A Mixed-Signal Near-Sensor Convolutional Imager SoC with Charge-Based 4b-Weighted 5-to-84-TOPS/W MAC Operations for Feature Extraction and Region-of-Interest Detection

Martin Lefebvre and David Bol

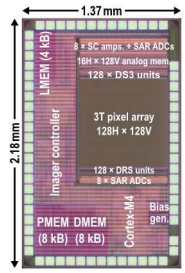
Université catholique de Louvain (UCLouvain), Belgium

The massive deployment of Internet-of-Things (IoT) sensor nodes has fostered the development of machine learning algorithms and hardware at the edge. Regarding vision sensors, edge devices must be able to solve low- to medium-level computer vision tasks, e.g., feature extraction (FE) and region-of-interest (RoI) detection, before transmitting meaningful data to the cloud. This must be achieved within a sub-mW power budget, as IoT sensor nodes are supplied by batteries or energy harvesting. Therefore, mixed-signal vision chips appear as a suitable solution as they outperform digital chips in terms of energy efficiency (EE) while maintaining reasonable processing accuracy. They can be divided into (i) in- [1] and (ii) near-sensor [2,3,4] vision chips, implemented with processing elements (PEs) inside and near the pixel array, and (iii) hybrid vision chips [5,6], combining elements from both schemes. On the one hand, in-sensor chips [1] offer a massive parallelism but suffer from a large pixel pitch due to the pixel-level PEs. They are limited to low-level tasks as they employ 1-to-1.5b filter weights, and exploit raw [1] or amplified [5] photocurrents subject to significant mismatch. In addition, connections between PEs are limited to neighboring pixels, making it difficult to scale processing to image-level features. On the other hand, near-sensor and hybrid chips generally allow to implement medium-level tasks, either by employing large-size Haar-like filters [2,5,6], or by increasing the filter weights resolution [3,4]. They employ conventional pixels such as DPS or 3T/4T APS, and operate at multiple spatial scales using image downsampling (DS) [1] or filter dilation [5]. An analog memory storing pixel values is often required, but entails power and/or area overheads [2]. Nonetheless, existing works fall short of preserving EE while simultaneously supporting low- and medium-level tasks, which require mismatch-compensated inputs, increased weights resolution, and multiscale operation.

In this work, we present a near-sensor convolutional (conv.) imager system-on-chip (SoC) named MANTIS, supporting FE and RoI detection. It embeds a mixed-signal conv. processor implementing charge-based multiply-and-accumulate (MAC) operations with switched-cap (SC) amplifiers. This processor uses pixel voltages computed in DS3 units by delta-reset sampling (DRS) and image DS. The conv. processor and DS3 units are the main contributions of this work as they provide an effective answer to the limitations of existing vision chips. MANTIS' modes of operations are illustrated in Fig. 1: (i) 8b imaging, (ii) FE, producing 1-to-8b output feature maps (fmaps), and (iii) RoI detection, producing a 1b map. These last two modes employ up to 32 programmable 16×16 filters, a filter stride (S) from 2 to 16, and can operate at three spatial scales corresponding to DS factors from 1 to 4×. In addition, the SoC includes a Cortex-M4 with direct memory access for efficient data transfer. The analog core consists of a 3T APS and a conv. pipeline detailed in Fig. 2(a). First, raw pixel voltages are read from the pixel array into column-wise DS3 units, which perform (i) DRS to mitigate fixed-pattern noise (FPN), (ii) DS to support multiscale operation, and (iii) voltage downshifting. The FPN-compensated pixel voltages are then stored in an analog memory and used as inputs for the conv. operations performed by MAC units and SC amplifiers. Finally, the conv. results are digitized by 1-to-8b SAR ADCs and stored in the output registers.

Fig. 2 provides a detailed view of the DS3 units. Raw pixel voltages are obtained by a partial settling readout. During the first phase, the DS3 units sample signal V_{SIG} on C_{SIG} ($\phi_{1,SIG}$=1), and reset V_{RST} on C_{RST} ($\phi_{1,RST}$=1), while the inverter-based OTA is in auto-zero (AZ) (ϕ_1=1). Then, during the second phase (ϕ_2=1), the output voltage V_{PIX} is computed based on an SC structure using the feedback capacitor (cap) C_{FB}, resulting in a voltage proportional to (V_{RST}-V_{SIG}). A voltage range adapted to the analog memory is achieved with V_{REF} = 0.6 V and C_S/C_{FB} = 0.4. Fig. 2(c) presents the image DS principle for a factor 4×. It is implemented by shorting input and output voltages of the OTAs in four adjacent DS3 units, and storing the average of four pixels of the same row on the additional cap C_H of one of the DS3 units. When four rows have been read, all the hold caps are connected together to generate the average of a 4×4 image patch. Finally, pixel voltages are stored in the analog memory.

