Implementation of a Retina Model on the SCAMP-5 Vision Sensor

Maciej Lewandowski
Department of EEE
University of Manchester, UK
maciej.lewandowski@postgrad.manchester.ac.uk

Prince Philip
Department of EEE,
Indian Institute of Science, Bangalore, India
princephilip@iisc.ac.in

Alexandre Marcireau
Department of Computer Science,
University of Manchester, UK
alexandre.marcireau@manchester.ac.uk

Chetan Singh Thakur
Department of EEE,
Indian Institute of Science, Bangalore, India
csthakur@iisc.ac.in

André van Schaik
Department of Computer Science,
University of Manchester, UK
andre.vanschaik@manchester.ac.uk

Piotr Dudek
Department of EEE,
University of Manchester, UK
p.dudek@manchester.ac.uk

Abstract

We present the first implementation of a multi-stage silicon retina model on the SCAMP-5 Pixel Processor Array (PPA), incorporating center-surround filtering, contrast gain control, and Leaky Integrate-and-Fire spiking directly at the focal plane. Unlike a simplified retina-inspired event-based encoding offered by Dynamic Vision Sensors (DVS), our on-sensor "silicon retina" model captures a larger amount of retina-like processing. We provide a GPU-based simulation framework, and compare a performance of the silicon retina model with the DVS, on a task of video saliency prediction. The silicon retina model achieves a 12% lower validation loss than a standard DVS baseline while generating 47% fewer events, demonstrating that more biologically plausible pre-filtering can produce more efficient representations for downstream semantic tasks.

1. Introduction

Conventional frame-based CMOS sensors are poorly suited for high-speed or low-power vision. Streaming full-resolution frames at kilohertz rates imposes severe bandwidth and latency constraints, often requiring energy-intensive GPUs. Dynamic Vision Sensors (DVS) address this by encoding visual input as sparse, asynchronous streams of brightness change events [5, 8], enabling high temporal resolution at low bandwidth. However, most commercial event cameras implement a simplified model of retinal processing: they primarily mimic thresholded photoreceptor derivatives and skip key biological features such as spatial filtering, contrast gain control and retinal adaptation mechanisms [2, 10]. Pixel Processor Arrays (PPAs) offer a different paradigm, integrating analog processing elements directly into the imaging array [1, 3]. Unlike DVS, PPAs are fully programmable in software, enabling execution of complex visual pipelines - including, as in this work, a multi-stage retinal model - at kilohertz frame rates with sub-watt power consumption.

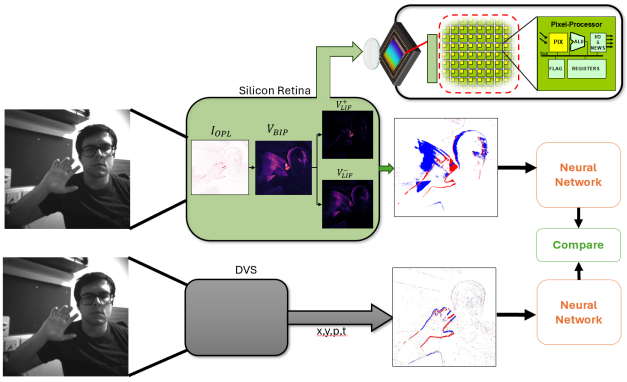

Figure 1. Visual summary of the work.

The work presented in this paper is summarized in Figure 1, and the key contributions of this work are:

1. The first implementation of a multi-stage silicon retina model on SCAMP-5, incorporating center-surround fil-

tering, contrast gain control, and LIF spiking.

2. A superpixel memory architecture and 5-bit fixed-point analog multiplication scheme that overcomes memory limitations of SCAMP-5 and makes the silicon retina implementation feasible.
3. Empirical evidence that retina events outperform standard DVS events on video saliency prediction (12% lower loss, 47% fewer events), with ablation studies validating the hardware-efficient approximations.

A live demonstration of the system running on SCAMP-5 at approximately 100-150 FPS under ambient lighting will be shown at the workshop.

