dual-frame-rate FD [21] can alleviate the issue, the sensor resolution is sacrificed and the minimum frame rate is still limited by the in-pixel capacitor leakage. In addition, FD is unsuitable for OS as only the edge regions can be detected, as shown in Fig. 2(a).

B. BS

Fig. 2(b) illustrates the BS operation, as described by (3). $M_{\mathrm{BS}n}$ is obtained by comparing the absolute difference between F_n and a BG frame (B_n) against TH. With the same sensor/noise, TH is the same for FD and BS.

$$M_{\mathrm{BS}n}(i,j)=\begin{cases}1, & |F_n(i,j)-B_n(i,j)|>TH\\ 0, & |F_n(i,j)-B_n(i,j)|\leq TH\end{cases} \tag{3}$$

As B_n represents the BG that contains no moving objects, the total number of motion events by BS (N_{BS}) is insensitive to V and frame rate, but only related to S

$$N_{\mathrm{BS}} \propto (S). \tag{4}$$

Thus, with a high frame rate to capture fast motions, BS can cover a wide range of V. In addition, BS typically exhibits higher sensitivity to S than FD for small object detection, as it detects the whole object silhouette, which is also suitable for OS. The main shortcomings of BS are the weak adaptability to

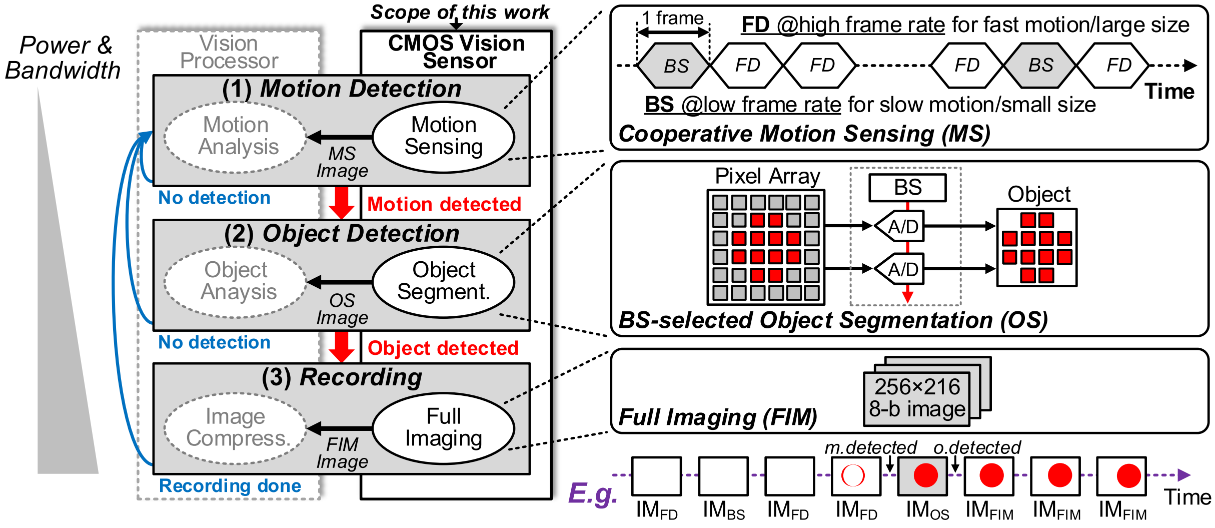


Fig. 3. Proposed mixed-signal multi-mode CMOS vision sensor with operation principle for the always-on and scene-adaptive edge computing in the WVSN.

sudden BG change and the relatively high circuit complexity. Different from FD's fast reaction to BG change, BS will continuously generate false triggers when the BG is influenced by the extraneous events. In addition, BS requires long-term B_n storage, which often relies on very large capacitors [23] or low-leakage CMOS-incompatible devices [24], as well as costly analog filters for B_n update [23]. Therefore, BS typically consumes more power and area than FD.

C. Mixed-Signal Multi-Mode CMOS Vision Sensor

The proposed mixed-signal multi-mode CMOS vision sensor is illustrated in Fig. 3. It supports three operation modes, namely MS, OS, and FIM, to interface with a vision processor (outside the scope of this article) for motion detection, object detection, and recording in scene-adaptive edge computing, respectively. Fig. 3 also shows the typical operation flow of the sensor in three operating states along with an example:

1) State 1: The vision sensor operates at the MS mode, capturing FD/BS images for the always-on motion detection, which consumes the least power and bandwidth. A cooperative MS scheme combining FD and BS is employed. FD operates at a high frame rate (sensor rate) to maintain a prompt response to fast and large objects. To make up for the FD shortcomings, low-frame-rate BS is inserted by replacing a portion of FD frames to sense slow and small objects. The frame rate of BS is adjustable but low frame rate is preferred to minimize the power overhead of BS while still providing accurate sensing of the slow and small targets. In this way, the robustness of MS can be improved while maintaining low power consumption.

With the vision sensor continuously providing MS images, motion trigger is determined by the vision processor. As in [17], [18], and [22], all motion events of the current FD/BS frame are summed up and compared against a threshold. Motion is detected when the total number exceeds the threshold, and then, the operating state is switched to State 2. The switching latency between the two states is one integration time as the pixels have to reconfigure and restart integration.

2) State 2: After motion triggering, the vision sensor will conduct OS for object detection. Based on the detected BS silhouette, analog-to-digital conversion is selectively performed to generate segmented object images [29]. This method can not only reduce the signal quantization and digital post-processing but also minimize the amount of data transmitted to the vision processor and facilitate the subsequent object analysis.

3) State 3: When an object of interest is detected, the current scene is recorded by the FIM for a predefined period. The vision processor then initiates wireless transmission of the compressed images and returns the sensor to State 1.

III. Sensor Architecture and Pixel Implementation

A. Sensor Architecture

The proposed vision sensor employs a column-parallel architecture, as shown in Fig. 4. It consists of a pixel array, column-parallel RPEs, row drivers, a digital controller (Ctrl), threshold/ramp generators (TH/Ramp), and input/output (I/O) channels. The pixel array is composed of 128 × 108 groups (each of 2 × 2 pixels). MS employs lower resolution (128 × 108) by pixel binning to reduce power while OS and FIM use full resolution (256 × 216) for better image quality. Each pixel group is fully dynamic and reconfigurable for different operation modes. In the column, each of 108 fully dynamic mixed-signal RPEs is shared by two pixel columns through a multiplexor (MUX). Each RPE can be dynamically configured as an FD PE, a BS PE, or an ADC by Ctrl based on the operation state. Multiple modes are achieved in a small area by combining the reconfigurable pixels and the RPEs. The 18 I/O channels serve for the result output and the BG input. The BG for BS is stored digitally in SRAM for long-term operation. Due to the limited tapeout area, the SRAM is not implemented on the prototype chip, but it has been considered in the performance evaluation.

B. Reconfigurable Pixel Group

As shown in Fig. 5, the reconfigurable pixel group is designed to support multiple operation modes. The pixel group

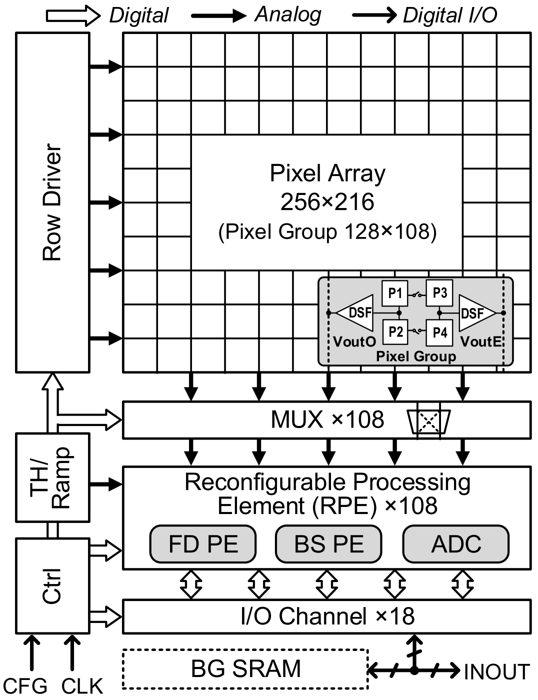


Fig. 4. Block diagram of the CMOS vision sensor.

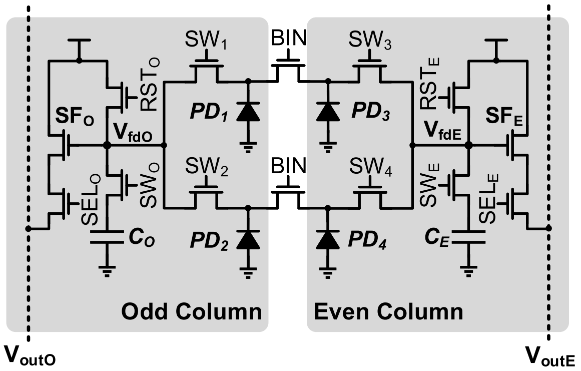


Fig. 5. Architecture of reconfigurable pixel group with 2 × 2 pixels.

consists of 2 × 2 pixels with two readout channels (V_{outO} and V_{outE}). Two pixels in the same column share a floating diffusion and a readout channel (SF). The transistor count is only 3.5 T/pixel to ensure a small pixel size. The two metal–insulator–metal (MIM) capacitors (C_{O}, C_{E}) function as in-pixel memories for FD. They are placed on the top of the active circuits, resulting in a 7.9-μm pixel with 33.4% fill factor (FF) (i.e., 12% FF loss). The loss can be further reduced in a process with relaxed design rules for MIM capacitors or completely eliminated if a backside-illuminated (BSI) CIS process is used.

