A Neuromorphic Visual System Using RRAM Synaptic Devices with Sub-pJ Energy and Tolerance to Variability: Experimental Characterization and Large-Scale Modeling

Shimeng Yu[1*], Bin Gao[1,2], Zheng Fang[4], Hongyu Yu[3], Jinfeng Kang[2], and H.-S. Philip Wong[1#]
[1] Department of Electrical Engineering and Center for Integrated System, Stanford University, Stanford, CA 94305, USA;
[2] Institute of Microelectronics, Peking University, Beijing 100871, China;
[3] South University of Science and Technology of China, Shenzhen 518055, China;
[4] Institute of Microelectronics, A*STAR, 117685 Singapore
Email: *simonyu@stanford.edu #hspwong@stanford.edu

Abstract – We report the use of metal oxide resistive switching memory (RRAM) as synaptic devices for a neuromorphic visual system. At the device level, we experimentally characterized the gradual resistance modulation of RRAM by hundreds of identical pulses. As compared with phase change memory (PCM) reported recently in [1,2], >100×-1000× energy consumption reduction was achieved in RRAM as synaptic devices (<1 pJ per spike). Based on the experimental results, we developed a stochastic model to quantify the device switching dynamics. At the system level, we simulated the performance of image orientation selectivity on a neuromorphic visual system which consists of 1,024 CMOS neuron circuits and 16,348 RRAM synaptic devices. It was found that the system can tolerate the temporal and spatial variability which are commonly present in RRAM devices, suggesting the feasibility of large-scale hardware implementation of neuromorphic system using RRAM synaptic devices.

I. Introduction

Biologically inspired neuromorphic computing is an emerging computation paradigm beyond von Neumann computing for the applications such as image processing. Neuromorphic computing capitalizes on massive parallelism and adaptivity to the varying and complex input information [3]. Energy efficiency is a key challenge for software implementation of neuromorphic computing [4]. A hardware-based implementation may overcome this challenge if the solid-state synaptic devices consume as low energy as the biological synapses (~fJ to 10 fJ per spike [5]). Recently, two-terminal emerging memories that show electrically-triggered resistive switching phenomenon have been proposed as plastic synaptic device that can be programmed in an analog fashion [6], e.g. $Ge_2Sb_2Te_5$ based PCM [1,2], Ag/a-Si [7] and Ag/Ag_2S [8] based conducting conductive bridge memory (CBRAM), TiO_x [9], WO_x [10], HfO_x [11] based RRAM. Among these candidates, oxide based resistive switching memory is attractive for large-scale system demonstration due to its compatibility with CMOS technology and relatively lower energy consumption as compared with present-day PCM synpatic devices, e.g. ~1500 pJ per spike in [1] and ~100 pJ per spike in [2] due to the requirement of thermally melting the materials. In this work, for the first time, we report metal oxide based RRAM synaptic devices with sub-pJ energy per spike and a gradual resistance modulation. We verified the robustness of the neuromorphic system against the RRAM variability, and demonstrated the key features of neuromorphic computing: the tolerance to training error caused by device variation.

