Fully Memristive SNNs with Temporal Coding for Fast and Low-power Edge Computing

Xumeng Zhang[†1,3], Zuheng Wu[†2], Jikai Lu[2,4], Jinsong Wei[2,4], Jian Lu[2,4], Jiaxue Zhu[2], Jie Qiu[4], Rui Wang[2], Kaihua Lou[2], Yongzhou Wang[2], Tuo Shi[2,4], Chunmeng Dou[2], Dashan Shang[2], Qi Liu[*1,3], and Ming Liu[*1,2,3]

[1]Frontier Institute of Chip and System, Fudan University, Shanghai, 200433, China; [2] Key Laboratory of Microelectronics Devices and Integrated Technology, Institute of Microelectronics, Chinese Academy of Sciences, Beijing100029, China; [3]School of Microelectronics, Fudan University, Shanghai, China, [4]Zhejiang Laboratory, Hangzhou 311122, China. [*]E-mail: liuqi@ime.ac.cn; liuming@ime.ac.cn.

2020 IEEE International Electron Devices Meeting (IEDM) | 978-1-7281-8888-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/IEDM13553.2020.9371937

Abstract- SNNs with temporal coding (TC), inspired by the human visual system, have a powerful ability to enable fast and low-power neuromorphic computing. Memristive devices show excellent performance on emulating spiking neurons and synapses in hardware. However, the neuron circuits used for implementing a fully memristive TC SNN are absent. In this work, for the first time, we demonstrate a LIF neuron based on a NbO_x device to meet the requirements for the hardware implementation of TC SNNs. The neuron fires at most one spike within an inference window, and its spiking latency inverse to the input current intensity. Using such a neuron, we further experimentally demonstrated a fully memristive TC SNN (256 × 5) to recognize the *Olivetti* face patterns. Attributing to the one-spike scheme, the TC SNN achieves a sparser spiking number (~ 72 × reductions), faster inference speed (> 1.5 × improvement), lower power (~ 53 × reductions) than what happens in rate-coding SNNs.

I. Introduction

In the IoT era, spiking neural networks (SNNs) with TC is one of the powerful platforms for enabling low-power edge computing due to their event-driven feature and temporal information processing (**Fig. 1**). Also, attributing to the one-spike scheme, the TC SNNs are hardware friendly and efficient [1]. Recently, emerging technologies, such as Resistive Memories (RRAMs) [2], Phase Change Materials (PCMs) [3], Ferroelectric FET (FeFET) [4], et al., have shown remarkable performance over CMOS technology on emulating spiking neurons and plastic synapses (Fig. 1(c)). However, the reported works with emerging neurons focus on implementing rate coding (RC) SNNs, which requires a large number of spikes to express the intensity of one input data (Fig. 1(d)). This implementation poses a significant challenge on device endurance and features less bio-plausibility. Also, the RC SNNs need a long inference time to average the spiking rates to make decisions, which induces more latency and power consumption. Thus, a study to construct a compact neuron circuit using emerging devices for implementing TC SNNs deserves more attention.

In this work, we designed a leaky integrate-and-fire (LIF) neuron based on NbO_x devices for constructing TC SNNs (**Table I**). The neuron could encode the input intensity into spiking latency and fires at most once within an inference window. The current-driven firing scheme avoids offset operations and supports a low inference voltage. Then, based on these neurons and RRAM synapses, we experimentally demonstrated a TC SNN for face recognition. Compared to the RC SNNs, the TC SNN shows lower latency, lower power, and sparser spiking number (indicating a longer lifetime of the NbO_x neuron). Moreover, the evaluated energy efficiency is up to 20.1 TMACS/W under a 10 ns time-step. These results demonstrate that our neuron could enable faster and low-power TC SNNs and have great potential on edge applications.

II. Mott Neurons and Memristive Synapses

A. NbO_x TS device and TaO_x/HfO_x synapses

Emerging devices with volatile threshold switching (TS) behavior is favorable for constructing spiking neurons [2],[5-7]. Among various TS devices, the NbO_x device features high uniformity, high endurance (> 10^{12}), and superior thermal stability (~808□) for on-chip integration [8]. Thus we select the NbO_x device to demonstrate the neuron circuit in this work. **Fig. 2** shows the SEM image of the fabricated array that consists of 32 discrete TiN(40nm)/NbO_x (50 nm)/TiN(40nm) devices. After forming, a stable TS behavior is observed (**Fig. 3**). Then, we tested the V_{TH} fluctuation under 10^4 cycles by an oscillator circuit [9], obtaining < 5% error (**Fig. 4),** and the distribution follows a gaussian curve. **Fig. 5** illustrates the compact V_{TH} distribution of five chosen devices, indicating the feasibility of emulating high-precision neurons. A 64 × 64 1T1R RRAM array with the structure of TiN/TaO_x/HfO_x/TiN was fabricated to serve as the synapses storing the weights.

