scientific reports

Check for updates

OPEN

Mnemonic-opto-synaptic transistor for in-sensor vision system

Joon-Kyu Han[1,3], Young-Woo Chung[1,2,3], Jaeho Sim[1], Ji-Man Yu[1], Geon-Beom Lee[1], Sang-Hyeon Kim[1] & Yang-Kyu Choi[1]✉

A mnemonic-opto-synaptic transistor (MOST) that has triple functions is demonstrated for an in-sensor vision system. It memorizes a photoresponsivity that corresponds to a synaptic weight as a memory cell, senses light as a photodetector, and performs weight updates as a synapse for machine vision with an artificial neural network (ANN). Herein the memory function added to a previous photodetecting device combined with a photodetector and a synapse provides a technical breakthrough for realizing in-sensor processing that is able to perform image sensing and signal processing in a sensor. A charge trap layer (CTL) was intercalated to gate dielectrics of a vertical pillar-shaped transistor for the memory function. Weight memorized in the CTL makes photoresponsivity tunable for real-time multiplication of the image with a memorized photoresponsivity matrix. Therefore, these multi-faceted features can allow in-sensor processing without external memory for the in-sensor vision system. In particular, the in-sensor vision system can enhance speed and energy efficiency compared to a conventional vision system due to the simultaneous preprocessing of massive data at sensor nodes prior to ANN nodes. Recognition of a simple pattern was demonstrated with full sets of the fabricated MOSTs. Furthermore, recognition of complex hand-written digits in the MNIST database was also demonstrated with software simulations.

The von Neumann architecture provides accurate calculations, however, it is not suitable for low power applications because of the data bottleneck between the memory and the processor[1]. In order to overcome the limitations of the von Neumann architecture, various artificial neuromorphic devices were explored to imitate functions of the brain. In details, two-terminal memristors such as resistive random-access memory (RRAM) and phase-change memory (PCM), and the three-terminal charge trap memory and electrochemical random-access memory (ECRAM) with separated reading and writing paths have been actively studied as synaptic devices for artificial neural networks (ANN)[2–6].

By the way, vision systems assisted by neural processing allow accurate object detection, pattern recognition, and real-time image processing for robotics, autonomous vehicles, and sensory electronics[7–11]. A conventional vision system separates image sensing and signal processing. Its performance is thus adversely limited owing to signal latency and power consumption that arises from a huge amount of data processing with the inclusion of redundant data passing through a converting circuit such as an analog-to-digital converter (ADC), as illustrated in Fig. 1a[12–14]. In contrast, a biological retina performs sensing and simultaneous pre-processing of visual information in order to extract key features from the input visual data[15–18]. By the elimination of redundant visual data, subsequent information processing in the brain such as object detection and pattern recognition can become faster with lower power consumption.

Recently, inspired by a biological vision system, various optoelectronic synaptic devices that can act as both a photodetector and a synapse used for an ANN by preprocessing of the data in a sensor have been demonstrated[9–11]. During the optical sensing, however, their synaptic weight is changed owing to an optically controllable synaptic weight. This optical weight update is useful for recognizing one pattern or similar patterns, but it is difficult to recognize various subsequent patterns because the synaptic weights are customized to a previous pattern. Therefore, repetitive reset operations are needed before accepting new patterns. Unlike the abovementioned optoelectronic synaptic devices, Wang et al. and Mennal et al. demonstrated vision sensors where repetitive reset operations were unnecessary due to the invariant synaptic weight during the optical sensing. They reported tunable photoresponsivity using a photodetecting device composed of two-dimensional (2D) materials, such as a phototransistor or a photodiode[18,19]. The tunable photoresponsivity in a photodetecting device corresponds to the controllability of weight update in a synapse, and it is a significant advantage for an in-sensor vision system, because photoresponsivity tunable photodetecting device can act as a synapse for an

[1]School of Electrical Engineering, Korea Advanced Institute of Science and Technology, (KAIST) 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea. [2]Foundry Division, Samsung Electronics, Yongin 17113, Republic of Korea. [3]These authors contributed equally: Joon-Kyu Han and Young-Woo Chung. ✉email: ykchoi@ee.kaist.ac.kr