The charge-based 4b-weighted conv. operation is explained in Fig. 3. First, Fig. 3(a) illustrates the striding of a filter over the analog memory, with connections between the memory and the eight SC amplifiers changing with the filter location, and inactive SC amplifiers being power-gated. When image DS is employed, the storage pattern in Fig. 2(d) increases throughput by enabling simultaneous computations of conv. results corresponding to different filter locations, with slightly more complex routing between DS3 units and analog memory. Furthermore, each conv. operation between an image patch and a filter is computed on row by row (Fig. 3(b)). Each row of the image patch is multiplied element-wise by the corresponding row of filter weights read from the local SRAM (LMEM) and controlling the input caps of the SC amplifier (Fig. 3(c)). Then, this partial sum (psum) is stored on a 16th of the SAR ADC's capacitive DAC. When all 16 psums have been stored in the DAC, it is disconnected from the SC amplifier and psums are averaged by charge sharing on V_{SH}. Fig. 3(c) depicts the operation of the MAC units and SC amplifier. First, the SC amplifier is enabled before any psum calculation. Then, pixel values multiplied by a positive weight are read from the analog memory and stored on input caps $C_{+,i}$ ($\phi_{1,MAC}$=1), while the feedback cap C_{FB} is in reset ($\phi_{1,SC}$=1). Finally, pixel values multiplied by a negative weight are read on input caps $C_{-,j}$ ($\phi_{2,MAC-}$=1) while the psum is computed ($\phi_{2,MAC+}$, $\phi_{2,SC}$=1). This SC structure is insensitive to the OTA offset, resilient to process variations due to its ratiometric nature, and robust against mismatch and noise thanks to the use of 7-fF unitary MoM caps.



Die microphotograph.

The SoC was fabricated in UMC 0.11-µm CMOS. Electrical performance is characterized in Fig. 4(a) in all sensor configurations. First, the fmap RMSE, measured over 10 filters for 10 images of the KODAK dataset, is comprised between 3 and 11.3%, and stems from the discrepancies between ideal and measured fmaps arising from the normalization. Examples of conv. results in Fig. 4(b) highlight a good matching between fmaps. Next, the EE of the MAC operations, normalized to 1b operations, reaches a peak value of 84.1 TOPS/W for S=2 and DS=4. It decreases with S as the ratio between processing and exposure times diminishes, and increases with DS as the number of MACs on the original image raises by 4-16× for DS=2-4. These trends are less visible when considering the SoC EE as they are amortized by the power consumption of the digital part. The power breakdown in several configurations (Fig. 4(c)) indeed reveals that >70% of the power consumption can be attributed to the digital logic, outlining a clear path for further optimizations. Furthermore, Fig. 5 illustrates the training pipeline of the face RoI detector, based on a quantization-aware extension of Keras, and achieving an 8.5-% false negative and 3.1-% false positive rates. The on-chip conv. layer reduces data volume by 12.8× compared to the 8b image, and features an 11.5-% false negative rate while discarding 81.3% of all image patches over 10 test images. Finally, we compare our work to the state of the art of mixed-signal vision chips in Fig. 6. Our design features the lowest 6.04-µm pixel pitch and highest 54-% fill factor, while implementing DRS to mitigate FPN. Our work is the only one to implement large-size 4b-weighted filters while supporting multiscale operation. Despite these additions, the 0.2-to-4.6-TOPS/W SoC EE is on par with existing works.

Acknowledgement:

The authors would like to thank Dr. Rémi Dekimpe for his help with the DMA, Eléonore Masarweh for the chip microphotograph, and ECS group members for their proofreading.