B. Neuron circuits design for TC SNNs

Based on the characteristics of the NbO_x device, we design a neuron circuit used for TC SNNs (**Fig. 6**). First, an RC oscillator neuron is built, then over which a D-flipflop and a transfer gate are introduced to make the neuron circuit fire only once. The D-flipflop generates a long refractory period signal after detecting the first fire signal. Moreover, the transfer gate blocks the post-fire inputs, as well as saving additional energy consumption. Attributing to the self-feedback connection of the D-flipflop, the neuron circuit is purely asynchronous and needs no clock signal. To further enable fast pattern-to-pattern inputs without much delay, we adopt an NMOS to accelerate the leaky process after firing (inset of Fig. 7 shows a leaky process without NMOS). **Fig. 7** illustrates the outputs of the neuron under a series of current pulses (100 ns width, 50 ns interval) with different intensity. Under each input intensity, the neuron circuit fires only once, and the spiking latency decreases with increasing the input intensity (**Fig. 8**). **Fig. 9** shows the comparison of the spiking behavior of our neuron and the RC neuron under a 470 pF

capacitor. The results show that the TC neuron could work under a wider (3 ×) synaptic current range, which means that the TC neuron could support a more massive synaptic array. Due to the one-spike scheme, our neuron features a better performance on both the latency and energy consumption (**Fig.10**). These gaps increase with increasing the synaptic input current, which means the promising merits of our neuron for implementing hardware SNNs.

To further evaluate the ultimate integration time and spike energy, we studied the spiking behavior of the neuron under different capacitors **(Fig. 11(a))**. Then, we extracted out the functional curve between the integration time and capacitance under different inputs current (**Fig. 11(b)**). The results show that when the integration capacitor is larger than 10 fF, the capacitor integration time ($T_{integration}$) dominates the total integration time ($T_{integration} + T_{switch}$). Otherwise, the switching time (T_{switch}) of the device (~1 ns) dominates. **Fig. 12** shows the spiking energy of the neuron under different input current and capacitors. A smaller capacitor allows a lower energy consumption per spike due to the short integration process. With decreasing the capacitors, the energy consumption will be dominated by the switching energy of the device (E_{switch}) and the D-flipflop (E_D) eventually. In that case, we obtain an ultra-low energy consumption of the neuron circuit (~500fJ/spike). Furthermore, the integration time also depends on the interval time between two input pulses (**Fig. 13**). Then, two input current trains with different temporal forms but equal total current intensity are poured into the neuron circuit (**Fig. 14**). Different spiking behavior is observed, indicating that the neuron circuit could successfully process the spatiotemporal signals.

Then, we propose an optimized LIF model for training the TC SNNs (**Fig. 15**), in which the V_{TH} stochasticity and nonlinear leaky process are introduced. Then, we train an SNN with this model. During training, a supervised temporal backpropagation algorithm is performed [10]. Fifty down-sampled *Olivetti* face patterns [11] (16 × 16) from five persons are used as the inputs (**Fig.16**). The results show that the performance of our neuron is comparable to the ideal IF neuron model **(Fig. 17). Fig. 18** presents the effect of cycle-to-cycle and neuron-to-neuron variations on network learning speed, in which our neuron is with high performance. The results indicate that the network has high robustness on both the cycle-to-cycle and neuron-to-neuron variations, especially on the device-to-device variation. This feature is much favorable for large scale applications of the NbO_x neurons.

