A Study on a Feature Extractable CMOS Image Sensor for On-Chip Image Classification

Shunsuke OKURA*, Yudai MORIKAKU*, Yu Osuka*, Ryuichi UJIIE†,
Daisuke MORIKAWA†, Hideki SHIMA†, and Kota YOSHIDA*
*Ritsumeikan University, 1-1-1 Noji-higashi, Kusatsu, Shiga, 525-8577 JAPAN
Phone: +81-77-599-3149, Email: sokura@fc.ritsumei.ac.jp
†Nisshinbo Micro Devices Inc.

***Abstract*—In the emerging IoT era, a CMOS image sensor (CIS) that can output features required for AI recognition is expected to reduce the power consumption of image recognition systems. In this paper, we propose an on-chip signal processing pipeline to extract two channels of feature data composed of 1-bit grayscale intensity and horizontal edges for a 2.0T pixel CMOS image sensor. Additionally, a tiny neural network model for on-chip AI processing is verified to examine image classification accuracy. According to simulation results, the image classification accuracy with the feature data reached 79.8%, even though the effective data volume of the feature data is reduced to only 0.7% of that of the 8-bit RGB color image. The memory occupation flow for a 2 Mpixel 60 fps CIS is also estimated to consider the feasibility of implementing an on-chip image classifier. Around 60 kB of peak memory and weight memory are additionally required for the processing, which is only 17 times the line memory required to serialize the RGB color image.**

***Index Terms*—Image classification, Artificial neural networks, CMOS image sensors, Feature extraction**

I. Introduction

In the emerging IoT era, artificial intelligence (AI) in cyberspace analyzes sensor data from the physical space and provides feedback to the physical domain. Since conventional image sensors often output redundant data that AI removes in feature extraction, image sensors that can output lightweight feature data are exptected to reduce the power consumption of image recognition systems. A log-gradient QVGA image sensor [1] outputs feature data for histograms of oriented gradients (HOGs). However, conventional RGB color photographic images cannot be output because the readout circuit is specific to HOG feature extraction. Feature-extractable image sensors [2], [3] can output RGB color images and convolution-filtered features. However, an in-pixel capacitor [2] or column counter [3] is additionally required for the feature extraction process. Event-based vision sensors (EVSs) [4] that suppress temporal redundancy enable operation in the high dynamic range using lightweight feature data. However, EVSs cannot detect stationary objects, and the pixel size is large due to an in-pixel circuit that detects temporal luminance changes.

Our research group also focuses on a CMOS image sensor (CIS) that can output a feature required for AI recognition, thereby reducing the power consumption of the image recognition system. As shown in Fig. 1, when a specific object is detected using the feature data and an on-chip AI, the sensor switches to output conventional RGB color images for further analysis by AI in cyberspace or for verification by humans. In previous works [5], [6], an on-chip signal processing pipeline for the CMOS image sensor has been proposed to extract 1-bit horizontal edge as feature data. Besides, a lightweight and low-power deep neural network (DNN) for feature data has been verified based on YOLOv7 running on a GPU for object detection tasks. In this paper, we propose to extract two channels of feature data composed of 1-bit grayscale intensity and horizontal edges. Besides, a tiny neural network model for on-chip AI processing is verified to examine the binary classification accuracy of detecting ”person” in an image. The conceptual CMOS image sensor which can extract the two channels of feature data is proposed in Sec. II. Simulation results of image classification are presented in Sec. III, followed by discussion and future work in Sec.IV. Section V summarizes this paper.

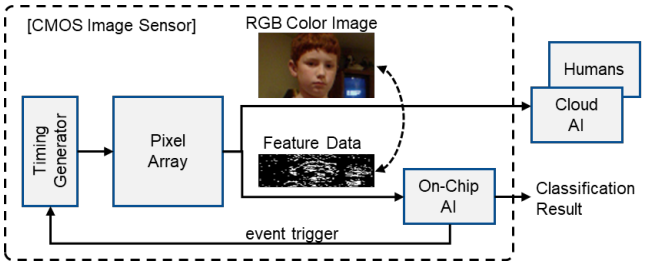


Fig. 1. Concept overview of an image recognition system with a feature extractable CMOS image sensor (Pictures adapted from [7]).