The pixel configuration and operation at different modes are illustrated in Fig. 6. For FD, four pixels within the same group are binned by the interpixel switches (BINs) and transfer switches ($\mathrm{SW}_{1\text{-}4}$), as shown in Fig. 6(a). To compensate for the larger capacitance, both $\mathrm{RST}_{\mathrm{O}}$ and $\mathrm{RST}_{\mathrm{E}}$ are used to reset the photodiodes (PDs). The two samples required by FD are sequentially obtained after two integrations and stored in C_{O} and C_{E}, respectively, through charge sharing with PDs. If charge sharing happens after integration, 40% voltage swing/dynamic range is lost because of the reduced conversion gain for the intrinsic full-well capacity of the PDs. To compensate for the signal loss, $C_{\mathrm{O}}/C_{\mathrm{E}}$ is connected to the PDs to extend the full-well capacity during integration, by activating $\mathrm{SW}_{1,2,\mathrm{O}}/\mathrm{SW}_{3,4,\mathrm{E}}$. Each sample is stored by turning off $\mathrm{SW}_{1,2,\mathrm{O}}/\mathrm{SW}_{3,4,\mathrm{E}}$ when integration is finished. After storing, the two samples are read out by SF_{O} and SF_{E}, respectively, for the FD operation detailed in Section IV-B. Although C_{O} and C_{E} can be reused to enhance the dynamic range as in [30] for other modes, it is not implemented to avoid increased complexity, and thus save power and area. As shown in Fig. 6(b), BG capture and BS employ the same pixel operation. After initial reset, the integrated signal is read out through SF_{O} by enabling $\mathrm{SW}_{1,2}$, followed by reset and the second sample readout. In Fig. 6(c), binning is disabled to achieve full sensor resolution for both OS and FIM. Each pixel operates as that in BS and four pixels are read out sequentially by the corresponding channels in a zigzag way.

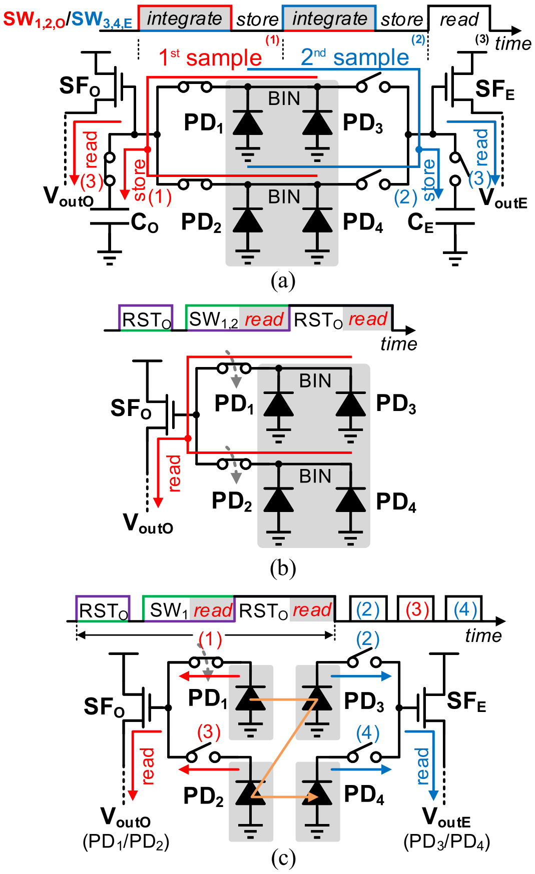


Fig. 6. Pixel configuration and operation at different modes (RST and SEL unshown for simplicity). (a) FD. (b) BG capture and BS. (c) OS and FIM.

C. Dynamic Pixel Readout

Incorporating column-level processing circuits can ensure a small pixel pitch and facilitate circuit reuse for different rows. However, voltage buffers are needed to read out the analog signals from the pixels to the columns. Conventionally, a static source follower (SSF) with a constant bias current (I_{b}) [15] is employed, as shown in Fig. 7(a), which consumes static power not only in the SSF but also in the bias generator. The power

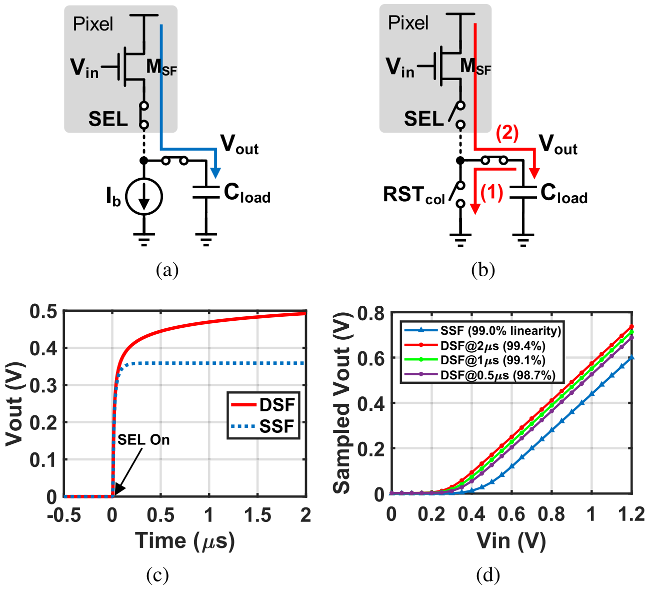


Fig. 7. Pixel readout. (a) SSF. (b) DSF. (c) Transient responses ($V_{\text{in}} = 0.9$ V). (d) Transfer curves of SSF and DSF at $V_{\text{dd}} = 1.2$ V and $C_{\text{load}} = 1$ pF.

TABLE I

PERFORMANCE COMPARISON OF SSF AND DSF READOUTS

Parameter		SSF	DSF @0.5μs
Input Range		0.68 V	0.81 V
Gain		0.79	0.79
Linearity		99.0%	98.7%
Noise		312 μV	288 μV
1σ Offset	w/o cancellation	13 mV	21 mV
	w/ cancellation	6 μV	72 μV
Energy/sample		1.32 pJ	0.46 pJ

problem gets even worse when a long duty cycle is required in the multi-sample readout (e.g., FD). In addition, the voltage headroom required by the current bias also limits the dynamic range especially at a low supply voltage.

To overcome the limitations of the SSF, this article exploits the dynamic source follower (DSF), as shown in Fig. 7(b). Compared with the SSF, the DSF simplifies the circuit implementation as it replaces I_{b} with a simple reset switch (RST_{col}). The dynamic readout operates as follows. First, the loading capacitance (C_{load}) is reset by RST_{col} before signal readout to remove the residue charge (make sure the initial V_{out} is the same and lower than $V_{\text{in}} - V_{\text{th}}$ for all input signals). Second, after turning off RST_{col}, the target pixel is selected by SEL to charge C_{load} through the output transistor (M_{SF}) toward the cutoff voltage ($V_{\text{in}} - V_{\text{th}}$). The difference in the behavior between the SSF and the DSF is depicted by the sample transient responses in Fig. 7(c). The SSF has a steady V_{out} after settling, while V_{out} of the DSF keeps increasing after M_{SF} enters the subthreshold region. From the transfer curves shown in Fig. 7(d), it is also observed that the DSF can provide a highly linear readout with a similar gain as the SSF, as well as a larger input range.

Table I compares the simulated performances of the SSF and the DSF when V_{out} is sampled after 0.5-μs settling, which is applied in all operation modes. The DSF and SSF have comparable gain and noise level, as also demonstrated by another kind of dynamic readout [31]. For 1-σ offset, the DSF is worse than the SSF due to the partial settling when cancellation is not applied. Nevertheless, the offset of both schemes can be suppressed to a small level through double sampling [32]. In addition to the simplified circuit implementation, the major benefits of the DSF are the improvements in the input range and energy efficiency. The input range of the DSF is more than 100 mV larger than that of the SSF thanks to the removed bias voltage headroom and reduced V_{gs} in the subthreshold region. In addition, the DSF consumes almost three times smaller energy than the SSF for one sampling, which will be more substantial if bias generation is also considered. In addition, as the DSF exhibits no energy overhead during the steady state, its advantage is even more prominent for in-sensor computation applications, which involve pixel-column combined processing with long pixel access time. From Fig. 7(d), it is noted that clock jitter will cause sampling noise due to the leaking effect in the sub-threshold region. The peak sampling error is linearly correlated with the clock jitter at a slope of 76 μV/ns. A 5-ns (1%) clock jitter will only induce 380-μV (0.16 $\text{LSB}_{8\text{ bit}}$) error when sampling at 0.5 μs, which is small enough for practical applications. The crosstalk between the adjacent channels is avoided by inserting a grounded shielding line between them.

IV. RECONFIGURABLE AND FULLY DYNAMIC MIXED-SIGNAL PROCESSING ARCHITECTURE

A. Fully Dynamic RPE

To cater for the three sensor operation modes with low power and small area, a fully dynamic mixed-signal RPE architecture is implemented, as illustrated in Fig. 8. Controlled by the global digital controller (Ctrl), the 108 column-parallel RPEs can function as FD PEs, BS PEs, or SAR-SS (successive-approximation-register and single-slope) ADCs. FD and BS are carried out by FD PE and BS PE, respectively. BS PE and ADC are combined to enable OS, while only ADC is used for FIM. The RPEs take one and two cycles to process one row at low and high resolutions, respectively. Each RPE is mainly composed of two differential capacitive digital-to-analog converters (CDAC_{p}, CDAC_{n}), a dynamic comparator (CMP), and a column digital controller. CDAC_{p} and CDAC_{n} employ an identical 6-bit binary-weighted architecture with two 3-bit split sections for area saving and use large capacitance ($C_{\text{unit}} = 49$ fF) to avoid calibration [33]. Depending on the operating mode, CDAC_{p} is directly controlled by Ctrl to connect either the global threshold or the ramp (i.e., $V_{\text{TH}}/V_{\text{Ramp}}$), while CDAC_{n} is operated by the column digital controller using the corresponding logic. The global signals can be gated during OS to disable the column operation. The pixel outputs (V_{outO}, V_{outE}) are multiplexed and bottom-plate sampled at CDAC_{p} or CDAC_{n}. The reference voltage (V_{ref}) is set at $V_{\text{dd}}/2$ (0.6 V) so as to match the V_{out} range of the DSF and avoid common-mode voltage generation. Apart from that, it also simplifies the threshold generation for FD and BS, as explained in Section IV-B. An energy-efficient two-stage dynamic comparator [34] is adopted for

