



Invited: Neuromorphic Vision Modalities in the NimbleAI 3D Chip

Xabier Iturbe
Ikerlan
Spain
xiturbe@ikerlan.es

Bernabé
Linares-Barranco
CSIC
Spain

Sio-Hoi Ieng
CSIC &
Sorbonne University
Spain, France

Arne Erdmann
Raytrix GmbH
Germany

Luca Peres,
Oliver Rhodes
University of Manchester
UK

Rafael Tornero
Universitat Politècnica
de València
Spain

Manolis Sifalakis,
Marcel v.d. Burgwal
Imec
Netherlands, Belgium

Amirreza
Yousefzadeh
University of Twente
Netherlands

Maha Kooli,
Riccardo Alidori
CEA-LIST Grenoble
France

Pavel Zaykov
Codalip s.r.o.
Czechia

ABSTRACT

This paper provides an overview of the ongoing work to enable novel modalities of passive monocular neuromorphic vision in the NimbleAI sensing-processing architecture; namely, foveated and light-field event-driven vision with selective visual attention. The latter vision modality encodes 3D visual surroundings as sparse visual events in a 4D spatiotemporal domain, adding depth to current representation of visual information delivered by Dynamic Vision Sensors (DVS). The NimbleAI architecture implements hardware support for efficient execution of mainstream computer vision algorithms and AI models using these visual inputs. The architecture is designed to harness the latest advancements in 3D silicon integration, making it possible to squeeze sensing and spiking circuitry, memory, and processing engines into a miniature silicon volume.

1 NIMBLEAI CONCEPT AND ARCHITECTURE

NimbleAI implements novel Dynamic Vision Sensing (DVS) modalities to perceive depth with extremely low-latency and low-power [1]. This novel technology has been recently highlighted in the Yole 2024 Neuromorphic report [2] as a pathway to achieve the "Neuromorphic 3D sensing" milestone by 2026. As opposed to frame-based vision sensors, DVS sense surroundings in a continuous-time manner, generating a flow of visual events that encode a spatiotemporal representation of the visual reality with extremely high temporal resolution (eq. to 1,000 fps) [3], capturing very fast motions without suffering from motion blur. Moreover, data volume produced by DVS compares very favourably with regular full frames, which include much redundant data. This translates into lower energy consumption and shorter latency, especially if the sparseness of the event output of the DVS is leveraged by the processing paradigm, e.g., using event-driven Spiking Neural Networks (SNNs).

The NimbleAI architecture implements an SNN-based ultra-efficient regulatory loop of DVS visual events that dynamically

configures the sensing and processing components to perform optimally in each detected visual context. Likewise, visual events are formatted into data structures compatible with widely used image processing kernels and industry standard Convolutional Neural Networks (CNNs) for visual inference, thus enhancing efficiency, productivity, and capabilities.

In the intended 3D-integrated sensing-processing chip, *DVS and sensor front-end layers* (see section 1.1) at the top of the 3D stack work closely coupled with an *SNN-powered early perception engine* (see section 1.2) that selects salient image features or Regions of Interest (ROIs). Visual event flows from selected ROIs are streamed to *processing and inference tiles* (see section 1.3) in the interior layers of the 3D stack through *visual pathways*, which are dynamically created (and destroyed) as new ROIs are detected. The visual pathways are independently configured to best serve the specific properties of each ROI (e.g., size, accuracy, latency, motion), taking advantage of the irregular distribution of visual information and uneven temporal dynamics in the scene. Physically, they are implemented using Silicon Through Vias (TSV) across vertically-stacked layers. Logically, they rely on a *3D Network-on-Chip (3D-NoC)*.

While developing the NimbleAI 3D-integrated architecture, system level trade-offs are being explored in a *board-level prototype* using 2D testchips of selected sensing and processing components. As shown in Fig. 1, this prototype includes DVS and SNN testchips, a Xilinx XCZU15EG MPSoC, and commercial AI processor chips (e.g., HAILO) connected via PCIe lanes. The prototype implements four dedicated DVS to SNN links to demonstrate visual pathways.