2. Silicon Retina Model

We implement the biologically plausible retina model proposed by Wohrer et al. [10], which extends standard Linear-Nonlinear architectures by incorporating a biologically plausible Contrast Gain Control mechanism. This functional model abstracts away specific cellular details but captures essential retinal dynamics, particularly adaptation to varying light conditions using regular image processing techniques. The model consists of three sequential stages, shown in Figure 2.

2.1. Outer Plexiform Layer (OPL)

The OPL mimics the synaptic interactions between photoreceptors, horizontal cells, and bipolar cells. It transforms the input luminance into a center-surround signal by computing the difference between a narrow "center" current and a wider "surround" current. Functionally, this stage behaves as a spatio-temporal band-pass filter: spatially, it acts as a Difference of Gaussians (DoG) to detect edges; temporally, the delay between center and surround integration allows the system to respond to motion

2.2. Contrast Gain Control (Bipolar Cells)

To adapt to sudden changes in light levels, this stage implements a local contrast gain control mechanism modeled as a non-linear feedback loop. The bipolar cell membrane potential is governed by a leaky integration, driven by the OPL current, with a variable leak conductance acting as a divisive feedback term to modulate the input gain. This conductance is derived from the spatially and temporally filtered state of the cell, passed through a non-linear parabolic function. In regions of high contrast, the function output increases, raising the leak conductance to reduce gain and prevent saturation. Conversely, in low-contrast areas, the conductance remains near its resting value to maintain high sensitivity.

2.3. Ganglion Cells (LIF Spiking)

The final stage converts the analog bipolar signals into an asynchronous stream of spikes, mimicking the output of Retinal Ganglion Cells. The non-spiking current from the previous stage passes through a temporal filter and a rectification function, which models nonlinear synaptic transmission to drive a Leaky Integrate-and-Fire (LIF) neuron. A spike is generated when the membrane potential crosses a specific threshold, after which the potential is reset. To make the output compatible with standard event-processing algorithms that rely on parallel ON and OFF streams, we extend this stage to use two complementary LIF neuron channels driven by oppositely rectified versions of the bipolar output.

3. Implementation on SCAMP-5

3.1. SCAMP-5 Hardware

SCAMP-5 is a 256×256 array of analog, tiny "processors", which we refer to as processing elements (PEs). Each of the PE has a dedicated photodiode [1, 3] with which it can interact, an analog ALU (add, subtract, divide-by-2), six general-purpose analog registers, seven single-bit digital registers, and local 4-neighbor communication. The chip operates at kilohertz frame rates at under 1 watt. Crucially, unlike a DVS [4], or other vision sensors that implement complex pixel-level functionality in hardware [7], SCAMP-5 is fully software-defined: the event generation rule is programmable, using a C++ based framework, not fixed in silicon.

3.2. Superpixel Memory Architecture

The retina model (Fig. 2) requires at least 7 persistent state variables per pixel (E_C, T_C, E_S, V_{BIP}, E_A, T_2, V_{LIF}), exceeding the 6 analog registers per PE. We resolve this by logically partitioning the 256×256 PE array into 2×2 blocks, creating a 128×128 array of *superpixels* (Fig. 3). In each block, one PE (`PROC`) handles active computation and image capture; the remaining three (`STO1`, `STO2`, `STO3`) act as extended analog memory, increasing available registers from 6 to 24 per logical "pixel" at the cost of halving spatial resolution.

For spatial operations such as Gaussian blur, the register is duplicated from `PROC` to all `STO` units before the operation, ensuring the resistive grid operates correctly across the full superpixel.

3.3. Fixed-Point Analog Multiplication

The retinal model requires both global scaling (by constants such as λ) and per-pixel modulation (by the spatially varying gain g_A). Since SCAMP-5 has no hardware multiplier, we approximate scalar multiplication using 5-bit fixed-point arithmetic. The product $\alpha \cdot A$ is computed by accumulating progressively halved versions of register A using the native `diva` instruction, gated by the binary coefficients b_i. For global constants, b_i are stored in the microcontroller

