Contents lists available at ScienceDirect

Engineering

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

Views & Comments

感存算一体化智能视觉芯片展望

潘汶[a], 郑纪元[b], 汪莱[a,b], 罗毅[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*

近年来，人工智能（AI）的应用变得越来越广泛，其发展也随着生物学、数学的进步而日渐成熟。然而，AI的发展也对系统的计算能力和能量效率提出了更高的要求，因此迫切需要新的硬件架构来满足AI的需求。AI的目标是使机器获得类似人的智能，当前的硬件体系在执行计算时仍然基于传统的冯·诺依曼架构。首先通过传感器得到物理信号，然后将信号传输至计算中心结合算法进行感知，这种信息处理的模式与人脑完全不同。以视觉为例，人类的视觉系统（包含视觉皮层）是高度紧凑和高效的，其中，视网膜上的数亿光敏神经元与预处理、控制神经元相连接，能够实现感光和信号预处理（增强图像、提取特征等）。一旦光敏神经元检测到冗余信号，视觉系统会将其弱化，仅将关键信息传输至大脑皮层进行深度处理。

目前常用的人工成像硬件系统的功能并不像人类视觉系统那样，如电荷耦合器件（CCD）阵列和互补金属氧化物半导体（CMOS）阵列，这类传感器通过总线将图像数据串行传输至存储器和处理单元进行交互运算（即冯·诺依曼架构）。尽管当前的成像硬件系统在传感单元密度、响应时间和波长范围方面优于人类视觉系统，但在执行复杂AI任务时，它们的功耗和延时变成了不可忽略的问题。在大多数图像处理任务中，超过90%的图像数据是冗余且无用的[1]，随着像素数量的快速增长，数据冗余量显著增加，给模数转换（ADC）和数据传输带来了严重负担，并限制了实时图像处理技术的发展[2]。因此，AI的发展会迅速消耗硬件资源，并产生对新型硬件系统的强烈需求。

受人类视觉系统的启发，部分研究尝试将一些处理任务转移至图像传感器内，从而实现原位计算，并且减少数据传输。在20世纪90年代，加州理工学院的Mead和Mahowald [3]提出了人工智能视觉芯片，他们构想了一种可以同时获取图像、处理图像的半导体芯片，这种芯片可以将获取的图像数据进行并行处理，最终输出处理结果。早期的视觉芯片旨在模仿视网膜的预处理功能，但只能实现简单图像处理，如图像滤波和边缘检测[2]，而后逐渐提出在传感器内部实现复杂图像处理，包括图像识别和分类，这也成为了人工智能视觉芯片的目标。此外，在2006年提出视觉芯片需要具备可编程功能，从而通过软件控制灵活地处理各种应用场景[4]。在2021年，Liao等[5]总结了生物视网膜的原理，并讨论了基于新兴器件的智能视觉传感器发展。Wan等[6]概述了用于神经拟态传感计算的电子、光学以及混合光电计算技术。

目前有两种主要的智能视觉芯片架构[2,4,7]。

（1）架构一：传感单元内部计算。这种架构的光电探测器被置于模拟存储器和计算单元中，以组成处理元件（PE）[4,8–9]，然后利用PE电路来实现原位传感功能，并处理传感器获得的模拟信号。这种架构如图1（a）所示，其优势在于具有高度并行处理速度。然而，模拟存储器和计算单元占用了较大的面积，使得PE电路比传统传感单元大得多，这导致像素填充因子较低，并限制了成像分辨率。

（2）架构二：传感单元附近计算。由于较低的填充因子，视觉芯片难以采用原位传感和计算相结合的架构。相

英文原文：Engineering 2022, 14(7): 19–21

引用本文：Wen Pan, Jiyuan Zheng, Lai Wang, Yi Luo. A Future Perspective on In-Sensor Computing. *Engineering*, https://doi.org/10.1016/j.eng.2022.01.009

