A Low-Power Convolutional Neural Network Face Recognition Processor and a CIS Integrated With Always-on Face Detector

Kyeongryeol Bong, *Student Member, IEEE*, Sungpill Choi, *Member, IEEE*, Changhyeon Kim, *Student Member, IEEE*, Donghyeon Han, *Student Member, IEEE*, and Hoi-Jun Yoo, *Fellow, IEEE*

***Abstract*— A Low-power convolutional neural network (CNN)-based face recognition system is proposed for the user authentication in smart devices. The system consists of two chips: an always-on CMOS image sensor (CIS)-based face detector (FD) and a low-power CNN processor. For always-on FD, analog–digital Hybrid Haar-like FD is proposed to improve the energy efficiency of FD by 39%. For low-power CNN processing, the CNN processor with 1024 MAC units and 8192-bit-wide local distributed memory operates at near threshold voltage, 0.46 V with 5-MHz clock frequency. In addition, the separable filter approximation is adopted for the workload reduction of CNN, and transpose-read SRAM using 7T SRAM cell is proposed to reduce the activity factor of the data read operation. Implemented in 65-nm CMOS technology, the 3.30 × 3.36 mm^2 CIS chip and the 4 × 4 mm^2 CNN processor consume 0.62 mW to evaluate one face at 1 fps and achieved 97% accuracy in LFW dataset.**

***Index Terms*— Always-on system, convolutional neural network (CNN), face recognition (FR), functional CMOS image sensor (CIS).**

I. Introduction

RECENTLY, face recognition (FR) based on always-on CMOS image sensor (CIS) has been investigated for the next generation UI/UX of wearable devices. A mobile FR system was developed as a life cycle analyzer [1] or a personal black box, constantly recording the people we meet, along with time and place information. Also, as the number of smart devices per each person increases [2], the FR with always-on capability becomes an attractive solution for user authentication since it can provide a non-intrusive unlock for multiple devices as shown in Fig. 1. Moreover, in the Internet of Things era, the always-on FR becomes an essential functionality for every device to be more intelligent and interact with their users.

Since wearable devices have a limited battery capacity for a small form factor, the FR system should have extremely low-power consumption for its always-on operation, while maintaining high recognition accuracy. Previously, a 23-mW FR accelerator [3] was proposed, but its accuracy was low due to its hand-crafted feature-based algorithm. In addition, its architecture based on a low-power image sensor chip [4] and the digital accelerator chip was inappropriate to achieve low-power consumption in the always-on FR because the entire image data generated from the image sensor should be streamed to the digital accelerator to check whether there is a face or not.

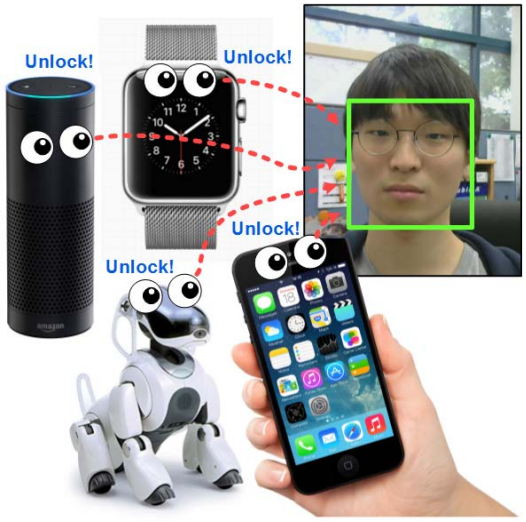


Fig. 1. Always-on FR for user authentication.

In this paper, we introduce a functional CIS integrated with an always-on Haar-like face detector (FD) and a convolutional neural network (CNN) processor (CNNP) [5]. The functional CIS can transmit only face images only when they exist. For low-power FD, the hybrid use of analog processing unit and digital processing unit is proposed. While using CNN, which is essential to achieve high accuracy in the FR [6], separable filter approximation (SFA) is adopted to reduce the large computational workload of the CNN [7]. In addition, transpose-read SRAM (T-SRAM) is presented to reduce the power consumption of the CNNP by enabling efficient SRAM access when using the SFA.

The rest of this paper is organized as follows. In Section II, the overall system architecture is described. Section III explains the detailed implementation of key building blocks, an analog–digital hybrid Haar-like FD, a PE of CNN processor,

Manuscript received May 10, 2017; revised July 31, 2017 and September 30, 2017; accepted October 15, 2017. Date of publication December 13, 2017; date of current version December 26, 2017. This paper was approved by Guest Editor Muhammad M. Khellah. *(Corresponding author: Kyeongryeol Bong.)*

The authors are with the School of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 305-701, South Korea (e-mail: krbong@kaist.ac.kr).

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

Digital Object Identifier 10.1109/JSSC.2017.2767705

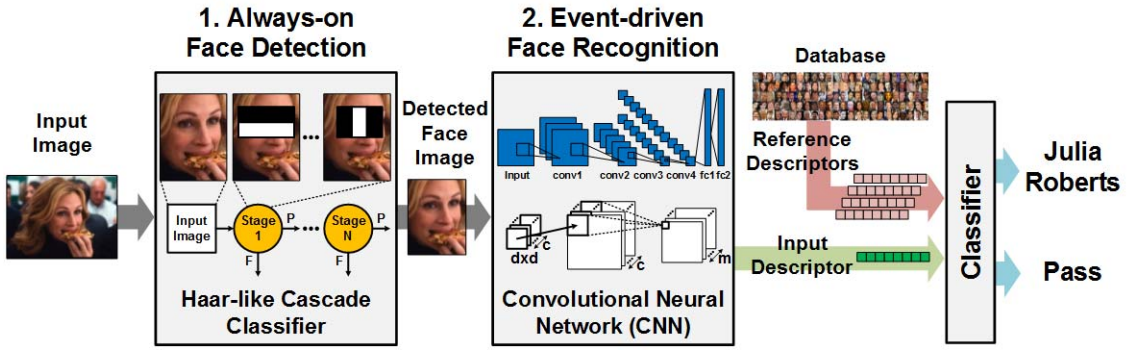


Fig. 2. Overall algorithm flow.