II. Proposed Image Classification System

Figure 2 shows the proposal of signal processing pipeline to extract two channels of feature data in a CIS, where it is noted that the RGB color image can be output with the intensity signal of each pixel at an 8-bit resolution, similar to conventional CIS. The feature data consists of two channels: horizontal edge in channel-1 and intensity in channel-2. The Bayer-patterned 2×2 pixel signal is converted into a grayscale signal, given by $R + 2G + B$, by pixel binning. Subsequently, vertically adjacent grayscale signals are subtracted by differential double sampling (DDS) to generate the horizontal edge, while the intensity signal is generated through correlated double sampling (CDS). By applying 1-bit quantization and horizontal binning, the data volume is

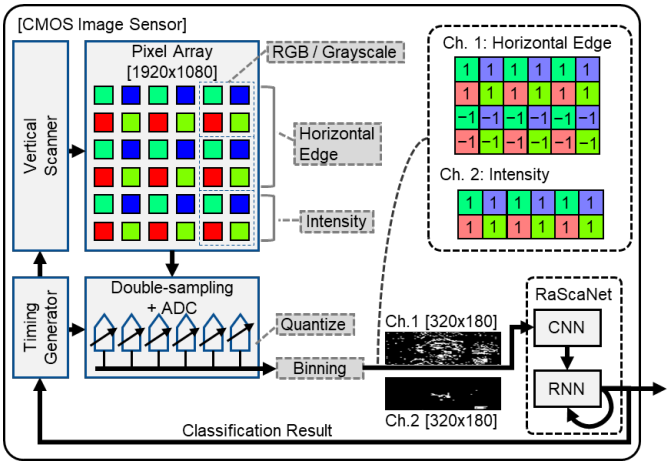


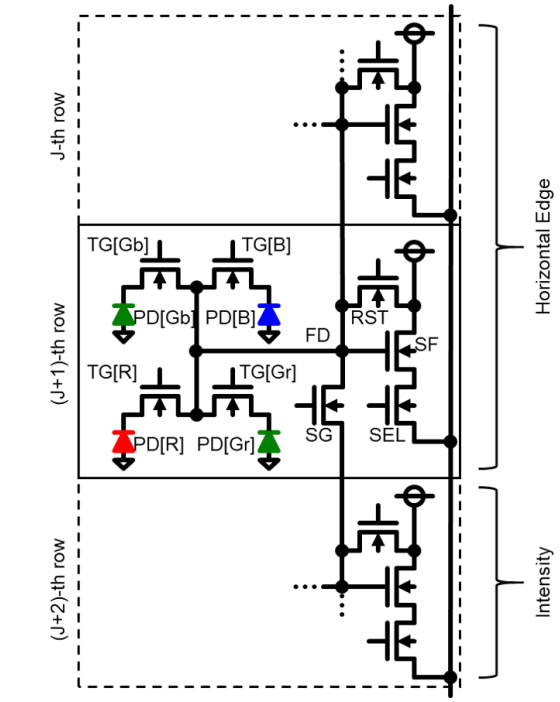
Fig. 2. Proposed on-chip image classification system (Pictures adapted from [7]).

reduced into $2 \times 320 \times 180 \times 1$-bit, that is 99.3% reduction compared to $1920 \times 1080 \times 8$-bit RGB color image, for a full HD CIS. To indicate the presence of a "person", the two-channel feature data is then directly fed into RaScaNet [8] which is composed of a convolutional neural network (CNN) and a recurrent neural network (RNN).