1.1 DVS sensor and front-end

NimbleAI explores two complementary DVS modalities: (1) *Digitally-Foveated DVS (DF-DVS)* allows dynamic resolution adjustment over the field of view to implement selective visual attention and visual processing inspired by retinotopy (see section 2); and (2) *Light-Field DVS (LF-DVS)* allows to perceive depth using event-driven lightweight stereopsis, implemented in a feed-forward manner to achieve sub-ms latency (see section 3). A 384×304 pixel DF-DVS testchip is being taped out using XFAB 180-nm, whereas the LF-DVS prototype uses a commercial Prophesee IMX636 DVS chip coupled with a custom Raytrix Micro-Lens Array (MLA). Both sensors are integrated in the board-level NimbleAI prototype.

The NimbleAI architecture includes a *DVS front-end (BEGI)* to: (1) control the incoming flow of events to avoid processing bottlenecks, especially in the early perception engine; (2) remove noise and

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

DAC '24, June 23 - 27, 2024, San Francisco, CA, USA

© 2024 Copyright is held by the owner/author(s). Publication rights licensed to ACM.

ACM 979-8-4007-0601-1/24/06...

<https://doi.org/10.1145/3649329.3689622>

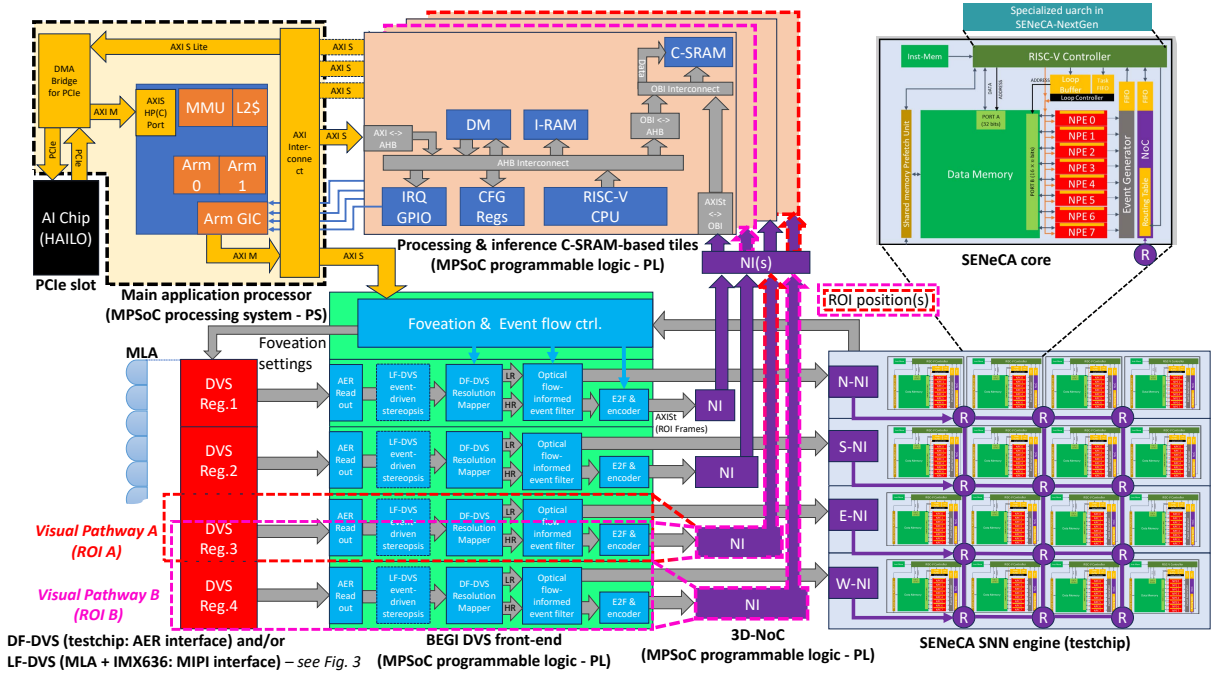


Figure 1: Simplified view of the NimbleAI architecture and prototype.

expose sparsity; (3) run event-driven light-field stereopsis; and (4) adapt the format of visual inputs to ensure compatibility with CNNs and image kernels running on the processing and inference tiles.

