An Ultra-Low-Power Analog-Digital Hybrid CNN Face Recognition Processor Integrated with a CIS for Always-on Mobile Devices

Ji-Hoon Kim, Changhyeon Kim, Kwantae Kim and Hoi-Jun Yoo
School of Electrical Engineering
Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, Republic of Korea
jihoon0708@kaist.ac.kr

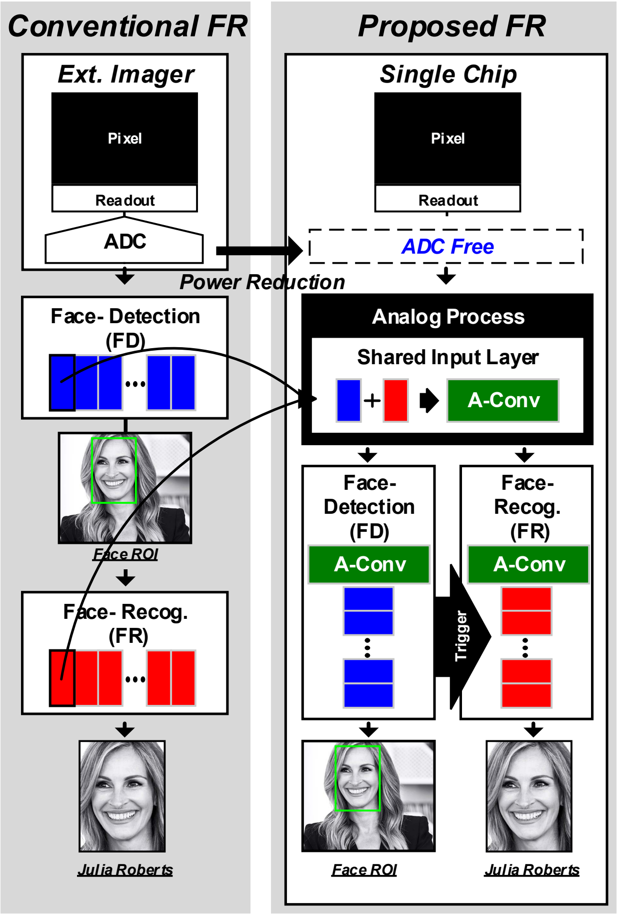
Abstract*— An ultra-low-power analog-digital hybrid always-on face recognition (FR) processor integrated with a CMOS image sensor (CIS) is proposed for the wearable mobile devices applications such as user authentication. The proposed processor is the first IC with full process of FR in a single chip. The processor adopts analog-digital hybrid convolution operation for efficient integration of CNN processor with CIS. The analog convolution processor is proposed for the computation of the 1st layer of CNN and the quantization operation without an ADC that can achieve 15.7% power reduction with 1.3% minimal accuracy loss. In addition, the analog weighted-sum unit with low power (< 20μW) and high efficiency (> 5.18TOPS/W) is proposed with switched-drain regulation (SDR) current mirror which can achieve less than 6% mirroring error. The processor is simulated in 65-nm CMOS technology, 15.84mm^2 area with 2.5V and 1.2V for analog domain and 0.77-1.1V for digital domain. It consumes 0.6198mW to evaluate one face at 1 fps and achieves 96.18% FR accuracy in LFW dataset.

***Keywords*—always-on, analog-digital hybrid, CNN, CMOS image sensor, face detection, face recognition, mixed-mode**

I. Introduction

Recently, always-on face recognition (FR) is becoming a promising modality for user authentication on mobile devices. Always-on FR can identify the user without a direct physical contact such as finger print scanning. Therefore, it is highly convenient for the users, and can be easily applied to any mobile devices [1]. In addition, as the number of smart devices has increased with the rise of the Internet of Things (IoT) era [2], a demand for always-on FR in smart devices is increasing. However, the integration of always-on FR into the mobile devices is challenging, because ultra-low-power consumption (< 1mW [3]) should be guaranteed due to the limited battery capacity. However, for highly accurate FR satisfying security issues, power-hungry deep neural network (DNN) based FR (> 95% [5]) is required.

