IN-SENSOR COMPUTING

Silicon photodiodes that multiply

Silicon-based dual-gate photodiodes with electrostatically controlled photocurrents can be used to create imaging systems that can compute incoming visual data.

Yang Chai

Image processing with artificial intelligence is of use in a range of applications from face recognition and authentication to autonomous vehicles and industrial manufacturing[1–3]. Most machine learning algorithms require high-performance hardware for such data-intensive applications, and local data from sensory terminals are typically transferred to sophisticated computation units or the 'cloud' for the necessary processing. This movement of data among image sensors, memory and processing units greatly increases power consumption and latency, creating challenges in power-constrained and widely distributed camera nodes and other Internet of Things (IoT) systems. New computing approaches are thus needed that can reduce data transfer, efficiently process image information inside sensors and execute machine learning algorithms.

In-sensor computing is an approach that allows multiple sensors to directly process information without the need to transfer data to external processing units[4,5]. Optoelectronic conversion is a key process in image sensing, which also includes information reconstruction and reorganization. By using different photoresponse characteristics (linear or nonlinear), researchers have been able to demonstrate feature enhancement or image recognition within sensor arrays[3,5]. Most of these demonstrations have used devices based on emerging materials that lack mature process technologies. However, achieving this in mature silicon sensors is challenging as they are chemically doped. Writing in *Nature Electronics*, Seongjun Park, Donhee Ham and colleagues now report in-sensor optoelectronic computing using silicon technology[6].

The researchers — who are based at Harvard University, Brookhaven National Laboratory, Samsung Electronics, Korea Institute of Science and Technology, and Pusan National University — fabricated a dual-gate silicon p–i–n diode in which the carrier polarity and density are electrostatically controlled by the voltages applied to the gate terminal. The photoresponsivity of the silicon p–i–n diode can be continuously modulated by the gate voltage, emulating the synaptic weight in an artificial neural network. Owing to the mature silicon processing technology employed, the resulting photocurrent shows high linearity as a function of light intensity, enabling high accuracy in machine learning applications[3,6].

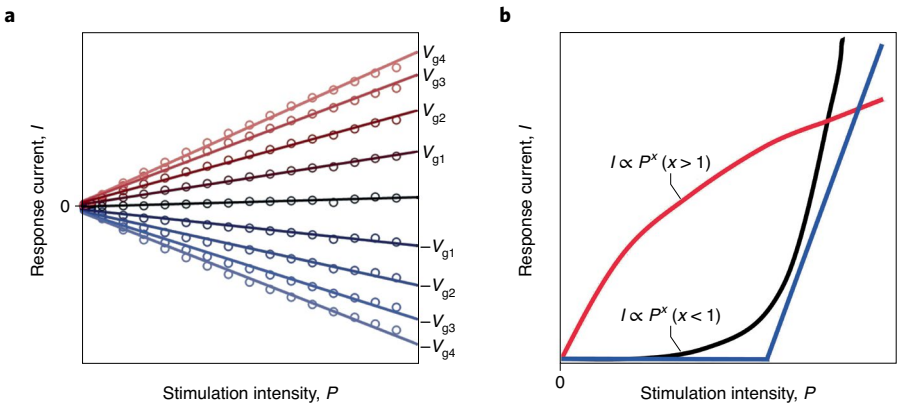

Fig. 1 | Response characteristics of sensors for in-sensor computing. a, Conventional sensors typically require linear response characteristics that can accurately represent external stimulation. For in-sensor computing with artificial neural networks, the responsivity can represent synaptic weights and can be modulated by gate voltages (V_{gi}). Therefore, the linear response characteristics can provide high precision in in-sensor computing with artificial neural networks. **b**, Nonlinear response characteristics of sensors, including superlinear (black), sublinear (red) and threshold (blue), can output intensity-dependent information, which can be used to encode spatiotemporal information and enrich computational functions at sensory terminals. Panel **a** adapted with permission from ref. [6], Springer Nature Ltd.