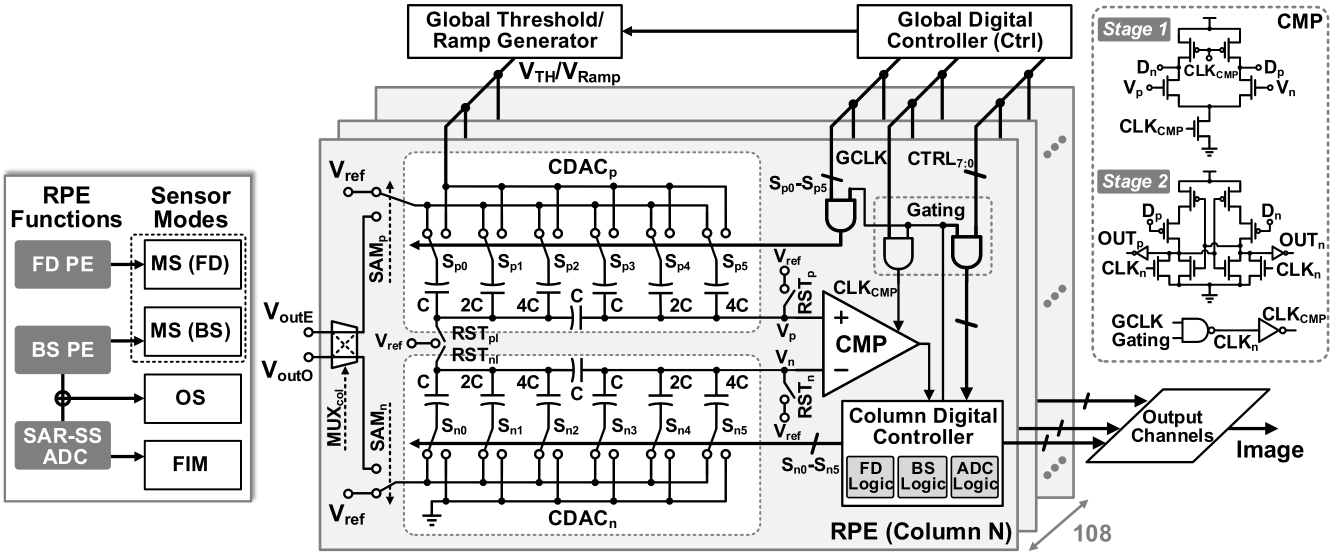
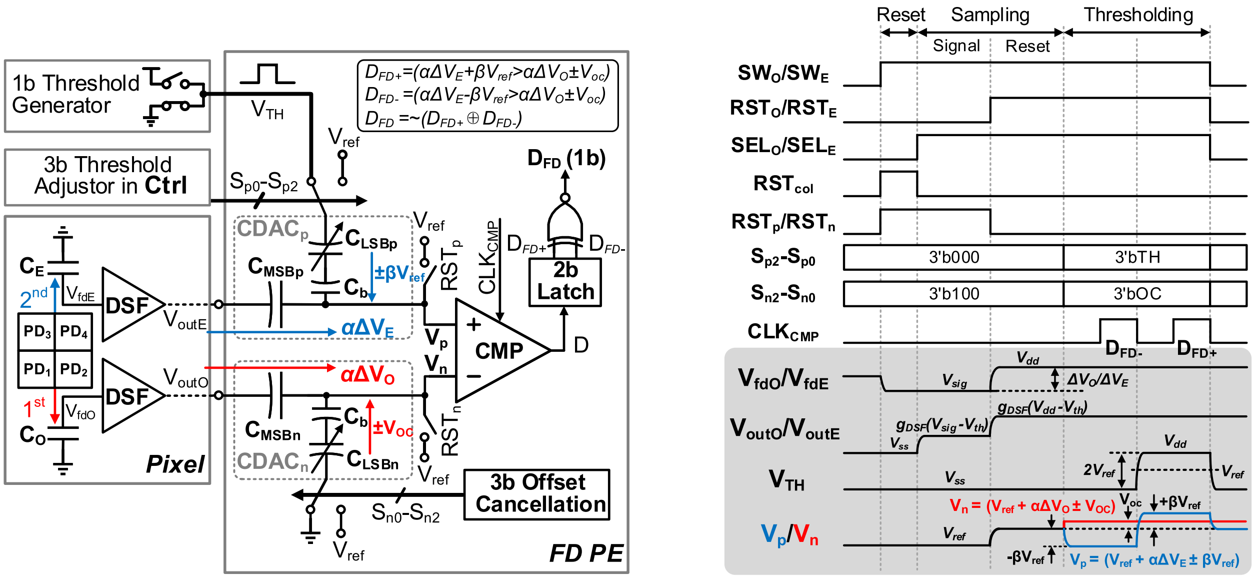


Fig. 8. Column-parallel mixed-signal RPE.



(a) (b)

Fig. 9. FD. (a) Configuration. (b) Operation timing diagram (including control signals for pixel operation).

all the processing functions. Thanks to the high circuit reusability, each RPE occupies 15.8 × 372 μm^2, incurring merely 15.5% area overhead when compared with a single column ADC.

B. FD Configuration and Operation

Fig. 9(a) shows the configuration with a fully dynamic structure for FD, which is formed by a pixel group and a column FD PE. The MSB sections of $CDAC_p$ and $CDAC_n$ (C_{MSBp} and C_{MSBn}) stay connected to V_{outE} and V_{outO}, respectively, during the whole operation, eliminating the need for two large column capacitors for reset voltage sampling to save power and area. The LSB section of $CDAC_p(C_{LSBp})$ couples the global threshold (V_{TH}) for comparison, as illustrated by the example shown in Fig. 10. During reset and sampling, the bottom plate of C_{LSBp} is clamped at $V_{ref}(V_{dd}/2)$.

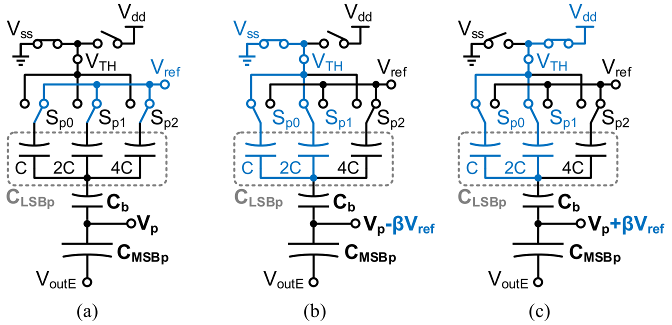


Fig. 10. Fully dynamic threshold generation (TH = 3 $LSB_{6\ bit}$). (a) Reset and sampling. (b) Negative threshold ($-\beta V_{ref}$). (c) Positive threshold ($+\beta V_{ref}$).

After that, by switching the C_{LSBp} bottom plate to V_{TH}(connected to V_{ss}), the negative threshold ($-\beta V_{ref}$) is generated at V_p. Next, V_{TH} is raised to V_{dd} for positive

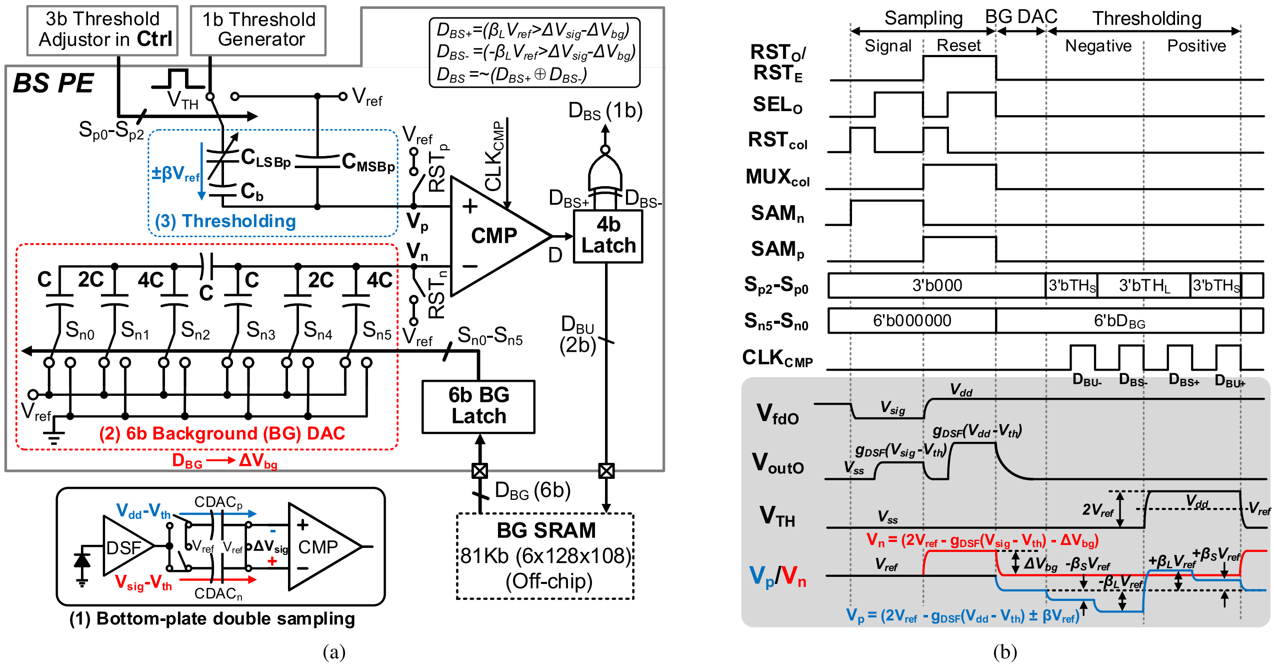


Fig. 11. BS. (a) Configuration. (b) Operation timing diagram (including control signals for pixel operation).

threshold ($+\beta V_{\mathrm{ref}}$) generation. The threshold ratio (β) is adjustable from 0/63 (0 $\mathrm{LSB}_{6\,\mathrm{bit}}$) to 7/63 (7 $\mathrm{LSB}_{6\,\mathrm{bit}}$) by controlling S_{p2}–S_{p0} through the adjustor in the Ctrl. With the local threshold adjustment and $V_{\mathrm{ref}} = V_{\mathrm{dd}}/2$, V_{TH} is either set at V_{ss} or V_{dd} for full-swing threshold generation ($\pm V_{\mathrm{ref}}$), eliminating the power-hungry reference generator and buffer. On the opposite side, C_{LSBn} functions as a 3-bit signed DAC to cancel the comparator offset. After clamped at 3'b100 during sampling, C_{LSBn} generates V_{OC} to compensate for an offset voltage within ± 3.5 $\mathrm{LSB}_{6\,\mathrm{bit}}$, which is sufficient for this article (± 2.2 $\mathrm{LSB}_{6\,\mathrm{bit}}$). Note that the offset of the double-tail dynamic comparator exhibits a high immunity to the common-mode voltage and temperature changes [35]. According to the simulation, the maximum offset drift due to the common-mode voltage change (0.5~1.2 V) and the temperature change (−40 °C~80 °C) is only 0.2 $\mathrm{LSB}_{6\,\mathrm{bit}}$. Therefore, the measurement of the comparator offset can be performed only once with the three compensation bits stored in each column so that negligible power and speed overhead are incurred for the normal FD operation.