In Fig. 1(a), the conventional FR system is described. It recognizes the face through a 3-step sequence. 1) An external imager converts the image into a digital domain by analog-to-digital converter (ADC), 2) Face detection (FD) processor obtains the face region of interest (RoI) from the entire image, and then 3) Identification result is obtained through a FR processor.



(a) (b)

Fig. 1. (a) Conventional face recognition system (b) Proposed face recognition system

Previously, several ASICs have been proposed for ultra-low-power FR [4-5]. However, an accuracy of [4] is low because of hand-crafted feature-based FR algorithm. In addition, [5] uses Viola-Jones face detector [7] for FD which fails to cope with environmental changes such as changes in face position or expressions and illumination [8]. Furthermore, both ASICs have a separate image sensor paired with a FR digital processor. Inevitably, they should consume additional power to transmit the data to a different chip, which is unsuitable for ultra-low-power operation.

In this work, we propose the first always-on analog-digital hybrid FR processor integrated with a CMOS image sensor (CIS) as shown in Fig. 1(b). It integrates the full process of FR in a single chip, including the front-end CIS and convolutional neural network (CNN) based analog-

This paper is supported by Future Non-Von Neumann Architecture Cluster Program of Samsung Electronics

978-1-7281-0397-6/19/$31.00 ©2019 IEEE

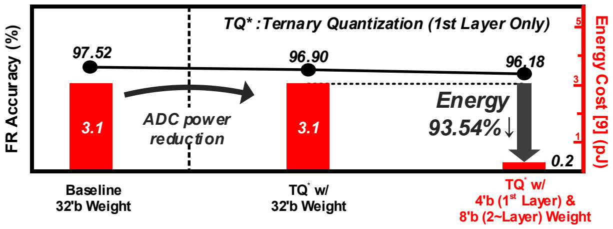


Fig. 2. Face recognition accuracy

digital hybrid FD & FR processor. For ultra-low-power consumption, it proposes 3 key features: 1) At the algorithm level, the 1st layer of the FD CNN and FR CNN is shared. By sharing the 1st layer, analog FD & FR can be performed with a single analog convolution hardware at the same time reducing latency and power consumption. 2) The analog convolution processor is proposed for the computation of the 1st layer of CNN and the quantization operation without an ADC, which occupies the majority (> 50%) of power in conventional CIS [6]. 3) The analog weighted-sum unit with low power (< 20µW) and high efficiency (> 5.18TOPS/W) is proposed. Additionally, to compensate the error of the analog convolution, the switched-drain regulation (SDR) current mirror is proposed which can achieve less than 6% current mirroring error for analog multiplication.

II. Analog-Digital Hybrid FD & FR

A. FD/FR CNN with Shared Input Layer

When integrating CIS and CNN based FD & FR, it is necessary to convert analog domain images into the digital domain. Previous architectures [4-5] use high-bit (8-12 bits) ADCs for these domain conversions. However, using ADCs that consume majority power in existing CIS (>50% [6]) is inefficient. Therefore, we replace conventional ADC to shared analog convolution layer that can perform both first layer of CNN and quantization. Additionally to solve the heavy power consumption of CNN, the bit-precision of weight is reduced to 8-bit that the computational energy is reduced by 93.54% with minimal accuracy loss under 1.5%.

Fig. 2 shows the simulated FR accuracy degradation in implementing proposed analog-digital hybrid CNN. The baseline FR accuracy using 32-bit floating point (FP) weight is 97.52%. First, to perform analog to digital domain conversion and to replace the power consuming ADC, we adopt ternary quantization in output of first layer. In that case, simulated FR accuracy is degraded to 96.9%. Then, to reduce the total power consumption in analog-digital hybrid CNN, weights of overall convolutional layers are reduced to 8-bit fixed point (FXP) and shared input layer adopts 4-bit FXP. As a results, with a slight loss of accuracy (Only 1.3%), the energy cost of digital multiplier can be dramatically reduced as 93.54% [9], and in analog convolution processor, we can perform analog computation in low power and quantize output of 1st layer without ADC.

