# A Convolutional Neural Network VLSI for Image Recognition Using Merged/Mixed Analog-Digital Architecture

Keisuke Korekado[1], Takashi Morie[1], Osamu Nomura[2], Hiroshi Ando[3],
Teppei Nakano[1], Masakazu Matsugu[2], and Atsushi Iwata[3]

[1] Graduate School of Life Science and Systems Engineering,
Kyushu Institute of Technology, Kitakyushu, 808-0196, Japan
{korekado-keisuke@edu., morie@,
nakano-teppei@edu.}brain.kyutech.ac.jp
http://www.brain.kyutech.ac.jp/˜morie
[2] Canon Research Center,
Atsugi, 243-0193, Japan
{nomura.osamu, matsugu.masakazu}@canon.co.jp
[3] Graduate School of Advanced Sciences of Matter, Hiroshima University,
Higashi-Hiroshima, 739-8526, Japan
{ando, iwa}@dsl.hiroshima-u.ac.jp

**Abstract.** Hierarchical convolutional neural networks are a well-known robust image-recognition model. In order to apply this model to robot vision or various intelligent vision systems, its VLSI implementation with high performance and low power consumption is required. This paper proposes a convolutional network VLSI architecture using a hybrid approach composed of pulse-width modulation (PWM) and digital circuits. We call this approach merged/mixed analog-digital architecture. The VLSI includes PWM neuron circuits, PWM/digital converters, digital adder-subtracters, and digital memory. We have designed and fabricated a VLSI chip by using a 0.35 $\mu$m CMOS process. The VLSI chip can perform 6-bit precision convolution calculations for an image of 100×100 pixels with a receptive field area of up to 20×20 pixels within 5 ms, which means a performance of 2 GOPS. Power consumption of PWM neuron circuits is estimated to be 20 mW. We have verified successful operations using a fabricated VLSI chip.

## 1 Introduction

For object detection or recognition from natural images, processing models for extracting image features should tolerate pattern deformations and pattern position shifts. Convolutional neural networks with a hierarchical structure, which imitate the vision nerve system in the brain, have such functions [1–3].

The operations required for implementing convolutional networks are multiplication by weights and nonlinear conversion, as usual neural network models. Because they require huge computational power, to execute these operations in real-time and with low power consumption for intelligent applications such as robot vision, efficient VLSI implementation is required. Various neural network VLSIs have actively been developed, and an analog VLSI processor suitable for convolutional networks was also reported [4].

On the other hand, we have already proposed a new circuit architecture, which is based on a pulse-width modulation (PWM) approach merging analog and digital approaches [5]. This architecture has various advantages of both approaches, especially it achieves low power consumption, and it is suitable for implementing neural networks.

In this paper, by combining this merged analog-digital architecture with the digital approach, we propose a convolutional network VLSI architecture that consists of PWM neuron circuits and digital memory. We also present the measurement results of a VLSI chip fabricated using a 0.35 $\mu$m CMOS process.

## 2 Hierarchical Convolutional Network Model

Figure 1 shows the principle of pattern detection using a convolutional network. The first layer of the hierarchical structure only receives images. The following layers consist of two sub-layers: a feature detection (FD) layer and a feature pooling (FP) layer. Each layer includes some feature classes, each of which has neurons that react the same image feature. The neurons are arranged in a 2-D array to maintain the feature position of the input image. Therefore, the feature class pixel size is equal to the input image pixel size, and each neuron corresponds to each pixel. All neurons are connected to the neurons in a predefined area near the same position of the previous layer, which is called a receptive field. The FP neurons are used to achieve recognition tolerant to pattern deformation and position shifts. The FD neurons operate for integrating a feature. By the hierarchically repetitive structure, local simple features (e.g., line segments) of the input image are gradually assembled into complex features.

Operations between layers are considered as a convolution because all neurons belonging to a feature class have a receptive field with the same weight distribution. The receptive field of the FP neuron is on the same feature class of the previous FD layer. All neurons of the FP layer have the same positive weight distribution, in which the weight is largest in the center of the receptive field and it decreases as the position is apart from the center. The shifts of feature positions in the FD layers are tolerated in the FP layers by this weight distribution. On the other hand, the receptive fields of the FD neurons are on all feature classes of the previous FP layer. The weights of the FD neurons are obtained by training.

## 3 Convolutional Network VLSI Architecture

We propose a VLSI architecture that implements the hierarchical convolutional networks. Because the number of processing circuits integrated in a chip is restricted, it is difficult to realize all connections of the hierarchical network by real processing circuits. Therefore, in our architecture, neuron circuits are repetitively used by time-sharing operation.