TABLE I
SPECIFICATION OF SUB-WINDOWS

	Scale 1	Scale 2	Scale 3
Downsizing Factor	8×	6×	4×
Original Sub-window Size	160×160	120×120	80×80
Stride	8	6	4
# of Sub-windows per Frame	200	660	2,400

and T-SRAM. Section IV shows the chip implementation results, and Section V provides the conclusion.

II. OVERALL SYSTEM ARCHITECTURE

A. Overall Algorithm Flow

Fig. 2 shows the overall algorithm flow of the FR. While an image sensor captures images, face detection (FD) is performed on the image frames, and the face verification (FV) is applied for the detected faces. In this paper, the Viola–Jones algorithm [8] is used for the FD, and a CNN is utilized for the FV.

When an image frame is given, sub-windows at different positions and scales are evaluated at the FD stage. With a given sub-window, the Viola–Jones algorithm processes a number of cascaded classifier stages sequentially until the sub-window is rejected at a certain stage or classified as a face by passing all stages. Each stage contains a number of Haar-like filters, which are computed to determine whether the sub-window is passed or rejected at that stage. A Haar-like filter specifies the positions and the shapes of black rectangular regions and white rectangular regions within a sub-window, and the intensity summations over the black regions and the white regions are compared. The cascaded classifier used in this paper consists of 23 stages with 4188 filters. As shown in Table I, three different sizes of sub-windows, 80 × 80, 120 × 120, and 160 × 160, are evaluated. The strides are 4, 6, and 8, and the number of windows per frames are 2400, 660, and 200, respectively.

For the detected face images, the FV stage performs the feedforward processing of the CNN. In this paper, SFA is adopted to reduce the large computational workload of

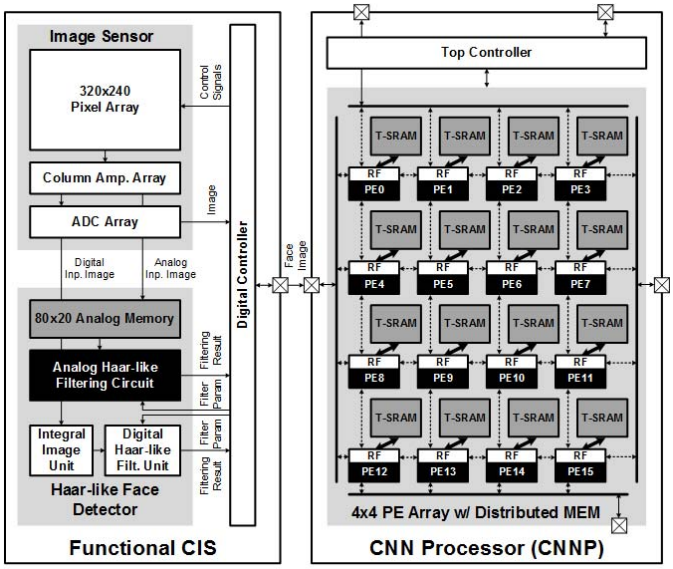


Fig. 3. Overall system architecture.

the CNN inference. Then, 2-D d × d filters in a convolutional layer can be approximated by the combination of 1-D d × 1 horizontal filters and 1 × d vertical filters.

The CNN is trained for the classification of CASIA-Webface dataset, which has 494 414 images of 10 575 subjects. After then, for the FV in the system, the last softmax layer is discarded, and the 256-D output vector of the previous fully connected layer is used as the descriptor of an input face image. With this descriptor, an additional classifier is employed to check whether any descriptor registered in the database matches the input descriptor. When a face image of a new user is added to the system, the descriptor of the face is computed, and the classifier is newly trained. Since a well-trained CNN has an ability to generate discriminable descriptors, a relatively simple classifier that is easy to train produces good results; a support vector machine (SVM) is adopted in this paper.

B. Overall System Architecture

Fig. 3 shows the overall system architecture of the proposed FR system. The system consists of two chips: a functional CIS for the FD and a CNNP for the FV. First, the functional

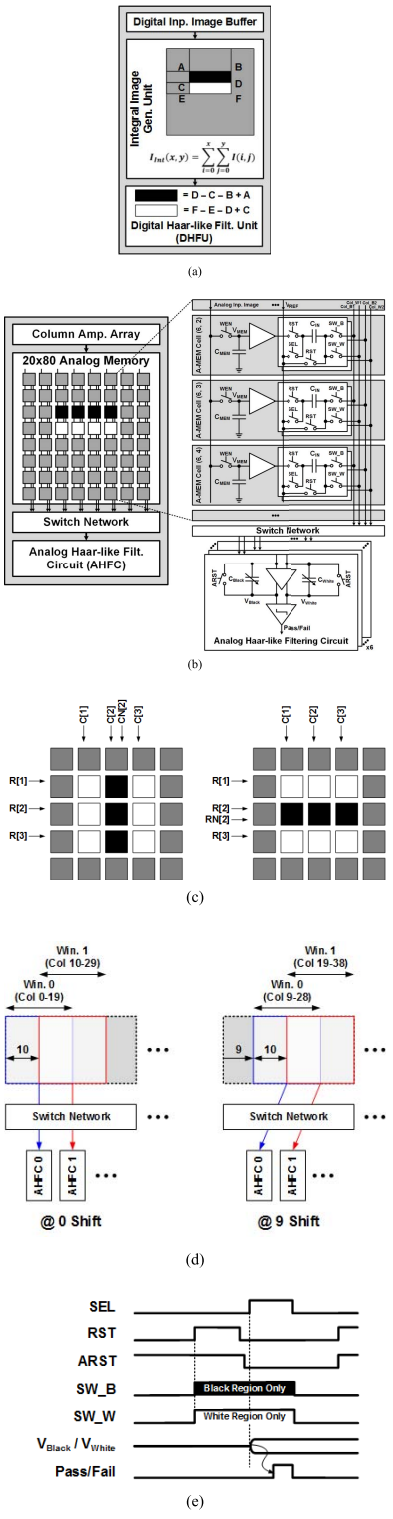


Fig. 4. (a) DHFU. (b) AHFC. (c) Analog memory cell selection for Haar-like filters. (d) Sub-windows allocation and sliding. (e) AHFC operation.