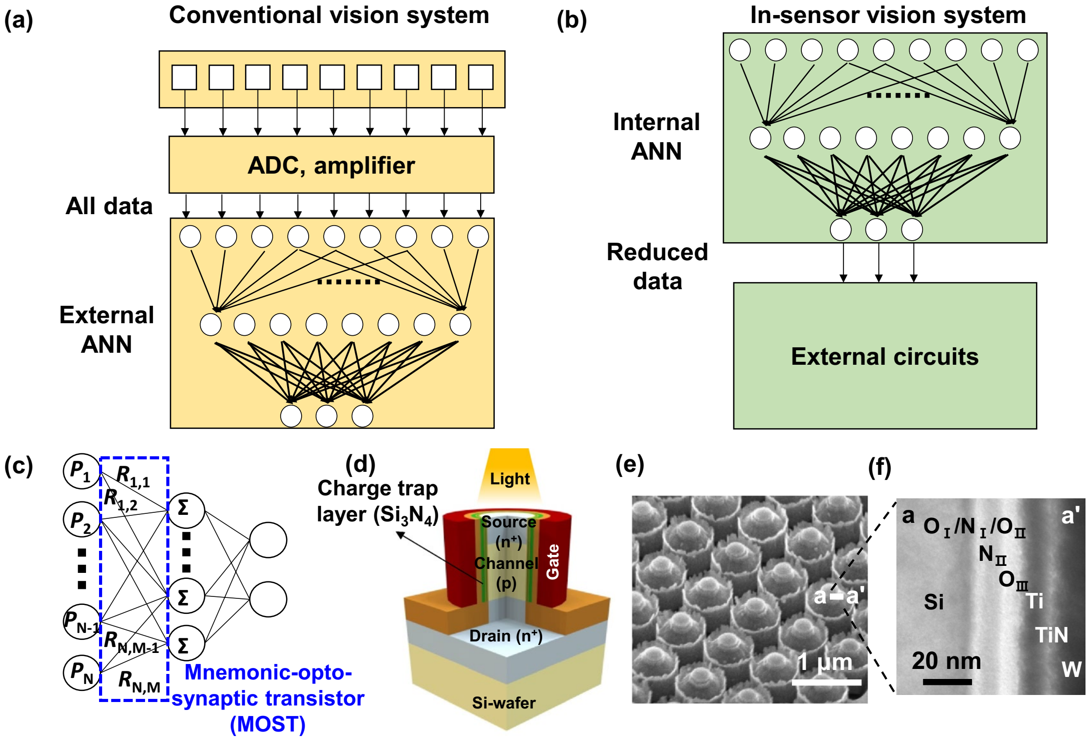

Figure 1. Schematic diagram of (**a**) conventional vision system and (**b**) in-sensor vision system. An internal artificial neural network (ANN) performs both sensing and preprocessing in a sensor for reduction of the signal latency and power consumption at converting circuits such as an analog-to-digital converter (ADC). (**c**) An internal ANN constituting an in-sensor vision system with the MOSTs, which can be located at the forefront of the ANN to simultaneously detect the optical signals and transmitting the preprocessed signal to the next layer. (**d**) Schematic of the mnemonic-opto-synaptic transistor (MOST). A charge trap layer (CTL) for tunable photoresponsivity and memory function is inserted into the gate dielectrics. (**e**) Scanning electron microscopy (SEM) image of the MOST array and (**f**) cross-sectional transmission electron microscopy (TEM) image of the gate region. Bandgap-engineered (BE) tunneling dielectrics ($O_I/N_I/O_{II}$) were adopted to reduce the operating voltage.

ANN as well as a photodetector for a sensor. Thus, the in-sensor processing with the inclusion of image sensing and signal processing allows real-time multiplication of the image with a memorized photoresponsivity matrix. Such an in-sensor vision system is attractive for reduction of signal latency and power consumption, which occur at converting circuits such as the ADC, as illustrated in Fig. 1b.

It is worth noting that the previous photodetecting device with tunable photoresponsivity requires external memory, which is indispensable for storing the value of gate voltage to tune the photoresponsivity[18,19]. This memory can impose a burden on accessing a designated memory cell with high speed and realizing a mobile vision system with a compact size for an all-in-one chip. Thus, signal latency and power consumption that arise from external memory become increasingly problematic. In addition, 2D materials cannot be easily integrated by microfabrication of a complementary metal–oxide–semiconductor (CMOS) based image sensor system with high throughput owing to less CMOS compatibility. For a large-scale vision system, a CMOS compatible photodetecting device such as a photodiode and a phototransistor is preferred; however, tunable photoresponsivity is not available. Each approach for tunable photoresponsivity without CMOS compatibility and CMOS compatibility without tunable photoresponsivity has its respective strengths and weaknesses. Therefore, it is very timely to explore another photodetecting device with tunable photoresponsivity, CMOS compatibility, and even more memorability.

In this work, a mnemonic-opto-synaptic transistor (MOST) is demonstrated in the form of a metal–oxide–semiconductor field-effect transistor (MOSFET). This MOSFET has a vertical pillar-shaped channel protruded from a silicon bulk substrate and a gate wraps a sidewall of the pillared channel completely with a gate-all-around structure. This vertical MOSFET is advantageous from the perspective of the footprint area and light absorption[20–22]. Moreover, by embedding a charge trap layer (CTL) of a nitride (Si_3N_4) to the gate dielectrics of the MOST for the memory function, individual control of photoresponsivity for each MOST is achieved and real-time multiplication of the image with a memorized photoresponsivity matrix is performed. Therefore, it can act as a photodetector and a synapse with non-volatile retention of learned weights in the ANN for the in-sensor vision system due to the intrinsic memory function of the intercalated CTL. It does not need repetitive reset operations because the synaptic weight is not changed during the optical sensing. This characteristic is attributed to fully electrical control of the synaptic weight. Furthermore, by virtue of 100% CMOS compatible

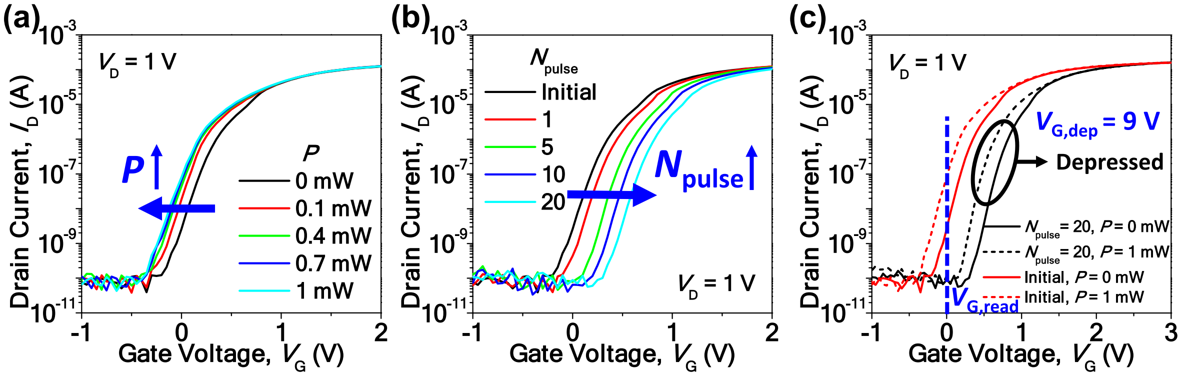

Figure 2. (**a**) Transfer characteristics (I_D–V_G) of the MOST for various light intensities (P). Leftward V_T shift with increased P that corresponds to a temporal response by the photo-carrier density. (**b**) I_D–V_G for various N_{pulse} with + V_G. Rightward V_T shift with increased N_{pulse} that corresponds to a semi-permanent response by the trapped electron density. This is a depression operation for reducing the weight of the synaptic device. (**c**) I_D–V_G at dark and 1 mW light illumination before and after depression. Photocurrent (I_{photo}) at read gate voltage ($V_{G,read}$) of 0 V is approximately 0.1 µA before the depression and 0.1 nA after the depression, respectively. In this way, the trapped electron density tunes the photoresponsivity.

fabrication, it can be integrated with a conventional large-scale CMOS image sensor system comprising numerous small-sized pixels. After optical and electrical characterization of the MOST, recognition of a simple pattern is performed using the sets of the fabricated devices, and recognition of a complex MNIST hand-written number is exploited using software simulations.