References:

[1] H. Xu *et al.*, "Senputing: An Ultra-Low-Power Always-On Vision Perception Chip Featuring the Deep Fusion of Sensing and Computing", TCAS-I, 2022.
[2] K. Bong *et al.*, "A Low-Power Convolutional Neural Network Face Recognition Processor and a CIS Integrated With Always-on Face Detector", JSSC, 2018.
[3] T.-H. Hsu *et al.*, "A 0.5-V Real-Time Computational CMOS Image Sensor With Programmable Kernel for Feature Extraction", JSSC, 2021.
[4] T.-H. Hsu *et al.*, "A 0.8 V Intelligent Vision Sensor With Tiny Convolutional Neural Network and Programmable Weights Using Mixed-Mode Processing-in-Sensor Technique for Image Classification", JSSC, 2023.
[5] M. Lefebvre *et al.*, "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", ISSCC, 2021.
[6] H. Song *et al.*, "A 5.1ms Low-Latency Face Detection Imager with In-Memory Charge-Domain Computing of Machine-Learning Classifiers", VLSI, 2021.

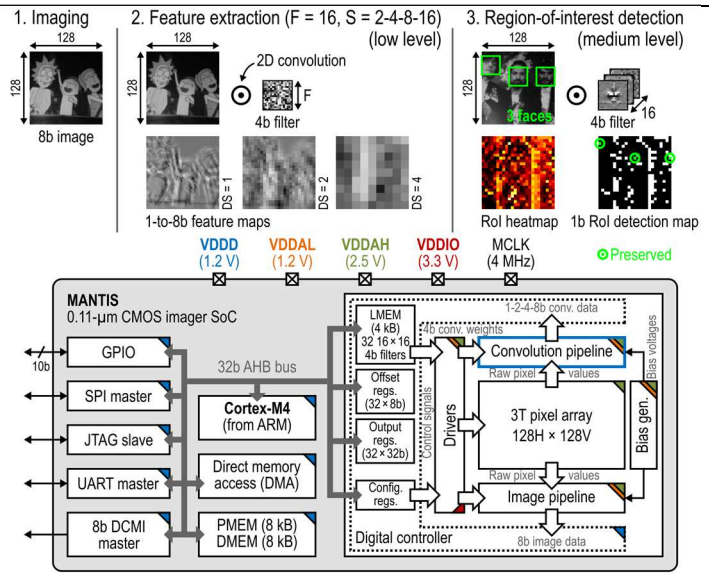


Fig. 1. Imager SoC modes of operation, and block diagram of the digital and image sensor analog cores with their power domains.

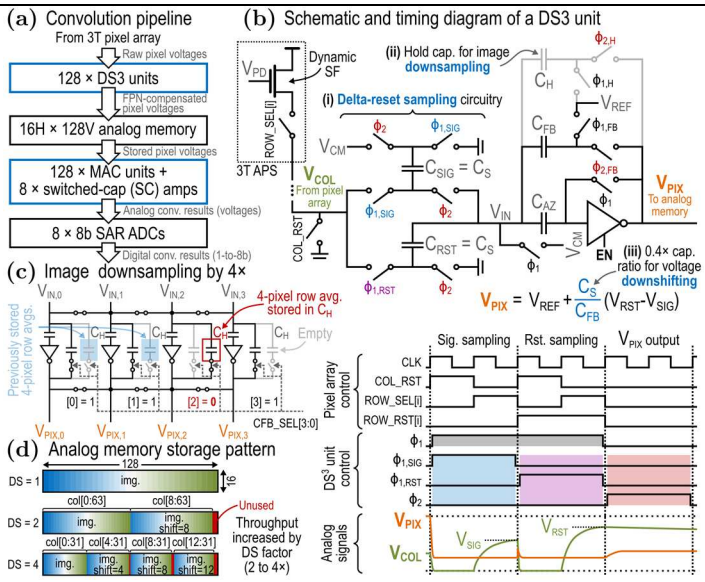


Fig. 2. (a) Block diagram of the conv. pipeline. (b) Schematic and timing of a DS3 unit. (c) Image DS and (d) storage in the memory.