The overall FD operation for one pixel is depicted by the timing diagram in Fig. 9(b). After the pixel integration shown in Fig. 6(a), the two samples in C_{O} and C_{E} are read out through the DSFs and stored in C_{MSBn} and C_{MSBp}, respectively. Then, both V_{fdO} and V_{fdE} are reset to V_{dd} and read out through C_{MSBs}, acting as double sampling to cancel out the readout channel mismatch. As the second readout continues to charge the same capacitor after the first readout, channel reset between two readouts is eliminated to save energy. After that, the integration signals ΔV_{O} and ΔV_{E} are obtained at V_{p} and V_{n} as $V_{\mathrm{ref}} + \alpha\Delta V_{\mathrm{E}}$ and $V_{\mathrm{ref}} + \alpha\Delta V_{\mathrm{O}}$, respectively, where $\alpha = g_{\mathrm{DSF}} \cdot C_{\mathrm{MSB}}/(C_{\mathrm{MSB}} + C_{\mathrm{LSB}})$. The DSF gain ($g_{\mathrm{DSF}}$) is almost constant, as discussed in Section III-C, and α is about 0.7. After the offset compensation, the thresholding operation is conducted with the comparisons in the following:

$$\begin{cases} D_{\mathrm{FD}+} = (\alpha\Delta V_E + \beta V_{\mathrm{ref}} > \alpha\Delta V_O \pm V_{\mathrm{OC}}) \\ D_{\mathrm{FD}-} = (\alpha\Delta V_E - \beta V_{\mathrm{ref}} > \alpha\Delta V_O \pm V_{\mathrm{OC}}). \end{cases} \tag{5}$$

The 2-bit comparison results are passed through an XNOR gate to produce the 1-bit FD output [i.e., $D_{\mathrm{FD}} = \sim(D_{\mathrm{FD}+} \oplus D_{\mathrm{FD}-})$].

C. BS Configuration and Operation

The proposed mixed-signal BS stores the BG in the digital memory (SRAM) and converts it back into a voltage for analog thresholding. This method ensures reliable long-term BG storage without charge leakage issues as in the existing analog methods [23], [24], while avoiding ADC and digital processing to achieve low-power and high-speed operation.

Fig. 11(a) shows the fully dynamic configuration for BS, in which the BS PE interfaces with the BG SRAM and the threshold generation circuits. Inside the BS PE, both $\mathrm{CDAC_p}$ and $\mathrm{CDAC_n}$ are employed for bottom-plate voltage sampling. In addition, $\mathrm{CDAC_n}$ also converts the digital BG into analog voltages, while $\mathrm{CDAC_p}$ generates the required thresholds similar to the case of FD, as shown in Fig. 10. Initially, a 128 × 108 6-bit BG image is obtained using an SAR-SS ADC and stored in the off-chip SRAM. The required energy of capturing a BG image is 354 nJ, about 1/4 of that of the FIM. During BS operation, 6-bit BG values ($\mathrm{D_{BG}}$[5:0]) are fetched from the SRAM row by row and loaded serially into the on-chip BG latch to control $\mathrm{CDAC_n}$. As studied in [36], 6 (instead of 8) bits/pixel are chosen to maintain accurate BS (only 0.28 $\mathrm{LSB}_{8\,\mathrm{bit}}$ noise increase), while reducing power and area, by ~4× in CDACs, and 1.32× and 1.23× in SRAM, respectively. Note that offset cancellation is unnecessary as

BS employs the same comparator and CDAC$_n$ as those for BG capture.

The corresponding mixed-signal BS operation for one pixel is illustrated by the timing diagram in Fig. 11(b). After pixel integration, as shown in Fig. 6(b), the signal voltage (V_{sig}) and the reset voltage (V_{dd}) are sequentially read out through the DSF and bottom-plate sampled at CDAC$_n$ and CDAC$_p$, respectively. Meanwhile, D$_{\mathrm{BG}}$[5:0] is loaded into the BG latch. After sampling, CDAC$_n$ is controlled by the BG latch to convert D$_{\mathrm{BG}}$[5:0] into the BG voltage (ΔV_{bg}), which is subtracted from the input signal (ΔV_{sig}). Through thresholding, the 1-bit BS result is obtained from $D_{\mathrm{BS}} = \sim (D_{\mathrm{BS+}} \oplus D_{\mathrm{BS-}})$ with

$$\begin{cases} D_{\mathrm{BS+}} = (\beta_L V_{\mathrm{ref}} > \Delta V_{\mathrm{sig}} - \Delta V_{\mathrm{bg}}) \\ D_{\mathrm{BS-}} = (-\beta_L V_{\mathrm{ref}} > \Delta V_{\mathrm{sig}} - \Delta V_{\mathrm{bg}}). \end{cases} \tag{6}$$

Meanwhile, two extra smaller thresholds ($\pm\beta_S V_{\mathrm{ref}}$) are introduced to provide two BG update bits ($D_{\mathrm{BU\pm}}$) based on the hardware-friendly Σ-Δ BG estimate [37]

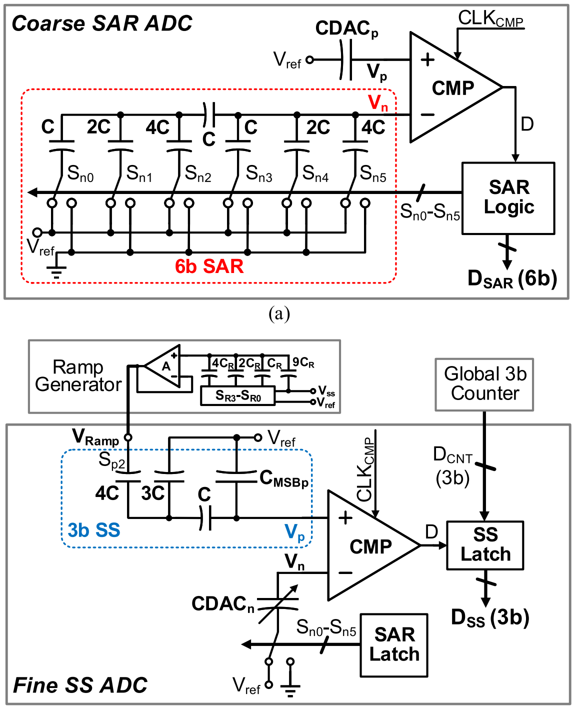
$$D_{\mathrm{BG}}(n+1) = \begin{cases} D_{\mathrm{BG}}(n), & D_{\mathrm{BS}} = 1 | D_{\mathrm{BU\pm}} = 2'b10 \\ D_{\mathrm{BG}}(n) + \Delta, & D_{\mathrm{BU+}} = 0 \\ D_{\mathrm{BG}}(n) - \Delta, & D_{\mathrm{BU-}} = 1. \end{cases} \tag{7}$$

The above model acts as a low-pass filter to adapt BG to slow scene change (e.g., lighting). The update rate is adjustable from zero (no update) to the BS frame rate. The maximum adaptable interframe change rate is ±Δ/frame. Abrupt BG change will continuously generate false motion triggers, requiring BG update. In such a case, the frame rate of FD can be lowered to confirm the BG change. If it is a real BG change, low-frame-rate FD will detect nothing. Otherwise, FD will see the moving object and deny the change. After confirmation by FD, the BG is updated by replacing the detected pixels with new values, which are obtained in the mixed-signal OS way discussed in Section IV-D. The whole process is controlled by the motion analysis block of the vision processor, which is outside the scope of this article, as indicated in Fig. 3.

The transition between $\beta_L V_{\mathrm{ref}}$ and $\beta_S V_{\mathrm{ref}}$ is achieved by switching S$_{p2}$−S$_{p0}$. To minimize the energy overhead, the thresholds are sequentially arranged as: $-\beta_S V_{\mathrm{ref}}$, $-\beta_L V_{\mathrm{ref}}$, $+\beta_L V_{\mathrm{ref}}$, and $+\beta_S V_{\mathrm{ref}}$ such that each BS operation only needs to charge V_{TH} from V_{ss} to V_{dd} once and uses minimum S$_{p2}$-S$_{p0}$ switching.

D. SAR-SS ADC, FIM, and OS

An 8-bit SAR-SS ADC is reconfigured from an RPE to support both OS and FIM. The coarse SAR ADC and the fine SS ADC resolve the MSBs and the LSBs, respectively, to achieve a good balance between area and power for CMOS image sensors [7], [9], [38]. The configuration of the SAR-SS ADC is shown in Fig. 12. In the SAR ADC, the BG latch in Fig. 11 is reused to form the optimum SAR logic [39], which controls the CDAC$_n$ and the comparator for the 6-bit SAR quantization when CDAC$_p$ holds the reference voltage. The SS ADC quantizes the residue after the SAR operation to 3 bits with one extra bit for error correction [38]. With CDAC$_n$ retaining the SAR voltage, CDAC$_p$ couples the ramping voltage (V_{Ramp}) for comparison. To balance between the voltage swing, and the noise and offset requirements of the ramp generator, only the 4C of C_{LSBp} is used to couple V_{Ramp}, resulting in a moderate voltage swing (225 mV) and a relaxed total noise and offset requirement (<18.8 mV) at a low supply voltage (1.2 V). Both the counter and the ramp generator are shared by all the columns to maintain a small column area. Implemented with a 3-bit capacitive DAC, the ramp generator consumes less power than the resistive ladder type [38] and achieves higher accuracy than the current-mode type [7], which requires calibration between the ramp slope and the clock speed. Each column ADC operates at 143k sample/s and consumes 10.8 pJ/conversion, excluding the power from the ramp generator. As dynamic pixel readout does not allow direct access to the ADCs, measurement of differential nonlinearity (DNL) and integral nonlinearity (INL) is unavailable. Post-layout Monte Carlo simulation has shown that the worst DNL and INL are 0.52 LSB$_{8\ \mathrm{bit}}$ and 0.88 LSB$_{8\ \mathrm{bit}}$, respectively, which are mainly contributed by the parasitics at the top plates of C_{LSBs}.