III. SPIKING NEURAL NETWORK

C. TC memristive SNNs for face recognition

For testing the performance of the SNNs, five neurons are connected to the RRAM array through a channel switch module, as shown in **Fig. 19**. In which an operational amplifier serves as the virtual ground to enable the neuron circuit to precisely integrate the synaptic current, as well as permitting a low input inference voltage. In actual operation, every input pattern (16 × 16) is divided into eight parts to feed the array separately. Then eight output current was summed up to drive the neurons. **Fig. 20** shows the mapped weight map read out from the array. During inference, 200 ns read pulses of -0.2/0.2 V (differential paired voltages) were applied through 256 time-steps according to the gray-level of the pixels (**Fig. 21**). **Fig. 22** shows the dynamic membrane potential outputs of five neurons during inference. Each output neuron only fires once. The input pattern is recognized as the neuron that fires first. After finishing the inference process, a "clear" signal applies to all neurons to prepare for the next pattern. When the first fire neuron is detected, the following inputs can cease, which further lowers the inference latency and energy consumption.

The firing outputs of five output neurons under fifty input face patterns are shown in **Fig. 23**, in which the red dots are the first fire neurons. **Fig. 24** illustrates the average firing time of the output neurons under different input face patterns. Due to at most one spike is required for all the input/output neurons, the spike number is much sparser than that in RC SNNs. **Table II** shows the evaluated comparison of our neuron and RC neurons on system performance. The SNN with our neurons shows a reduction of ~72 × spiking number, ~ 1.5 × inference latency, and ~ 53 × power consumption. Also, when using 10 ns input pulses and 100fF capacitors, the energy efficiency (including the RRAM array and TS neurons) is evaluated to be 20.1 TMACS/W, ~ 18 × of the RC SNNs.

IV. CONCLUSION

In this work, for the first time, a LIF neuron based on a highly stable NbO_x device is equipped for constructing TC SNNs in hardware. The neuron fires at most one spike and supports a higher input synaptic current. Based on such a neuron, we trained a TC SNNs for face recognition, obtaining a comparable result compared to the ideal IF neurons. Then, we demonstrate a 256 × 5 fully memristive TC SNN, achieving 100% inference accuracy on five persons from the Olivetti datasets. Attributing to the one-spike scheme, the TC SNNs features sparser spiking than that in RC SNNs and achieves up to 20.1 TMACS/W energy efficiency. These results demonstrate that our neuron could serve as a promising candidate to build high-efficient TC SNNs in hardware for edge applications.

Acknowledgment

This work was supported by the National Key R&D Program of China under Grant No. 2018YFA0701500 and 2017YFB0405600, the National Natural Science Foundation of China under Grant Nos. 61825404, 61732020, and 61851402, and the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant XDB44000000, Major Scientific Research Project of Zhejiang Lab (No. 2019KC0AD02).

REFERENCES

[1]. M. Pfeiffer et al., *Front. Neurosci*., 12(774), 2018. [2] Z. Wang et al., *Nat. Electron*., 1(2), 2018. [3] T. Tuma, et al., *Nat. Nanotechnol.*, 11, 2016. [4] J. Luo, et al., *IEDM*, 6.4.1-6.4.1, 2019. [5] M. H. Wu et al. *VLSI*, T34-T35, 2019. [6] M. Jerry et al., *VLSI*, T186-T187, 2017. [7] X. Zhang, et al., *IEDM*, 6.7.1-6.7.4, 2019. [8] https://irds.ieee.org/editions/2018. *IRDS*, 2018. [9] X. Zhang et al. *Nat. Commun*., 11(1), 2020. [10] S. R. Kheradpisheh, et al., *IJNS*, 30(6), 2020. [11] http://www.cs.nyu.edu/~roweis/data.html.

Temporal coding SNNs and requirements of spiking neurons

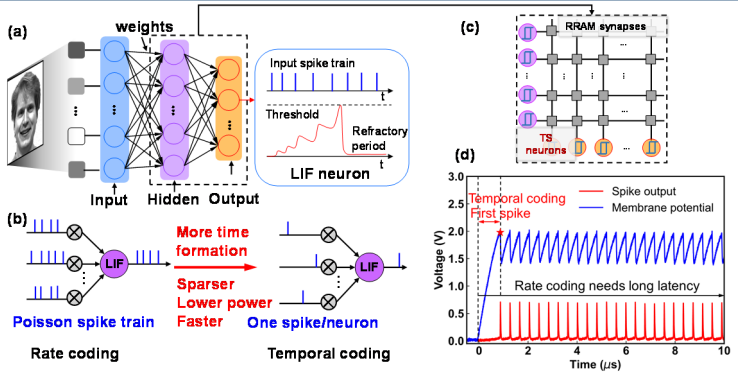

Fig. 1. (a) Schematic of the SNNs. (b) Coding methods of the SNNs: rate-coding (left) and temporal coding (right). (c) The fully memristive SNNs, including memristive synapses and neurons. (d) Spiking behavior of a RC oscillator neuron. TC SNNs only cares about the first spike.