Results and discussion

Figure 1c represents the ANN for the in-sensor vision system using the MOSTs. The MOSTs are located at the forefront of the ANN for detecting the light intensity and transmitting pre-processed weights with a reflection of optical signals to the next layer. The photocurrent (I_{photo}) summed from each neuron at the next layer is produced by the multiplication of the memorized photoresponsivity matrix and the light intensity of each pixel. When the vision system has N pixels and M neurons at the next layer, current summed in the mth neuron of the next layer (I_m) can be represented by the following equation: $I_m = \sum_{n=1}^{N} I_{photo} = \sum_{n=1}^{N} R_{mn}P_n$, where $n = 1, 2, \ldots, N$ and $m = 1, 2, \ldots, M$ denote the indices of the pixel and the neuron at the next layer, respectively. R_{mn} represents the memorized photoresponsivity matrix and P_n represents the light intensity of each pixel. In this way, the in-sensor processing with the inclusion of image sensing and signal processing allows real-time multiplication of the image with the memorized photoresponsivity matrix[19].

Figure 1d shows a schematic of an n-channel MOST with a vertical pillar structure. n^+ heavily doped source (S) and drain (D) are located at the top and the bottom of each pillar in the array of MOSTs shown in Fig. 1e, which protrudes from a bulk-silicon wafer, respectively. Between the S and D, there is a p-type channel. As gate dielectrics, quintuple-layers ($O_I/N_I/O_{II}/N_{II}/O_{III}$) composed of triple-layered tunneling dielectrics ($O_I/N_I/O_{II}$), the aforementioned CTL nitride (N_{II}), and a blocking oxide (O_{III}) wrap around a sidewall of the pillared channel, as shown in Fig. 1f. The triple layers of the $O_I/N_I/O_{II}$ were adopted to reduce the operating voltage by barrier engineering (BE) of the tunneling dielectrics[23,24]. Each thickness of the gate dielectrics is 1.3 nm/1.3 nm /1.6 nm/5.6 nm/6.3 nm in the order of $O_I/N_I/O_{II}/N_{II}/O_{III}$, respectively. A triple-layered metal gate composed of titanium, titanium nitride, and tungsten (Ti/TiN/W) also surrounds the sidewall exterior of the gate dielectrics and pillar. When the light is illuminated, the carriers are generated and flown in the channel in the form of I_{photo} that drives the photodetector. I_{photo} is actually the drain current (I_D) flowing between the source and the drain, which is controlled by the gate voltage (V_G) and drain voltage (V_D). The gate electrode makes the photoresponsivity tunable by charging and discharging the CTL of N_{II} (hereafter simply abbreviated as 'CTL') and controls the memory function. Note that N_I in the tunneling dielectrics cannot serve as a CTL because O_I is too thin to block tunneling of the trapped charges. Fabrication details of the MOST are described in Figure S1.

In the MOST, threshold voltage (V_T) can be adjusted by two factors, photo-carriers controlled by light illumination and trapped electrons modulated by the V_G in the CTL. Figure 2 shows the transfer characteristic curve of I_D versus V_G (I_D–V_G) according to the light intensity (P) and the number of gate pulses (N_{pulse}). This N_{pulse} determines the level of I_D at each state in the synaptic operation, i.e., the number of states. As an example, N_{pulse} of 0 is the initial state with the highest I_D due to the lowest V_T, and N_{pulse} of 31 is composed of 31 gate pulses that produce the lowest I_D due to the highest V_T in the depression for multi-states of 32. In this work, a variable pulse number with an identical pulse amplitude and width is used for a potentiation–depression (P–D) operation. An LED (SOL 3.0, Fiber Optic Korea Co., Ltd.) was used as a white light source. The P indicated in Fig. 2 is the measured value in a blue region with a wavelength of 405 nm. It was quantified by a power meter that has a detection spot area of 0.785 cm^2. Figure 2a shows a leftward V_T shift. This is caused by the photo-carrier generation, which arises from light illumination[25]. In contrast, Fig. 2b exhibits a rightward V_T shift. It is attributed to electron trapping in the CTL by applied positive depression gate voltage ($V_{G,dep}$); i.e., it suppresses inversion at the channel surface. This is analogous to the depression operation to reduce the synaptic weight in an artificial synapse[26–28]. The magnitude of $V_{G,dep}$ is 9 V and its pulse width is 10 µs. It should be noted that the