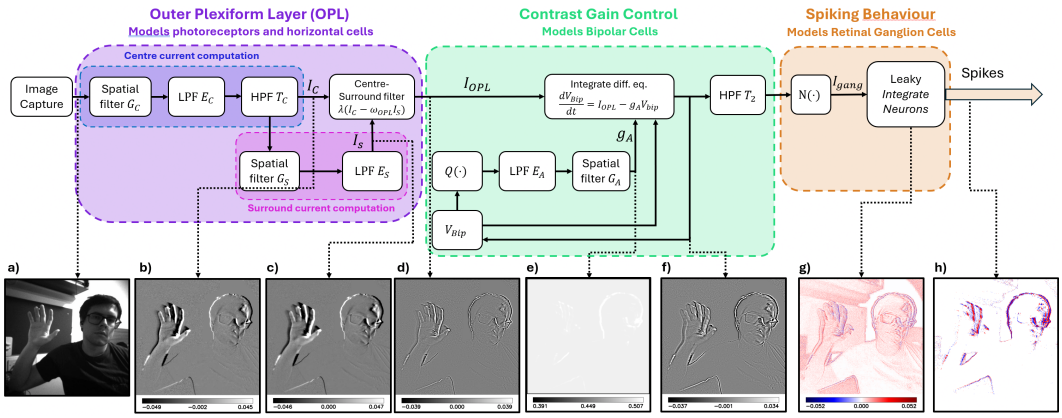

Figure 2. Multi-stage retinal processing pipeline showing intermediate outputs: center current I_C, surround current I_S, OPL difference I_{OPL}, bipolar potential V_{BIP}, adaptive gain g_A, LIF membrane V_{LIF}, and final ON/OFF spike output.

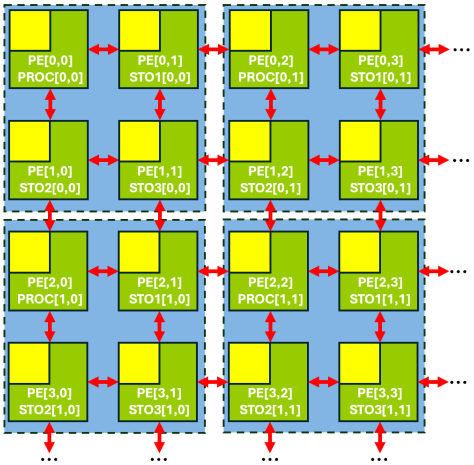

Figure 3. Superpixel architecture. The 256×256 physical array is partitioned into 2×2 blocks. `PROC` performs computation; `STO1--STO3` provide extended analog memory, nearly quadrupling the available register count per effective pixel.

(MCU). For spatially varying maps, the coefficients are digitized into four single-bit DREG planes per pixel, enabling parallel per-pixel multiplication. Empirically, 5 bits (in format 1.4f) offers the best tradeoff: fewer bits reduce range; more bits accumulate excessive analog noise.

3.4. Hardware Approximations

Mapping the full model to analog hardware required several approximations:

- **High-pass filtering:** IIR formulation was replaced with simple frame differencing, $A[t] = B[t] - B[t-1]$, to avoid noise accumulation across sequential analog division operations. This was most beneficial in the Outer Plexiform Layer stage: applying LPF and HPF sequentially leads to massive degradation of speed and signal quality.
- **Q-function:** quadratic form V_{BIP}^2 was replaced with $|V_{BIP}|$, avoiding the SCAMP-5 squarer circuit which exhibits large spatial non-uniformity.
- $N(\cdot)$ **nonlinearity:** approximated as $\max(0,\, x) + I_{bias}$. The formulation is a significant simplification of the function provided by Wohrer [10], however, in our evaluation setup, we discovered it still allows for good performance on saliency.
- **Differential equations:** solved via explicit forward Euler integration.

3.5. Performance

Internally (no readout), the model executes at approximately 11,000 FPS. When streaming binary spike maps off-chip, throughput is limited by the digital interface to approximately 580 FPS. Under practical indoor ambient lighting without artificial illumination, the system operates at 100–200 FPS with sub-watt power consumption.