Current state of emerging neurons

Spiking neurons based on emerging devices							
References	[2]	[3]	[4]	[5]	[6]	[7]	This work
Materials	RRAM	PCM	FeFET	MRAM	VO_2	NbO_x	NbO_x
Coding method	RC	RC	RC	RC	RC	RC	TC
Model	LIF	IF	IF	IF	LIF	LIF	LIF
Operation stability	> 10^8	> 10^9	#	#	> 10^9	> 10^{12}	> 10^{12}
Without Offset	Yes	#	#	No	No	No	Yes
Inference time independence	Yes	Yes	Yes	No	No	No	Yes
Low inference voltage	No	#	#	Yes	No	Yes	Yes
System demonstration	Expt	Expt	Sim	Sim	Sim	Expt	Expt

Table I. Comparison between our neuron and reported emerging neurons. The neuron is designed base on a LIF model and supports TC SNNs and low voltage inference, no offset voltage or current is required.

NbO_x-based TS device and electrical characteristics

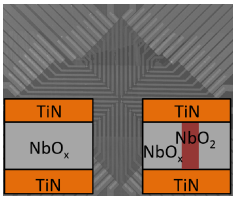

Fig. 2. SEM image of the fabricated NbO_x array and the working mechanisms of the device.

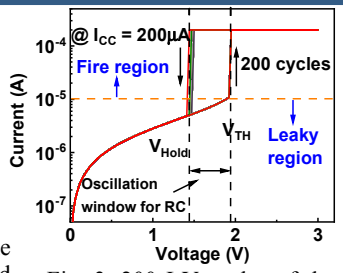

Fig. 3. 200 I-V cycles of the device after forming.

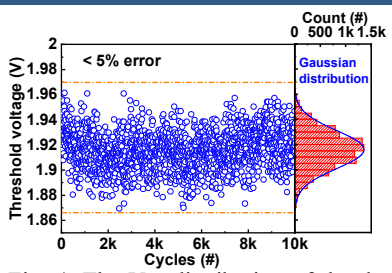

Fig. 4. The V_{TH} distribution of the device under 10^4 cycles. Tested with an oscillator circuit, withstanding > 10^{12} events.

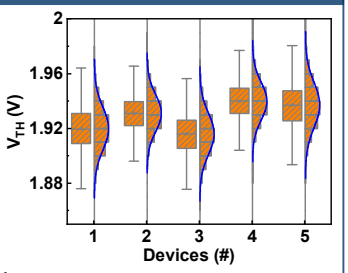

Fig. 5. The V_{th} distribution of five different devices.

Neuron circuits design for temporal coding SNNs

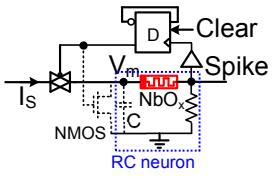

Fig. 6. The proposed neuron circuit for TC SNNs. The capacitor could be the parasitic or an external capacitor.

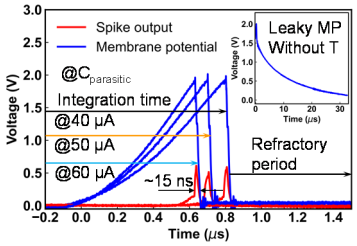

Fig. 7. The spiking behavior of the neuron circuit under different synaptic current pulses. Inset: the leaky process of the membrane potential (MP) after firing when without the NMOS.

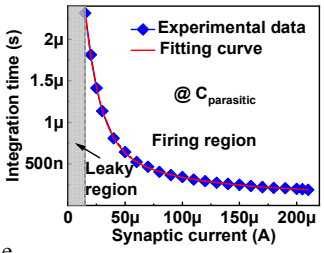

Fig. 8. The integration time as a function of input synaptic current. Including a leaky region and a firing region.

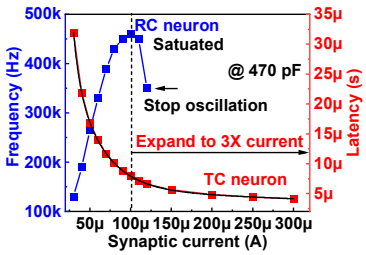

Fig. 9. The TC neuron achieves more than 3 × working current range than RC neuron that has a saturated current.