反，将像素阵列和处理电路物理分离，但仍然保持片上并行连接[4,7]，这使得二者可以根据系统要求进行独立设计和优化。这种架构如图1（b）所示，首先通过总线从像素阵列中获取传感数据（模拟），并转换成数字信号，然后在附近的处理单元内进行计算。这种架构具有广域图像处理、高分辨率和大规模并行处理的优势，并且可以在数字处理电路中结合现有的AI算法（包括人工神经网络等）。

目前，视觉芯片的神经元规模只有10^2~10^3个，远少于视网膜和皮层（10^{10}个），因此，感存算一体化智能视觉芯片需要更大规模的集成。其中一种方法是通过片上光学卷积神经网络（CNN）和光学脉冲神经网络（SNN）实现大规模高并行运算，以显著提高计算效率。另一种方法是采用三维（3D）集成技术，使用硅通孔（TSV）垂直集成空间中的功能层（传感器、存储器、计算、通信等）[10]。在2017年，索尼提出了一种3D集成视觉芯片，像素分辨率为1296×976，处理速度达到1000 fps [11]。部分研究人员认为，3D集成芯片已经成为一种必然趋势，但在架构设计和引线互连等方面仍然需要更深入的研究。研究证明，虽然短互连可以降低功耗和延迟，但由于层间距离较短可能会导致散热难题[12–13]。因此，解决3D集成的可靠性问题和提高性能至关重要。

近些年来，在AI发展需求的驱动下，涉及新型材料和先进器件的技术不断涌现，也为感存算一体化智能成像系统提供了新方案。

（1）具有探测和记忆功能的材料（DAM）。光电突触器件[15–20]被视为构建感存算一体化智能成像系统的一种方式，并有望促进视网膜仿生技术的发展。研究发现，一些金属氧化物（氧化物半导体、二元氧化物等）、氧化物异质结和二维（2D）材料[15]在实现光电突触器件方面有巨大的潜力。光电突触具有临时记忆能力和突触可塑性，如短时程可塑性（STP）和长时程可塑性（LTP），可以通过光信号进行调制以完成实时图像处理。这类器件有许多优点，它提供了一种非接触式的写入方法（光写入），权重写入过程具有高速并行的特点。然而，这类器件仍然面临一些挑战，包括脉冲写入下电导非线性变化和由于较大的照明强度而导致的高能耗。在写入过程中，光刺激用于实现增强突触活性[21]，而电刺激用于抑制突触活性[21]。具体来说，器件的电导在光脉冲作用下逐渐增加，而在负电脉冲作用下则逐渐减小，这类似于生物突触中的增强和抑制，器件的电导变化对应突触的活性变化。此外，研究指出负光响应或者光刺激用于抑制突触活性[15,22]可以实现全光调制的复杂神经功能。目前大多数研究侧重于在器件中模拟突触行为[如兴奋性突触后电流（EPSC）、成对脉冲易化（PPF）、STP、LTP等]，因为模仿人眼视网膜神经元仍然是一大挑战。为了模仿视网膜，光电突触器件的大规模集成有待进一步研究。在DAM材料中，基于二元氧化物（如ZnO、HfO_2、AlO_x等）的器件具有结构简单和CMOS兼容性的优点，这是大规模集成的关键因素。相反，与集成电路（IC）基础结构不兼容的材料可以通过采用异质集成[23]、异质外延[24]、键合[25]和三维异质集成[13]等技术来实现。

（2）结合传感器与存储器的器件结构。近些年来，随着半导体器件的发展，部分研究提出采用先进器件代替PE电路，如新型存储器件[如阻变存储器（RRAM）和其他忆阻器等] [26–28]。例如，两类器件通过串联的方式来实现固有特性的结合[26]，使传感器阵列具有可编程性，并且将光学图像转变为易于识别的信息。这种结构将单个像素的占地面积显著降低到$4F^2$的理论极限（F是工艺的特征尺寸），可以实现高填充因子的集成方式。然而，与CCD不同的是，该阵列不显示破坏性读出，也不显示任何积分行为。在该阵列中，乘加运算（MAC）可以通过模拟域中的基尔霍夫定律直接实现[2,29]。然而，大规模集成引起的串