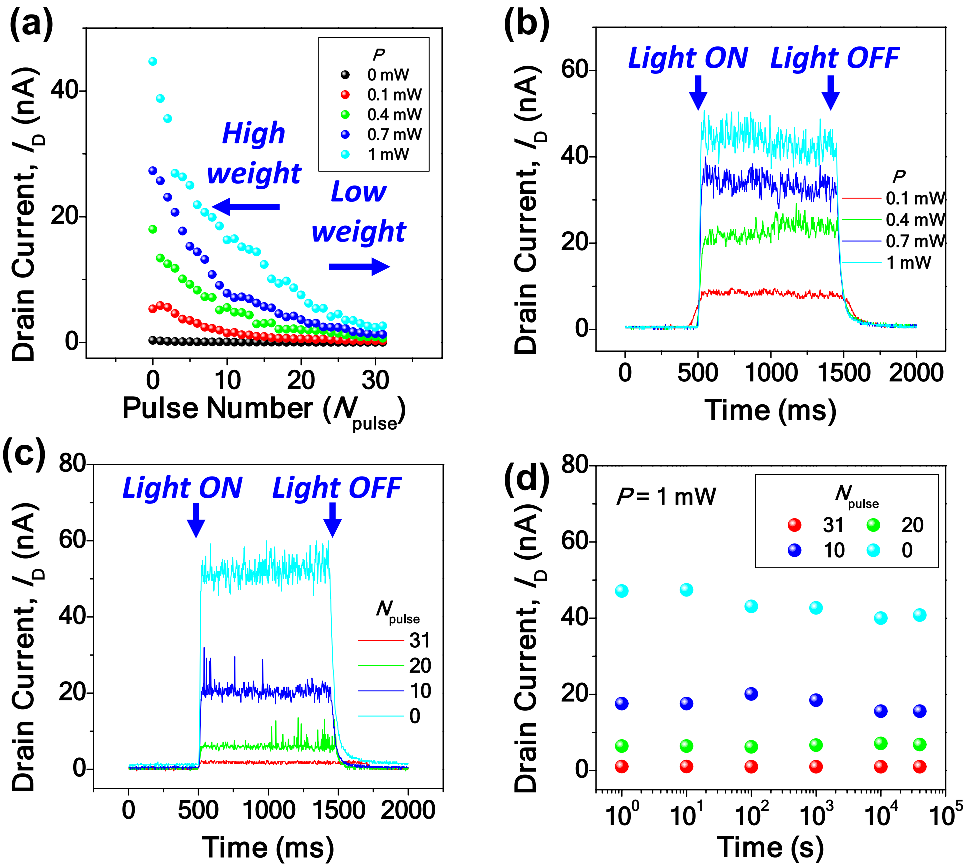

Figure 3. (**a**) I_D versus N_{pulse} for various P. The photoresponsivity is reduced by the depression. (**b**) Real-time I_D for various P when N_{pulse} is zero. I_D is increased as P increases. (**c**) Real-time I_D for various N_{pulse} when P is 1 mW. I_D is decreased as N_{pulse} increases. (**d**) Retention characteristics of the MOST for various N_{pulse} when P is 1 mW.

rightward V_T shift by the electron trapping is semi-permanent and the leftward V_T shift by the light illumination is temporal. In other words, the V_T shift is returned to a pristine state when the light illumination is removed. Figure 2c superimposes I_D–V_G with the photo-carrier generation by incident light and the electron trapping by the applied $V_{G,dep}$ in one graph. The ratio (η) of photoresponsivity without charge trapping to that with charge trapping by V_G is approximately 800 at a $V_{G,read}$ of 0 V. In this way, photoresponsivity can be modulated effectively by controlling the trapped electrons in the CTL. Therefore, the MOST acts as a photodetector by sensing I_{photo} with light, a synapse by updating a weight with V_G, and a non-volatile memory by holding a weighted state with trapped charges for the in-sensor vision system. This tunable photoresponsivity is utilized as a controllable synaptic weight in the ANN. Unlike the previously reported photodetecting device, extra memory is no longer needed because the MOST itself harnesses an inherent non-volatile memory function[18,19].

Figure 3a shows the depression where I_D was decreased by an increased N_{pulse} for various P. Herein N_{pulse} is varied from 0 to 31; i.e., there are 32 states. The magnitude of $V_{G,dep}$ is 9 V and its pulse width is 1 µs. This result shows that the photoresponsivity was finely tunable with multi-states. For a typical synaptic operation, the potentiation that increases the synaptic weight should be available, similar to the depression that decreases the synaptic weight. Figure S2(a) represents the P-D characteristics for various P, i.e., with light illumination. The conductance (G) is defined as I_D/V_D, which can be simplified to I_D because the applied V_D was 1 V. The photoresponsivity was finely tunable during the potentiation as well as the depression. The magnitude of potentiation gate voltage ($V_{G,pot}$) is – 10 V and its pulse width is 200 µs. Figure S2(b) shows another P-D characteristic in a dark environment, i.e., without light illumination. From Figure S2(b), the nonlinearity parameters (α) were extracted using the following equation:

$$G = \begin{cases} \left(\left(G_{max}^{\alpha} - G_{min}^{\alpha}\right) \times w + G_{min}^{\alpha}\right)^{1/\alpha} & if\, \alpha \neq 0, \\ G_{min}^{\alpha} \times (G_{max}/G_{min})^{w} & if\, \alpha = 0. \end{cases} \quad (1)$$

where G_{max} is the maximum conductance, G_{min} is the minimum conductance, α is a nonlinear parameter, and w is an internal variable that ranges from 0 to 1[29]. The extracted α_{pot} and α_{dep} were – 0.02 and – 0.58, respectively. These parameters are used for the subsequent software simulations. It is well known that a large number of states is preferred to enhance the performance of pattern recognition in a synaptic device[26–28]. In this context, it was