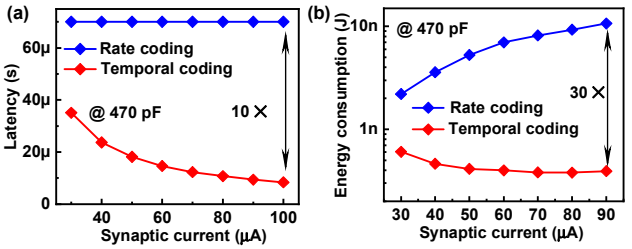

Fig. 10. Latency and energy consumption vs. input current under TC and RC. (a) 10× latency improvement and (b) 30× energy improvement of the TC neuron. The RC neuron is evaluated under a 70 μs inference time.

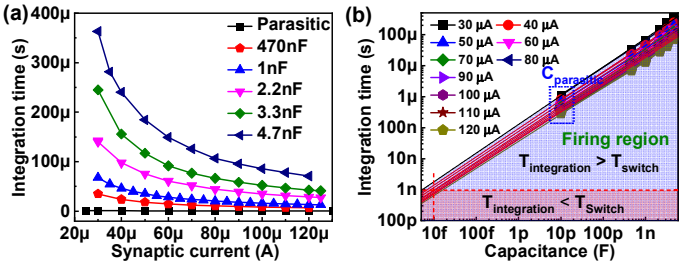

Fig. 11. Integration time vs. (a) input current under different capacitance, and (b) capacitance under different input current. An extending curve indicated that the ultimate integration time will be limited by the switching time (T_{switch}) of the device.

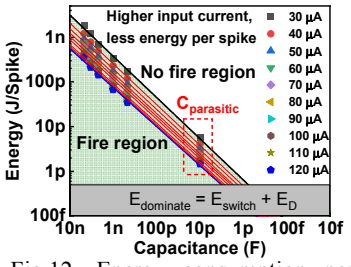

Fig.12. Energy consumption per spike under different integration capacitors. Higher input current induces a lower energy due to the shorter integration time.

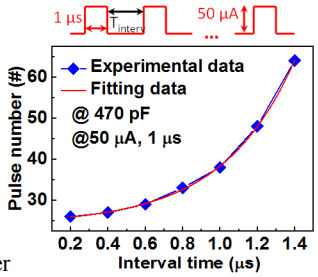

Fig.13. Required pulse number to fire as a function of the input pulse interval time.

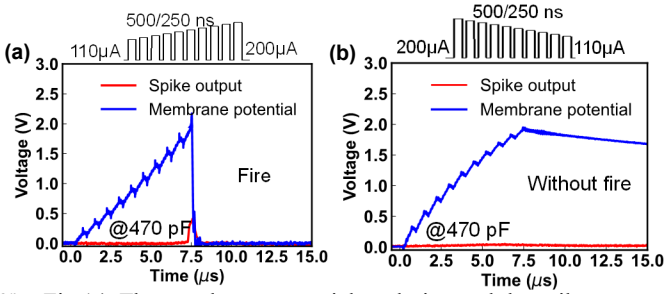

Fig.14. The membrane potential evolution and the spike output under two different temporal coded input current pulse trains. Indicating the capability of the neuron to process spatiotemporal signals.

$$\tau_m \frac{du}{dt} = -u(t) + RI(t)$$
$$if\ u(t) \geq u_{th} \pm 0.5 * rand(0,1)$$
$$u(t) \rightarrow 0$$
$$R = 532041 - 8335 * \exp(u/1.34)$$

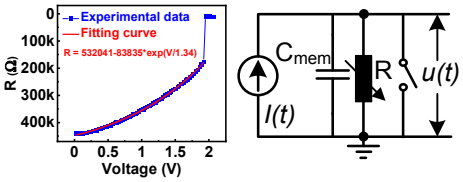

Fig. 15. Optimized LIF model based on the device characteristics. The V_{TH} stochasticity and the nonlinear leaky process are introduced.

Fig. 16. Part of the face patterns from five persons. Total of ten patterns for each person.

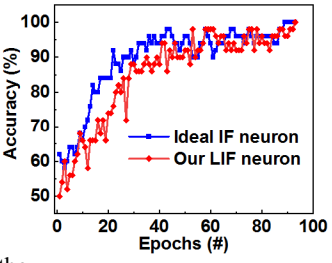

Fig. 17. Accuracy vs. training epochs using the ideal IF neuron and our LIF neuron, respectively.