CIS performs always-on imaging and the FD. Once a face is detected, the functional CIS transmits only the face image to the CNNP, and then the CNNP processes the FV.

The functional CIS consists of 320 × 240 pixel array, read-out circuits, analog Haar-like filtering circuits (AHFC) with 80 × 20 analog memory, a digital Haar-like filtering unit (DHFU) with an integral image unit, and a top controller.

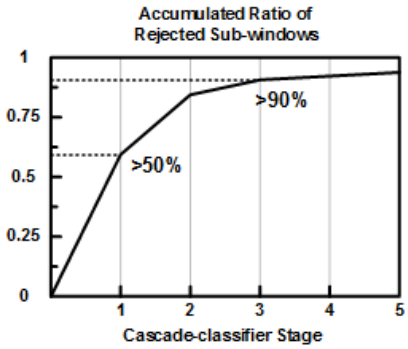


Fig. 5. Accumulated ratio of rejected sub-windows along cascade-classifier stages.

The CNNP is composed of 4 × 4 PE array, and each PE has its own local T-SRAM. The PE array is interconnected with a mesh-type network, and the boundary PEs are connected to the external interfaces. There is a top controller with instruction buffer. The CNNP is programmed by using global instructions. With a given cycle, the 16 PEs perform same operation but each PE can be enabled or gated for this operation. To utilize the 16 PEs as much as possible, the feature maps and the weights need to be distributed to each PE's own SRAM and weight buffer to process different part of required MAC workloads.

III. Detailed Building Blocks

A. Analog–Digital Hybrid Haar-Like Face Detector

As shown in Fig. 4(a), the DHFU follows the conventional processing method of Viola–Jones algorithm using an integral image [8], whose value at some position is given by integrating all the values of the original image inside the rectangle that encompasses the top left corner to the current position. Although computing the integral image requires a significant processing, it helps the processing of Haar-like filters to be very simple, since the intensity summation over some rectangular region can be calculated by 3 add/sub operations. Hence, the use of the integral image achieves good energy efficiency when the input sub-windows are not rejected across many stages and need to operate on a large number of Haar-like filters. However, in our test shown in Fig. 5, more than 50% and 90% of the input sub-windows are rejected after the first and the third stage, respectively, and the energy efficiency is decreased when using the integral image for these early-rejected windows.

To improve the energy efficiency, the AHFC is proposed to process Haar-like filters with direct summation of pixel intensities without such an initial processing. Fig. 4(b) shows the detailed circuit diagram and the operating sequence of the AHFC.

While a row of the pixel array is read out following the rolling shutter operation of the image sensor, downsizing is performed, and the voltage intensities of the downsized image are stored in analog memory [9]. The analog memory is designed to keep 20 rows so that it can provide all the intensities of 20 × 20 input sub-windows at the same time.

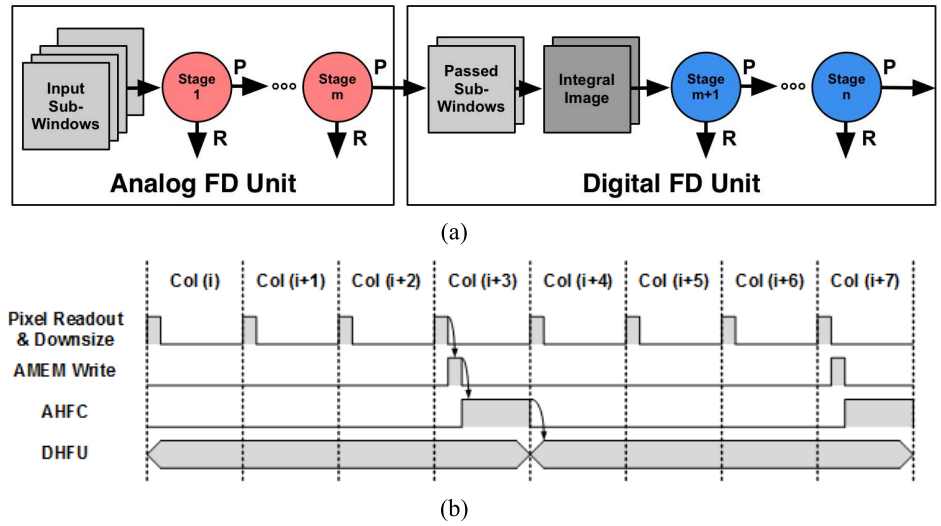


Fig. 6. (a) Analog–digital hybrid FD. (b) Operation sequence in the functional CIS.

The number of columns, 80, is decided by the width of the pixel array, 320, and the smallest downsizing factor 4×.

The analog memory cell consists of a sampling capacitor C_{MEM}, two input capacitors C_{IN}, a unity-gain buffer, combinational logic for cell selection, and several switches. A 174-fF MOS capacitor and a 55.6-fF MIM capacitor are used for C_{MEM} and C_{IN}, respectively. With two input capacitors, each analog memory cell can drive two different AHFCs at the same time. When processing a Haar-like filter, the switches SW_B and SW_W in each analog memory cell are configured according to the shape of the filter; either the SW_B or the SW_W is turned on depending on whether the cell is in the black area or in the white area, while both are turned off for the rest area. Fig. 4(c) shows the cell selection scheme. The input signals R and RN are shared by a row of analog memory while the signals C and CN are shared by a column. When the R and the C are both high for an analog memory cell, and the RN and the CN are low, the digital logic of the cell turns on the SW_W. If one of the RN and the CN is high, the logic turns on the SW_B. To configure two input capacitors, each cell is connected to two sets of R, RN, C, and CN.