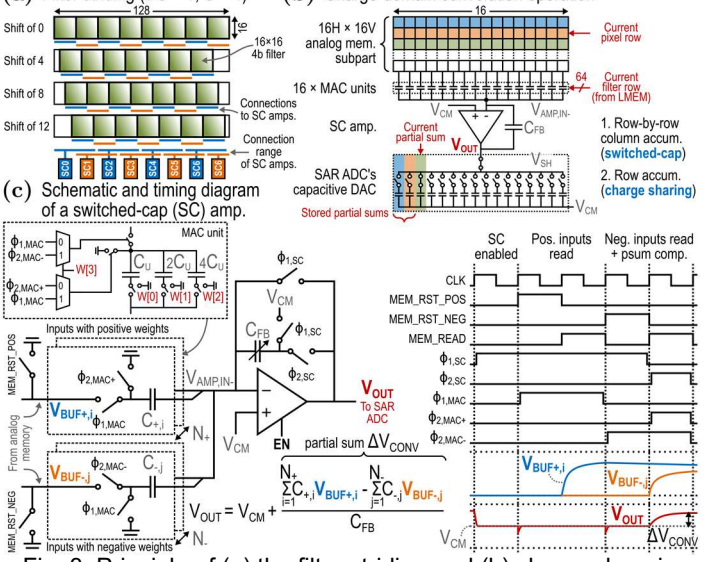
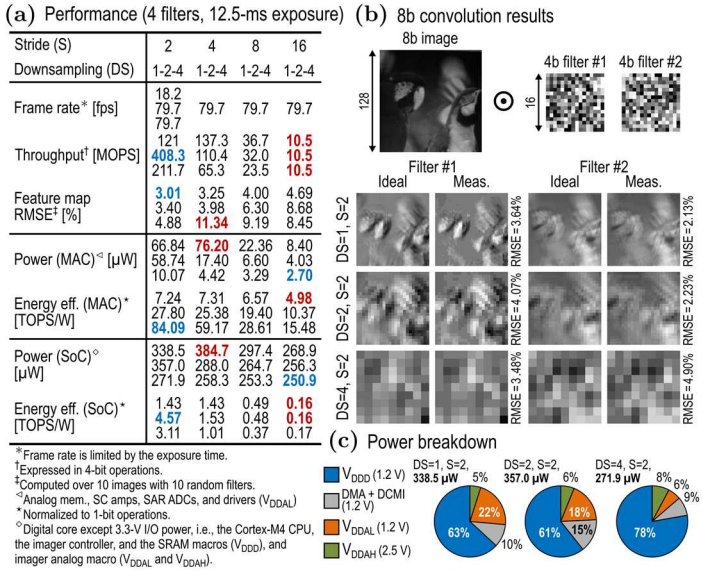


Fig. 3. Principle of (a) the filter striding and (b) charge-domain conv. operation. (c) Schematic and timing of the SC amplifier.



Stride (S)	2	4	8	16
Downsampling (DS)	1-2-4	1-2-4	1-2-4	1-2-4
Frame rate* [fps]	18.2 79.7 79.7	79.7	79.7	79.7
Throughput‡ [MOPS]	121 **408.3** 211.7	137.3 110.4 65.3	36.7 32.0 23.5	**10.5** **10.5** **10.5**
Feature map RMSE‡ [%]	**3.01** 3.40 4.88	3.25 3.98 **11.34**	4.00 6.30 9.19	4.69 8.68 8.45
Power (MAC)◁ [µW]	66.84 58.74 10.07	**76.20** 17.40 4.42	22.36 6.60 3.29	8.40 4.03 **2.70**
Energy eff. (MAC)* [TOPS/W]	7.24 27.80 **84.09**	7.31 25.38 59.17	6.57 19.40 28.61	**4.98** 10.37 15.48
Power (SoC)◇ [µW]	338.5 357.0 271.9	**384.7** 288.0 258.3	297.4 264.7 253.3	268.9 256.3 **250.9**
Energy eff. (SoC)* [TOPS/W]	1.43 **4.57** 3.11	1.43 1.53 1.01	0.49 0.48 0.37	**0.16** **0.16** 0.17

*Frame rate is limited by the exposure time.
†Expressed in 4-bit operations.
‡Computed over 10 images with 10 random filters.
◁Analog mem., SC amps, SAR ADCs, and drivers (V_{DDAL})
*Normalized to 1-bit operations.
◇Digital core except 3.3-V I/O power, i.e., the Cortex-M4 CPU, the imager controller, and the SRAM macros (V_{DDD}), and imager analog macro (V_{DDAL} and V_{DDAH}).

Fig. 4. (a) Measured performance of the image SoC in conv. mode, (b) ideal and measured fmaps, and (c) power breakdown.