4. Evaluation

4.1. Experimental Setup

To evaluate the performance of different event representations, we train neural networks on downstream tasks using both standard DVS and our silicon retina streams. We use the video upsampling pipeline of Gehrig et al. [6] to generate high-framerate synthetic sequences, which then drive either a standard DVS simulator or our CUDA-based retina simulator (Fig. 4). In this work, we focus our evaluation on the task of Video Saliency Prediction. Sensor noise is deliberately excluded from both simulators to isolate the algorithmic contribution of the retina model.

4.2. Video Saliency Prediction

We train a lightweight CNN (≈100k parameters, modified FireNet) on a 541-video subset of the DHF1K dataset [9]

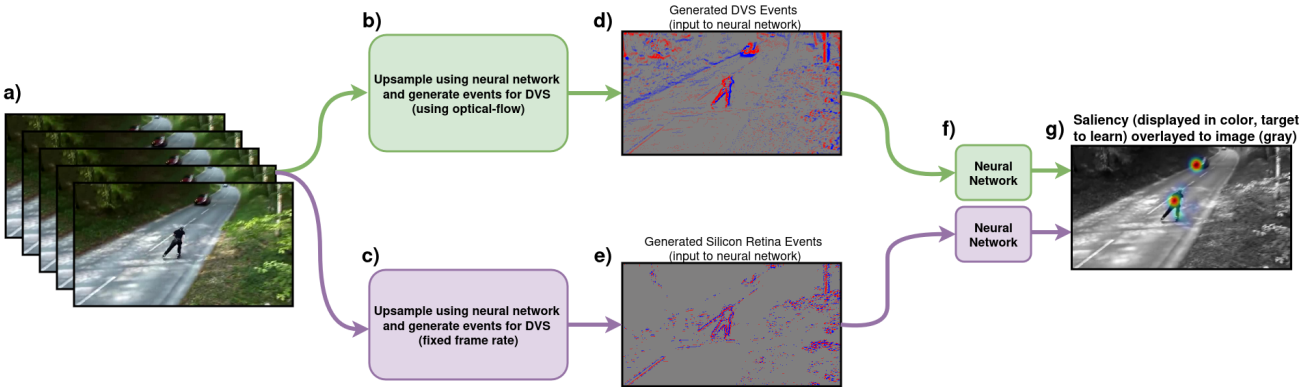

Figure 4. Experimental pipeline. (a) Input video sequence. (b,c) Frame upsampling: the DVS pipeline (b) uses an optical-flow-based approach [6]; the Silicon Retina pipeline (c) uses a network to upsample to a fixed 200 FPS. (d,e) Event generation for the standard DVS model (d) and the proposed Silicon Retina (e). (f) The representation is evaluated using identical network architectures and training procedures. (g) Ground truth target for the *Video Saliency Prediction*, where the model predicts human gaze attention maps [9].

to predict human gaze maps from event streams. The loss function is a weighted combination of KL divergence and Pearson Correlation Coefficient.

Results are shown in Fig. 5. The Retina model achieves a best validation loss of 0.701 versus 0.799 for DVS (a 12% relative improvement) while generating significantly fewer events: average map occupancy of 5.7% (Retina) versus 10.8% (DVS). We hypothesize that this biologically inspired pre-filtering suppresses redundant motion and noise, producing a sparser, more semantically relevant event stream that the downstream network exploits more efficiently. Ablation studies further confirm that the hardware-efficient absolute-value approximation of the gain control feedback does not degrade downstream performance, and that the Retina consistently outperforms DVS across all tested event density levels (1–15% occupancy); details will be reported in a full paper.

