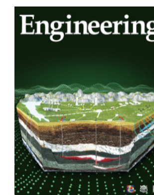
Engineering 14 (2022) 19–21

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng



Views & Comments

A Future Perspective on In-Sensor Computing

Wen Pan [a], Jiyuan Zheng [b], Lai Wang [a,b], Yi Luo [a,b]

[a] *Department of Electronic Engineering, Tsinghua University, Beijing 100084, China*
[b] *Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China*

The use of artificial intelligence (AI) is escalating rapidly in most applications nowadays, thanks to breakthroughs in biology and mathematics. Novel hardware systems are greatly needed to meet the requirements of AI, which include computing capacity and energy efficiency. One of the major aims of AI is to mimic the functions of the human brain, which are enabled by the massive interconnection of neurons. For example, the visual cortex is the region of the brain that processes visual information. The human vision system, which includes the visual cortex, is highly compact and energy efficient. The retina contains hundreds of millions of light-sensitive neurons interconnected by preprocessing and control neurons to enhance image quality, extract features, and recognize objects. Once light-sensitive neurons have detected trivial signals, they are disabled thereafter, and only the critical information is transferred to the cortex for deep processing.

The artificial imaging hardware systems that are commonly used at present, however, do not function like the human visual system. Sensors such as charge-coupled device (CCD) arrays and complementary metal oxide semiconductor (CMOS) arrays are interconnected serially with memory and processing units, through bus lines (i.e, Von Neumann architecture). Although current imaging hardware systems have an advantage over human brains in sensing unit density, response time, and sensitive wavelength range, their power consumption and processing latency are becoming problematic when a complex AI mission is being conducted. In most imaging processing applications, more than 90% of the data generated by sensors is redundant and useless [1]. As the number of pixels increases rapidly, the volume of unnecessary data multiplies, imposing a severe burden on analog-to-digital conversion (ADC) and data movement, and limiting the development of real-time image processing technology [2]. As a result, AI rapidly uses up hardware resources. Thus, there is strong demand for a breakthrough in hardware systems, which will surely emerge shortly.

Inspired by the human vision system, researchers have attempted to shift some processing tasks to sensors, thereby allowing *in situ* computing and reducing data movement. For example, Mead and Mahowald [3] at the California Institute of Technology proposed the AI vision chip in the 1990s. They envisioned a semiconductor chip that could capture images, directly carry out the parallel processing of visual information, and eventually output the processing results. Early vision chips aimed to imitate the retina's preprocessing function but could only achieve low-level processing, such as image filtering and edge detection [2]. Gradually, low-level processing was found to be insufficient, and high-level processing, including recognition and classification, became the goal for AI vision chips. Moreover, researchers proposed the development of programmable vision chips around 2006, with the goal of flexibly dealing with various processing scenes through software control [4]. In 2021, Liao et al. [5] summarized the principle of the biological retina and discussed developments in neuromorphic vision sensors in emerging devices. Wan et al. [6] provided an overview of the technology of electronic, optical, and hybrid optoelectronic computing for neuromorphic sensory computing.

There are currently two significant types of vision chip architecture [2,4,7].

(1) **Architectures with computing inside sensing units.** In this type of architecture, the photodetector is placed directly into the analog memory and computing unit to form a processing element (PE) [4,8,9]. The PEs are then developed to possess *in situ* sensing and to deal with the analog signals obtained by the sensors. This type of architecture, which is illustrated in Fig. 1(a) [10], has the advantage of highly parallel processing speed. However, the analog memory and computing unit takes up a large volume, which makes the PEs much larger than the sensor; this results in a low pixel fill factor and limits the image resolution.

(2) **Architectures with computing near the sensing units.** Most vision chips cannot incorporate *in situ* sensing and computing architecture due to the low fill factor issue. Instead, the pixel array and processing circuits are separated physically while still being connected in parallel on a chip [4,7], which makes independent design possible according to the system's requirements. This type of architecture is illustrated in Fig. 1(b) [10]. The sensing data (analog) is first extracted from the sensor array through the bus line and converted into a digital signal, which is then dealt with in the nearby processing unit. This architecture has the specific capabilities of wide-area image processing, high resolution, and large-scale parallel processing. In addition, AI algorithms, including artificial neural networks, can be conducted in this architecture in the digital process circuits.

https://doi.org/10.1016/j.eng.2022.01.009
2095-8099/© 2022 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