(a)

(b)

Fig. 12. SAR-SS ADC configuration. (a) 6-bit coarse SAR ADC. (b) 3-bit fine SS ADC.

The operation of FIM is illustrated in Fig. 13. After the pixel-level operation shown in Fig. 6(c), V_g and V_{rst} are read out and bottom-plate sampled as that in BS. The SAR phase takes six comparison cycles to search the MSBs (D_{SAR}). In the SS phase, S$_{p2}$ connects the 4C of C_{LSBp} to V_{Ramp}, which increases from (1−1/16) to (1+5/16)V$_{\mathrm{ref}}$ with an increment in (1/16)V$_{\mathrm{ref}}$ to quantize the residue to 3 LSBs (D_{SS}). The final 8-bit ADC result is obtained as follows:

$$D_{\mathrm{ADC}} = D_{\mathrm{SAR}} \times 2^2 + D_{\mathrm{SS}} - 2. \tag{8}$$

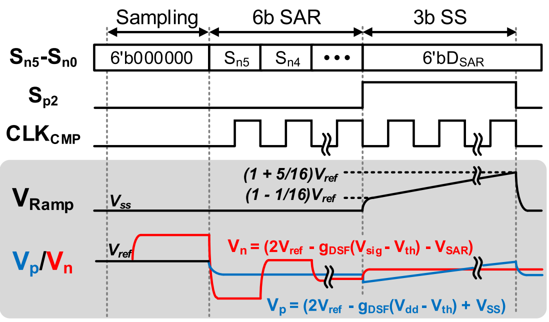


Fig. 13. Operation timing diagram of FIM.

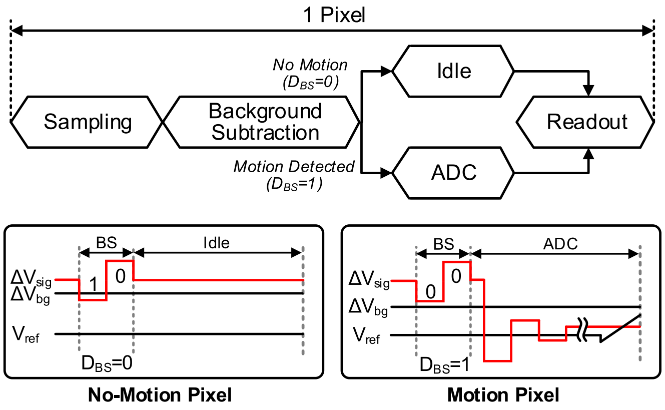


Fig. 14. Operation timing diagram of OS.

The mixed-signal OS is realized by combining the BS and the SAR-SS ADC, as illustrated in Fig. 14. For each pixel, BS is performed after double-sampling to obtain the motion detection result (D_{BS}), as shown in Fig. 11. Upon detection of a motion ($D_{\mathrm{BS}} = 1$), ΔV_{bg} is first removed, followed by the selected ADC quantization of the corresponding signal (ΔV_{sig}). Otherwise ($D_{\mathrm{BS}} = 0$), the column circuits remain idle by gating the global controls and zeroing the pixel value. In this way, the OS image is directly generated after ADC.

V. Experimental Results

The proposed multi-mode vision sensor has been fabricated in a 0.18-µm CMOS process. The chip microphotograph is shown in Fig. 15, where the core and chip areas are 2.43 × 1.96 mm^2 and 2.95 × 2.50 mm^2, respectively. With the compact architecture, the RPEs only occupy 11% of the core area. Estimated from the memory compiler of the same process, the SRAM for BG storage (128 × 108 × 6 bits) consumes an extra area of 1.74 × 0.36 mm^2 (13% core area increase). The chip performance is summarized in Table II. The 7.9-µm pixel pitch is mainly limited by the design rules of the MIM capacitor in this process. Without comparator offset cancellation, the sensor exhibits a fixed-pattern noise (FPN) of 1.34%$_{\mathrm{rms}}$, which can be suppressed to 0.26%$_{\mathrm{rms}}$ after dark frame subtraction. The temporal noise level is at 0.55%$_{\mathrm{rms}}$, which is sufficient for typical vision analysis that requires PSNR > 30 dB (noise < 3.1%$_{\mathrm{rms}}$) [40]. Experimental results show that setting the threshold at 2 LSB$_{6\ \mathrm{bit}}$(3.2%) is sufficient to filter out both the temporal noise and FPN for MS.

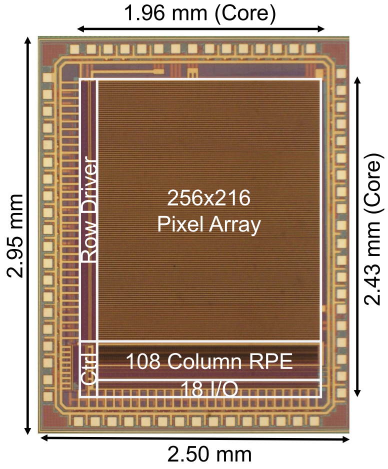


Fig. 15. Microphotograph of the prototype CMOS vision sensor.

TABLE II
Chip Performance Summary

Process	0.18 µm 1P6M CMOS				
Supply	1.2 V (analog) / 0.8 V (digital)				
Core size	2.43 mm × 1.96 mm				
Pixel size	7.9 µm × 7.9 µm				
Fill factor	33.7%				
Pixel array	256×216				
Sensitivity	1.6 V/lux•s				
FPN	1.34%$_{\mathrm{rms}}$ (w/o dark sub.) / 0.26%$_{\mathrm{rms}}$ (w/ dark sub.)				
Temporal noise	0.55%$_{\mathrm{rms}}$				
Dynamic range	45.2 dB				
Frame rate	15 fps (max: FD 672, BS 431, OS 60, FIM 93)				
Operation modes	FD	BS	Cooperative (14FD+1BS)	OS	FIM
Power* (µW @15 fps)	2.20	3.16	2.26	16.08~25.05	21.14
Energy/Frame (µJ)	0.15	0.21	0.15	1.07~1.67	1.41
FoM (pJ/pixel·frame)	10.6	15.2	10.9	19.4~30.2	25.5

* Only sensor power, SRAM power not included in BS and OS.

Fig. 16 shows the sample images at different operation modes taken by the prototype CMOS vision sensor along with transitions. As shown in Fig. 16(a) and (b), the sensor can capture 1-bit MS images by FD or BS at a lower resolution (128 × 108). The FD image can successfully record motion at the object edge regions, while the BS image can capture the whole object silhouette, demonstrating their respective specialities. With $\Delta = 1$ LSB$_{6\ \mathrm{bit}}$, the $\Sigma - \Delta$ BG update is performed every 10 s (or even longer) due to the stable indoor lighting in our experiments. Fig. 16(c) shows the 8-bit BS-selected OS image after motion triggering, where the object is completely extracted from the BG. The OS image size can be cropped to match the object size based on D_{BS} such that only necessary amount of data is passed to the vision processor for further processing. The 8-bit FIM image in Fig. 16(d) can provide complete information for scene recording if object of interest is detected. As studied in [41], the extent of rolling shutter and blurring effects is measured or quantized in terms

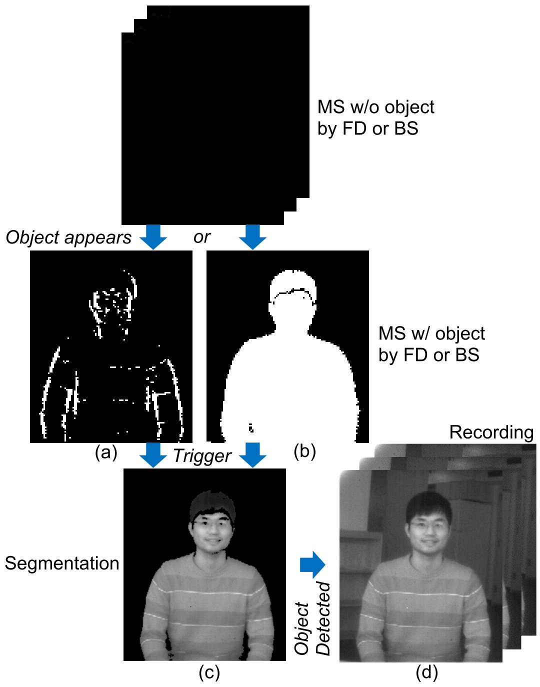


Fig. 16. Sample images. (a) FD image (1 bit, 128 × 108). (b) BS image (1 bit, 128 × 108). (c) OS image (8 bit, 256 × 216). (d) FIM image (8 bit, 256 × 216).

of the exposure time and the readout time of the vision sensor, which are 33.3 ms and 21 μs, respectively, in this design.

Similar to [21], a moving object with an adjustable size and speed is displayed on the screen to characterize the FD and BS capabilities of the proposed CMOS vision sensor. Fig. 17(a) illustrates the measurement setup and the captured FD and BS sample images by the sensor (TH = 2 $\mathrm{LSB}_{6\ \mathrm{bit}}$ and $T_{\mathrm{int}} = 33.3$ ms). FD and BS with different frame rates are tested by changing the sensor frame rate. All the detected motion events are summed up and normalized by the object size to be the performance index. Fig. 17(b) shows the influence of object speed on the total events. As expected, FD is sensitive to the object speed variation, since the total FD events are proportional to the object speed. Reducing the FD frame rate can increase the object displacement to boost the total number of FD events, but at an expense of reduced system response time. However, the detection performance is still limited in the slow-motion region. In contrast, BS exhibits stable total events at different object speeds and frame rates, as discussed in Section II. The more than 100% detection at higher speeds is due to motion blurring. BS maintains a large total event number for slow motion and, therefore, can compensate for the low sensitivity of FD. Fig. 17(c) shows the relation between the total events and the object size. As the object size increases, the total events of both FD and BS also increase proportionally. At small object sizes, BS still shows sufficient total events, while high-frame-rate FD becomes incapable of detecting the object. Low-frame-rate FD can mitigate the issue but is less effective than the BS.