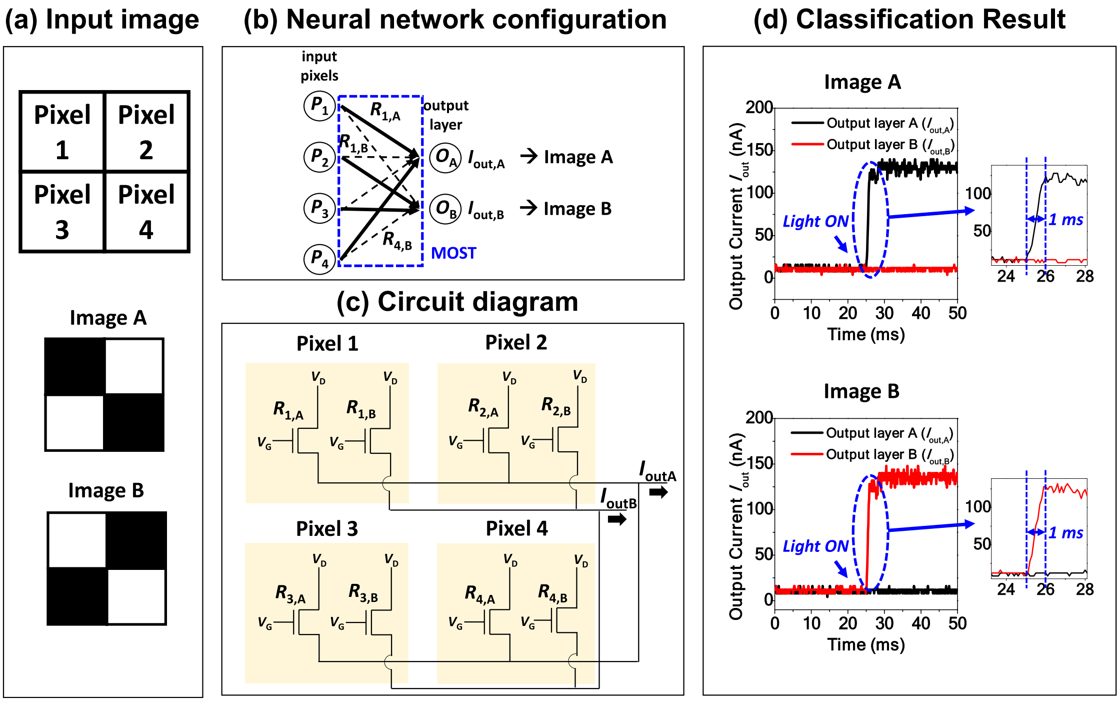

Figure 4. Demonstration of hardware-based pattern recognition. (**a**) Two input images 'A' (off-diagonal) and 'B' (diagonal), which are composed of 2 × 2 black-and-white pixels. (**b**) Neural network and (**c**) Circuit diagram for the 2 × 2 pattern recognition. They are composed of eight MOSTs with each tunable photoresponsivity represented as '*R*'. (**d**) Measured classification data of the off-diagonal and diagonal patterns. By comparing two output currents ($I_{\mathrm{out,A}}$ and $I_{\mathrm{out,B}}$), fast classification within 1 ms was achieved with low power consumption under 150 nW.

also confirmed that the P–D characteristics for N_{pulse} of 64 and 128 were achievable by delicately tuning the gate pulse, as shown in Figure S3.

Figure 3b, c show the real-time I_D for various P and N_{pulse}, respectively, when the light is turned on and off. At a fixed N_{pulse}, I_D was increased as P increased. At a fixed P, I_D decreased as the N_{pulse} increased. It is worth noting that I_D returned to the initial state when the light was off. This feature assures that the synaptic weight is not changed during the optical sensing and repetitive reset operations are not needed. As shown in Fig. 3d, I_D was sustained even after 40,000 s owing to the superior retention characteristics of the CTL-based memory. This attribution has been proven by commercial flash memory adopting the CTL. It should be recalled that good retention characteristics of a synaptic device are crucial for reliable operation over time[28].

Figure S4 shows the P–D characteristics of the MOST for various wavelengths (λ). Measurements were performed by using a blue (B), red (R), and infrared (IR) light source. Each λ of B, R, and IR light is 405 nm, 638 nm, and 1550 nm, respectively. As shown in Figure S4, tunable photoresponsivity was observed for visible light of B and R, whereas it was not for the IR light. This is because the B and R light can generate photo-carriers to increase I_{photo}. However, the IR light cannot create them owing to a small photon energy of 0.80 eV compared to the silicon energy bandgap of 1.12 eV[30,31]. It should also be noted that the photoresponsivity of the B light was smaller than that of the R light because the penetration depth is decreased with shorter λ[32]. The demonstrated wavelength dependency as well as the intensity dependency of the tunable photoresponsivity can help in recognizing a color mixed pattern[33,34].

As mentioned above, BE tunneling dielectrics composed of the triple layers renamed BE layers were adopted to reduce the operating voltage. In order to confirm this effect, simplified MOSTs were fabricated as a control group. The BE layers of $O_I/N_I/O_{II}$ were replaced by a single layer of thermal oxide (O_{single}). Other structures were set to be the same. As plotted in Figure S5(a), the measured transfer characteristics of the fabricated MOST with $O_{\mathrm{single}}/N_{II}/O_{III}$ showed similar photoresponsivity compared to those with $O_I/N_I/O_{II}/N_{II}/O_{III}$. This is because the gate dielectric has no effect on the photo-carrier generation by light. Whereas V_T was shifted rightward by a $V_{\mathrm{G,dep}}$ of 9 V in the case of the $O_I/N_I/O_{II}/N_{II}/O_{III}$ (Fig. 2), it was not changed by that in the case of the $O_{\mathrm{single}}/N_{II}/O_{III}$, as shown in Figure S5(b). A $V_{\mathrm{G,dep}}$ larger than 11 V should be applied to change the V_T and update the synaptic weight, as shown in Figure S5b. As a consequence, the P–D characteristics in Figure S5(c) show that synaptic weight update is impossible with the same $V_{\mathrm{G,dep}}$ in the case of the $O_{\mathrm{single}}/N_{II}/O_{III}$. Therefore, it is confirmed that the gate dielectric structure of $O_I/N_I/O_{II}/N_{II}/O_{III}$ is more attractive than that of $O_{\mathrm{single}}/N_{II}/O_{III}$ for low-power neuromorphic hardware.