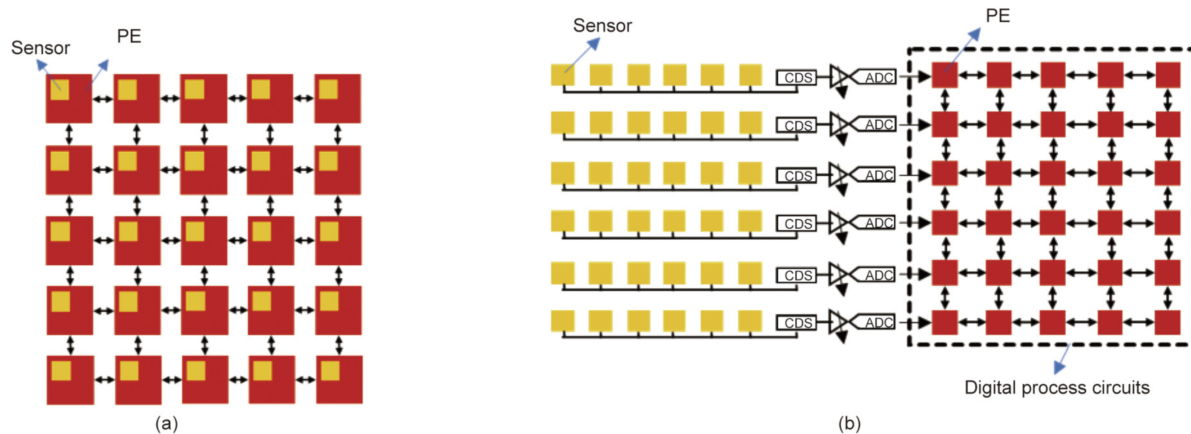


Fig. 1. Vision chip architecture. (a) Computing inside the sensing unit; (b) computing near the sensing unit. CDS: correlation double sampling. Reproduced from Ref. [10] with permission of IEEE, ©2014.

The current vision chip only has a neuron scale of 10^2–10^3, which is much smaller than those of the retina and cortex (10^{10}). Therefore, larger scale integration technology is needed to achieve a greater neuron scale for in-sensor computing. One such method is implemented by convolutional neural networks (CNNs) and spiking neuron networks (SNNs) to significantly improve the processing efficiency. The other method is to adopt three-dimensional (3D) integration technology to vertically integrate the functional layers (sensor, memory, computing, communication, etc.) in space using through-silicon vias (TSVs) [11]. In 2017, Sony proposed a 3D integrated vision chip with a pixel resolution of 1296 × 976 and a processing speed reaching 1000 frames per second (fps) [12]. Some researchers believe that the 3D integrated chip has become an inevitable trend. However, further development of 3D integration technology is still necessary in areas such as architecture design and interconnections. It has been demonstrated that, although short interconnects could lower power consumption and latency, they could introduce thermal problems due to the short distance between layers [13,14]. Thus, it is crucial for the reliability issues of 3D integration to be solved and for the performance to be improved.

Driven by the need for AI development, technologies involving novel material systems and advanced devices have recently been emerging.

(1) **Detect-and-memorize (DAM) materials.** Photonic synaptic devices [15–20] have been proposed as a means of constructing in-sensor computing systems and are expected to facilitate the evolution of retina-mimicking technologies. It has been found that some metal oxides (oxide semiconductors, binary oxides, etc.), oxide heterojunctions, and two-dimensional (2D) materials [15] hold great potential as DAM materials for the realization of photonic synaptic devices. Photonic synapses possess temporary memory and synaptic plasticities, such as short-term plasticity (STP) and long-term plasticity (LTP), which can be modulated by light signals to implement real-time image processing. These devices have the advantages of ultrahigh propagation speed and high bandwidth; they also provide a noncontact writing method. However, some issues remain to be addressed, including nonlinear writing and high energy consumption due to the relatively large illumination intensity. Potentiation is achieved under optical stimuli during the writing process, while electric stimuli are utilized for habituation [21]. To be specific, the conductance of devices increases gradually upon a series of photonic pulses due to the photogenerated electrons and holes, and decreases gradually under negative electric pulses, which is similar to the potentiation and depression in a biological synapse. Hence, it is expected to obtain a negative photoresponse and achieve habituation under optical stimulation [15,22]. Most studies focus on mimicking synaptic behaviors (excitatory postsynaptic current (EPSC), paired-pulse facilitation (PPF), STP, LTP, etc.) in devices, as imitating the retinal neurons in the human eye remains a major challenge. In order to imitate the retina, the scaling-up of photonic synaptic devices requires further study. Among DAM materials, devices based on binary oxides (e.g., ZnO, HfO_2, AlO_x, etc.) have the advantages of a simple device structure and CMOS compatibility, which are the decisive factors for scaling-up. In contrast, materials that are incompatible with an integrated circuit (IC) infrastructure can be used by adopting technologies such as heterogeneous integration [23], heteroepitaxy [24], bonding [25], and 3D heterogeneous integration [14].