As shown in Fig. 18, motion detection maps are generated with a threshold of ten events to compare detection performance among FD, BS, and cooperative MS. The FD map shows that high-rate FD fails to detect slow and small objects. In contrast, low-frame-rate BS succeeds in detecting slow and small objects, but misses the high-speed objects due to high latency. Here, assume that the maximum allowable moving distance is 54 pixels for 1-s latency. Finally, the cooperative MS scheme can provide accurate motion detection, for fast and large objects with high-frame-rate FD and for slow and small objects using low-frame-rate BS.

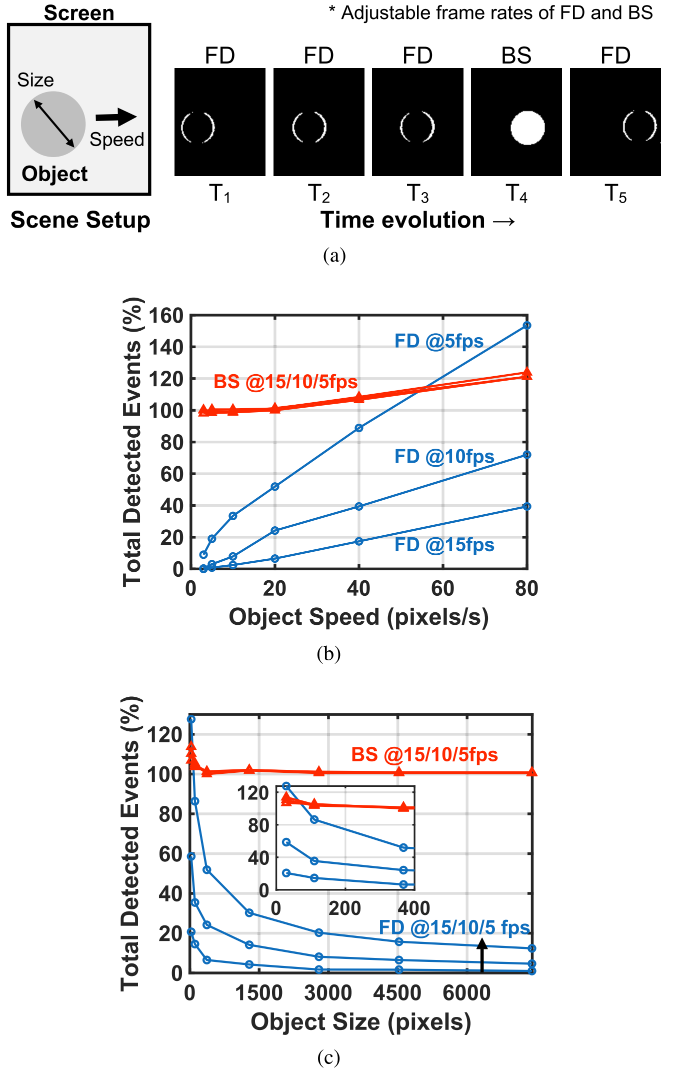
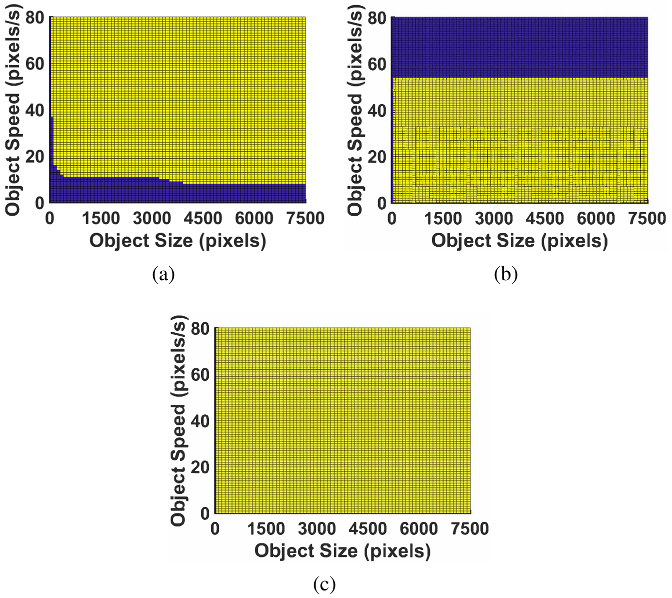


Fig. 17. MS characterization. (a) Scene setup and sample motion images. (b) Total detected events versus object speed (object size = 368 pixels). (c) Total detected events versus object size (object speed = 20 pixels/s).

The power consumptions of the prototype CMOS vision sensor (including 0.43 pJ/bit from IO channels) at different operation modes are summarized in Table II. They are measured under 1.2 $\mathrm{V}_{\mathrm{analog}}$ and 0.8 $\mathrm{V}_{\mathrm{digital}}$ at 15 frames/s. The maximum frame rates are mainly limited by the I/O speed of 8 Mb/s/channel. Fig. 19 shows the analog power improvement by employing the DSF at different modes, demonstrating 6.7×, 2.3×, and 2.0× improvements for FD, BS, and FIM when compared with using the SSF, respectively. FD saves more power than the others, because the pixel readout remains active during processing. Excluding the SRAM power, FD and BS consume only 150 and 210 nJ/frame, respectively.



(a)

(b)

(c)

Fig. 18. Motion detection maps (yellow: pass; blue: fail). (a) FD map (15 frames/s). (b) BS map (1 frames/s). (c) Cooperative MS map (14 FD + 1 BS).

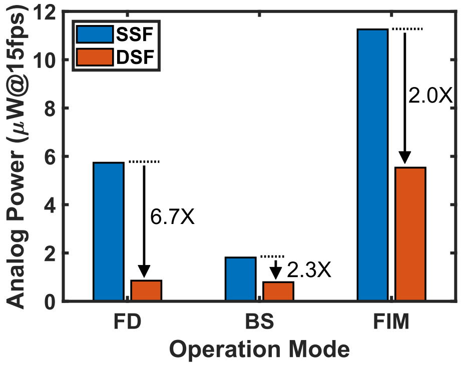


Fig. 19. Analog power comparison between the DSF and the SSF.

The power of cooperative MS linearly depends on the frame rate of BS, which is adjustable based on the maximum allowable sensing latency for slow and small objects. Taking a combination of 14 FD frames and 1 BS frame (1 frame/s) as an example, the cooperative MS consumes only 2.26 μW. By using the memory compiler, the SRAM power for BS is estimated to be 6.6 pJ/pixel (9.1% leakage) for a 6-bit D_{BG} under 0.8-V and 144-Mb/s read/write rate. Considering the SRAM power, the BS power is increased to 302 nJ/frame, resulting in 2.36 μW for the cooperative MS, corresponding to merely 7% increase when compared with the FD-only MS. Fig. 20 compares the power consumptions of the proposed mixed-signal BS and the conventional digital BS using the same bit depth. The digital architecture includes the powers of imaging, digital BS, and SRAM operation, while the proposed mixed-signal design only includes the SRAM power and the mixed-signal BS power. The SRAM power is the same for both architectures, since they have the same memory operation. However, the overall power is reduced by 1.96× for the mixed-signal BS architecture. As for OS, the power is linearly proportional to the object occupancy,

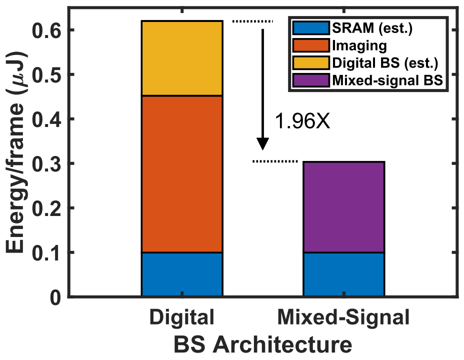


Fig. 20. Power comparison between the digital and mixed-signal BS architectures (estimated SRAM and digital BS powers).

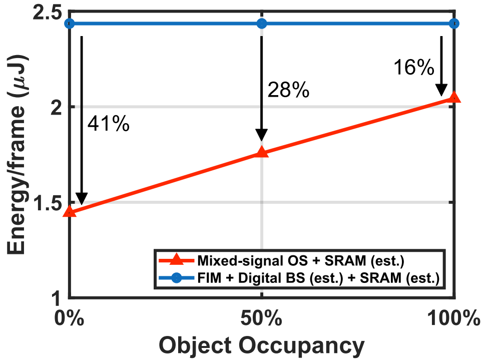


Fig. 21. Power comparison between the mixed-signal OS and the digital OS at different object occupancies (estimated SRAM and digital BS powers).

as shown in Fig. 21. Together with the estimated SRAM power, the energy consumed per frame is increased linearly from 1.44 to 2.04 μJ, as the occupancy increases from 0% to 100%, corresponding to an energy consumption of 10.8 pJ for each ADC operation. With the energy-efficient processing and reduced ADC operation, the proposed mixed-signal OS (including SRAM) achieves 16%∼41% power improvement in comparison with the digital OS, which consists of the FIM, digital BS, and SRAM operation.