When processing the scale with downsizing factor 4×, the entire area of 20 × 80 analog memory is valid, and 60 sub-windows are evaluated whenever one row of the memory is updated. In this paper, six AHFCs are integrated to process the 60 sub-windows in ten iterations. As shown in Fig. 4(d), for each iteration, six sub-windows are selected to be ten columns apart from each other. The overlap between two neighboring sub-windows can be handled by connecting two input capacitors of each analog memory cells to two neighboring AHFCs, respectively. During the iteration, the cell selections for the Haar-like filter are shifted by reconfiguring the C and the CN, and the connections between the analog memory and the AHFCs are shifted by reconfiguring the switch network. In the switch network, one AHFC is connected to 29 output columns of the analog memory so that it can cover ten different 20×20 input areas during the iterations with different sliding numbers. To reconfigure the switch network, the controller set the sliding number, and this number is decoded by combinational logic to generate the enable signals for each switch in the network.

Fig. 4(e) shows the operation of the AHFC. After initialized with V_{REF} at the reset phase, each analog cell in the black and white area transfers charge proportional to $(V_{\mathrm{MEM}} - V_{\mathrm{REF}})$, and they are accumulated at C_{Black} or C_{White}, respectively. Then, the voltage V_{Black} and V_{White} become the summation of the pixel intensities for the Haar-like filter, and the output comparator gives the "Pass" or "Fail" result. The C_{Black} and the C_{White} are designed to be configurable from 55.6 to 64×55.6 fF with 6-b signal. This is to decrease the voltage gain when the area of a Haar-like filter is large so that the V_{Black} and the V_{White} can be saturated.

As the AHFC performs direct summation and does not require any initial processing, it has good energy efficiency for early-rejected windows. However, due to the bias current when operating the analog memory and the AHFC, it is not suitable for long-lasting windows.

In this paper, the AHFC and the DHFU are combined to achieve better energy efficiency in the FD. As shown in Fig. 6(a), this hybrid approach first utilizes the AHFC to filter out early-rejected windows, and the only passed sub-windows are transferred to the DHFU for the processing of the rest stages. Fig. 6(b) shows the operating sequence of the functional CIS. Preserving the per-row latency for the rolling shutter operation, the entire rows are scanned. Whenever one row of downsized image is generated, the analog memory is updated, and the AHFC is utilized. After then, if there exist passed sub-windows, the DHFU processes the rest stages.

Fig. 7(a) shows the accumulated energy consumption of the hybrid method when processing the first three stages with 52 Haar-like filters at the AHFC. In the functional CIS, the number of stages processed by the AHFC is configurable, and with the cascaded classifier model [12] adopted in this paper, it is decided as stage 3 to optimize the energy consumption. Due to filtering out at the AHFC, the energy consumption of

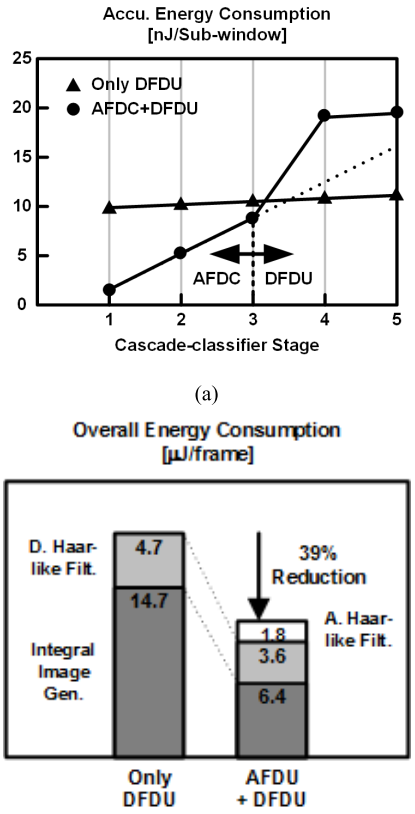


Fig. 7. (a) Accumulated energy consumption with an input sub-window along cascade-classifier stages. (b) Overall energy consumption of the FD.

the hybrid method is decreased for the first three stages but is increased for the rest stages, compared to the case using only the DHFC. However, because the portion of the sub-windows that are rejected until the third stage is much larger than the other, the workload of the DHFC is reduced significantly, and the overall energy consumption is improved by ~40% as shown in Fig. 7(b). In simulation, the FD achieved 90% true positive rate with 0.5% false alarm rate for the Labeled Faces in the Wild (LFW) dataset.

B. CNNP Processing Element

Fig. 8 shows the block diagram of the CNNP PE. The PE consists of local SRAM, a weight buffer, a register file, four convolutional units for MAC operations, and two ALU for other operations. The local SRAM can fetch 32 word/cycle to the register file. The convolutional unit of the PE consists of input registers, 16-way SIMD-type MAC datapath, and accumulation registers. The CNNP with 4×4 PE array can access 512 word/cycle from the local SRAM array and support 1024 MAC operations/cycle.

When processing a convolutional layer, the input registers of the convolutional unit is updated by loading a row of an input feature map from the local SRAM or by shifting the existing row by one column at each cycle. When the input register is shifted, the word at the headmost column can be sent to the neighboring PE. This makes multiple neighboring PEs be connected and work together when processing a row

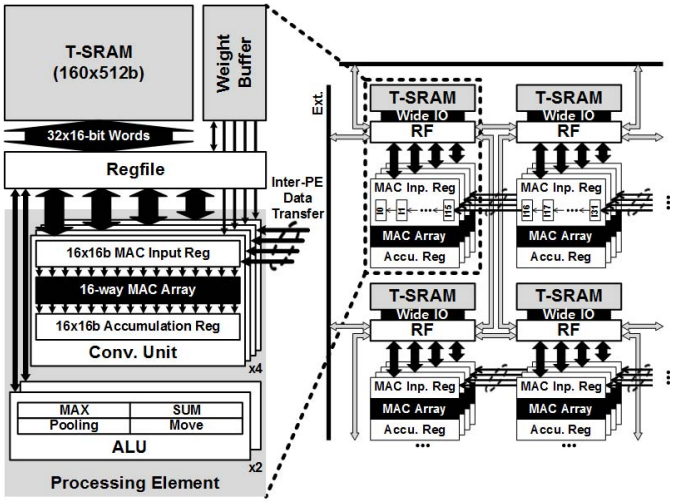


Fig. 8. CNNP PE block diagram.