Time-sharing operation in the convolutional network is shown in Fig. 2. The feature class size and the receptive field size are assumed $N \times N$ and $m \times m$ pixels, respectively. The outputs of $N$ neurons belonging to one column of a feature class are inputted to

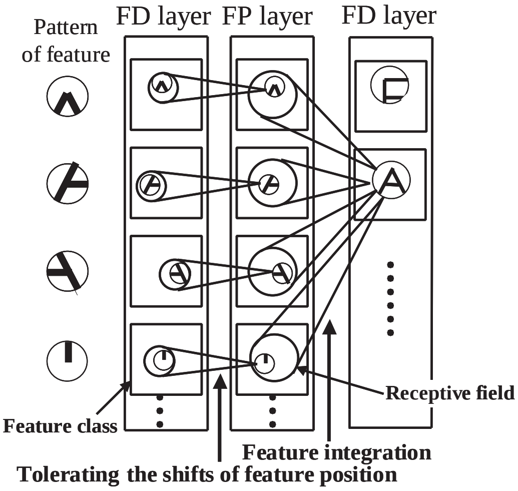

**Fig. 1.** Principle of pattern recognition using a convolutional network (an example of letter recognition)

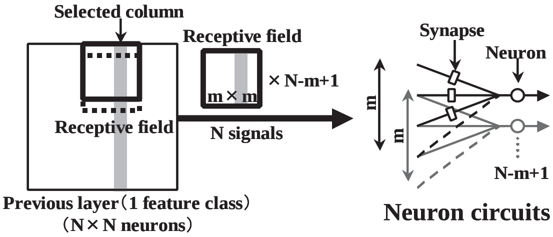

**Fig. 2.** Time-sharing operation in the convolutional network

$(N - m + 1)$ neuron circuits simultaneously. The neuron operations for one column of the receptive fields are performed in parallel. For $m$ rows in a receptive fields, the above neuron operations are repeated by $m$ times, and furthermore each of them is repeated twice for positive and negative weighting. Thus, the number of repetitions in $(N - m + 1)$-parallel neuron operations for convolution between feature classes is $(N - m + 1) \times m \times 2$.

The block diagram of our convolutional network circuit is shown Fig. 3. By utilizing the advantage of small circuit size in the PWM approach, $m$-input PWM neuron circuits are integrated. To achieve time-sharing operation, the partial accumulation results of neuron operation are temporarily held in the neuron circuit. These partial results are accumulated and stored in an SRAM through the PWM/digital converter (WDC) and the digital adder-subtracter (DAS). The WDC converts PWM signals output from a neuron circuit into digital signals. The DAS is used in time-sharing operations for one column of the receptive field and for the positive and negative weighting.

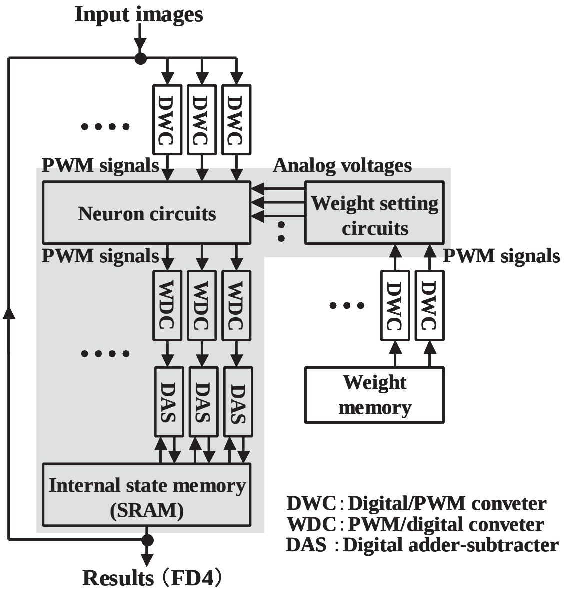

**Fig. 3.** Block diagram of our convolutional network circuit. The components in the shaded region are included in our VLSI chip

Although we assumed that the number of inputs of the neuron circuits is $m \times m$, convolution with a smaller receptive field size can be calculated by setting the extra inputs at zero. Convolution with a larger receptive field size can also be calculated by time-sharing operation.