II. Experiments and Modeling of RRAM Synaptic Devices

Fig. 1 shows a vision of future neuromorphic computing system consists of RRAM cross-point array. In this work, TiN/TiO_x/HfO_x/TiO_x/HfO_x/Pt multi-layer RRAM stacks with cell area (5 μm)^2 were fabricated. The detailed fabrication processes were reported in [12]. Fig. 2 shows the typical DC I-V curves of the fabricated RRAM synaptic devices. Multilevel resistance states in this measurement were achieved by varying the reset stop voltages. The SET process is abrupt while the RESET process is gradual. To emulate the biological synapses, the devices should exhibit plasticity, i.e. the conductance can gradually change according to the input stimuli [6]. Thus we utilized the RESET process for adaptive learning since it is more gradual than the SET process. We started from two initial resistance states (~500 Ω and ~20 kΩ), and applied 400 consecutive pulses (-1.1V/10ns and -1.3V/10ns), and the results collected from 5 different devices are shown in Fig. 3. During this training process, the resistance gradually increases. It is noted that the spatial (device to device) and temporal (pulse to pulse) variations are remarkable, which are commonly observed in RRAM devices due to the inherent scholastic switching dynamics [13]. But as we will see in the next section, neuromorphic computing is capable of overcoming the device variations at the system level. To test the device endurance, we repeated such 400-pulse training process for 1000 cycles (Fig. 4). The training processes at several milestones are shown in Fig. 5. No significant degradation was found after 1000 cycles. As mentioned earlier, the energy per spike is critical. Fig. 6 shows the training process starting from 3 different initial states. The higher the initial states, the lower the energy per spike consumes. For a starting resistance of ~20 kΩ, the energy per spike drops below 1 pJ. Therefore, the fabricated RRAM synaptic devices in this work outperform PCM synaptic devices in [1,2] by 100×-1000× in terms of energy per spike. To model the gradual RESET process, we employed a 1D filament model [14] (Fig. 7), so it can be applied to large scale simulations. The conductance is exponentially dependent on the tunneling gap distance (g) (Eq. 1). I-V curves obtained in the experiments are reproduced by the model (Fig. 8). The resistance change is attributed to the electric field and temperature-enhanced oxygen ion migration; and the dynamics of the evolution of the gap distance, dg/dt, is given by Eq. 2. The parameters were fitted with the experiments starting from states ~500 Ω (Fig. 9) and ~20 kΩ (Fig. 10) with various pulse amplitudes. The RRAM variability is introduced by a Gaussian random number δg in the gap dynamics as Eq. 3. The simulated training process matches the experiments for temporal variation (Fig. 11) and spatial variation (Fig. 12) with a measured variability δR/R ~9 %.

III. Simulation of Neuromorphic Visual System

A neuromorphic system that emulates the primary visual cortex [15] in the brain was simulated. Fig. 13 shows the winner-take-all system architecture [16] implemented by integrate-and-fire CMOS neuron circuits with RRAM synaptic devices (Fig. 14). In the simulation, 32×32 neurons in the 1st layer (retina) are connected with 4×4 neurons in the 2nd layer (cortex) through 16,348 excitatory RRAM synapses. The competitive unsupervised learning is employed [16]: if one neuron in the 2nd layer fires first (becomes the winner), it inhibits all the other neurons (takes all), and a feedback spike (-1.3V/10ns pulse as in above experiments) is transmitted to the synapses that connect the non-firing neurons in the 1st layer to suppress their conductance. During the training phase, 1000 gray-scale images of a 2D Gaussian bar with random center position and random orientation (Fig. 15) were fed into the 1st layer neurons. After the training, specific neuron in the 2nd layer would respond to specific orientation of the input image [15]. Initially, the resistances of all RRAM synaptic devices were randomized to a value around ~20 kΩ as shown in the normalized conductance map between the 1st layer neurons and the 2nd layer neurons (Fig. 16). As the training progresses, the resistance begins to diverge (Fig. 17): the conductance of synapses connecting the winner in the 2nd layer neuron and the 1st layer neurons with a higher firing rate was least suppressed. After the training, the normalized conductance map becomes orientation selective (Fig. 18). To examine the orientation selectivity, standard images of 24 orientations (0 to 180°) were used, during the testing phases. The tuning curve (normalized response strength in the sum of the input current) of the first neuron in the 2nd layer is shown in Fig. 19. The selectivity is defined the contrast between the 1st peak and the 2nd peak in the tuning curve as Eq. 4. And the selectivity of all neurons in the 2nd layer is shown in Fig. 20 as a 3D plot. To study the impact of RRAM variability on the robustness of the system, we increase the variability (δR/R) in the simulation: at the measured experimental variability level (~9%), there is no degradation of the system performance, and further increasing the variability results in a slight degradation (Fig. 21).