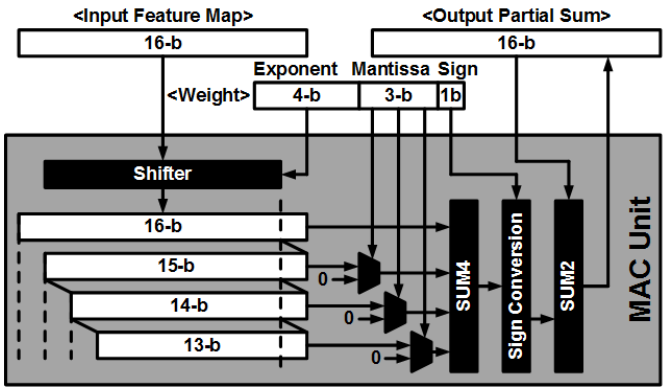


Fig. 9. MAC unit block diagram.

larger than 16 words. The input weight for the 16-way MAC datapath is shared for each convolutional unit, and the output partial sums are accumulated in the accumulation registers at the same column.

For the MAC operation, 16-bit fixed-point number and 8-bit floating-point number are used for feature maps and weights, respectively. Fig. 9 shows the block diagram of the proposed MAC unit. With weights represented by the floating point, which is composed of 1-bit sign, 4-bit exponent, and 3-bit mantissa, the MAC unit is implemented by using a shifter and adders instead of a multiplier. Compared to the conventional MAC unit with a multiplier for two fixed-point inputs, it achieved 43% and 21% reduction in its energy consumption and area, respectively, as shown in Fig. 10.

C. Separable Filter Approximation and Transpose-Read SRAM

The CNNP adopts the SFA to reduce the workload of convolutional layers. As shown in Fig. 11, the SFA can replace the convolution of 2-D $d \times d$ filters with two convolution stages of 1-D $d \times 1$ vertical filters and $1 \times d$ horizontal filters. In this paper, with an already trained CNN with 2-D filters, new 1-D vertical and horizontal filters are trained to approximate the original values of the 2-D filters by using the gradient

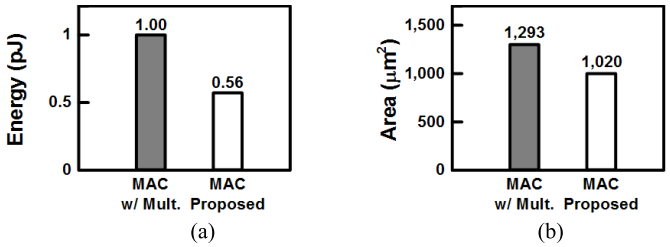


Fig. 10. MAC unit. (a) Energy consumption (b) Area.

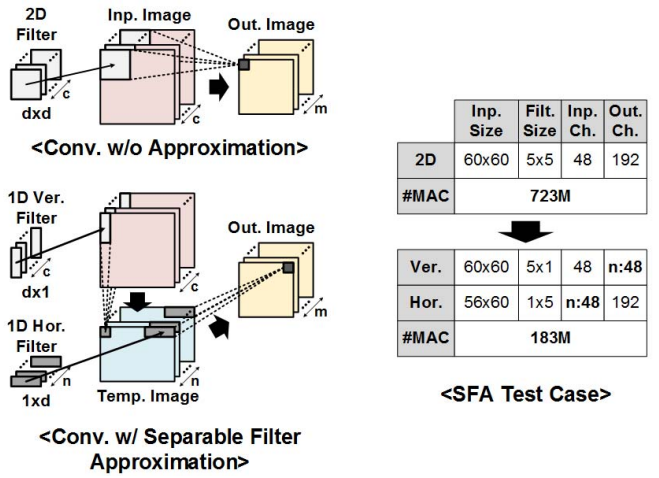


Fig. 11. SFA.

descent method [6]. In the convolutional layer shown in Fig. 11, the number of MAC operations required is decreased from 723 to 183 M, which corresponds to 4× reduction. The overall workload reduction is described in Fig. 17. The channel size of the temporal images n is set to the same value as the channel size of the input images c, but it can be selected as a smaller value to reduce the workload further while it can cause more accuracy degradation. In addition, the SFA can be independently applied to the multiple convolutional layers of a given CNN. In this paper, the overall workload of the CNN is reduced by 2×–3× with the help of the SFA while the accuracy degradation is managed to be less than 1% for the FV in LFW dataset [10].

Fig. 12 shows the processing of a convolutional layer with the SFA. When processing the horizontal filters, a single SRAM access can complete the convolution operation since the direction of the horizontal filters is matched to the direction of the row feature vector. However, SRAM cannot read the column feature vector at once, whose elements are connected to the same bit line, and the vertical filtering must fetch it through multiple SRAM accesses. Although the latency is decreased by the SFA as shown in Fig. 13(a), the activity factor of the SRAM in Fig. 13(b), hence the dynamic power consumption is increased due to redundant memory accesses during the vertical filtering.

To resolve this issue, the T-SRAM, which has two read modes: normal read mode to access a row vector and transpose-read mode to access a column vector, is presented. As shown in Fig. 14, the T-SRAM is based on 7T SRAM

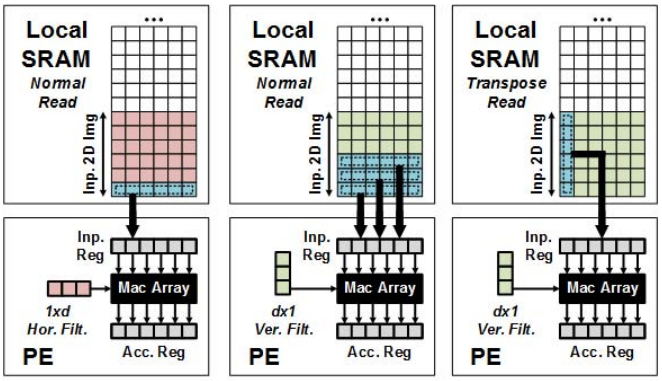


Fig. 12. SFA processing with normal SRAM read and transpose SRAM read.