Using a full set of the fabricated MOSTs, simple pattern recognition was performed using a single-layer perceptron (SLP). As illustrated in Fig. 4a, two images, 'A' of an off-diagonal pattern and 'B' of a diagonal pattern, were prepared. Each pattern comprises 2 × 2 black-and-white pixels. Classification of the two patterns was attempted. A neural network was composed of four input pixels labeled P_1, P_2, P_3, and P_4 and two nodes in the

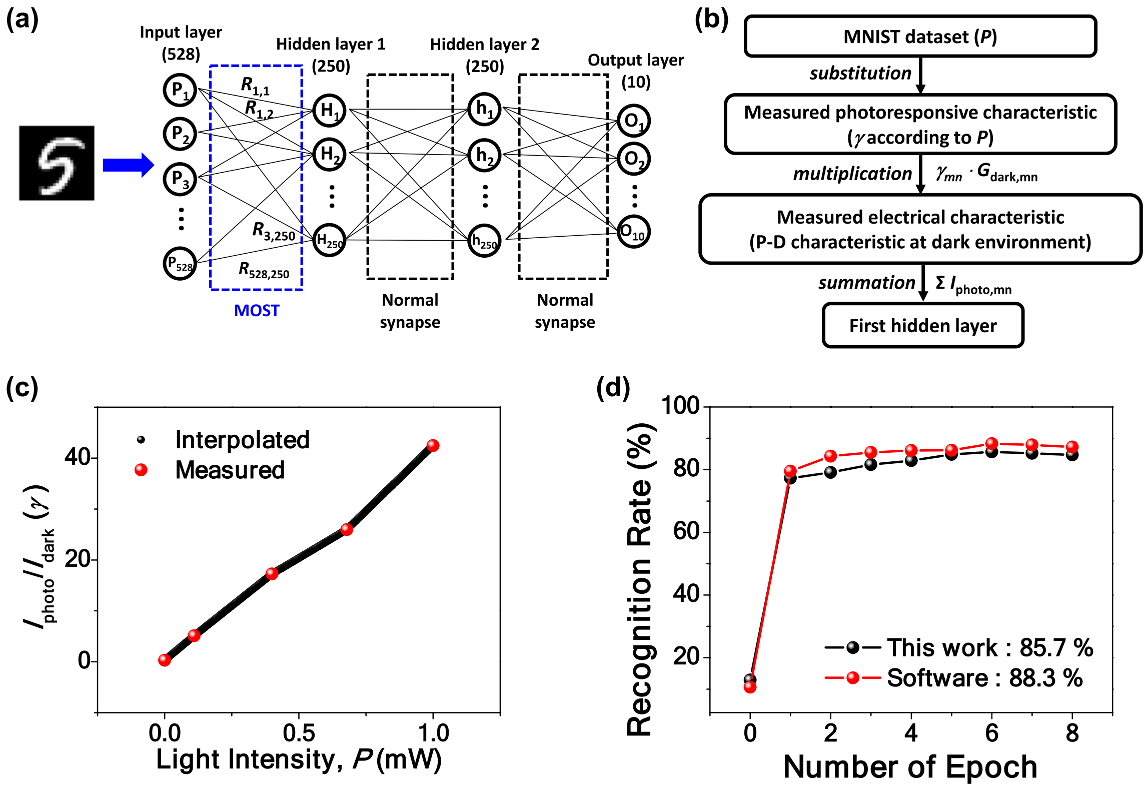

Figure 5. Demonstration of software-based pattern recognition. (**a**) Neural network for recognition of hand-written numbers in the MNIST dataset. Photoresponsive optical characteristics (sensory function) and non-volatile electrical characteristics (mnemonic and synaptic function) measured from the fabricated MOSTs are reflected at the forefront of the network. Measured electrical characteristics are reflected in normal synapses that are connected to the first hidden layer and the second hidden layer, or the second hidden layer and the output layer. (**b**) Software-based simulation sequence to reflect the measured characteristics of the MOST. (**c**) γ ($\equiv I_{photo}/I_{dark}$) as a function of the light intensity (P). Linear interpolation from the measured data is utilized to create extra data. (**d**) Simulated recognition accuracy according to the number of training epochs. Recognition rate of 85.7% is achieved, which is close to the upper limit of 88.3% by an ideal software-based algorithm.

output layer labeled O_A and O_B, as depicted in Fig. 4b. By detecting the output current of the MOSTs connected to each output node, each pattern was recognized. The photoresponsivity that corresponds to the synaptic weight was preset with a binary value, the maximum photoresponsivity and the minimum photoresponsivity, from the data of Fig. 3a. The solid lines and the dashed lines in Fig. 4b represent the device with the maximum photoresponsivity and the minimum photoresponsivity, respectively. Each photoresponsivity is represented as '*R*' in the neural network configuration. This in-sensor processing with the inclusion of image sensing and signal processing performs real-time multiplication of the image with a memorized photoresponsivity matrix[19]. Figure 4c shows the circuit diagram to construct the neural network of Fig. 4b. V_G and V_D were set as 0 V and 1 V, respectively. Each output was measured in the form of the output current: $I_{out,A}$ and $I_{out,B}$; i.e., $I_{out,A}$ was measured in the output node O_A for the input image of 'A' and $I_{out,B}$ was measured in the output node O_B for the input image of 'B', as shown in Fig. 4d. As a result, inference for the simple pattern was experimentally verified. It is worth comparing the required components to distinguish the abovementioned two simple patterns. This work that is applicable to an in-sensor vision system demands only eight MOSTs without extra photodetectors, ADCs or synaptic devices. In contrast, a conventional approach that is suitable for a conventional vision system may need four photodetectors, an ADC, and eight synaptic devices. Thanks to this in-sensor vision system, rapid classification within 1 ms was achieved with low power consumption under 150 nW. This is very small compared to the power consumption of an ADC used for a conventional vision system, which ranges from a few tens of μW to a few mW[35,36].