IV. Conclusion

Key achievements in this work include: 1) RRAM synaptic devices with sub-pJ energy per spike were experimentally characterized; 2) hundreds of resistance states could be gradually modulated by using identical pulses and this gradual resistance modulation behavior is useful for learning in the presence of variation/error; 3) a stochastic model was developed to quantify the gradual resistance modulation; 4) the neuromorphic visual system was found to be robust against the RRAM device variability at the system level. With the rapid progress of RRAM technology, we can envision a large-scale neuromorphic computing system using RRAM synaptic devices in the near future.

Acknowledgements

This work is supported in part by the NRI of the SRC through the NSF/NRI Supplement to the NSF NSEC CPN, the Stanford NMTRI, the C2S2 Center of the FCRP, an SRC subsidiary, and 973 Program (2011CBA00602). H. Y. Yu acknowledges the 1000-plan funding. B. Gao is supported in part by the Stanford School of Engineering China Research Exchange Program. S. Yu is additionally supported by the Stanford Graduate Fellowship.

References

[1] M. Suri, O. Bichler, D. Querlioz, O. Cueto, L. Perniola, V. Sousa, D. Vuillaume, C.Gamrat, B. DeSalvo, *IEDM* 2011, p. 79.

[2] D. Kuzum, R. G. D. Jeyasingh, H.-S. P. Wong, *IEDM* 2011*,* p. 693.

[3] C. Mead, *Proc. IEEE* **78**, p. 1629, 1990.

[4] C.-S. Poon, K. Zhou, *Front. Neurosci.* **5**, p. 108, 2011.

[5] E. R. Kandel, J. H. Schwartz, *Principles of Neural Science*, Elsevier, 1985.

[6] G. S. Snider, *Nanotechnology* **18**, 365202, 2007.

[7] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, W. Lu, *Nano Lett.* **10**, p. 1297, 2010.

[8] T. Ohno, T. Hasegawa, T. Tsuruoka, K. Terabe, J. K. Gimzewski, M. Aono, *Nat. Mater.* **10**, p. 591, 2011.

[9] K. Seo, I. Kim, S. Jung, M. Jo, S. Park, J. Park, J. Shin, K. P. Biju, J. Kong, K. Lee, B. Lee, H. Hwang *Nanotechnology* **22**, 254023, 2011.

[10] T. Chang, S.-H. Jo, W. Lu, *ACS Nano* 5, p. 7669, 2011.

[11] S. Yu, Y. Wu, R. Jeyasingh, D. Kuzum, H.-S. P. Wong, *IEEE Trans. Electron Devices* **58**, p. 2729, 2011.

[12] Z. Fang, H. Y. Yu, X. Li, N. Singh, G. Q. Lo, D. L. Kwong, *IEEE Electron Device Lett.* **32**, p. 566, 2011.

[13] S. Yu, X. Guan, H.-S. P. Wong, *IEDM* 2011, p. 413.

[14] X. Guan, S. Yu, H.-S. P. Wong, *IEEE Electron Device Lett.* **33**, 2012, DOI: 10.1109/LED.2012.2210856

[15] C. Zamarreno-Ramos, L. A. Camunas-Mesa, J. A. Perez-Carrasco, T. Masquelier, T. Serrano-Gotarredona, B. Linares-Barranco, *Front. Neurosci.* **5**, p. 26, 2011.

[16] K. Mehrotra, C. K. Mohan, S. Ranka, *Elements of Artificial Neural Networks*, MIT Press, 1996.

Eq. 1 $\boldsymbol{I = I_0 \cdot exp(-g/g_0) \cdot sinh\,(V/V_0)}$

g is average tunneling gap distance, I_0 (~1 mA), g_0 (~0.25 nm) and V_0 (~0.25 V) are fitting parameters by experiments in Fig. 8.