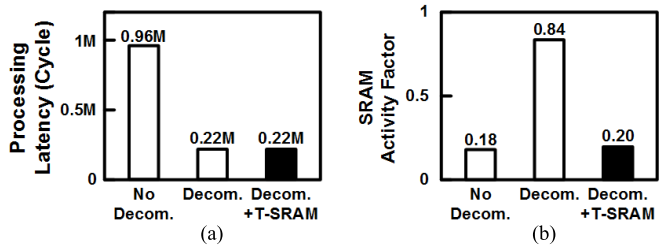


Fig. 13. (a) Latency and (b) SRAM activity factor with SFA and T-SRAM.

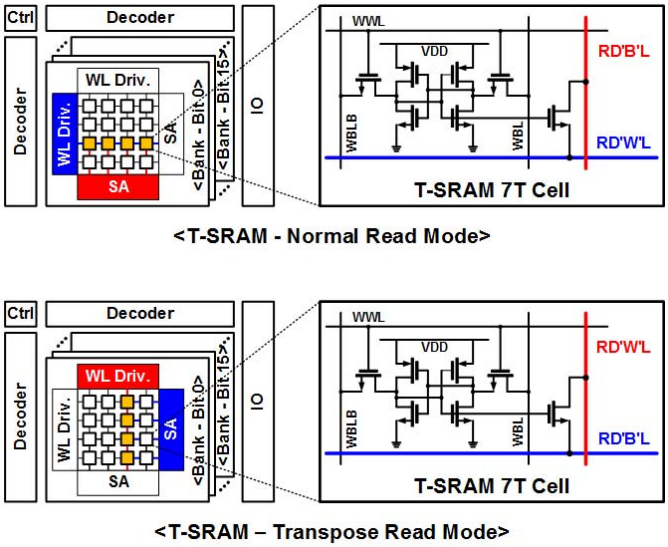
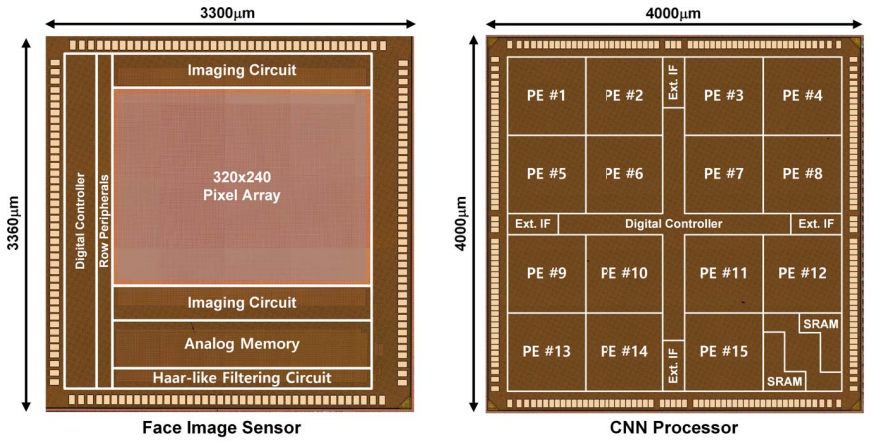


Fig. 14. Normal read mode and transpose-read mode of T-SRAM.

cell [11], which has one decoupled read MOS connected to a read bitline (RDBL) and a read wordline (RDWL). In order to provide two different read modes, the RDBL and the RDWL are designed to change the role of each other using two sets of WL drivers and sense amplifiers. In the normal read mode, the horizontal line becomes the RDWL to access cells in a row, while the vertical line becomes the RDWL to access cells in a column when using the transpose-read mode. In the T-SRAM, each bit of a 16-bit word is placed in different banks because a 16-bit word cannot be read at once in both normal and transpose-read modes if it is stored vertically or horizontally in 16 cells in the same bank.



Technology	65nm 1P8M Logic CMOS
Die Area	3300x3360μm^2
Pixel Array	320x240
Pixel Size	7x7μm^2
Supply Voltage	Analog: 2.5V, Digital: 0.5V ~ 0.8V
Clock Frequency	Analog: 50MHz, Digital: 5MHz ~ 100MHz
Power Consumption	24μW ~ 96μW @ 1fps

Technology	65nm 1P8M Logic CMOS
Die Area	4000x4000μm^2
Supply Voltage	0.46V ~ 0.8V
Clock Frequency	5MHz ~ 100MHz
Peak Power Consumption	211mW (0.8V, 100MHz)
	5.3mW (0.46V, 5MHz)
Energy Efficiency	2.11nJ/cycle (0.8V, 100MHz)
	1.06nJ/cycle (0.46V, 5MHz)

Fig. 15. Chip photograph and performance summary.

With the help of the T-SRAM, redundant memory accesses during the vertical filtering are resolved by the transpose-read mode. Compared to the case when it processes the approximated convolutional layer without the transpose-read mode, the CNNP with the transpose-read mode achieved 47% reduction in the power consumption. In total, the SFA and the T-SRAM achieve 78% reduction in the overall energy consumption.

IV. Implementation Results

Fig. 15 shows the chip photograph and the performance summary of the functional CIS and the CNNP fabricated by 65-nm 1P8M CMOS process. The functional CIS and the CNNP occupy 3.3 × 3.36 mm^2 and 4 × 4 mm^2, respectively. The 320 × 240 array of 7-μm pitch pixels is integrated in the functional CIS. The functional CIS operates at 2.5 V with 50 MHz for the analog domain and 0.5–0.8 V supply voltage with 5–100 MHz for the digital domain. At 1-fps framerate, the functional CIS consumes 24–96 μW for imaging and the FD depending on the number of faces and the face area in a given input scene. When ~20 faces are given to activate the readout for the whole frame, the CIS and the controller consume about 61 μW while the rest portion is used for FD units. As shown in Fig. 16(a), the CNNP operates at 0.46–0.8 V supply voltage with 5–100 MHz, and it can process 3.8 to 77 faces per second. The peak power consumption with maximum PE utilization is 5.3 and 211 mW while the corresponding energy efficiency is 1.06 and 2.11 nJ/cycle, respectively, as shown in Fig. 16(b).