(2) **Device structures that combine sensor and memory.** Researchers have proposed that PEs be replaced by advanced devices, such as storage elements (i.e., resistive random-access memory (RRAM) and other memristors) [26–28]. For example, combining these device-intrinsic features in a serial connection of both elements [26] makes the sensor array programmable and converts the light image into information that can be easily recognized. This structure significantly reduces the footprint of a single pixel down to the theoretical limit of $4F^2$ (F is the feature size of the process), allowing integration with a high fill factor. Unlike CCD, however, this array does not show a destructive read-out and does not exhibit any integrating behavior. In this array, multiply-and-accumulation (MAC) operations can be directly implemented through Kirchhoff's law in the analog domain [2,29]; however, crosstalk caused by large-scale integration is an urgent problem that remains to be solved. Researchers have also proposed a system comprised of single-photon avalanche diodes (SPADs) and memristors [30,31] to process information in the form of spike events, which would allow real-time imaging recognition.

New architectures or even algorithms must be introduced to accommodate the emerging materials and device technologies. For example, applying deep learning algorithms (deep neural networks (DNNs), CNNs, SNNs, etc.) to in-sensor computing is an urgent issue. SNNs provide a promising solution to enhance efficiency by encoding and processing time-encoded neural signals in parallel [2].

This paper presented a summary of two different kinds of architecture (i.e., with computing inside or near the sensing units) utilized in in-sensor computing and then discussed future development directions (including architecture matching with algorithms, 3D integration technology, novel material systems, and advanced devices). In sum, the ultimate goal for in-sensor computing is to achieve efficient AI hardware that has low power

consumption, high speed, high resolution, high accuracy recognition, and large-scale integration, while being programmable. To commercialize in-sensor computing technology, further research is needed in physics, materials, computer science, electronics, and biology.

Acknowledgments

The authors highly appreciate Professor Supratik Guha from the University of Chicago for his useful discussion to improve the paper. This work is funded by the National Key Research and Development Program of China (2021YFA0716400), the National Natural Science Foundation of China (61904093, 61975093, 61991443, 61974080, 61927811, 61822404, 62175126, and 61875104), the Key Lab Program of BNRist (BNR2019ZS01005), the China Postdoctoral Science Foundation (2018M640129 and 2019T120090), and the Collaborative Innovation Center of Solid-State Lighting and Energy-Saving Electronics the Ministry of Science and Technology of China (2021ZD0109900 and 2021ZD0109903).