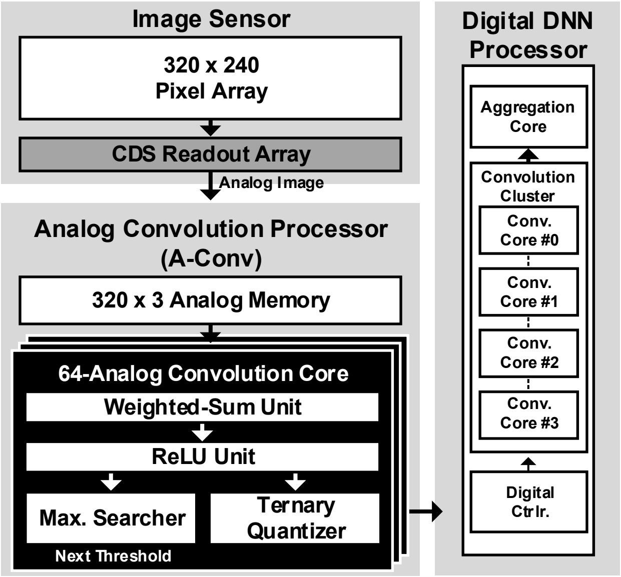


Fig. 4. Overall architecture

Fig. 3 shows final algorithm flow of proposed analog-digital hybrid FD & FR system. In this paper, both FD and FR are implemented with CNN. In addition, by sharing the 1st layer of FD & FR CNN, it is possible to conduct both FD & FR process in a single readout. It has following sequence. First, the image sensor captures the image and it conducts the shared analog-CNN input layer process. This 1st layer feature outputs are stored in ternary form, and the FD is performed through the always-on FD CNN. If the face is detected, the RoI part of 1st layer feature output is reused to perform the FR CNN.

B. Overall Architecture

Fig. 4 shows the overall architecture of the proposed FR system. The entire system is integrated with a CIS and a CNN-based analog-digital hybrid FD & FR processor in a single chip. The architecture consists of three main components: 1) CMOS image sensor 2) Analog convolution processor (A-Conv) 3) Digital DNN processor. The CIS consists of a 320×240 pixel array and a CDS readout circuits. A-Conv has 320×3 analog memory and 64-analog convolution core. Each analog convolution core consists of weighted-sum unit which can calculate partial sum of 3×3 weight kernel, ReLU unit, and ternary quantizer. Additionally, maximum value searcher (Max. searcher) exists to pre-determine the quantization and threshold value. The digital DNN processor consists of the convolution