To demonstrate recognition of more complex patterns such as hand-written digits in the MNIST dataset, a multi-layer perceptron (MLP) network composed of two hidden layers was constructed, as illustrated in Fig. 5a. An input layer corresponds to 528 input pixels, which were cropped from the 28 × 28 pixels, and an output layer corresponds to the 10 numbers from 0 to 9. Each hidden layer is composed of 250 neurons. The MOSTs were located at the forefront of the network for detecting the light intensity and transmitting pre-processed weights with a reflection of optical signals to the first hidden layer. Each device has its own photoresponsivity corresponding to the synaptic weight, which is represented as '*R*' in the neural network configuration. This simultaneous image sensing and signal processing allow real-time multiplication of the image with a memorized photoresponsivity matrix[19]. The measured photoresponsive and P-D characteristics from the fabricated MOSTs in a dark environment were reflected in the software simulations. Figure 5b shows a flow chart that summarizes the simulation sequence to reflect the measured photoresponse characteristics and electrical characteristics of the fabricated MOST. I_{photo} is the drain current with light illumination ($I_{\mathrm{D,light}}$) and I_{dark} is the referenced drain current without light illumination ($I_{\mathrm{D,dark}}$). Except light-on and light-off, all other conditions are the same. Herein the ratio of $I_{\mathrm{photo}}/I_{\mathrm{dark}}$, i.e., $I_{\mathrm{D,light}}/I_{\mathrm{D,dark}}$ is defined as γ, which is extracted from the experimental results. Prior to the simulation, γ was extracted for various light intensities (P) by linear interpolation, as shown in Fig. 5c. For improvement of the simulation accuracy, this step was repeated for each synaptic state. γ of each pixel was extracted by substituting the MNIST dataset into the interpolated curve, because the MNIST dataset represents the pixel intensity. Afterwards, the conductance of each synapse in a dark environment (G_{dark}), which was extracted from the P–D characteristic of Figure S2(b), was multiplied by γ. Because the applied V_{D} of the MOST is 1 V, G_{dark}, defined as $I_{\mathrm{dark}}/V_{\mathrm{D}}$, is simplified to the I_{dark}. The multiplication thus results in I_{photo}. Finally, I_{photo} that contains information of the pixel intensity and the photoresponsivity of the synapse is transmitted to the first hidden layer for summation at each neuron. In detail, current summed in the mth neuron in the first hidden layer (I_{m}) can be represented by the following equation:

$$I_m = \sum_{n=1}^{528} I_{photo,mn} = \sum_{n=1}^{528} \gamma_{mn} G_{dark,mn},$$

where $n = 1, 2, \ldots, 528$ and $m = 1, 2, \ldots, 250$ denote the indices of the pixel and the neuron at the first hidden layer, respectively.

For a normal synapse between the first hidden layer and the second hidden layer or between the second hidden layer and the output layer, only the electrical characteristics (e.g., P–D characteristics at dark environment) were reflected because they could not respond to the light owing to deficiency of a photo-effect. The sigmoid activation function was adopted and supervised learning with back propagation was employed for the learning process to update the synaptic weight of the MOST and a normal synapse. Figure 5d shows the simulated recognition accuracy according to the number of training epochs and the saturated recognition rate was 85.7%. This recognition rate is comparable to an upper limit of 88.3%, which is achievable by software-based pattern recognition simulations that directly multiply the MNIST dataset by the conductance of each synapse, which has ideal P-D characteristics of perfect linearity and symmetry; i.e., $\alpha_{\mathrm{pot}} = 1$ and $\alpha_{\mathrm{dep}} = 1$.

Conclusions

In summary, a mnemonic-opto-synaptic transistor (MOST) was demonstrated for an in-sensor vision system by embedding a non-volatile memory function into a photodetecting device. Because the threshold voltage of the MOST was controlled both by light illumination and by an electrical pulse, the photoresponsivity was tunable by changing the trapped electrons in the charge trap layer (CTL) that enable the non-volatile memory function. Thereby it performed triple functions: photoresponsivity memorizing as a memory cell, light-sensing as a photodetector, and weight updating as a synapse. At the forefront of the ANN, the MOST simultaneously detects light and generates a pre-processed signal to perform real-time multiplication of an image with a memorized photoresponsivity matrix in sensors. More advantageously, it does not require repetitive reset operations because of the invariant synaptic weight during the optical sensing (Table S6). Furthermore, it does not require external memory because of the inherent memory function of the CTL. In addition, the MOST can be integrated with a conventional CMOS image sensor composed of numerous small-sized pixels because it was fabricated with 100% CMOS compatible microfabrication.

Data availability

Scientific Reports requires the inclusion of a data availability statement with all submitted manuscripts, as this journal requires authors to make available materials, data, and associated protocols to readers.

Received: 1 December 2021; Accepted: 7 January 2022
Published online: 02 February 2022

References

1. Zidan, M. A., Strachan, J. P. & Lu, W. D. The future of electronics based on memristive systems. *Nat. Electron.* **1**, 22–29 (2018).
2. Moon, K. *et al.* RRAM-based synapse devices for neuromorphic systems. *Faraday Discuss.* **213**, 421–451 (2019).
3. Van de Burgt, Y. *et al.* A non-volatile organic electrochemical device as a low-voltage artificial synapse for neuromorphic computing. *Nat. Mater.* **16**, 414–418 (2017).
4. Nikam, R. D. *et al.* Ionic sieving through one-atom-thick 2D material enables analog nonvolatile memory for neuromorphic computing. *Small* **17**, 2103543 (2021).
5. Fuller, E. J. *et al.* Parallel programming of an ionic floating-gate memory array for scalable neuromorphic computing. *Science* **364**, 570–574 (2019).
6. Lee, C., Kwak, M., Kim, S. & Hwang, H. Impact of electrolyte density on synaptic characteristics of oxygen-based ionic synaptic transistor. *Appl. Phys. Lett.* **119**, 103503 (2021).
7. Wan, C. *et al.* Artificial sensory memory. *Adv. Mater.* **32**, 1–22 (2020).
8. Lee, G. J., Choi, C., Kim, D. & Song, Y. M. Bioinspired artificial eyes: Optic components, digital cameras, and visual prostheses. *Adv. Funct. Mater.* **28**(1705202), 1–17 (2018).
9. Qiu, W. *et al.* Optoelectronic In-Ga-Zn-O memtransistors for artificial vision system. *Adv. Funct. Mater.* **30**, 2002325 (2020).
10. Zhou, F. *et al.* Optoelectronic resistive random access memory for neuromorphic vision sensors. *Nat. Nanotechnol.* **14**, 776–782 (2019).
11. Ahmed, T. *et al.* Fully light-controlled memory and neuromorphic computation in layered black phosphorus. *Adv. Mater.* **33**, 2004207 (2021).
12. Moini, A. *Vision Chips* Vol. 526 (Springer, Berlin, 2012).