## 4 PWM Neuron Circuit

### 4.1 Connection Model

In the general feedforward networks, internal state $u_i$ and output $o_i$ of postsynaptic neuron $i$ are given by the following equations, respectively;

$$u_i = \sum_j w_{ij} o_j \ , \tag{1}$$

$$o_i = f(u_i) \ , \tag{2}$$

where $w_{ij}$ is the connection weight from presynaptic neuron $j$ to postsynaptic neuron $i$, and $f$ is the nonlinear conversion function.

In the conventional model, the synapse part multiplies $o_j$ by $w_{ij}$ and the neuron (soma) part executes summation and nonlinear conversion $f(u_i)$. From eqs. (1) and (2), the output of postsynaptic neuron $i$ is given by

$$o_i = f(\sum_j w_{ij} o_j) \ . \tag{3}$$

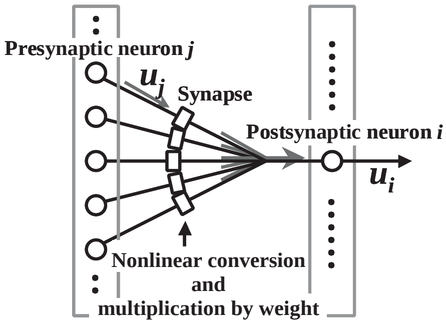

**Fig. 4.** Connection model of our neuron circuit

However, in our circuit model, both nonlinear conversion and multiplication are performed by the synapse part, and the neuron part executes summation and outputs the internal state, as shown in Fig. 4. Thus, from eqs. (1) and (2), the internal state which is the output of neuron $i$ is given by

$$u_i = \sum_j w_{ij} f(u_j) \ . \qquad (4)$$

Equations (3) and (4) are the equivalent operations in hierarchical networks.

### 4.2 Circuit Design

Our PWM neuron circuit is shown in Fig. 5. Its operation is as follows: (1) A PWM signal $P_i$ that corresponds to the internal state of the presynaptic neuron is transmitted to the synapse part; (2) the input PWM signals are converted with the nonlinear function, and weighted summation is performed by converting the PWM signals into charges stored in capacitor $C_N$; (3) the voltage between the nodes of the capacitor, $V_N$, is converted into a PWM signal by comparing it with linearly-ramped voltage signal $V_{ref}$.

In this circuit, nonlinear conversion and multiplication are performed by two MOSFETs, M1 and M2, at the same time. The nonlinear function is applied to all synapses by changing analog voltage $V_F$. The connection weighting is achieved by applying analog DC voltage $V_W$. When $V_F$ slightly exceeds the threshold voltage of M1, both MOSFETs M1 and M2 operate in the saturation region, and thus current $I_S$ flowing to capacitor $C_N$ mainly depends on $V_F$. When $V_F$ becomes lower, M1 operates in the triode region, and M2 still operates in the saturation region, thus current $I_S$ mainly depends on $V_W$. Thus, weighting and nonlinear conversion $w_{ij} \cdot f(u_j)$ are achieved at the synapse part.

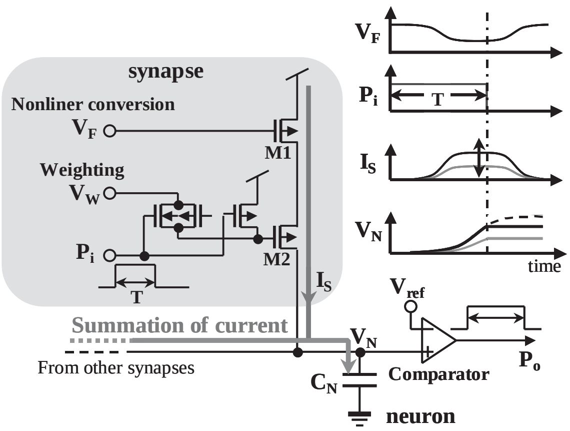

**Fig. 5.** PWM neuron circuit

## 5 Experimental Results Using a Fabricated VLSI Chip

We fabricated a convolutional network VLSI by using a 0.35 $\mu$m CMOS process.[4] Figure 6 shows a micro-photograph of the fabricated chip. The VLSI chip includes 81 neuron circuits with 20 synaptic inputs, 81 PWM/digital converters, 81 digital adder-subtracters, 39 kb SRAM, and 20 weight setting circuits. Therefore, this chip can implement a convolutional network with $N = 100$ and $m = 20$. By using this VLSI chip with external feedback control, we can construct a hierarchical convolutional network.

We measured the PWM input-output relationship of neuron circuits when all of 20 PWM input signals per neuron are identical. The measurement results are shown in Fig. 7 with the corresponding circuit simulation (HSPICE) results. The measurement results agree well with the simulation results, and it is demonstrated that weighting and nonlinear conversion are achieved simultaneously.