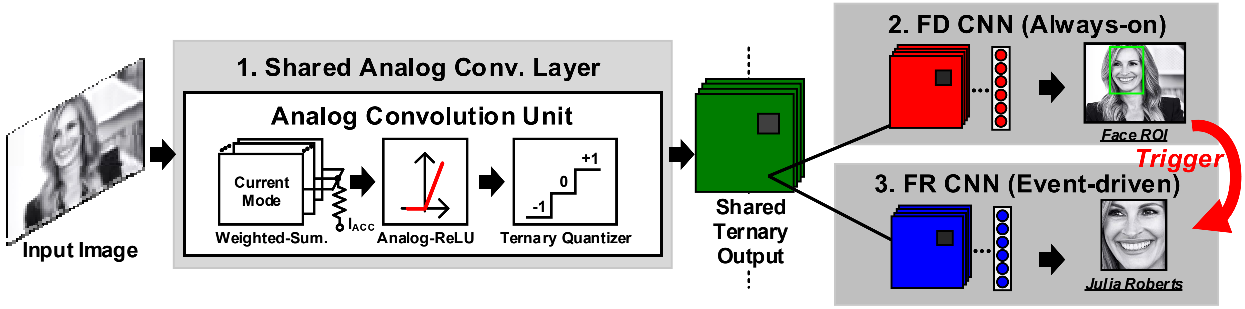


Fig. 3. Overall algorithm flow of analog-digital hybrid CNN

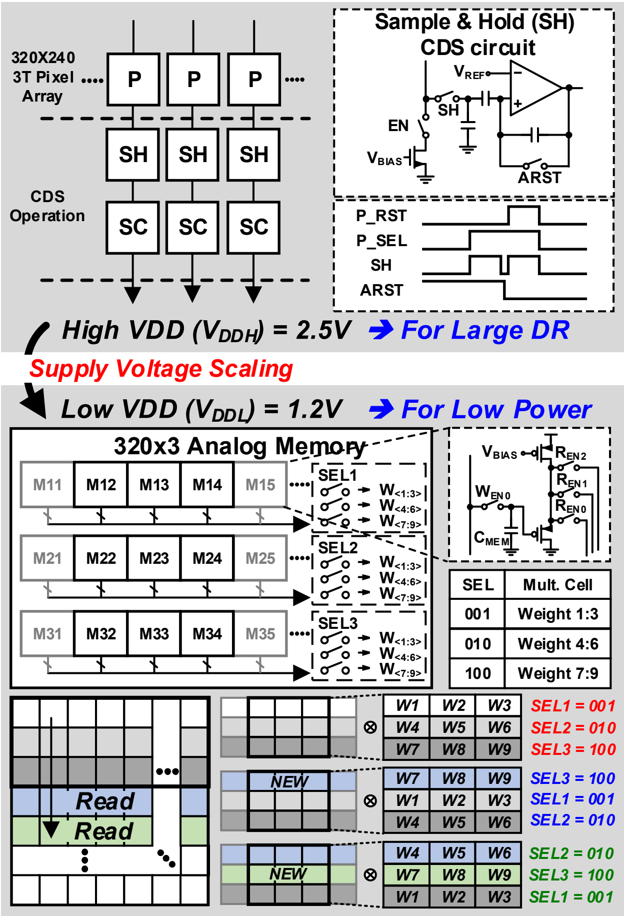


Fig. 5. CDS-readout array circuit & analog memory architecture

cluster, aggregation core and digital controller. Convolution cluster performs convolution operations with 4 convolution cores and transfers the accumulation results into aggregation core. In addition, digital controller exists for controlling each processor.

III. Detailed Building Blocks

A. CDS-Readout Array Circuit and Analog Memory Architecture

Fig. 5 shows the correlated double sampling (CDS) operation and corresponding sample & hold (SH) and switched capacitor (SC) circuit for readout of image data from the 3T pixel CIS. The implemented CDS circuit in each column reads the image data according to the timing diagram as shown in Fig. 5. To support the 3×3 weight kernel operation, the source follower based analog memory is implemented with 320×3. The 3×3 data stored in the analog memory are moved to weight multiplier for convolution operation sequentially. In this case, each row of memory is connected to the corresponding multiplier cell by the control of SEL signal. Then the convolution operation can be performed by changing the REN signal same as the 3×3 weight kernel sliding. Also, The overall analog memory operates as a circular queue memory. The CIS pixel array and the CDS circuit operate in 2.5V for wide dynamic range of CIS, and the analog convolution circuit operates in 1.2V of low supply for low power operation.