Table III summarizes the performance comparison with the state-of-the-art multi-mode vision sensors with in-sensor MS. The MS mode enables the sensors to operate at a much lower power consumption than a conventional image sensor (2170 pJ/pixel·frame) [42], which features high speed and high resolution. Although digital designs [17], [18] can integrate both FD and BS for robust MS, they suffer from large power consumption. Prior mixed-signal designs [16], [21], [23], [24] demonstrate much lower MS power, but they do not have good sensing robustness due to the reliance on one single MS technique. The proposed CMOS vision sensor demonstrates the first mixed-signal design to support both FD and BS for robust motion detection. In BS, employing SRAM to store BG avoids the large pixel size and the CMOS-incompatible process in the analog storage methods [23], [24] and provides flexibility to use other embedded memories such as MRAM and RRAM. With the first fully dynamic processing architectures, FD and BS achieve 1.3× and 120.1× higher energy efficiencies than the state of the art, respectively. By combining high-frame-rate FD and low-frame-rate BS, the cooperative scheme can

TABLE III

PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART MULTI-MODE VISION SENSORS WITH IN-SENSOR MS

Reference	This work			G. Kim ISSCC'13 [21]	J. Choi JSSC'14 [16]	N. Cottini JSSC'13 [23]	T. Ohmaru JSSC'16 [24]	O. Kumagai ISSCC'18 [17]	K. Choo ISSCC'19 [18]
Process	0.18 μm 1P6M CMOS			0.13 μm 1P8M CMOS	0.18 μm 1P4M CMOS	0.35 μm 2P3M CMOS	0.5 μm CAAC-IGZO FET+ 0.18 μm Si-FET	90 nm CIS+ 40 nm CMOS	65 nm CIS
Supply A/D (V)	1.2/0.8			1.2/0.6	1.3/0.8	3.3	3.3/1.8	1.8/1.8/1.0	-
Core size (mm^2)	2.43×1.96			1.5×1.6	2.35×3.18	/	6.5×6.5	5.0×4.4	4.3×3.8
Pixel size (μm^2)	7.9×7.9			6.4×6.4	5.9×5.9	26×26	20×20	1.5×1.5	1.5×1.5
Fill factor (%)	33.7			38	30	12	27.5	100	
Pixel array (MS resolution)	256×216 (128×108)			128×128 (48×16)	256×256 (128×128)	64×64	160×240 (160×1)	2560×1536 (16×5)	792×528 (32×20)
FPN$_{rms}$	1.34% (w/o dark sub.) 0.26% (w/ dark sub.)			2.3%	0.05%	/	2.4% (w/o CDS) 0.29% (w/ CDS)	/	/
Dynamic range (dB)	45.2			38.5	54.8	52	43.8	96	64.3
Frame rate (fps)	15			5/19	15	13	60	10/60	170/5.6
MS technique	FD	BS	Cooperative 14FD+1BS	FD	FD	BS	BS	FD+BS	FD
Processing domain	Mixed			Mixed	Mixed	Mixed	Mixed	Digital	Digital
MS power (μW)	2.20	4.53*	2.36*	0.47 @5fps	3.31	33	25.3	1100 @10fps	288 @170fps
MS FoM$^{\#}$ (pJ/pixel·frame)	**10.6**	**21.8***	**11.4***	121.6	13.5	619.7	2635	1375000	2650
Imaging power (μW)	OS		FIM	29 @19fps	51.06	/	3600	95000 @60fps	392 @5.6fps
	21.56~30.53*		21.14						
Imaging FoM$^{\#}$ (pJ/pixel·frame)	**26.0~36.8***		**25.5**	93.2	51.9	/	1562.5	403	167.5

* SRAM power (estimated) included. $^{\#}$ FoM=Power/(Pixel number•Frame rate)

improve the MS robustness while still maintaining the highest FoM of 11.4 pJ/pixel·frame. Furthermore, the proposed sensor is also capable of mixed-signal OS to facilitate object analysis and reduce power and bandwidth, which is not found in other designs. In addition to the low MS power, the proposed sensor also achieves the highest imaging FoM of 25.5 pJ/pixel·frame, which is comparable with those ultra-low-power imaging-only sensors [5], [6]. Overall, the proposed CMOS vision sensor features extended processing capabilities with the highest energy efficiencies, thanks to the highly reconfigurable and fully dynamic mixed-signal processing architectures.

VI. CONCLUSION

A low-power multi-mode CMOS vision sensor is presented for energy-constrained WSNs. It is reconfigurable for MS, OS, and FIM to support scene-adaptive edge computing for power and bandwidth optimization. High-frame-rate FD and low-frame-rate BS are combined as a cooperative scheme to improve the MS robustness while maintaining low power consumption. With the fully dynamic mixed-signal processing architectures, the cooperative MS composed of 14 FD frames and 1 BS frame consumes only 2.36 μW at 15 frames/s, achieving a state-of-the-art FoM of 11.4 pJ/pixel·frame. The mixed-signal OS is also enabled for the first time by selecting analog-to-digital conversion based on the BS results. With a linear dependence on the object occupancy, the power consumption of OS ranges from 1.44 to 2.04 μJ/frame, corresponding to 41%∼16% power improvement when compared with the conventional digital method. FIM consumes only 1.41 μJ/frame, which is comparable with the start-of-the-art low-power CMOS image sensors. With the multi-mode capabilities, compact area, and low power, the proposed CMOS vision sensor is suitable for wireless IoT applications.

REFERENCES

[1] T. Karnik *et al.*, "A cm-scale self-powered intelligent and secure IoT edge mote featuring an ultra-low-power SoC in 14nm tri-gate CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2018, pp. 46–48.

[2] S. Paul *et al.*, "A sub-cm^3 energy-harvesting stacked wireless sensor node featuring a near-threshold voltage IA-32 microcontroller in 14-nm Tri-Gate CMOS for always-on always-sensing applications," *IEEE J. Solid-State Circuits*, vol. 52, no. 4, pp. 961–971, Apr. 2017.

[3] Q. Wan and P. K. T. Mok, "A 14 nA quiescent current inductorless dual-input-triple-output thermoelectric energy harvesting system based on a reconfigurable TEG array," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Apr. 2018, pp. 1–4.

[4] X. Meng, X. Li, Y. Yao, C. Tsui, and W. Ki, "An indoor solar energy harvester with ultra-low-power reconfigurable power-on-reset-styled voltage detector," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2018, pp. 1–5.

[5] N. Couniot, G. de Streel, F. Botman, A. K. Lusala, D. Flandre, and D. Bol, "A 65 nm 0.5 V DPS CMOS image sensor with 17 pJ/frame.pixel and 42 dB dynamic range for ultra-low-power SoCs," *IEEE J. Solid-State Circuits*, vol. 50, no. 10, pp. 2419–2430, Oct. 2015.

[6] A. Y. Chiou and C. Hsieh, "An ULV PWM CMOS imager with adaptive-multiple-sampling linear response, HDR imaging, and energy harvesting," *IEEE J. Solid-State Circuits*, vol. 54, no. 1, pp. 298–306, Jan. 2019.

[7] D. G. Chen, F. Tang, M. K. Law, and A. Bermak, "A 12 pJ/pixel analog-to-information converter based 816 × 640 pixel CMOS image sensor," *IEEE J. Solid-State Circuits*, vol. 49, no. 5, pp. 1210–1222, May 2014.

[8] H. Zhu, M. Zhang, Y. Suo, T. D. Tran, and J. V. der Spiegel, "Design of a digital address-event triggered compressive acquisition image sensor," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 63, no. 2, pp. 191–199, Feb. 2016.

[9] X. Zhong, B. Zhang, A. Bermak, C. Y. Tsui, and M. K. Law, "A low-power compression-based CMOS image sensor with microshift-guided SAR ADC," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 65, no. 10, pp. 1350–1354, Oct. 2018.

[10] S. Chen, W. Tang, X. Zhang, and E. Culurciello, "A 64 × 64 pixels UWB wireless temporal-difference digital image sensor," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 20, no. 12, pp. 2232–2240, Dec. 2012.

[11] G. de Streel *et al.*, "Sleeptalker: A ULV 802.15.4a IR-UWB transmitter SoC in 28-nm FDSOI achieving 14 pJ/b at 27 Mb/s with channel selection based on adaptive FBB and digitally programmable pulse shaping," *IEEE J. Solid-State Circuits*, vol. 52, no. 4, pp. 1163–1177, Apr. 2017.

[12] K. Bong, S. Choi, C. Kim, D. Han, and H. Yoo, "A low-power convolutional neural network face recognition processor and a CIS integrated with always-on face detector," *IEEE J. Solid-State Circuits*, vol. 53, no. 1, pp. 115–123, Jan. 2018.

[13] H. Kim, S. Hwang, J. Chung, J. Park, and S. Ryu, "A dual-imaging speed-enhanced CMOS image sensor for real-time edge image extraction," *IEEE J. Solid-State Circuits*, vol. 52, no. 9, pp. 2488–2497, Sep. 2017.

[14] 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*, Jun. 2018, pp. 247–248.

[15] C. Young, A. Omid-Zohoor, P. Lajevardi, and B. Murmann, "A data-compressive 1.5b/2.75b log-gradient QVGA image sensor with multi-scale readout for always-on object detection," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 98–100.

[16] J. Choi, S. Park, J. Cho, and E. Yoon, "A 3.4μW object-adaptive CMOS image sensor with embedded feature extraction algorithm for motion-triggered object-of-interest imaging," *IEEE J. Solid-State Circuits*, vol. 49, no. 1, pp. 289–300, Jan. 2014.

[17] O. Kumagai *et al.*, "A 1/4-inch 3.9Mpixel low-power event-driven back-illuminated stacked CMOS image sensor," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2018, pp. 86–88.

[18] K. D. Choo *et al.*, "Energy-efficient low-noise CMOS image sensor with capacitor array-assisted charge-injection SAR ADC for motion-triggered low-power IoT applications," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 96–98.

[19] Y. M. Chi, U. Mallik, M. A. Clapp, E. Choi, G. Cauwenberghs, and R. Etienne-Cummings, "CMOS camera with in-pixel temporal change detection and ADC," *IEEE J. Solid-State Circuits*, vol. 42, no. 10, pp. 2187–2196, Oct. 2007.

[20] B. Zhao, X. Zhang, S. Chen, K. Low, and H. Zhuang, "A 64 × 64 CMOS image sensor with on-chip moving object detection and localization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 4, pp. 581–588, Apr. 2012.

[21] G. Kim *et al.*, "A 467nW CMOS visual motion sensor with temporal averaging and pixel aggregation," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2013, pp. 480–481.

[22] X. Liu, M. Zhang, and J. V. der Spiegel, "A low-power multifunctional CMOS sensor node for an electronic facade," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 61, no. 9, pp. 2550–2559, Sep. 2014.

[23] N. Cottini, M. Gottardi, N. Massari, R. Passerone, and Z. Smilansky, "A 33 μW 64 × 64 pixel vision sensor embedding robust dynamic background subtraction for event detection and scene interpretation," *IEEE J. Solid-State Circuits*, vol. 48, no. 3, pp. 850–863, Mar. 2013.