1.2 SNN-powered early perception engine

The NimbleAI early perception engine is based on the Imec SENeCA multi-core processor [4], and runs attentional SNNs to select ROIs (see section 2.1). Each SENeCA core implements a hardware event loop and comprises a lightweight RISC-V controller along with an energy-efficient vector pipeline of parallel Neuron Processing Elements (NPEs) for event-based dataflow processing and 2 Mb local data memory for storing neuron states and synaptic parameters. When an event occurs, the controller processes it by fetching the corresponding microcodes from the local memory and dispatches these to the NPE co-processors. Hence, each core works independently of the others and consumes energy only when an event is received. The next-generation SENeCA cores (SENeCA-NextGen) will implement a specialized microarchitecture to boost efficiency and performance. We expect to increase processing throughput from approx. 5 Me/s per (SENeCA) core to approx. 500 Me/s per (SENeCA-NextGen) core. A testchip with 16 SENeCA and 16 SENeCA-NextGen cores, running at 500 MHz, is planned for tape-out using GlobalFoundries 12-nm for integration in the board-level NimbleAI prototype. The cores will be connected via a mesh-NoC with four inputs (N,S,E,W) to receive visual event flows from DVS regions.

1.3 Processing and inference C-SRAM tiles

Image kernels and user CNNs run on processing and inference tiles, which are supported by software programming and compilation flow compatible with popular AI frameworks. Each tile is composed of a customized embedded 32-bit Codasip RISC-V CPU, Instruction

and Data Memories (IRAM and DRAM), and CEA computational SRAM (C-SRAM) [5]. The C-SRAM is a vector co-processor based on a near-memory computing architecture that integrates SRAM memory tightly coupled with a vector processing unit that decodes specific instructions and performs operations on any vector-line of the SRAM. Hence, the RISC-V CPU runs the application control part, while the C-SRAM is used both as a programmable vector co-processor and as a low-latency SRAM to reduce energy consumption by minimizing data movement. Tiles with varying C-SRAM capacities can be seamlessly integrated into the 3D-NoC-based NimbleAI architecture for specialized processing of visual data in ROIs with different properties, closing the visual pathways. The current NimbleAI prototype implements two tiles, each with 2 Mb C-SRAM. We are investigating data encoding schemes and methods to expose and leverage the inherent sparsity of visual events and thus minimize memory, energy, and latency in the C-SRAMs.

2 DIGITALLY-FOVEATED DVS (DF-DVS)

The NimbleAI DF-DVS pixels are based on the design reported in [6] and can be separated into two parts: (1) photoreceptor and pre-amplification, noted as 'p'; and (2) difference detection and communication with periphery, noted as 'c'. By default, the full DF-DVS is set to Low-Resolution (LR) and physical pixels are grouped into macro-pixels. The 'p' parts of all physical pixels in macro-pixels are combined and all their 'c' stages are disabled, except for one. The latter 'c' part computes the relative variation of the sum of the photocurrents detected by the individual photosensors grouped in the macro-pixel. As soon as a macro-pixel is selected to be part of a foveated ROI, all their pixels are set to High-Resolution (HR) and their individual 'c' parts enabled.

For each pixel with an active 'c' part that produces an event, a readout circuitry sets its absolute coordinate and polarity using

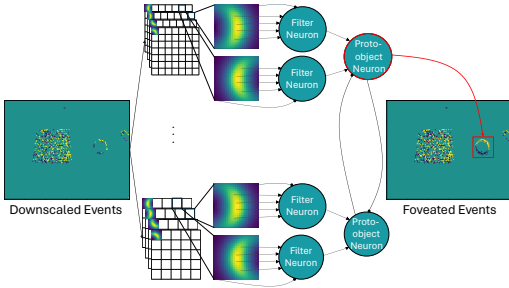


Figure 2: Saliency-based SNN structure and operation.