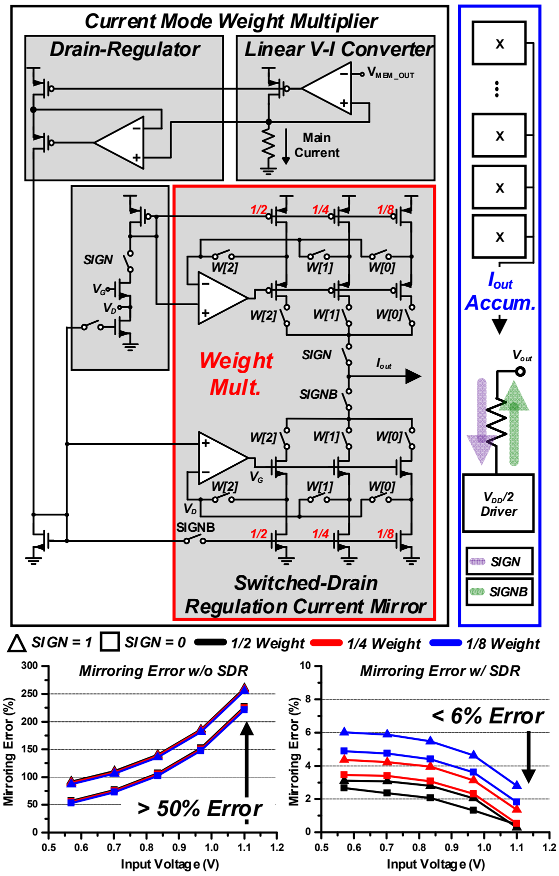


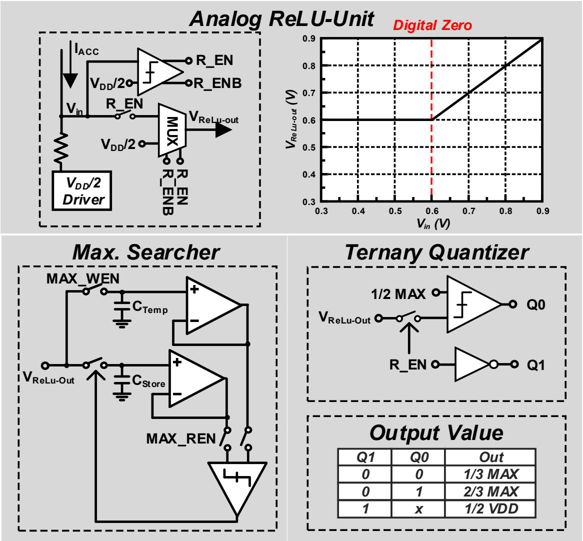
Fig. 6. Weighted-sum unit with switched-drain regulation current mirror

B. Weighted-Sum Unit with Swithced-Drain Regulation (SDR) Current Mirror

Fig. 6 shows the proposed current mode weighted-sum unit to perform the convolution operation. The output voltage from the memory is converted to current by linear V-I converter. Therefore, the output voltage of the memory is multiplied with the weight using a switched drain regulation (SDR) current mirror. The SDR current mirror is designed to have the same aspect ratio for the weight switches, cascode and mirroring MOSFET in the same branch for minimizing mirroring error. Overall SDR current mirror consists of not only PMOS mirror that actives at sign=1, but also the NMOS mirror that actives at sign=0. As shown in Fig. 6, as the drain voltage is regulated by the negative feedback loop, mirroring error is reduced to less than 6% even with many mirroring branches, which is negligible compared to the simple current mirror technique (>50% mirroring error). After the multiplication, the output current is accumulated with 9 multiplication currents corresponding to 3×3 kernel. The accumulation without an additional analog adder is the advantage of current mode multiplication. The resulted weighted-sum current output is converted to voltage domain by the bias resistor which is driven by VDD/2.