Ham and colleagues fabricated 4,900 dual-gate p–i–n diodes on a 4-inch silicon wafer. In an individual photodiode, optoelectronic conversion provides the multiplication function $I_{ph} = R \times P$, where I_{ph} is the photocurrent, R is the photoresponsivity and P is the power intensity of light stimulation. In the photodiode array, when connected in series, the photocurrent can be summed according to Kirchhoff's current law. Thus, machine learning algorithms can be executed via multiplication-and-accumulation operations based on the physical properties of the hardware. In this way, the photodiode array can simultaneously capture and process an image without transferring the data outside the array, exhibiting a high computation speed at the level of nanoseconds. The team chose a 3 × 3 photodiode array as an image filter kernel. This photodiode array can effectively recognize the edge of an image pattern by displaying different polarities of the photocurrent. Furthermore, the researchers adopted different filter kernels and extended in-sensor processing to 256 × 256 pixel images, demonstrating high parallelism in the in-sensor computing scheme.

The use of silicon technology reduces the device-to-device variations that are typically associated with emerging materials and makes it feasible to monolithically integrate in-sensor computing units with peripheral control circuits. These advantages could accelerate the hardware implementation of large-scale in-sensor computing approaches. However, to

achieve this, further developments will be required. To start, and at the device level, a fundamental understanding of the response characteristics of individual sensors could help the design of different computational functions. Conventional sensors usually require high linearity (Fig. 1a) to faithfully record external stimulation. Depending on how they are used, high-linearity sensors can also aid high accuracy in in-sensor computing that uses artificial neural networks[3,6]. Nonlinear sensors (Fig. 1b) concisely restructure spatiotemporal information and can dynamically adapt to unknown and complicated stimuli[7,8], such as the modulated threshold and responsivity of nociceptors according to their history states and external stimulation[9,10]. Threshold response characteristics also provide a potential way to achieve more energy-efficient spike encoding. Such nonlinear characteristics can enrich the functions of in-sensor computing. Therefore, understanding the fundamental device physics of these sensors is essential, and the development of relevant compact models will help the design of large-scale in-sensor computing systems.

To extend existing small-scale demonstrations to large-scale foundry-grade processes, careful design of a compact layout that will fit silicon processing is needed. A large neural network with more output nodes increases the complexity of the interconnections between individual sensors, thus requiring specific consideration of interconnect routing and layout design. For example, if two-terminal sensors are arranged in a crossbar array, the sneak current from neighbouring cells will lead to computation errors. Therefore, selectors need to be integrated into the large-scale array. Work with in-sensor computing has, to date, typically focused on relatively simple functions. Heterogeneous or monolithic integration of in-sensor computing modules with other computation units will allow the execution of more complicated tasks.

In terms of functionality, most in-sensor computing demonstrations focus on the recognition of static scenarios. However, the real world is dynamic and complicated. It is important to develop in-sensor computing approaches that can effectively respond to dynamic objects and still reduce data volume. Spike coding provides a promising solution to efficiently process dynamic scenarios, but it remains difficult to implement spike coding within the sensor. Hardware and software co-design of in-sensor computing could be used though to develop new functionalities and be extended to multimodal sensory signals. ❐

Yang Chai ✉

Department of Applied Physics, the Hong Kong Polytechnic University, Kowloon, Hong Kong, P. R. China.

✉*e-mail: ychai@polyu.edu.hk*

Published online: 23 August 2022
https://doi.org/10.1038/s41928-022-00822-x

References

1. Mead, C. *Proc. IEEE* **78**, 1629–1636 (1990).
2. Kyuma, K. et al. *Nature* **372**, 197–198 (1994).
3. Mennel, L. et al. *Nature* **579**, 62–66 (2020).
4. Chai, Y. *Nature* **579**, 32–33 (2020).
5. Zhou, F. & Chai, Y. *Nat. Electron.* **3**, 664–671 (2020).
6. Jang, H. et al. *Nat. Electron.* https://doi.org/10.1038/s41928-022-00819-6 (2022).
7. Zhou, F. et al. *Nat. Nanotechnol.* **14**, 776–782 (2019).
8. Liao, F. et al. *Nat. Electron.* **5**, 84–91 (2022).
9. Yoon, J. et al. *Nat. Commun.* **9**, 417 (2018).
10. Kumar, M., Kim, H. S. & Kim, J. *Adv. Mater.* **31**, 1900021 (2019).

Competing interests
The author declares no competing interests.