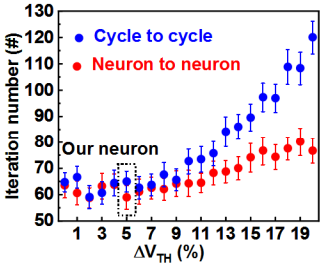

Fig. 18. The robustness of the network during training process under different V_{TH} variations, including the cycle-to-cycle and the neuron-to-neuron variations.

Temporal coding memristive SNNs for face recognition

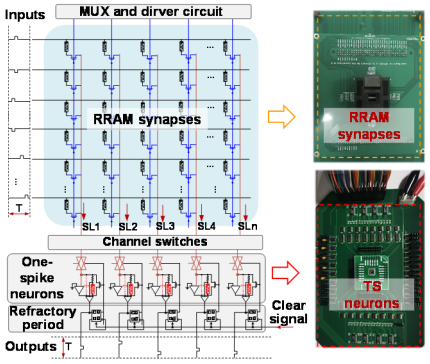

Fig. 19. Schematic of the constructed fully memristive TC SNNs (256 × 5) and the hardware implementation.

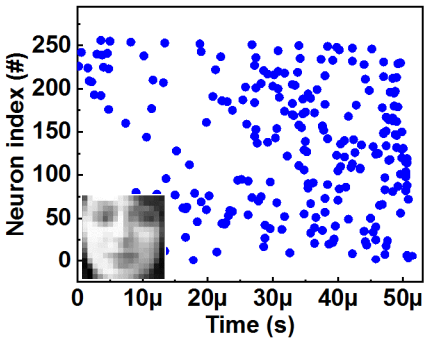

Fig. 21. The input raster plot of the inset face. Each input neuron only generates one spike. Every dot corresponds to a -0.2V/0.2V (200ns) differential paired pulses.

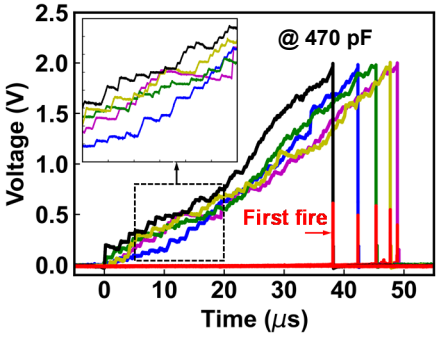

Fig. 22. Membrane potential evolution and spike output of five output neurons when the face pattern in Fig. 21 is applied. Inset: zoom-in view of the part in black box. The face pattern is recognized as the first fire neuron.

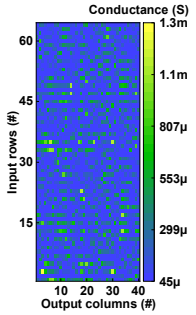

Fig. 20. Weight map of the RRAM synapses.

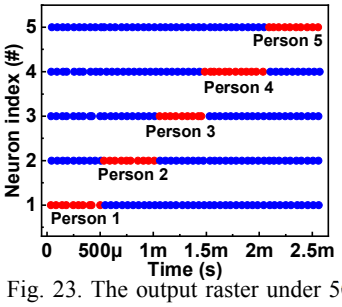

Fig. 23. The output raster under 50 test patterns from five persons, each pattern is with 51.4 μs, the red dot indicates the first fire neuron.

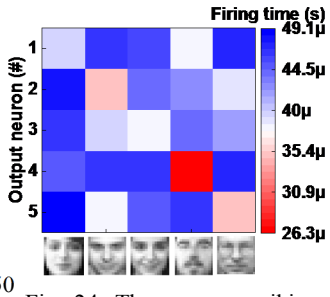

Fig. 24. The average spiking time of the network under different face patterns.

Table II. Comparison of our neuron and RC neurons on system performance with a 256 × 5 network, for face recognition.

	RC-neurons	Our neuron	Ratio
Accuracy	100%	100%	Equal
Total Spike number (input + output)	> 18868	< 261	↓> 72X
Latency (μs)	51.2	~ 35	↓~ 1.5X
Power (μW)	~ 2600	~ 49	↓~ 53X
Energy efficiency (TMACS/W)	1.1 (under 2.5 μs inference window)	20.1 (@10 ns time step)	↑~ 18X