Address-Event-Representation AER (x,y,p) . The DF-DVS sensor implements four AER channels, with an estimated throughput of 30-100 Mev/s per channel. Events are accessed by the BEGI DVS front-end, classified as HR or LR, and provided with appropriate coordinates for further processing. BEGI delivers LR events to the SENECA-based early perception engine for ROI detection and creates ROI frames using HR events. These frames are delivered to the processing and inference tiles through the 3D-NoC.

2.1 Bottom-up visual attention

The early perception engine harnesses the foveation capabilities of the DF-DVS to optimize HR visual information acquisition (and subsequent processing in the NimbleAI architecture).

A promising bottom-up attention mechanism to detect ROIs consists of building saliency maps [7] from the visual field and selecting the most salient area through a winner-take-all approach. To this extent, a two-layered SNN is being designed using Leaky Integrate-and-Fire (LIF) neuron models. The SNN scans the LR visual scene through a spiking convolutional approach looking for *proto-objects*, that is, elements generally resembling an object or portions of it. This is achieved using filters drawn by the Von Mises distribution (VM) [7]. The first SNN layer identifies occurrences of such filters, which are then combined to neighbouring kernels having opposite convexity in the second layer to build a saliency map. A winner-take-all approach on the second layer (i.e., proto-object neurons that inhibit each other) allows to select the most salient proto-object, which will be nominated as ROI. Fig. 2 shows the schematic of the SNN, which is presented with a LR scene composed of two objects (the circle and cylinder on the right) and cluttering (the square on the left) using 2×2 DF-DVS macro-pixels. The SNN selects the circle as ROI to be set up in HR (red square).

Due to the limited resources of the SENECA processor, the SNN is being constrained to be correctly mapped: (1) to avoid saturation of the local data memory, the neuron fan-in is constrained by considering only three different filter sizes: 70×70 , 50×50 , 30×30 pixels; and (2) to limit the number of neurons, the convolution stride of the SNN is reduced to a filter overlap of 70%.

In addition to reducing the amount of information to be processed, ROI detection is also interesting for real-time object tracking. To this extent, BEGI computes the optical flow of ROIs event-by-event to extract information about object movement, including direction and speed. This helps optimize the event filter settings per region, provide feedback on improving the estimated ROIs, and allow to estimate processing deadlines for them. Ultimately, BEGI updates the foveation settings of the DF-DVS, minimizing the number of configuration transactions when tracking objects.

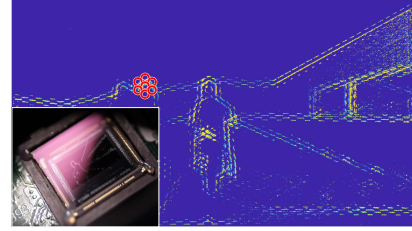


Figure 3: LF-DVS prototype and a slice of the time-surface.

2.2 Retinotopic processing with DF-DVS

As the NimbleAI DF-DVS delivers simultaneously LR and HR visual event flows, it allows to perceive depth and motion using retinotopic principles when the sensor is moving; e.g., when used in a flying drone. More specifically, the apparent size of visual objects decreases inversely with distance, such that a motion forward generates a zoom. Projected on a foveation-based multi-resolution map (LR and HR), this zoom will correspond to a translation that can be used to estimate depth and optical flow, as shown in [8].

3 LIGHT-FIELD DVS (LF-DVS)

As shown in Fig. 3, in the current LF-DVS prototype, a Raytrix MLA is positioned in front of a Prophesee IMX636 DVS to direct incoming light rays onto different pixels, enabling the sensor to record both intensity and directionality. The IMX636 features a $1,280 \times 720$ pixel array with a pixel size of $4.86 \mu\text{m}$. A diameter of 26 pixels per micro-lens was selected in the MLA with a lens pitch of $126 \mu\text{m}$. In the context of light-field imaging, a reduction in lateral resolution is observed due to inherent redundancies, resulting in an effective reconstructed resolution of 640×360 pixels. A hexagonal lattice, characterized by each micro-lens being surrounded by six neighbors, was selected for the MLA as it has the highest fill factor. Lens focal length of 8 mm was chosen and the MLA's aperture set to $f/1.8$, resulting in an operational range of around 5 m and depth accuracy in the range of mm. An $f/1.0$ MLA is currently being designed to accommodate longer ranges, expecting to reach 12 m. The LF-DVS prototype has been validated using adapted light-field algorithms [9] (originally designed for frame-based sensors) that process time-surfaces generated from incoming LF-DVS raw events (see Algorithm 1) [10]. Output LF-DVS events are delivered in AER format (x,y,z,p) .