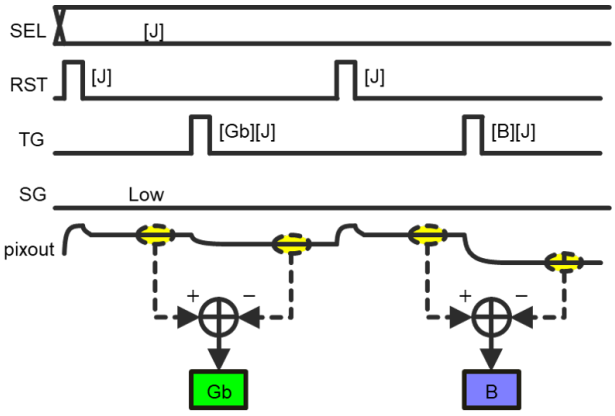
Figure 3 shows schematic and timing diagrams for the pixel array of the feature extractable CIS. The pixels are grouped by three cells composed of 2×2 pixels in a Bayer array which forms a 2.0T shared pixel as shown in Fig. 3(a). In the conventional RGB color mode shown in Fig. 3(b), TG gates for Red, Green, and Blue PDs are turned on sequentially, and the integrated photoelectrons in each PD are read out by CDS of the reset and signal. In the feature extraction mode shown in Fig. 3(c), all TG gates for Red, Green, and Blue PDs are turned on simultaneously, and the integrated photoelectrons in the PDs are summed in the FD node as a grayscale signal given by $R + 2G + B$. The horizontal edge is extracted by DDS of grayscale signals from the J- and (J+1)-th row pixel cells. It is noted that the switching gate SG in the J-th row is turned on to read the grayscale signal from the same SF transistor in the J-th row. The intensity is extracted by CDS of the reset and signal of the (J+2)-th row pixel cell. It is also noted that the SG in the (J+2)-th row is turned on to make the charge-to-voltage conversion gain coincident with that for the horizontal edge.

III. Simulation Results

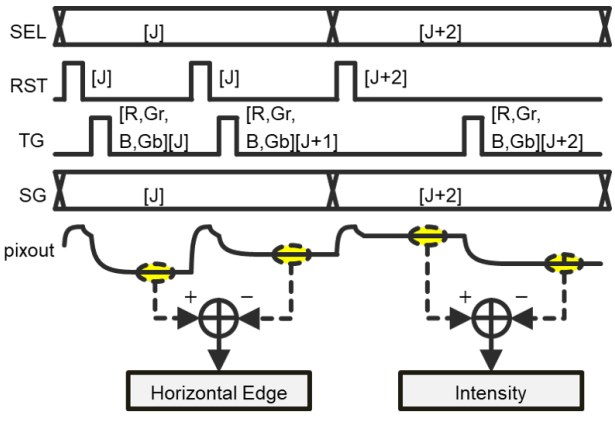
Figure 4 shows the simulation flow of image classification with the two-channel feature. The VWW dataset [7], which is mainly composed of 640×480 pixel RGB color image, is resized and then masked to generate 2 Mpixel pseudo RAW data. The pseudo RAW data is processed to simulate the proposed on-chip feature extraction processing pipeline and to generate a feature dataset. Estimated pixel reset noise residued in the DDS operation is imposed on the horizontal edge. A lightweight neural network model, RaScaNet [8], is



(a) schematic diagram



(b) timing diagram of RGB color mode



(c) timing diagram of feature extraction mode

Fig. 3. 1×3 pixel for the feature extractable CIS.

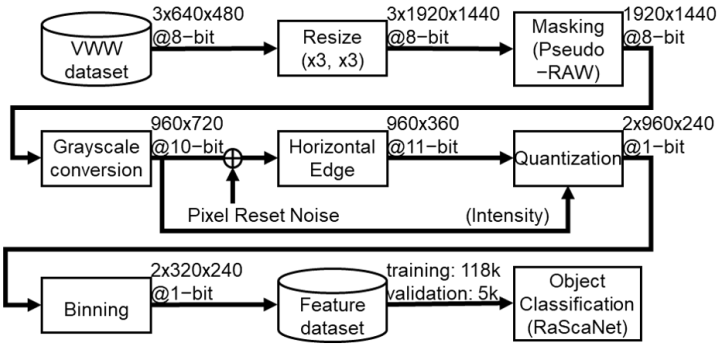


Fig. 4. Simulation flow of image classification with the two-channel feature.