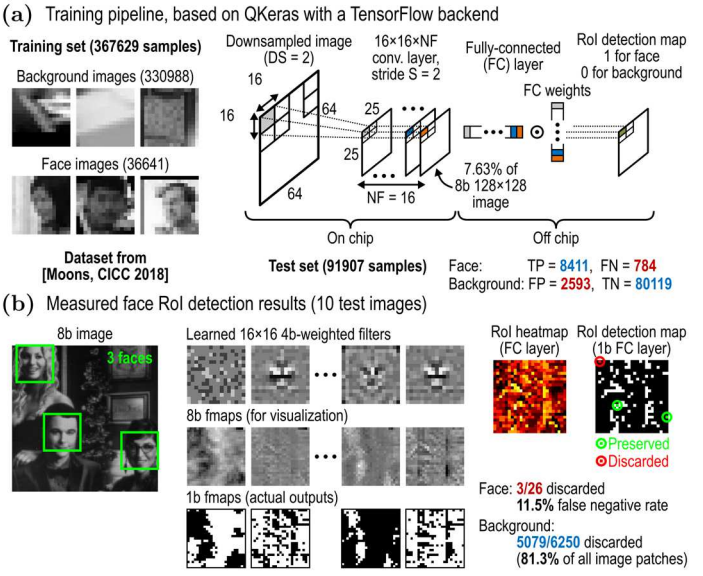


Fig. 5. (a) Training pipeline of the face RoI detector with QKeras. (b) Measured face RoI detection results for the 10 test images.

	In-sensor	Hybrid		Near-sensor			
	Xu TCAS-I 2022	Lefebvre ISSCC 2021	Song VLSI 2021	Bong JSSC 2018	Hsu JSCC 2021	Hsu JSSC 2023	**This work** CICC 2024
Technology	0.18µm CMOS	65nm CMOS	0.18µm CIS	65nm CMOS	0.18µm CMOS	0.18µm CMOS	0.11µm CMOS
Area [mm²]	N/A	2 × 2	4.1 × 5.2	3.3 × 3.6	2.46 × 2	2.46 × 2.18	1.37 × 2.18
Supply voltage [V]	0.8-1.8	0.8/1 (Dig.) 0.95/1.05 (Ana.)	1.8 (Dig.) 2.5 (Ana.)	0.5-0.8 (Dig.) 2.5 (Ana.)	0.5	0.8 (Dig.) 0.8 (Ana.)	1.2 (Dig.) 1.2/2.5/3.3 (Ana.)
Image resolution	32 × 32	160 × 128	240 × 240	320 × 240	128 × 128	128 × 128	128 × 128
Double sampling	No	No	CDS	No	No	No	DRS
Frame rate [fps]	156	24 - 268	120	1	480	50 - 250	18.2 - 79.7
Pixel pitch [µm]	35	9	9.8	7	7.6	7.6	6.04
Pixel complexity	61T APS	40T log + 1 MIM cap.	14T PWM + 4 caps.	3T APS	4T PWM	4T PWM	3T APS
Fill factor [%]	9.1	12.9	20.1	N/A	36	36	54
MAC type	Current-based	Current-based	Charge-based	Charge-based	Current-based	Current-based	Charge-based
Multiscale	No	6 scales (conv.) 3 scales (Haar)	Arbitrary scales	3 scales	No	No	3 scales
Feature type	32×32 kernels	2×2 to 64×64 kern. Lin. Haar filts.	Log. Haar filts.	Lin. Haar filts.	3×3 kernels	3×3 kernels	16×16 kernels
Weight resolution	1b	1.5b	1.5b	1.5b	4b	4b	4b
Feature resolution	1b	1-8b	1b	1b	1 to 8b	1 to 8b	1-2-4-8b
Throughput [MOPS]	5.1	15.1 - 252.1	N/A	N/A	137.2	63.5	10.5 - 408.3
Power (MAC) [µW]	0.15 - 0.54	N/A	N/A	N/A	N/A	N/A	2.7 - 76.2
Energy eff. (MAC) [TOPS/W]	9.52 - 34.77	N/A	N/A	N/A	N/A	N/A	4.98 - 84.09
Power (SoC) [µW]	8.5	42 - 206	2900	24 - 96	117	80.4 - 134.5	250.9 - 384.7
Energy (SoC) [pJ/(pix.frame.filter)]	3.34	2.5 - 103.9	16.76	N/A	14.88	4.24 - 12.66	45.9 - 254.3
Energy eff. (SoC) [TOPS/W]	0.60	0.23 - 5.46	N/A	N/A	4.67	0.63 - 1.89	0.16 - 4.57

Fig. 6. Comparison to the state of the art of mixed-signal conv. imager SoCs. EE is normalized to 1b operations.