Fig. 17 shows the CNN evaluation results. At the minimum energy point (MEP), 0.46-V supply voltage and 5-MHz operating frequency, the CNNP consumes 0.6 mJ to evaluate one face image with the target CNN that requires 1.26GMAC,

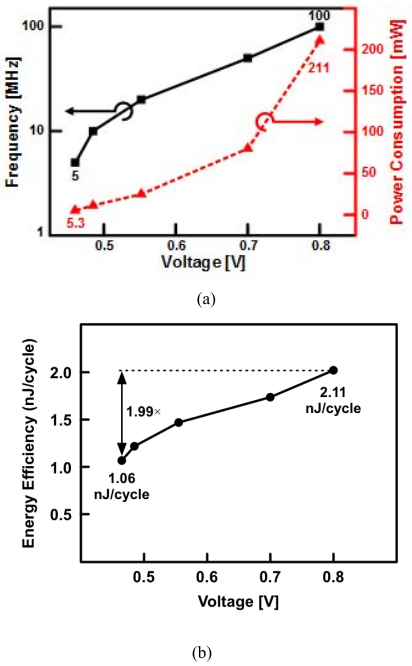
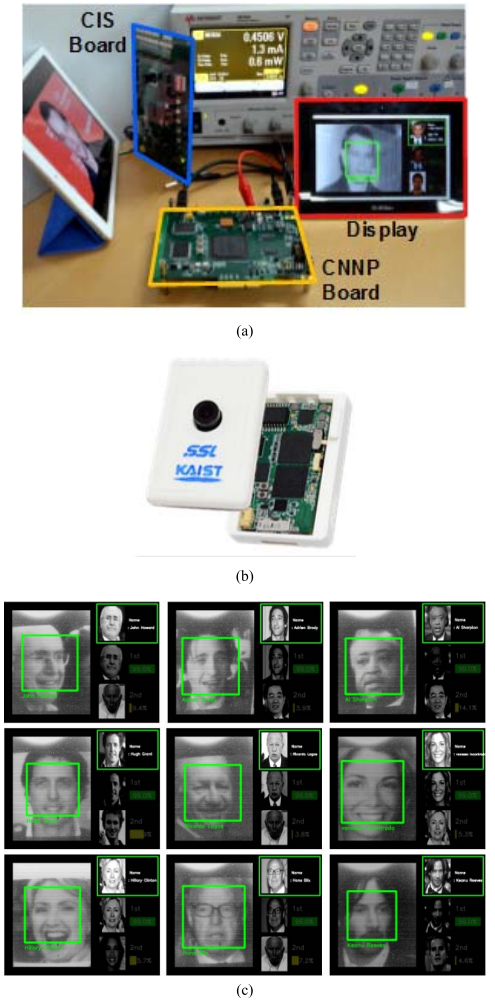


Fig. 16. (a) Voltage and frequency scaling in the CNNP. (b) Energy efficiency with voltage scaling in the CNNP.

and it takes 0.26 s. In this evaluation, the target CNN is approximated by the SFA as explained in Fig. 11 to have 0.72GMAC operations, which is 57% of the original workload. The SFA is applied to only the second convolutional layer, and it causes 0.2% accuracy degradation and achieves 97.4% accuracy at LFW dataset. The energy efficiency is 1.2TOPS/W, whereas the effective energy efficiency considering the original workload is 2.1TOPS/W.

Layer	Filter Size	Input Size	Input Channel	Output Channel	Filters	Maxout Groups	#MAC	Active #MAC	Latency (ms)	Power (mW)
Conv1	9x9	128x128	1	48	96	2	110M	896	44	1.96
Conv2	5x5	60x60	48	96	192	2	**720M/180M**	896	**217 / 50**	**2.49 / 2.50**
Conv3	5x5	28x28	96	128	256	2	350M	768	128	2.33
Conv4	4x4	12x12	128	192	384	2	64M	576	32	2.26
FC1	.	5x5	192	256	512	2	2.5M	64	9	1.70
Tot./Avg.							**1255M/715M**	64	**430 / 263**	**2.35 / 2.27**

Fig. 17. CNN evaluation results.



(a)

(b)

(c)

Fig. 18. (a) Experimental setup. (b) Prototype system. (c) FR results with images from LFW dataset.

The experimental setup is shown in Fig. 18(a). When the tablet screen shows input images, the CIS board captures it and performs the FD. Then, the detected face images are sent to the CNNP board to process the FV. The display shows the FR result, and the power supply measures the power consumption. Based on the proposed processing flow, the system prototype that includes the CIS, the CNNP, a field-programmable gate array, an external SRAM, a flash, and communication chips for USB and Bluetooth is shown in Fig. 18(b). Fig. 18(c) shows the FR results with images from LFW dataset.

TABLE II
PERFORMANCE COMPARISON

	SOVC '15 [3]	This Work
Technology	TSMC 40nm	Samsung 65nm
Algorithm	FD: Haar-like Cascade Classifier FR: PCA + SVM	FD: Haar-like Cascade Classifier FR: CNN
Accuracy	81% top-1 accuracy @ 32-class in LFW	97% for pairwise verification 83.9% top-1 accuracy @ whole LFW
Resolution	HD	QVGA
Framerate	5.5fps	1fps
Power	23mW @ 600mV, 100MHz	0.62mW

Table II shows the performance comparison with the previous work [3]. Despite the large workload of the CNN, this paper maintains low-power consumption, 0.62 mW to evaluate one face at 1 fps. While the previous work shows 81% accuracy with 32 classes in LFW dataset, the use of the CNN in this paper enables 97% accuracy with whole classes in LFW dataset.

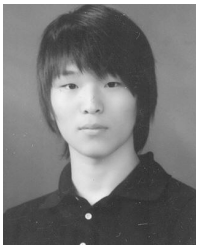
V. CONCLUSION

The functional CIS integrated with always-on FD and the CNNP are proposed to realize always-on FR for the user authentication of smart devices. In the functional CIS, the energy efficiency of the FD is improved by 39% with the help of the two-stage FD using the AHFC and the DHFU. The CNNP adopts the SFA to reduce the workload of convolutional layers by 2×–3× times with less than 1% accuracy degradation, and the T-SRAM enables efficient local SRAM access to lower the power consumption. While the functional CIS dissipates 24 to 96 μW at 1-fps framerate, the CNNP achieves 0.6-mJ per one face evaluation at its MEP. In conclusion, the low-power always-on FR based on high-accuracy CNN is implemented to realize more interactive and intelligent smart devices.