Eq. 2 $\boldsymbol{dg/dt = -v_0 \cdot exp(-E_a/kT) \cdot sinh(\gamma \cdot a_0/L \cdot qV/kT)}$

Ea is activation energy (~0.6 eV), a0 is atom spacing (~0.25 nm), L is oxide thickness (~12 nm), q is electron charge, k is Boltzmann constant, and v0 is fitting parameter (~10 nm/ns). The local field acceleration parameter γ is introduced due to the strong polarizability in high-k dielectrics, and it is empirically fitted to be g dependent: $\gamma = \gamma_0 - \beta \cdot g^3$ where γ0 (~16) and β (~0.8) are fitting parameters. In this way, field decreases as the tip of filaments gets farther away from the electrode. T is local temperature due to Joule heating $T = T_0 + V \cdot I \cdot R_{th}$, T_0=298K, R_{th} is equivalent thermal resistance (~2000 K/W). Fitting are done by experiments in Fig. 9 & 10.

Eq. 3 $\boldsymbol{dg = dg(ideal) + \delta g}$

dg(ideal) is the gap distance calculated by Eq. 2, δg is a Gaussian random number that accounts for the randomness of oxygen ion migration, the relative resistance variability is given by $\delta R/R = \delta g/g_0$, to match the experimentally measured variability, δg ~0.0224 nm, δR/R ~9%.

Eq. 4 $\boldsymbol{S = (I_1 - I_2)/(I_1 + I_2)}$

S is the image orientation selectivity. I_1 is maximum of the 1st peak of tuning curve, and I_2 is maximum of the 2nd peak. In the case of single peak, I_2 is the height at the peak angles ±30°, see Fig. 19 for illustration.

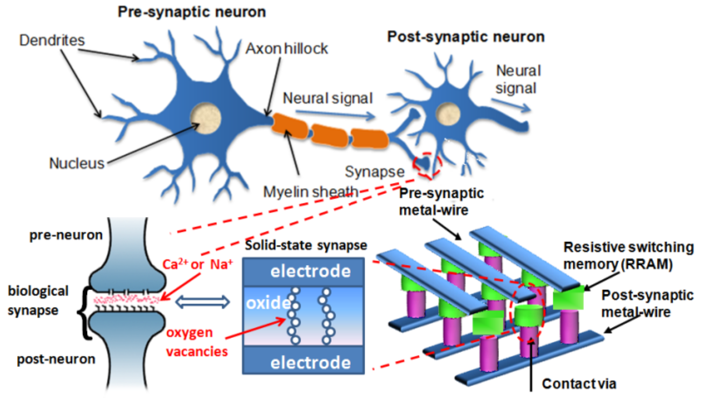

Fig. 1 An analogy between the connections of biological neural system and its artificial counterpart: the cross-point array with metal oxide RRAM synaptic devices at the junctions.

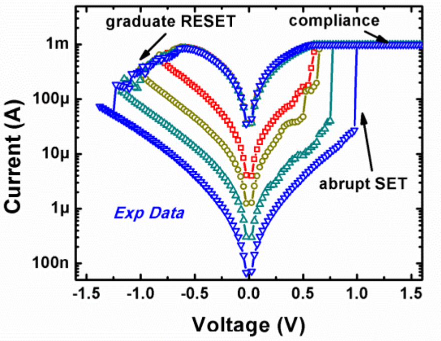

Fig. 2 Typical DC I-V bipolar switching curves of RRAM synaptic devices. The multilevel states in this measurement were achieved by varying the RESET stop voltages.

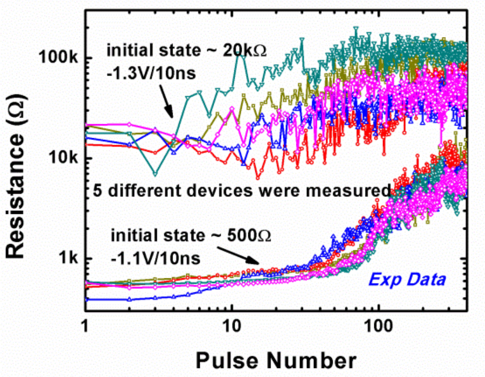

Fig. 3 Gradual RESET training process by 400 identical consecutive pulses. Initial states were ~500Ω and ~20kΩ. 5 different devices were measured. Variations were observable.