C. Analog ReLU-Unit and Ternary Quantizer

The output of the weighted-sum unit is converted to the digital domain through ReLU-unit and ternary quantizer. Fig. 7 shows the analog-ReLU unit. It determines the output value based on VDD/2 which indicates the digital zero. If the output of weighted-sum unit (V_{in}) is larger than VDD/2, the implemented comparator and multiplexer (MUX) passes V_{in}. In opposite, the output of MUX is driven to VDD/2. After



Q1	Q0	Out
0	0	1/3 MAX
0	1	2/3 MAX
1	x	1/2 VDD

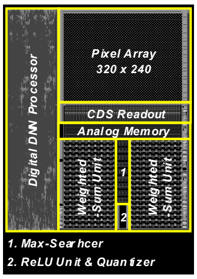
Fig. 7. Analog ReLU-unit & ternary quantizer

that, the output of ReLU-unit is quantized by ternary quantizer. The threshold value of quantizer is determined by 1/2×MAX and ReLU-enable signal (R_EN). The MAX is pre-determined value which is the maximum value of previous frame and R_EN indicates that the output of ReLU-unit is VDD/2 or not. By this MAX value, the output quantized value (1/3×Max, 2/3×Max) is determined and this analog value is converted to digital bits aforetime by using existed ADC which is implemented for other image processing in mobile devices. Therefore, the final ternary quantizer output is digital bits of 1/3×Max, 2/3×Max and VDD/2. In order to use the maximum value of the previous frame, the max. searcher is implemented to store ReLU output data. The analog output of ReLU unit is stored in the C_{temp} temporarily and compared with the C_{store} which has the previous maximum value. When the output of comparator is high, C_{store} is updated to the current ReLU output.

IV. Implementation Results

Fig. 8 shows the layout photograph of the proposed analog-digital hybrid FD & FR processor, which is simulated in 65nm 1P8M CMOS technology occupying 15.84mm^2 area. At 1-fps framerate, analog domain consumes 0.0588mW for imaging and analog convolution operation and digital domain consumes 0.561mW for FD & FR processing at 0.77V supply voltage with 50MHz.

Fig. 9 shows the FD & FR result and power breakdown of the proposed architecture. In the case of a positive image that contains a human face, processor extracts the RoI and in case of FR, 96.18% accuracy with whole classes in LFW dataset can be achieved. Also, it shows the estimated power breakdown when using proposed analog convolution layer instead of conventional ADC. In this case, the power of ADC is predicted by [6]. The overall power reduction is 15.7% when using the proposed analog-digital hybrid CNN, assuming that both process has same 320×240 image during 1-fps.



Technology	65nm 1P8M Logic CMOS
Die Area	3.6 × 4.4mm^2
Pixel Size	7.0 × 7.0μm^2
Supply Voltage	Analog: 2.5V, 1.2V Digital: 0.77 ~1.1V
Clock Frequency	Analog: 100MHz Digital: 50~200MHz
Power Consumption	Analog: 0.0588mW Digital: 0.561mW
Frame Rate	1 fps

Fig. 8. Layout Photograph & Summary

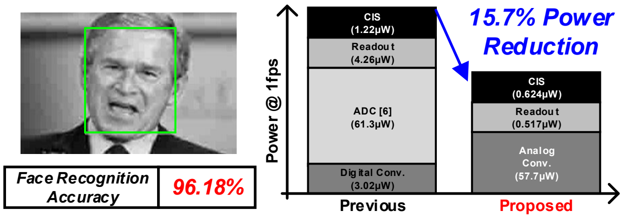


Fig. 9. FD & FR results & estimated power break down (@1-fps)

TABLE I. System Comparison table

	JSSC'17 [4]	JSSC'18 [5]	This Work
Technology	TSMC 40nm	Samsung 65nm	Samsung 65nm
Algorithm	FD: Haar-like FR: PCA+SVM	FD: Haar-like FR: CNN	*FD&FR : Analog-digital hybrid CNN*
Accuracy	81% @ 32-class in LFW	97% @ whole LFW	*96.18% @ whole LFW*
Resolution	HD	QVGA	QVGA
Power	23mW	0.62mW	*0.6198mW*

Table I shows the comparison with previous FR systems integrated with CIS [4-5]. Despite the large workload of the CNN both in FD and FR, this paper achieves low-power consumption 0.6198mW in evaluating one face at 1-fps. In addition, it proposes new analog-digital hybrid method in CNN and achieves 96.18% accuracy with whole classes in LFW dataset which is only 1.3% accuracy degradation. Table II shows the comparison with previous analog MAC for feature extraction [10-11]. The proposed weighted-sum unit achieves the lowest power consumption and additionally, it achieves 5.18-9.06 TOPS/W energy efficiency which is quite comparable to previous works.