Since we defined the operation cycle time as 1.6 $\mu$s, the whole convolution operation requires about 5 ms. This chip achieves an operation performance of 2 GOPS[5] by parallel operations for 81$(= N-m+1)$ neurons and 1620$(= (N-m+1)m)$ synapses. We have estimated a power consumption of PWM neuron circuits to be 20 mW although the digital circuit block consumes 190 mW.

We have verified that all circuit components operate successfully. The whole operation for a convolutional network has also been verified.

[4] The VLSI chip has been fabricated in the chip fabrication program of VLSI Design and Education Center(VDEC), the University of Tokyo, Japan with the collaboration by Rohm Corporation and Toppan Printing Corporation.

[5] Giga Operations (multiplications and summations) Per Second.

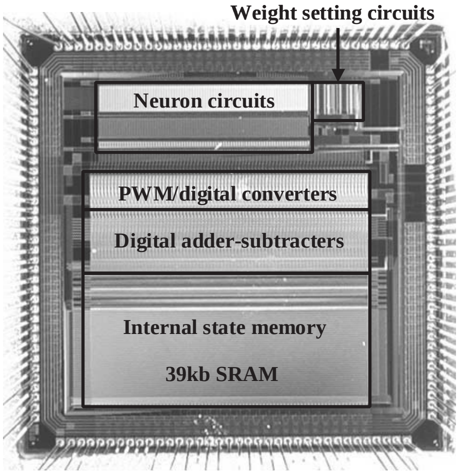

**Fig. 6.** Chip micro-photograph

## 6 Conclusion

We proposed a merged/mixed analog-digital VLSI architecture for convolutional neural networks using PWM and digital circuit techniques.

A neuron circuit with 20 synapses was designed. Nonlinear conversion and multiplications by connection weights are realized by two MOSFETs, thus a very small layout area and low power consumption of the synapse part were achieved. Since the connections between layers have the same weight distribution, hierarchical networks can be constructed by feedback and time-sharing operations using the convolutional network VLSI.

We designed and fabricated a convolutional network VLSI with an operation performance of 2 GOPS, and verified successful operations of all circuit components.

## References

1. Fukushima, K., Miyake, S.: Neocognitron: A New Algorithm for Pattern Recognition Tolerant of Deformations and Shifts in Position. Pattern Recognition **15** (1982) 455–469
2. Lawrence, S., Giles, C.L., Tsoi, A.C., Back, A.D.: Face Recognition: A Convolutional Neural-Network Approach. IEEE Trans. Neural Networks **8** (1997) 98–113
3. Matsugu, M., Mori, K., Ishii, M., Mitarai, Y.: Convolutional Spiking Neural Network Model for Robust Face Detection. In: Proc. Int. Conf. on Neural Information Processing (ICONIP) (2002) 660–664
4. Boser, B.E., Säckinger, E., Bromley, J., Le Cun, Y., Jackel, L.D.: An Analog Neural Network Processor with Programmable Topology. IEEE J. Solid-State Circuits **26** (1991) 2017–2025
5. Iwata, A., Morie, T., Nagata, M.: Merged Analog-Digital Circuits Using Pulse Modulation for Intelligent SoC Applications. IEICE Trans. Fundamentals. **E84-A** (2001) 486–496

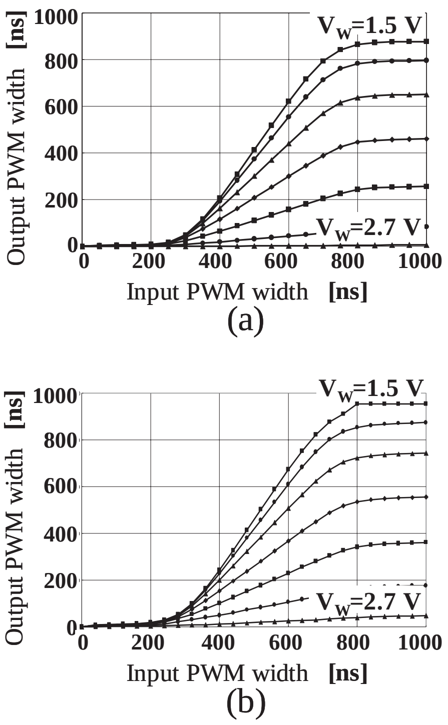

**Fig. 7.** PWM input-output relationship: (a) circuit simulation (HSPICE) results, and (b) measurement results