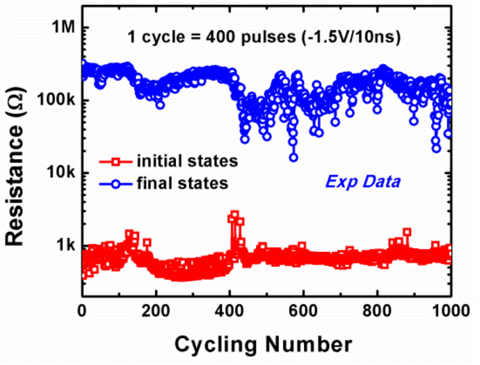

Fig. 4 Endurance test of the training process. In one cycle, 400 identical consecutive pulses were applied. The initial states and the final states were shown during the 1000-times cycling.

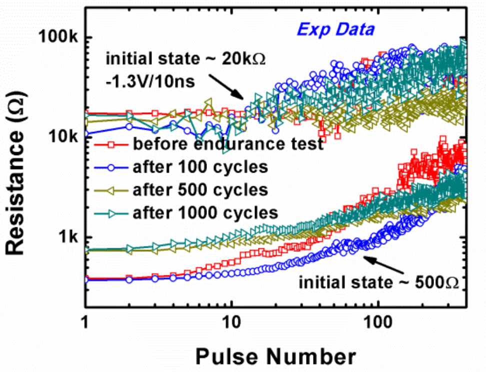

Fig. 5 Training process by 400 pulses at several milestones during the endurance cycling test. Gradual RESET process was reproducible after the 1000-times cycling.

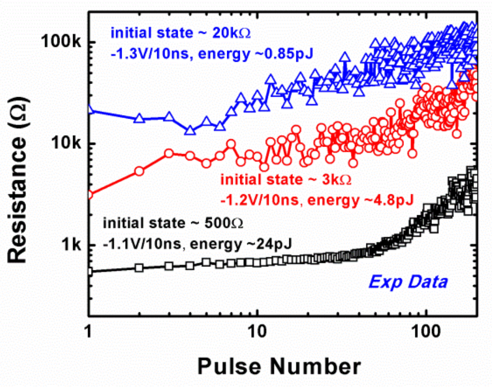

Fig. 6 Dependence of energy per spike on the initial resistance states for the training process. If starting ~ 20 kΩ, energy per spike drops below 1 pJ.

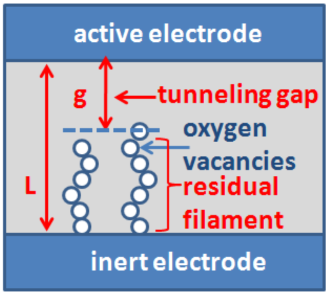

Fig. 7 Schematic of conductive filament with oxygen vacancies. The tunneling gap distance g determines the device resistance (Eq. 1), and g evolves due to field and thermally driven ion migration (Eq. 2)

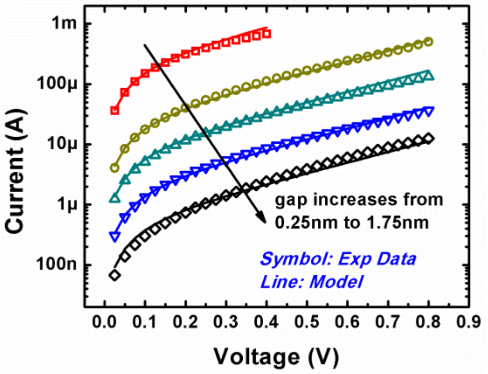

Fig. 8 I-V fitting of multilevel resistance states by varying the gap tunneling distance. Parameters in Eq. 1 were fitted.

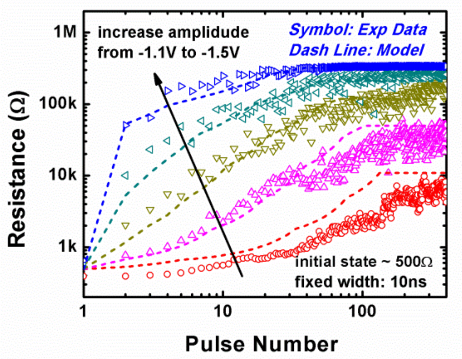

Fig. 9 Measured gradual training process staring from ~500 Ω by varying the pulse amplitudes. Parameters in Eq. 2 were fitted.