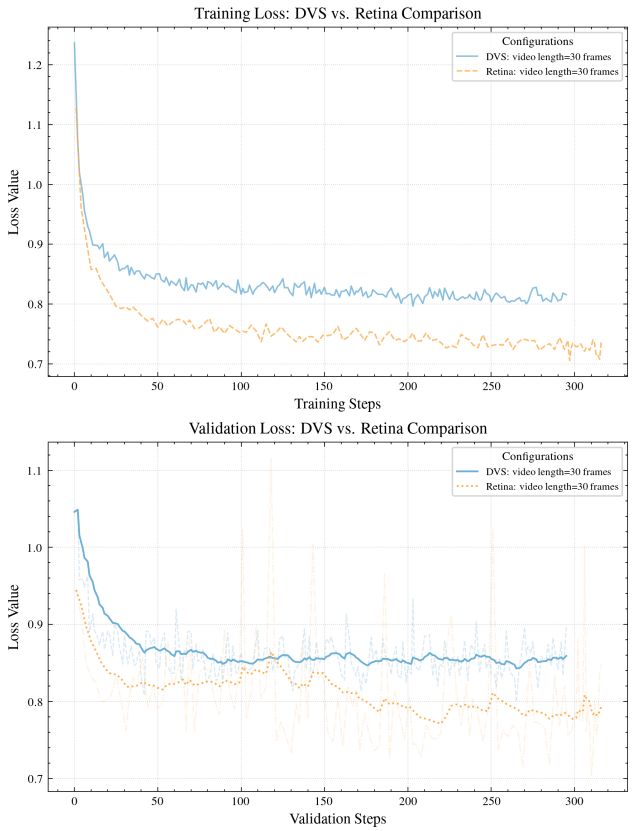

Figure 5. Video Saliency Prediction: training and validation loss over epochs for DVS (gray) vs. Silicon Retina (blue). The Retina achieves lower loss despite generating ≈47% fewer events.

5. Conclusion

We presented the first implementation of a multi-stage silicon retina model on the SCAMP-5 Pixel Processor Array, enabled by a superpixel memory architecture and fixed-point analog multiplication scheme. The retina model achieves 12% lower saliency prediction loss than a standard DVS baseline while generating 47% fewer events. These results suggest that incorporating biologically-inspired spatial filtering and gain control directly at the sensor can yield more efficient event representations for semantic vision tasks.

References

[1] Stephen J. Carey, Alexey Lopich, David R.W. Barr, Bin Wang, and Piotr Dudek. A 100,000 fps vision sensor with embedded 535gops/w 256×256 SIMD processor array. In *2013 Symposium on VLSI Circuits*, pages C182–C183. 1, 2

[2] Misha Mahowald Carver Mead. Silicon retina. 1

[3] Piotr Dudek, Thomas Richardson, Laurie Bose,

Stephen Carey, Jianing Chen, Colin Greatwood, Yanan Liu, and Walterio Mayol-Cuevas. Sensor-level computer vision with pixel processor arrays for agile robots. 7(67):eabl7755. 1, 2

[4] Thomas Finateu, Atsumi Niwa, Daniel Matolin, Koya Tsuchimoto, Andrea Mascheroni, Etienne Reynaud, Pooria Mostafalu, Frederick Brady, Ludovic Chotard, Florian LeGoff, Hirotsugu Takahashi, Hayato Wakabayashi, Yusuke Oike, and Christoph Posch. 5.10 a 1280×720 back-illuminated stacked temporal contrast event-based vision sensor with 4.86µm pixels, 1.066geps readout, programmable event-rate controller and compressive data-formatting pipeline. In *2020 IEEE International Solid-State Circuits Conference - (ISSCC)*, pages 112–114, 2020. 2

[5] Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew Davison, Joerg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. 44(1):154–180. 1

[6] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. 3, 4

[7] Prince Philip, Kapil Jainwal, André van Schaik, and Chetan Singh Thakur. Tau-cell-based analog silicon retina with spatio-temporal filtering and contrast gain control. *IEEE Transactions on Biomedical Circuits and Systems*, 18(2):423–437, 2024. 2

[8] Christoph Posch, Teresa Serrano-Gotarredona, Bernabe Linares-Barranco, and Tobi Delbruck. Retinomorphic event-based vision sensors: Bioinspired cameras with spiking output. 102(10):1470–1484. 1

[9] Wenguan Wang, Jianbing Shen, Jianwen Xie, Ming-Ming Cheng, Haibin Ling, and Ali Borji. Revisiting video saliency prediction in the deep learning era. 43 (1):220–237. 3, 4

[10] Adrien Wohrer and Pierre Kornprobst. Virtual retina: A biological retina model and simulator, with contrast gain control. 26(2):219–249. 1, 2, 3