REFERENCES

[1] MyMe. OrCam, NY, USA. Accessed: Nov. 8, 2017. [Online]. Available: http://www.orcam.com/myme/
[2] D. Evans, "The Internet of Things—How the next evolution of the internet is changing everything," White Paper, Cisco IBSG, Apr. 2011.
[3] D. Jeon *et al.*, "A 23-mW face recognition processor with mostly-read 5T memory in 40-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 52, no. 6, pp. 1628–1642, Jun. 2017.

[4] J. Choi, J. Shin, D. Kang, and D.-S. Park, "Always-on CMOS image sensor for mobile and wearable devices," *IEEE J. Solid-State Circuits*, vol. 51, no. 1, pp. 130–140, Jan. 2016.
[5] K. Bong, S. Choi, C. Kim, S. Kang, Y. Kim, and H.-J. Yoo, "A 0.62 mW ultra-low-power convolutional-neural-network face-recognition processor and a CIS integrated with always-on Haar-like face detector," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2017, pp. 248–249.
[6] Y. Taigman *et al.*, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1701–1708.
[7] M. Jaderberg *et al.* (May 2014). "Speeding up convolutional neural networks with low rank expansions." [Online]. Available: https://arxiv.org/abs/1405.3866
[8] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Dec. 2001, pp. 511–518.
[9] K. Bong, I. Hong, G. Kim, and H.-J. Yoo, "A 0.5° error 10 mW CMOS image sensor-based gaze estimation processor," *IEEE J. Solid-State Circuits*, vol. 51, no. 4, pp. 1032–1040, Apr. 2016.
[10] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep. 07-49, Oct. 2007.
[11] M.-F. Chang *et al.*, "A sub-0.3 V area-efficient L-shaped 7T SRAM with read bitline swing expansion schemes based on boosted read-bitline, asymmetric-V_{TH} read-port, and offset cell VDD biasing techniques," *IEEE J. Solid-State Circuits*, vol. 48, no. 10, pp. 2558–2569, Oct. 2013.
[12] OpenCV. [Online]. Available: http://opencv.org



Kyeongryeol Bong (S'12) received the B.S. and M.S. degrees from the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2012 and 2015, respectively, where he is currently pursuing the Ph.D. degree.

His current research interests include low-power vision system-on-chips, especially focused on deep neural network accelerators and functional CMOS image sensors.

Sungpill Choi (S'13–M'15) received the B.S. and M.S. degrees from the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2012 and 2014, respectively, where he is currently pursuing the Ph.D. degree.

His current research interests include low-power and memory-efficient architecture design for mobile vision system-on-chip, especially object segmentation, stereoscopic, and deep learning.

Changhyeon Kim (S'16) received the B.S. and M.S degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2014 and 2016, respectively, where he is currently pursuing the Ph.D. degree.

His current research interests include low-power system-on-chip design, especially focused on parallel processor for artificial intelligence and machine learning algorithms.

Donghyeon Han (S'17) received the B.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2017, where he is currently pursuing the M.S. degree.

His current research interests include low-power system-on-chip design, especially focused on deep neural network accelerators and hardware-friendly algorithms for deep learning.

Hoi-Jun Yoo (M'95–SM'04–F'08) graduated from the Electronic Department, Seoul National University, Seoul, South Korea, in 1983, and the M.S. and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 1985 and 1988, respectively.

Since 1998, he has been the faculty of the Department of Electrical Engineering, KAIST, where he is currently a Full Professor. From 2001 to 2005, he was the Director of the Korean System Integration and IP Authoring Research Center. From 2003 to 2005, he was the full-time Advisor to Minister of the Korea Ministry of Information and Communication and the National Project Manager for system-on-chip (SoC) and computer. In 2007, he founded the System Design Innovation and Application Research Center, KAIST. Since 2010, he has been the General Chair of the Korean Institute of Next Generation Computing. He has co-authored *DRAM Design* (South Korea: Hongrung, 1996), *High Performance DRAM* (South Korea: Sigma, 1999), *Future Memory: FRAM* (South Korea: Sigma, 2000), *Networks on Chips* (Morgan Kaufmann, 2006), *Low-Power NoC for High-Performance SoC Design* (CRC Press, 2008), *Circuits at the Nanoscale* (CRC Press, 2009), *Embedded Memories for Nano-Scale VLSIs* (Springer, 2009), *Mobile 3-D Graphics SoC from Algorithm to Chip* (Wiley, 2010), *Bio-Medical CMOS ICs* (Springer, 2011), *Embedded Systems* (Wiley, 2012), and *Ultra-Low-Power Short-Range Radios* (Springer, 2015). His current research interests include computer vision SoC, body area networks, and biomedical devices and circuits.

Prof. Yoo served as a member for the Executive Committee of ISSCC, Symposium on VLSI, and A-SSCC, the TPC Chair for the A-SSCC 2008 and ISWC 2010, the IEEE Distinguished Lecturer from 2010 to 2011, the Far East Chair for the ISSCC from 2011 to 2012, the Technology Direction Sub-Committee Chair for the ISSCC in 2013, the TPC Vice Chair for the ISSCC in 2014, and the TPC Chair for the ISSCC in 2015. He received the Electronic Industrial Association of Korea Award for his contribution to DRAM technology in 1994, the Hynix Development Award in 1995, the Korea Semiconductor Industry Association Award in 2002, the Best Research of KAIST Award in 2007, the Scientist/Engineer of this month Award from the Ministry of Education, Science and Technology of Korea in 2010, the Best Scholarship Awards of KAIST in 2011, and the Order of Service Merit from Ministry of Public Administration and Security of Korea in 2011, and has been a co-recipient of the ASP-DAC Design Award 2001, the Outstanding Design Awards of 2005, 2006, 2007, 2010, 2011, and 2014 A-SSCC, the Student Design Contest Award of 2007, 2008, 2010, and 2011 DAC/ISSCC.