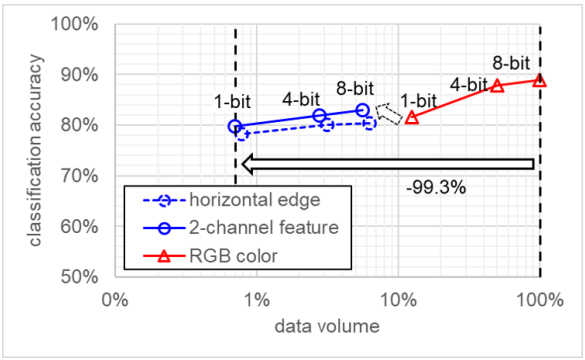


Fig. 6. Data volume and image recognition accuracy

(a) 1-bit horizontal edge

(b) 1-bit intensity

(c) Two channel feature data

(d) 8-bit RGB color image

Fig. 5. Sample images generated from [7]

TABLE I
WEIGHT PARAMETERS SIZE OF TRAINED NEURAL NETWORK

	This work (RaScaNet)	DNN (YOLOv7 [9])
Papameter size	31.8 kB	71.3 MB

trained with $118,000$ training feature data and then validated for accuracy with $5,000$ feature data.

Figure 5 shows sample images of the feature data, wherein Fig. 5(d) presents the original RGB color image [7]. The combination of the horizontal edge and intensity, shown in Fig. 5(c), enhances object visibility compared to using the channels separately, as shown in Fig. 5(a) and 5(b). In the 1-bit horizontal edge, the outlines of persons can be distinguished, while the noise in the dark background is considerable large. On the other hand, in the 1-bit intensity, the bright foreground and dark background can be distinguished, while the outlines of persons are not clear. Thus, improvement in image classification accuracy is expected with the combination of the horizontal edge and intensity, even with aggressive quantization to 1-bit.

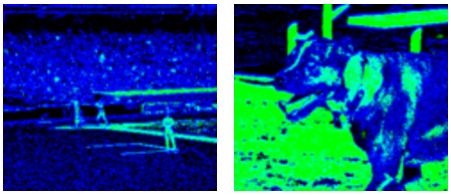
Figure 6 presents simulation results for image classification, where the feature data are quantized to 8-bit, 4-bit, and 1-bit. For reference, the RGB color images are also tested. Additionally, the single-channel 480×360 pixel horizontal edge [5] is tested, whose data volume is 50% larger than that of the two channel 320×180 pixel combination of the horizontal edge and intensity. When the two channels of feature data are quantized to 1-bit, the effective data volume input to RaScaNet is reduced to only 0.7% of that of the 8-bit RGB color image. Despite this significant reduction, the image classification accuracy reached 79.8%. Furthermore, the accuracy curve of the feature data demonstrates greater robustness to data quantization and exhibits a higher accuracy trend compared to that of the RGB color image. It is also confirmed that the image classification accuracy is improved by 1.5% at 1-bit with the addition of intensity to the horizontal edge.

Table I shows the weight parameter size. The weight memory of the trained RaScaNet is 31.8 kB, which is around 0.04% of a popular Deep Neural Network (DNN), YOLOv7 [9].

IV. DISCUSSION

Figure 7 shows example images for each classification result, where true positive (TP) and true negative (TN) indicate that the result is correct, and false positive (FP) and false negative (FN) indicate that the result is incorrect. Persons in Fig. 7(a) and a bear in Fig. 7(d) are correctly classified as "person" and "not," respectively, while persons in a bus shown in Fig. 7(c) and a dog in Fig. 7(b) are misclassified. Regarding the persons in a bus, it is supposed that the object size is too small for classification with the feature data. Regarding the dog, it is supposed that RaScaNet might have been trained to classify the dog as a person because many training data include dogs accompanied by persons.

Figure 8 assumes the memory occupation flow for a 2 Mpixel 60 fps CIS. In the 8-bit RGB color image mode, pixel data is stored in 1.9 kB of one line memory and is then output through a high-speed serial interface at around 1Gbps, as shown in Fig. 8(a). In the feature extraction mode shown in Fig. 8(b), the spacial resolution is reduced to 2 channels of 320×180 pixels. A 0.4 kB memory is sufficient for image classification with the RaScaNet which processes every 5 line



(a) True positive (confidence score: 1.00)

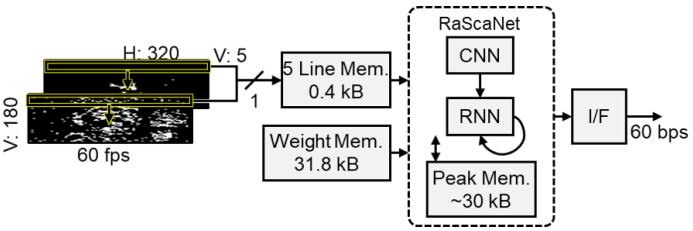
(b) False positive (confidence score: 0.660)