13. Posch, C., Serrano-Gotarredona, T., Linares-Barranco, B. & Delbruck, T. Retinomorphic event-based vision sensors: Bioinspired cameras with spiking output. *Proc. IEEE* **102**, 1470–1484 (2014).
14. Chu, M. *et al.* Neuromorphic hardware system for visual pattern recognition with memristor array and CMOS neuron. *IEEE Trans. Ind. Electron.* **62**, 2410–2419 (2015).
15. Eldred, K. C. *et al.* Thyroid hormone signaling specifies cone subtypes in human retinal organoids. *Science* **362**, eaau66348 (2018).
16. Hong, G. *et al.* A method for single-neuron chronic recording from the retina in awake mice. *Science* **360**, 1447–1451 (2018).
17. Gollisch, T. & Meister, M. Eye smarter than scientists believed: Neural computations in circuits of the retina. *Neuron* **65**, 150–164 (2010).
18. Wang, C. Y. *et al.* Gate tunable van der Waals heterostructure for reconfigurable neural network vision sensor. *Sci. Adv.* **6**, eaba6173 (2020).
19. Mennel, L. *et al.* Ultrafast machine vision with 2D material neural network image sensors. *Nature* **579**, 62–66 (2020).
20. Ng, H. T. *et al.* Single crystal nanowire vertical surround-gate field-effect transistor. *Nano Lett.* **4**, 1247–1252 (2004).
21. Hu, L. & Chen, G. Analysis of optical absorption in silicon nanowire arrays for photovoltaic applications. *Nano Lett.* **7**, 3249–3252 (2007).
22. Wei, W. *et al.* Direct heteroepitaxy of vertical InAs nanowires on Si substrates for broad band photovoltaics and photodetection. *Nano Lett.* **9**, 2926–2934 (2009).
23. Lue, H. T. *et al*. BE-SONOS: A Bandgap Engineered SONOS with Excellent Performance and Reliability. *IEDM Tech. Dig*. 547–550 (2005).
24. Lue, H. T. *et al*. Reliability Model of Bandgap Engineered SONOS (BE-SONOS). *IEDM Tech. Dig*. 495–498 (2006).
25. Wasapinyokul, K., Milne, W. I. & Chu, D. P. Origin of the threshold voltage shift of organic thin-film transistors under light illumination. *J. Appl. Phys.* **109**, 084510 (2011).
26. Seo, M. *et al.* First demonstration of a logic-process compatible junctionless ferroelectric FinFET synapse for neuromorphic applications. *IEEE Electron Device Lett.* **39**, 1445–1448 (2018).
27. Yu, J.-M. *et al.* All-solid-state ion synaptic transistor for wafer-scale integration with electrolyte of a nanoscale thickness. *Adv. Funct. Mater.* **31**, 2010971 (2021).
28. Yu, S. Neuro-inspired computing with emerging nonvolatile memory. *Proc. IEEE* **106**, 260–285 (2018).
29. Park, J. *et al.* TiOx-based Rram synapse with 64-levels of conductance and symmetric conductance change by adopting a hybrid pulse scheme for neuromorphic computing. *IEEE Electron Device Lett.* **37**, 1559–1562 (2016).
30. Schroder, D. K. Transparent gate silicon photodetectors. *IEEE Trans. Electron Devices* **25**, 90–97 (1978).
31. Membel, S. E. H., Membel, N. W. W. & Membel, W. W. M. S. Development of low noise, back-side illuminated silicon photodiode arrays. *IEEE Trans. Nucl. Sci.* **44**, 443–447 (1997).
32. Darmont, A. *Spectral response of silicon image sensors.* 4, 1–13 (White Paper, 2009).
33. Seo, S. *et al.* Artificial optic-neural synapse for colored and color-mixed pattern recognition. *Nat. Commun.* **9**, 5106 (2018).
34. Kim, S. K. *et al.* Photo-responsible synapse using Ge synaptic transistors and GaAs photodetectors. *IEEE Electron Device Lett.* **41**, 605–608 (2020).
35. Cottini, N., Gasparini, L., De Nicola, M., Massari, N. & Gottardi, M. A CMOS ultra-low power vision sensor with image compression and embedded event-driven energy-management. *IEEE J. Emerg. Sel. Top. Circuits Syst.* **1**, 299–307 (2011).
36. Fayyazi, A., Ansari, M., Kamal, M., Afzali-Kusha, A. & Pedram, M. An ultra low-power memristive neuromorphic circuit for internet of things smart sensors. *IEEE Internet Things J.* **5**, 1011–1022 (2018).

Acknowledgements

This work was supported by the National Research Foundation (NRF) of Republic of Korea (2018R1A2A3075302, 2019M3F3A1A03079603, and 2020M3F3A2A01082592), and in part by IC Design Education Center (EDA Tool and MPW).

Author contributions

J.-K.H. and Y.-W.C. equally contributed to this work. Y.-K.C. and S.-H.K. conceived, supervised, and led the project. J.-K.H., Y.-W.C., J.-M.Y. and Y.-K.C. designed the experiment and found a mechanism to propose MOSTs operation. J.S. conducted experiments on various light sources. J.-K.H. and Y.-W.C. conducted the simulation and wrote the manuscript. All the authors interpreted data, contributed reviewing the manuscript, and approved the final version of the article.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-05944-y.

Correspondence and requests for materials should be addressed to Y.-K.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.