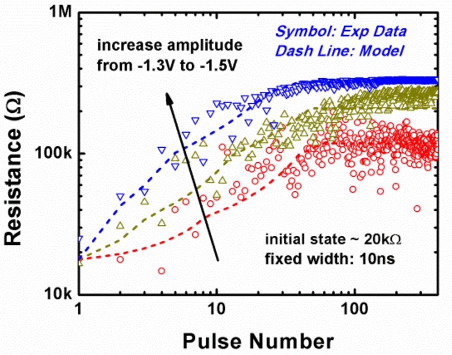

Fig. 10 Measured gradual training process staring from ~20 kΩ by varying the pulse amplitudes. Parameters in Eq. 2 were fitted.

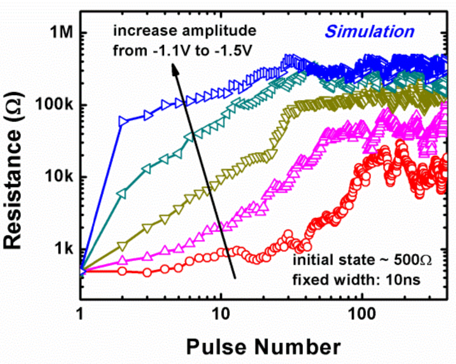

Fig. 11 Simulated gradual training process staring from ~500 Ω by varying the pulse amplitudes, similar to the Exp Data in Fig. 9.

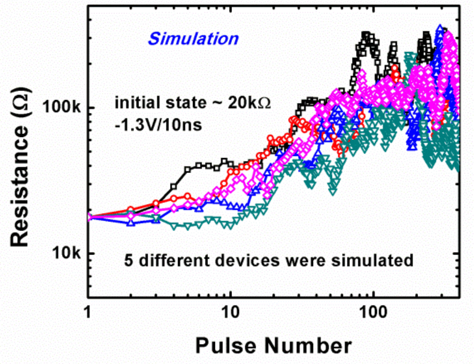

Fig. 12 Simulated gradual training process staring from ~20 kΩ for 5 different devices, similar to the Exp Data in Fig. 3.

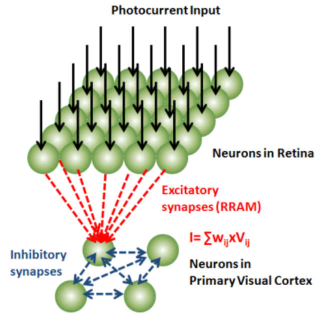

Fig. 13 Neuromorphic visual system based on winner-take-all architecture. The 1st layer and the 2nd layer are connected with 16,348 RRAM synaptic devices.

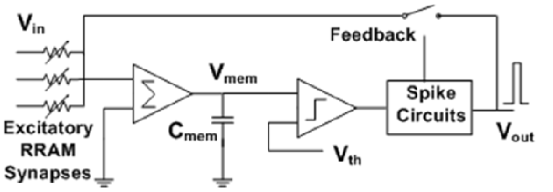

Fig. 14 An integrated-and-fire CMOS neuron circuit with RRAM synaptic devices. The membrane capacitor sums and integrates the input current. Once membrane voltage exceeds the threshold voltage, the spike generator gets fired. Meanwhile, it feeds back the spike to synaptic devices connected to the 1st layer neurons (learning) and triggers the discharge of membrane capacitors of the 2nd layer neurons (inhibits other neurons from firing).

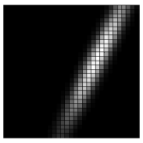

Fig. 15 Example of training image (32 × 32 pixels) showing 2D Gaussian bar with random center and random orientation.