(c) False negative (confidence score: 0.999)

(d) True negative (confidence score: 0.991)

Fig. 7. Sample images for classification results (Pictures from [7])

(a) 8 bit RGB color image



(b) 1-bit feature data

Fig. 8. Memory usage flow (Pictures adapted from [7]).

feature data. Around 60 kB of peak memory and weight memory are additionally required for the processing, which is only 17 times the line memory for the RGB color mode. Based on this simple estimation of memory usage, implementing an on-chip image classifier is considered feasible.

As future work, we will reduce the additional memory to a few kB. This is because the 32-bit depth of the weight parameter is likely redundant for 1-bit feature data and can be quantized. The peak memory is also expected to be reduced because the quantized weight parameter and 1-bit feature data result in simple combinational logic circuits for multiply-accumulate (MAC) operations in the CNN and RNN. The design of a low-power readout circuit for the 1-bit feature data and a low-power interface (I/F) circuit for 60 bps classification results are also part of our future work.

V. Summary

We have proposed the on-chip signal processing pipeline for the CIS to extract two channels of feature data composed of 1-bit grayscale intensity and horizontal edge. The pixel array is composed of a simple 2.0T shared pixel, where the conventional RGB color image and the two channels of feature data are read out by changing the pixel control pulse. Simulation results to examine image classification accuracy with the proposed feature data shows that the accuracy reached 79.8%, even though the data volume of the feature data is reduced to only 0.7% of that of the 8-bit RGB color image. According to the feasibility study of memory occupation flow for a 2 Mpixel 60 fps CIS, around 60 kB of peak memory and weight memory are additionally required for the image classification, which is only 17 times the line memory required to serialize the RGB color image.

References

[1] C. Young, A. Omid-Zohoor, P. Lajevardi, and B. Murmann, "A data-compressive 1.5/2.75-bit log-gradient qvga image sensor with multi-scale readout for always-on object detection," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 11, pp. 2932–2946, 2019.

[2] S. Yoneda, Y. Negoro, H. Kobayashi, K. Nei, T. Takeuchi, M. Oota, T. Kawata, T. Ikeda, and S. Yamazaki, "Image sensor capable of analog convolution for real-time image recognition system using crystalline oxide semiconductor fet," in *International Image Sensor Workshop (IISW)*, pp. 322–325, 2019.

[3] M.-J. Park and H.-J. Kim, "A real-time edge-detection cmos image sensor for machine vision applications," *IEEE Sensors Journal*, vol. 23, no. 9, pp. 9254–9261, 2023.

[4] T. Finateu, A. Niwa, D. Matolin, K. Tsuchimoto, A. Mascheroni, E. Reynaud, P. Mostafalu, F. Brady, L. Chotard, F. LeGoff, H. Takahashi, H. Wakabayashi, Y. Oike, and C. Posch, "5.10 a 1280×720 back-illuminated stacked temporal contrast event-based vision sensor with 4.86µm pixels, 1.066geps readout, programmable event-rate controller and compressive data-formatting pipeline," in *2020 IEEE International Solid-State Circuits Conference - (ISSCC)*, pp. 112–114, 2020.

[5] Y. Morikaku, R. Ujiie, D. Morikawa, H. Shima, K. Yoshida, and S. Okura, "On-chip data reduction and object detection for a feature-extractable cmos image sensor," *Electronics*, vol. 13, no. 21, 2024.

[6] K. Kuroda, Y. Morikaku, Y. Osuka, R. Iegaki, K. Yoshida, and S. Okura, "Lightweight object detection model for a cmos image sensor with binary feature extraction," in *2024 IEEE SENSORS*, pp. 1–4, 2024.

[7] A. Chowdhery, P. Warden, J. Shlens, A. Howard, and R. Rhodes, "Visual wake words dataset," 2019.

[8] J. Yoo, D. Lee, C. Son, S. Jung, B. Yoo, C. Choi, J.-J. Han, and B. Han, "Rascanet: Learning tiny models by raster-scanning images," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13668–13677, 2021.

[9] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7464–7475, 2023.