[24] T. Ohmaru *et al.*, "A 25.3 μW at 60 fps 240×160 pixel vision sensor for motion capturing with in-pixel nonvolatile analog memory using CAACIGZO FET," *IEEE J. Solid-State Circuits*, vol. 51, no. 9, pp. 2168–2179, Sep. 2016.

[25] Y. Zou, M. Gottardi, D. Perenzoni, M. Perenzoni, and D. Stoppa, "A 1.6 mW 320×240-pixel vision sensor with programmable dynamic background rejection and motion detection," in *Proc. IEEE SENSORS*, Oct. 2017, pp. 1–3.

[26] R. T. Collins *et al.*, "A system for video surveillance and monitoring," Robot. Inst., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-RI-TR-00-12, 2000.

[27] A. Rabinovich, A. Vedaldi, and S. J. Belongie, "Does image segmentation improve object categorization?" Dept. Comput. Sci. Eng., Univ. California, San Diego, La Jolla, CA, USA, Tech. Rep., 2007.

[28] J. Choi, S. Han, S. Kim, S. Chang, and E. Yoon, "A spatial-temporal multiresolution CMOS image sensor with adaptive frame rates for tracking the moving objects in region-of-interest and suppressing motion blur," *IEEE J. Solid-State Circuits*, vol. 42, no. 12, pp. 2978–2989, Dec. 2007.

[29] X. Zhong, B. Zhang, and A. Bermak, "A background subtraction based column-parallel analog-to-information converter for motion-triggered vision sensor," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2016, pp. 1426–1429.

[30] S. Sugawa, N. Akahane, S. Adachi, K. Mori, T. Ishiuchi, and K. Mizobuchi, "A 100 dB dynamic range CMOS image sensor using a lateral overflow integration capacitor," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2005, pp. 352–603.

[31] K. Kagawa, S. Shishido, M. Nunoshita, and J. Ohta, "A 3.6pW/frame·pixel 1.35 V PWM CMOS imager with dynamic pixel readout and no static bias current," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2008, pp. 54–595.

[32] S. Ji, J. Pu, B. C. Lim, and M. Horowitz, "A 220pJ/pixel/frame CMOS image sensor with partial settling readout architecture," in *Proc. IEEE Symp. VLSI Circuits*, Jun. 2016, pp. 1–2.

[33] D. G. Chen, F. Tang, M. Law, X. Zhong, and A. Bermak, "A 64 fJ/step 9-bit SAR ADC array with forward error correction and mixed-signal CDS for CMOS image sensors," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 61, no. 11, pp. 3085–3093, Nov. 2014.

[34] M. van Elzakker, E. van Tuijl, P. Geraedts, D. Schinkel, E. A. M. Klumperink, and B. Nauta, "A 10-bit charge-redistribution ADC consuming 1.9 μW at 1 MS/s," *IEEE J. Solid-State Circuits*, vol. 45, no. 5, pp. 1007–1015, May 2010.

[35] D. Schinkel, E. Mensink, E. Klumperink, E. van Tuijl, and B. Nauta, "A double-tail latch-type voltage sense amplifier with 18ps setup+hold time," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2007, pp. 314–605.

[36] C. Dupoiron, A. Verdant, and G. Sicard, "Trade-off between the number of bits per pixel and motion detection quality for a low power image sensor," *Electron. Imag.*, vol. 2016, no. 12, pp. 1–6, 2016.

[37] A. Manzanera and J. C. Richefeu, "A new motion detection algorithm based on Σ–Δ background estimation," *Pattern Recognit. Lett.*, vol. 28, no. 3, pp. 320–328, 2007.

[38] M. K. Kim, S. K. Hong, and O. K. Kwon, "An area-efficient and low-power 12-b SAR/single-slope ADC without calibration method for CMOS image sensors," *IEEE Trans. Electron. Devices*, vol. 63, no. 9, pp. 3599–3604, Sep. 2016.

[39] T. O. Anderson, "Optimum control logic for successive approximation analog-to-digital converters," *Deep Space Netw. Prog. Rep.*, vol. 13, pp. 168–176, Oct. 1972.

[40] S. Dodge and L. Karam, "Understanding how image quality affects deep neural networks," in *Proc. IEEE Int. Conf. Quality Multimedia Exper. (QoMEX)*, Jun. 2016, pp. 1–6.

[41] C. Liang, L. Chang, and H. H. Chen, "Analysis and compensation of rolling shutter effect," *IEEE Trans. Image Process.*, vol. 17, no. 8, pp. 1323–1330, Aug. 2008.

[42] T. Takahashi *et al.*, "A 4.1Mpix 280fps stacked CMOS image sensor with array-parallel ADC architecture for region control," in *Proc. Symp. VLSI Circuits*, Jun. 2017, pp. C244–C245.

Xiaopeng Zhong (Student Member, IEEE) received the B.E. degree (Hons.) from Zhejiang University (ZJU), Hangzhou, China, in 2013, and the Ph.D. degree from the Hong Kong University of Science and Technology (HKUST), Hong Kong, in 2019.

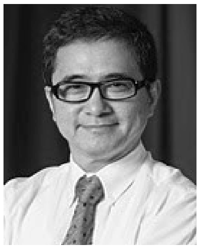
He was a Research Assistant with the Smart Sensory Integrated System (S2IS) Laboratory, Integrated Circuit Design Center (ICDC), HKUST, from 2013 to 2019. He was also a Research Fellow with the Prof. Chandrakasan's Group, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, from 2018 to 2019. Since November 2019, he has been a Senior Engineer with Qualcomm Technologies, Inc., San Diego, CA, USA.

Dr. Zhong was a recipient of the Hong Kong Ph.D. Fellowship Scheme (HKPFS) and HKUST Overseas Research Award.

Man-Kay Law (Senior Member, IEEE) received the B.Sc. degree in computer engineering and the Ph.D. degree in electronic and computer engineering from The Hong Kong University of Science and Technology (HKUST), Hong Kong, in 2006 and 2011, respectively.

In February 2011, he joined HKUST as a Visiting Assistant Professor. He is currently an Associate Professor with the State Key Laboratory of Analog and Mixed-Signal VLSI, Institute of Microelectronics and Faculty of Science and Technology, University of Macau, Macau, China. He has developed ultra-low power CMOS temperature/image sensing systems, as well as fully integrated high-efficiency solar/ultrasound energy harvesting solutions for implantable applications. He has authored or coauthored over 80 technical publications and holds six U.S./Chinese patents. His research interests are on the development of ultra-low power CMOS sensing/readout circuits and energy harvesting techniques for wireless and biomedical applications.

Dr. Law was a member of the Technical Program Committee of the International Solid-State Circuit Conference (ISSCC) and a Review Committee Member of the IEEE International Symposium on Circuits and Systems (ISCAS), the Biomedical Circuits and Systems Conference (BioCAS), and the International Symposium on Integrated Circuits (ISIC). He serves as a Technical Committee Member of the IEEE CAS Committee on Sensory Systems, as well as Biomedical and Life Science Circuits and Systems. He is currently a Technical Program Committee Member of ISSCC, as well as a Distinguished Lecturer of the IEEE CASS. He was a co-recipient of the ASQED Best Paper Award in 2013, the A-SSCC Distinguished Design Award in 2015, the ASPDAC Best Design Award in 2016, and the ISSCC Silkroad Award in 2016. He also received the Macao Science and Technology Invention Award (Second Class) by Macau Government—FDCT, in 2014 and 2018. He was the University Design Contest Co-Chair of the Asia and South Pacific Design Automation Conference (ASP-DAC) and the Asia Symposium on Quality Electronic Design (ASQED).



Chi-Ying Tsui (Senior Member, IEEE) received the B.S. degree in electrical engineering from The University of Hong Kong, Hong Kong, and the Ph.D. degree in computer engineering from the University of Southern California, Los Angeles, CA, USA, in 1994.

He joined the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, in 1994, where he is currently a Full Professor and also the Head of the Division of Integrative Systems and Design. He has authored over 250 refereed publications and holds 11 U.S. patents on power management, VLSI, and multimedia systems. His current research interests include designing very large-scale integration (VLSI) architectures for low-power wireless and artificial intelligence applications and developing energy harvesting and power management circuits and techniques for ultra-low-power embedded systems.

Dr. Tsui received the Best Paper Awards from the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS in 1995, the IEEE International Symposium on Circuits and Systems in 1999, the IEEE/Association for Computing Machinery (ACM) International Symposium on Low Power Electronics and Design (ISLPED) in 2007, and the IEEE DELTA in 2008 and CODES/ISSS in 2012. He also received the Design Awards from the IEEE ASP-DAC University Design Contest in 2004 and 2006, respectively. He was the General Chair of the 19th IFIP/IEEE VLSI-SOC 2011. He is an Associate Editor of *Integration*, the International VLSI Journal, and the IEEE TRANSACTIONS ON MULTI-SCALE COMPUTING SYSTEMS (TMSCS). He was the Co-Founder of Perception Digital Limited, a listed-company in Hong Kong.

Amine Bermak (Fellow, IEEE) received the M.Eng. and Ph.D. degrees in electronic engineering from Paul Sabatier University, Toulouse, France, in 1994 and 1998, respectively.

He was with LAAS-CNRS, French National Research Center, Microsystems and Microstructures Research Group, where he developed a 3-D VLSI chip for artificial neural network classification and detection applications. He joined the Advanced Computer Architecture Research Group, York University, York, U.K., where he held a post-doctoral position on VLSI implementation of CMM neural network for vision applications in a project funded by the British Aerospace. In 1998, he joined Edith Cowan University, Perth, WA, Australia, as a Research Fellow in smart vision sensors, and a Senior Lecturer with the School of Engineering and Mathematics. He served as a Professor with the Electronic and Computer Engineering Department, The Hong Kong University of Science and Technology, Hong Kong, where he also served as the Director of computer engineering and the Director of the M.Sc. degree program in integrated circuit design. He is currently a Professor with the College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar. His research interests include VLSI circuits and systems for signal, image processing, sensors, and microsystems application. He is currently serving on the Editorial Board of the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS and the IEEE TRANSACTIONS ON ELECTRON DEVICES. He is also an Editor of *Scientific Reports* (Nature). He is an IEEE Distinguished Lecturer.