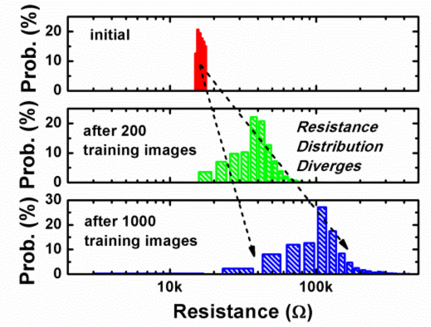

Fig. 17 Simulated resistance evolution of 16 kb RRAM synaptic devices. As training progresses, the distribution diverges: the conductance of synapses connecting the 1st layer neurons with higher firing rate was least suppressed.

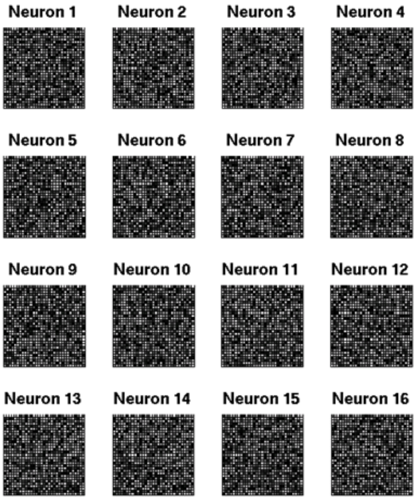

Fig. 16 Simulated initial normalized synapse conductance map between the 1st layer and the 2nd layer (before training)

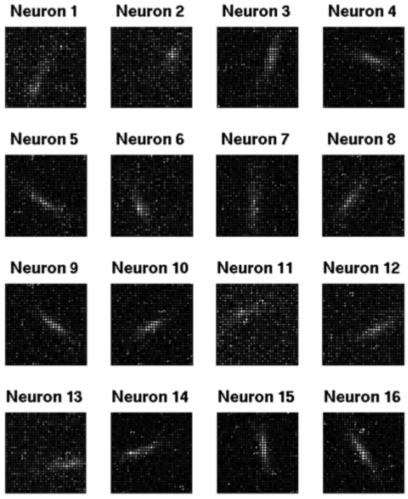

Fig. 18 Simulated final normalized synapse conductance map (after training with 1000 images). The noisy pixels are due to the RRAM variation.

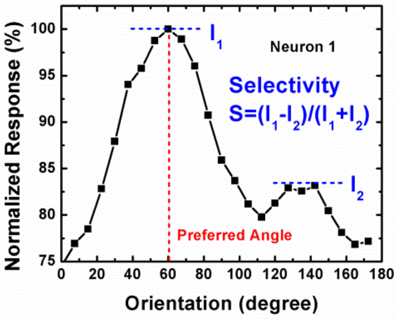

Fig. 19 Tuning curve (normalized response strength vs. standard image orientation) of neuron 1 in the 2nd layer. Inset: Selectivity definition.

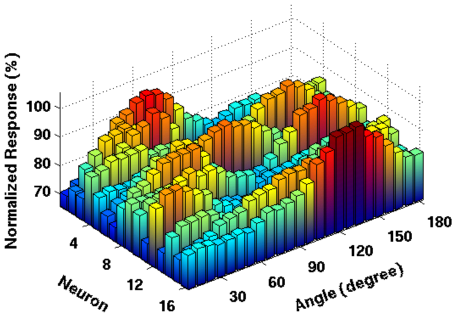

Fig. 20 Tuning curves of all the 16 neurons in the 2nd layer plotted as a 3D plot. Different neurons have different preferred angles.

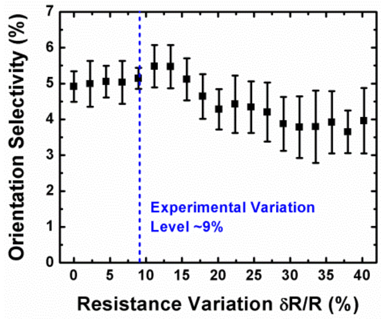

Fig. 21 Simulated orientation selectivity vs. RRAM variability δR/R. At each error bar, 20 independent simulations were carried out. Experimentally δR/R~ 9%