3.1 Event-driven stereopsis for 3D perception

NimbleAI is designing novel event-driven stereopsis algorithms to reduce the computational workload, latency, and power consumption w.r.t. frame-based light-field algorithms [9], while preserving (as much as possible) their depth estimation accuracy.

Depth perception with the LF-DVS prototype is not fundamentally different from classical stereo vision when focussing on the neighborhood of each micro-lens. Hence, the depth estimation is mainly decomposed into three major steps: (1) detect features using keypoint detectors such as HARRIS, SIFT, FAST; (2) match detected features in each micro-lens with similarity score constrained by epipolar geometry to reach a semi-dense matching; and (3) measure disparity information from matched features to recover depth.

We leverage the fact that the LF-DVS is natively a feature detector where the same 3D point seen by neighbor micro-lenses

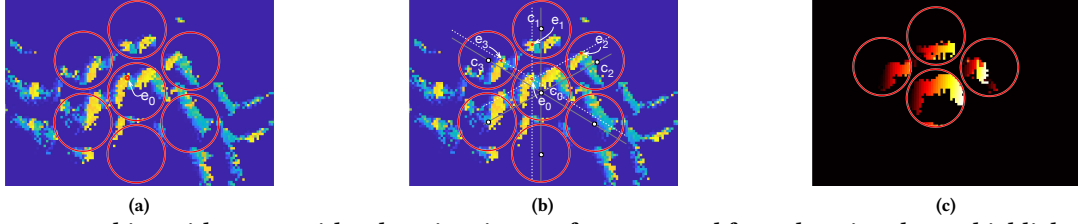


Figure 4: Stereo-matching with events with a decaying time-surface – zoomed from the micro-lenses highlighted in Fig. 3.

triggers spatially and temporally close events, allowing us to replace the feature detection techniques mentioned in (1) by an almost computationally-free coincidence detection of events produced by neighbor micro-lenses. Previous works prove the efficiency of this approach in event-based vision [11]. As shown in Algorithm 1, we use exponentially decaying time-surfaces [10] to improve spatiotemporal pattern recognition.

Algorithm 1 Update time-surface

Ensure: Initialize a $1,280 \times 720$ buffer TS with 0.
if event e_0 **then**
 for all $TS(x_k, y_k) \neq 0$ and $t_k \leq t_0$ **do**
 $TS(x_k, y_k) \leftarrow \exp((t_k - t_0)/\lambda)$
 end for
end if

Assuming that the most recent event in the decaying time-surface shown in Fig. 3 is $e_0 = (x_0, y_0, t_0)^T$, a zoom of its neighborhood in Fig. 4 illustrates the matching process with previous events produced in the six neighbor micro-lenses using Algorithm 2. (a) e_0 triggers the update of the time-surface; the pixel value at (x_0, y_0) is matched against previous events e_i using the time-surface that encodes spatiotemporal distances between events. (b) Since not all micro-lenses see the same 3D point that triggered the events, only three neighbor micro-lenses provide plausible matches: e_1, e_2, e_3 . These potential matches are located along the epipolar lines, shown as dashed lines parallel to the lines defined by each pair of micro-lens centers, represented by white dots. (c) A denser match can be (optionally) done on time-surface pixels using e_1, e_2, e_3 as seeds to propagate the matching process with the epipolar constraint. In this representation, colors encode matched events.

The disparity between matched pixels on the time surface due to events e_k and e_0 is defined as $d_{0,k} = \|(x_0, y_0)^T - (x_k, y_k)^T\|$; whereas the distance of micro-lenses centers C_0 to C_k defines the stereo baseline $B_{0,k} = \|C_0 - C_k\|$. Hence, assuming that all micro-lenses have the same focal length f , the depth estimated from the matched events is given as: $Z_{0,k} = f \frac{B_{0,k}}{d_{0,k}}$, as shown in Algorithm 2.