References

[1] Chai Y. In-sensor computing for machine vision. Nature 2020;579(7797):32–3.
[2] Zhou F, Chai Y. Near-sensor and in-sensor computing. Nat Electron 2020;3(11):664–71.
[3] Mead CA, Mahowald MA. A Silicon model of early visual processing. Neural Netw 1988;1(1):91–7.
[4] Liu L, Wu N. Artificial intelligent vision chip. Micro/nano Electron Intell Manuf 2019;1:12–9. Chinese.
[5] Liao F, Zhou F, Chai Y. Neuromorphic vision sensors: principle, progress and perspectives. J Semicond 2021;42(1):013105.
[6] Wan T, Ma S, Liao F, Fan L, Chai Y. Neuromorphic sensory computing. Sci China Inf Sci 2022;65:141401.
[7] Wu N. Neuromorphic vision chips. Sci China Inf Sci 2018;61:060421.
[8] Komuro T, Kagami S, Ishikawa M. A dynamically reconfigurable SIMD processor for a vision chip. IEEE J Solid-State Circuits 2004;39(1):265–8.
[9] Jendernalik W, Blakiewicz G, Jakusz J, Szczepanski S, Piotrowski R. An analog sub-miliwatt CMOS image sensor with pixel-level convolution processing. IEEE Trans Circuits Syst I Regul Pap 2013;60(2):279–89.
[10] Shi C, Yang J, Han Y, Cao Z, Qin Q, Liu L, et al. A 1000 fps vision chip based on a dynamically reconfigurable hybrid architecture comprising a PE array processor and self-organizing map neural network. IEEE J Solid-State Circuits 2014;49(9):2067–82.
[11] Feng P, Liu L, Wu N. Photoelectric and 3D integrated artificial intelligent vision chip. Micro/nano Electron Intell Manuf 2019;1:75–84. Chinese.
[12] Yamazaki T, Katayama H, Uehara S, Nose A, Kobayashi M, Shida S, et al. 4.9 A 1 ms high-speed vision chip with 3D-stacked 140GOPS column-parallel PEs for spatio-temporal image processing. In: Proceedings of 2017 IEEE International Solid-State Circuits Conference (ISSCC); 2017 Feb 5–9; San Francisco, CA, USA. New York: IEEE; 2017. p. 82–3.
[13] Amir MF, Ko JH, Na T, Kim D, Mukhopadhyay S. 3D stacked image sensor with deep neural network computation. IEEE Sens J 2018;18(10):4187–99.
[14] Lie D, Chae K, Mukhopadhyay S. Analysis of the performance, power, and noise characteristics of a CMOS image sensor with 3D integrated image compression unit. IEEE Trans Compon Packaging Manuf Technol 2014;4(2):198–208.
[15] Zhang J, Dai S, Zhao Y, Zhang J, Huang J. Recent progress in photonic synapses for neuromorphic systems. Adv Intell Syst 2020;2(3):1900136.
[16] Dai S, Wu X, Liu D, Chu Y, Wang K, Yang B, et al. Light-stimulated synaptic devices utilizing interfacial effect of organic field-effect transistors. ACS Appl Mater Interfaces 2018;10(25):21472–80.
[17] Gao S, Liu G, Yang H, Hu C, Chen Q, Gong G, et al. An oxide Schottky junction artificial optoelectronic synapse. ACS Nano 2019;13(2):2634–42.
[18] Hu DC, Yang R, Jiang L, Guo X. Memristive synapses with photoelectric plasticity realized in ZnO_{1-x}/AlO_y heterojunction. ACS Appl Mater Interfaces 2018;10(7):6463–70.
[19] Kumar M, Abbas S, Kim J. All-oxide-based highly transparent photonic synapse for neuromorphic computing. ACS Appl Mater Interfaces 2018;10(40):34370–6.
[20] Lee M, Lee W, Choi S, Jo JW, Kim J, Park SK, et al. Brain-inspired photonic neuromorphic devices using photodynamic amorphous oxide semiconductors and their persistent photoconductivity. Adv Mater 2017;29(28):1700951.
[21] He HK, Yang R, Zhou W, Huang HM, Xiong J, Gan L, et al. Photonic potentiation and electric habituation in ultrathin memristive synapses based on monolayer MoS_2. Small 2018;14(15):e1800079.
[22] Wu JY, Chun YT, Li S, Zhang T, Wang J, Shrestha PK, et al. Broadband MoS_2 field-effect phototransistors: ultrasensitive visible-light photoresponse and negative infrared photoresponse. Adv Mater 2018;30(7):1705880.
[23] Matsuo S. Heterogeneously integrated III–V photonic devices on Si. Semicond Semimetals 2019;101:43–89.
[24] Teichert C. Self-organization of nanostructures in semiconductor heteroepitaxy. Phys Rep 2002;365(5–6):335–432.
[25] Benaissa L, Di Cioccio L, Beilliard Y, Coudrain P, Dominguez S, Balan V, et al. Next generation image sensor via direct hybrid bonding. In: Proceedings of 17th IEEE Electronics Packaging and Technology Conference (EPTC); 2015 Dec 2–4; Singapore. New York: IEEE; 2015. p. 1–3.
[26] Nau S, Wolf C, Sax S, List-Kratochvil EJ. Organic non-volatile resistive photo-switches for flexible image detector arrays. Adv Mater 2015;27(6):1048–52.
[27] Wang H, Liu H, Zhao Q, Ni Z, Zou Y, Yang J, et al. A retina-like dual band organic photosensor array for filter-free near-infrared-to-memory operations. Adv Mater 2017;29(32):1701772.
[28] Wang H, Zhao Q, Ni Z, Li Q, Liu H, Yang Y, et al. A ferroelectric/electrochemical modulated organic synapse for ultraflexible, artificial visual-perception system. Adv Mater 2018;30(46):e1803961.
[29] Mennel L, Symonowicz J, Wachter S, Polyushkin DK, Molina-Mendoza AJ, Mueller T. Ultrafast machine vision with 2D material neural network image sensors. Nature 2020;579(7797):62–6.
[30] Shawkat MSA, Sayyarparaju S, McFarlane N, Rose GS. Single photon avalanche diode based vision sensor with on-chip memristive spiking neuromorphic processing. In: Proceedings of 2020 IEEE 63rd International Midwest Symposium on Circuits and Systems (MWSCAS); 2020 Aug 9–12; Springfield, MA, USA. New York: IEEE; 2020. p. 377–80.
[31] Sayyaparaju S, Weiss R, Rose GS. A mixed-mode neuron with on-chip tunability for generic use in memristive neuromorphic systems. In: Proceedings of 2018 IEEE Computer Society Annual Symposium on VLSI (ISVLSI); 2018 Jul 8–11; Hong Kong, China. New York: IEEE; 2018. p. 441–6.