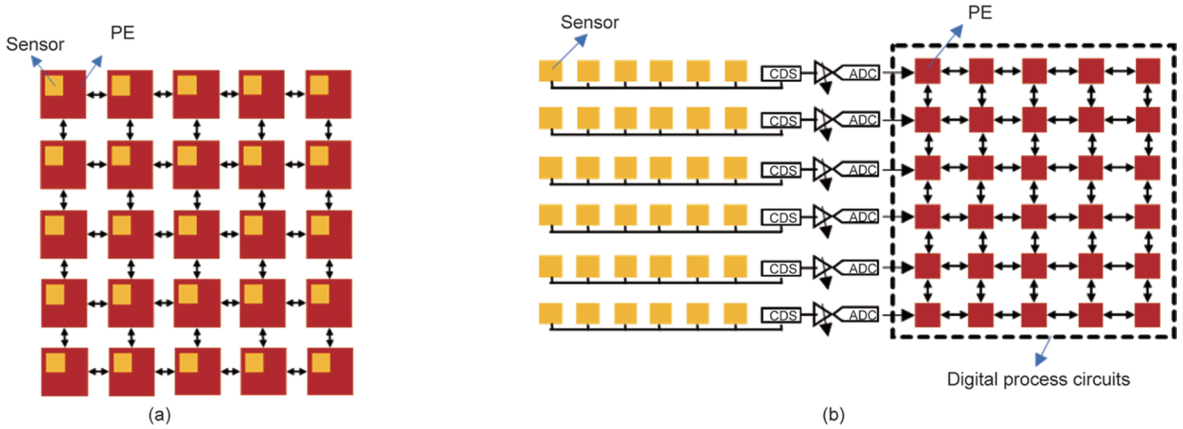

图1. 视觉芯片架构。（a）传感单元内部计算；（b）传感单元附近计算。CDS：双精度采样[14]。

扰是一个亟待解决的问题。有研究报道了一种由单光子雪崩二极管（SPAD）和忆阻器[30–31]组成的系统，用于处理脉冲事件形式的信息，从而完成实时成像识别。

随着新型材料与器件的发展，感存算一体化智能成像系统也同样需要新的架构和算法来适配其应用。例如，深度学习算法（如DNN、CNN、SNN等）是目前较为成熟的图像处理技术，而如何将其应用于感存算一体化智能成像系统是一个亟待解决的难题。SNN通过对时间并行编码的神经信号进行编码和处理，为提高计算效率提供了一种很有前景的解决方案[2]。

本文总结了感存算一体化智能成像系统中使用的两种不同类型的架构（即在传感单元内或附近进行计算），然后讨论了未来的发展方向（包括与算法匹配的架构、3D集成技术、新型材料和先进器件）。总之，感存算一体化智能成像系统的最终目标是实现更高效的AI硬件，该硬件系统具有低功耗、高速、高分辨率、高识别准确率和大规模集成的特点，同时具有可编程性。为了将感存算一体化智能成像系统商业化，需要在物理学、材料学、计算机科学、电子学和生物学等领域进行更深入的研究。

致谢

作者十分感谢芝加哥大学的Supratik Guha教授为改进论文所作的有益讨论。这项工作得到了国家重点研发计划（2021YFA0716400）、国家自然科学基金（61904093、61975093、61991443、61974080、61927811、61822404、62175126和61875104）、BNRist重点实验室项目（BNR2019ZS01005）、中国博士后科学基金（2018M640129和2019T120090）以及固态照明和节能电子协同创新中心（2021ZD0109900和2021ZD0109903）的资助。

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 MoS2. Small 2018;14(15):e1800079.
[22] Wu JY, Chun YT, Li S, Zhang T, Wang J, Shrestha PK, et al. Broadband MoS2 fieldeffect 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 photoswitches 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; Springfielf, 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 IEEEComputer Society Annual Symposium on VLSI ISVLSI); 2018 Jul 8 – 11; Hong Kong, China. New York: IEEE; 2018. p. 441 – 6. W. Pan, J. Zheng, L. Wanget al. Engineering 14 (2022) 19–21 21.