Algorithm 2 Event-based stereo matching and depth estimation

Ensure: $t_k \leq t_0$
for each event e_0 **do**
 Update the time-surface with Algorithm 1
 Find the six closest micro-lenses centers C_0, \dots, C_6
 for all pair of centers C_0, C_k **do**
 Find the best match e_k in the line passing by e_0 and parallel to (C_0, C_k)
 end for
 (Optionally: apply semi-dense matching using seeds e_k along epipolar lines.)
 $Z_{0,k} \leftarrow f \frac{B_{0,k}}{d_{0,k}}$
end for

3.2 3D scene flow estimation using LF-DVS

While state-of-the-art optical flow techniques primarily rely on projecting 3D motion onto 2D images, new methods that naturally exploit 4D LF-DVS event representations could enable accurate and instantaneous scene flow estimation [12]. This holds a great potential for robot navigation tasks among other applications.

4 CONCLUSIONS

This paper sums up the NimbleAI sensing-processing architecture that enables new modalities of event-based vision and 3D perception with extremely low energy consumption and processing latency. As we complete the exploration and technology feasibility study for integrating the NimbleAI components in a 3D chip, a board-level prototype is being built where vision pipelines of early adopters can be tested. LF-DVS technology has been proven to be feasible using adapted frame-based light-field algorithms. An optimized silicon implementation of an accelerator for event-driven stereopsis, enabling sub-ms 3D perception using only tens of mW, is expected by 2026.

ACKNOWLEDGMENTS

NimbleAI has received funding from the EU’s Horizon Europe programme (GA no. 101070679), and by the UK Research and Innovation (UKRI) under the UK government’s Horizon Europe funding guarantee (GA no. 10039070). See: <https://www.nimbleai.eu>

REFERENCES

- [1] Xabier Iturbe et al. NimbleAI: Towards neuromorphic sensing-processing 3D-integrated chips. In *DATE*, 2023.
- [2] Yole. *Neuromorphic Computing, Memory and Sensing Report*. 2024.
- [3] Patrick Lichtsteiner et al. A 128×128 120 dB 15 μ s latency asynchronous temporal contrast vision sensor. *IEEE J. Solid-State Circuits*, 43(2):566–576, 2008.
- [4] Amirreza Yousefzadeh et al. SENeCA: Scalable energy-efficient neuromorphic computer architecture. In *AICAS*, 2022.
- [5] Maha Kooli et al. Towards a truly integrated vector processing unit for memory-bound applications based on a cost-competitive computational sram design solution. *ACM J. Emerg. Technol. Comput. Syst.*, 18(2), 2022.
- [6] Teresa Serrano and Bernabé Linares. A 128×128 1.5% contrast sensitivity 0.9% FPN 3 μ s latency 4 mW asynchronous frame-free dynamic vision sensor using transimpedance preamplifiers. *IEEE J. Solid-State Circuits*, 48(3):827–838, 2013.
- [7] Massimiliano Iacono et al. Proto-object based saliency for event-driven cameras. In *IROS*, 2019.
- [8] Jean-Nicolas Jérémie et al. Retinotopic mapping enhances the robustness of convolutional neural networks, 2024.
- [9] Christian Perwaß and Lennart Wietzke. Single lens 3D-camera with extended depth-of-field. *SPIE Human Vision and Electronic Imaging XVII*, (8291), 2012.
- [10] Xavier Lagorce et al. HOTS: A hierarchy of event-based time-surfaces for pattern recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(7):1346–1359, 2017.
- [11] Marc Osswald et al. A spiking neural network model of 3D perception for event-based neuromorphic stereo vision systems. *Scientific Reports*, 7, 2017.
- [12] Sio-Hoi Ieng et al. Event-based 3D motion flow estimation using 4D spatio-temporal subspaces properties. *Frontiers in Neuroscience*, 10, 2017.