V. Conclusion

In conclusion, an analog-digital hybrid CNN face recognition processor is proposed with a high accuracy 96.18% and 0.6198mW ultra-low-power consumption. The proposed processor is the first IC integrated full process up to FR using analog-digital hybrid CNN. The analog convolution processor is proposed for the computation of the 1st layer of CNN and the quantization operation without an ADC that can achieve 15.7% power reduction with 1.3% minimal accuracy loss. Additionally, the analog weighted-sum unit with low power (< 20μW) and high efficiency (> 5.18 TOPS/W) is proposed with switched-drain regulation (SDR) current mirror which can achieve less than 6% mirroring error. Also, this paper suggests a new method for implementing analog-digital hybrid CNN and confirms that it can achieve low power operation with minimal accuracy loss.

TABLE II. Analog-MAC Comparison table

	ISSCC'15 [10]	ISSCC'16 [11]	This Work
Technology	130nm	40nm	65nm
Clock	20kHz	1GHz	5.4MHz
Weight Precision	4-bit	3-bit	*4-bit*
Supply	1.2V	1V	1.2V
Total Power	663nW (Only Mult.)	228μW	*10.17 – 18.75 μW*
Efficiency	0.0603 TOPS/W	8.77 TOPS/W	*5.18 – 9.06 TOPS/W*

REFERENCES

[1] E. Fernandez and D. Jimenez, “Face Recognition for Authentication on Mobile Devices,” Image and Vision Computing, vol 55, pp. 31-33, November 2016

[2] D. Evans, “The Internet of Things – How the Next Evolution of the Internet Is Changing Everything,” White Paper, Cisco IBSG, pp. 1-11, April 2011

[3] G. Andrews, L. przywara, and Cadence, “Keeping Always-On Systems On for Low-Energy Internet-of-Things Applications.”

[4] D. Jeon, Q. Dong, Y. Kim, X. Wang, S. Chen, H. Yu, D. Blaauw, and D. Sylvester, “A 23-mW Face Recognition Processor with Mostly-Read 5T Memory in 40nm CMOS,” IEEE J. Solid-State Circuits, vol. 52, pp. 1628-1642, June 2017

[5] 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, pp. 115-123, January 2018

[6] M. Shin, J. Kim, M. Kim, Y. Jo, and O. Kwon, “A 1.92-Mega pixel CMOS Image Sensor with Column-Parallel Low-Power and Area-Efficient SA-ADCs,” IEEE Transactions on Electron Devices, vol. 59, pp. 1693-1700, June 2012

[7] P. Viola, and M. Jones, “Rapid Object Detection using a Boosted Cascade of Simple Features,” Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, vol. 1, pp. I-511-I-518, 2001

[8] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, “A Convolutional Neural Network Cascade for Face Detection,” 2015 IEEE Conference on Computer Vision and Patter Recognition (CVPR), pp. 5325-5334, 2015

[9] M. Horowitz, “Computing’s Energy Problem (and what we can do about it),” 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), pp. 10-14, 2014

[10] J. Zhang, Z. Wang, and N. Verma, “A Matrix-Multiplying ADC Implementing a Machine-Learning Classifier Directly with Data Convesion,” 2015 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), pp. 332-333, 2015

[11] E. Lee, and S. Wong, “A 2.5GHz 7.7TOPS/W Switched-Capacitor Matrix Multiplier with Co-designed Local Memory in 40nm,” 2016 IEEE International Solid-State Circuits Conference (ISSCC), pp. 418-419, 2016