Pixel Processor Arrays For Low Latency Gaze Estimation

Laurie Bose[1] Jianing Chen[1] Stephen J. Carey[1] Piotr Dudek[2]
[1]Pixelcore Research Ltd, Manchester, United Kingdom, [2]The University of Manchester, Manchester, United Kingdom

***Abstract*—We demonstrate the use of a Pixel Processor Array (PPA) vision sensor for achieving rates of over 10,000 Hz, with processing latency below 0.1 ms, in a gaze tracking application, performing image processing directly upon the sensor itself with minimal external computation. Each pixel processing element of the PPA sensor is capable of light sensing, data storage and computation, allowing various visual processing to be performed efficiently at the point of light capture. We demonstrate how information such as pupil location, size, and LED reflections upon the eye's surface can be extracted by the PPA, along with a simple gaze tracker using pupil location. By extracting such information directly on-sensor, the data needing to be transferred from sensor to external processing is reduced from entire images to a hand-full of contextual bytes, this provides significant saving in terms of power, time and allows for frame-rates far exceeding traditional camera sensors.**

I. Introduction And Overview

When developing for mobile platforms, constraints such as computational power and battery life often provide significant bottlenecks to overcome. Mobile Virtual Reality (VR) and Augmented Reality (AR) headsets are a particularly challenging area as visual rendering and 6DOF tracking tasks must be performed rapidly and consistently in order to provide a comfortable user experience, often placing hard constraints on the fidelity and complexity of content developed for such devices. It is speculated that gaze tracking is poised to become a core technology in VR and AR devices, as not only does it enable interaction via gaze leading to a more immersive experience, but also can be utilized to achieve significantly more efficient rendering via techniques such as foveated rendering [1].

Pixel Processor Array (PPA) vision sensors embed processing, memory and light capture into every "pixel" processing element. This programmable sensor architecture enables a wide variety of visual tasks to be performed directly "on-sensor" using highly efficient parallel computation, such as neural network based digit recognition [2], and visual odometry [3]. This work describes a demonstration of a high-speed gaze tracking implementation using a SCAMP-5 PPA sensor [4]. The PPA sensor enables contextual information necessary for determining the users gaze to be immediately and efficiently extracted on-sensor. This small amount of data is then transferred to connected standard processing for final calculation of the user's gaze at a rate of 10,000 Hz, with accuracy comparable to commercial tracking solutions.

This approach is significantly more efficient than that of using a traditional camera sensor, where entire images must be transferred off-sensor incurring significant power and time costs, and greater computation needing to be conducted on the connected external processing to extract the user's gaze. Other gaze tracking demonstrations using non standard sensors [5] have been shown to operate at extreme speeds, but still require all computation to be performed externally from the sensor.

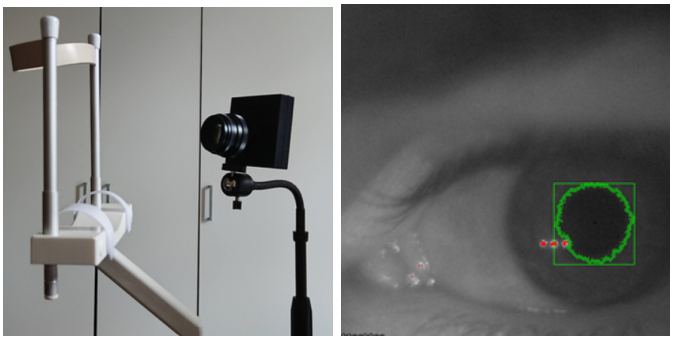

Fig. 1. Left: Prototype SCAMP-5 PPA sensor setup to observe a close up image of a user's eye when resting their head against the chin-rest. Right: captured image of users eye along with extracted features, Pupil region, Pupil bounding box (Green) and LED reflections (Red)

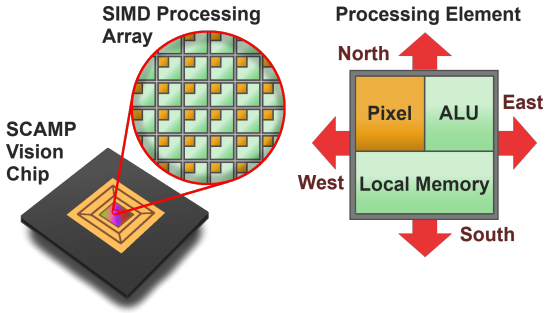

Fig. 2. Overview of the SCAMP vision chip, a massively-parallel SIMD processor array of 256x256 Processing Elements, capable of capturing and processing visual data directly on sensor.

II. Methods

The prototype setup is shown in Figure 1. It makes use of the SCAMP-5 PPA chip which comprises a 256×256 array of Processing Elements (PEs), each containing photosensors, analog and digital memory registers, circuitry for various processing, and data transfer capability between neighbouring elements as shown in Figure 2. The PE array receives instructions from a single controller unit, responsible for the overall program

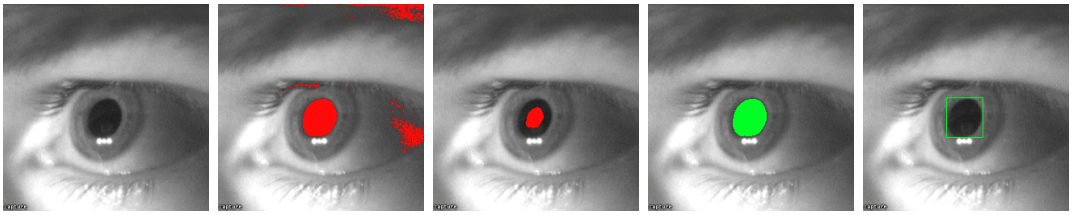

Fig. 3. The process of segmenting the pupil using parallel operations upon the SCAMP-5 PPA. From left to right, an image of the eye is acquired and stored in analog registers of the PE array. A threshold image is generated, and then eroded. This eroded image is then used as the flooding source in a parallel flood fill operation, performed upon the original threshold image. This operation results in flooding only the region of the pupil. Finally the bounding box of this pupil region is extracted, which then forms the only output of the PPA when visualizations are not required.

flow, with all PEs simultaneously executing microinstruction from it, operating as a SIMD processor. Further details of the SCAMP-5 implementation can be found in [4].

Given an appropriate observation of a subject's eye (such as that taken under NIR illumination as shown in Figure 3), the pupil will be responsible for the largest dark region within the image. Under this assumption a short sequence of operations upon SCAMP-5 can be performed to generate a binary mask image of the subject's pupil, from which it is simple to extract various information relevant for gaze estimation, and then transfer said information off-sensor. These steps involve first thresholding the image to create a binary mask image for any dark regions (the threshold level is adaptively adjusted outside the main processing loop). Erosion is then performed, until only a single region remains (that belonging to the pupil). Finally a flood fill operation is performed upon the previous un-eroded image, using this eroded image as the flooding source. This final step effectively floods the region belonging to the pupil as seen in Figure 3. A bounding box of the pupil's region is then extracted and output from the PPA sensor. All steps involved can be performed entirely upon the SCAMP-5 PPA with highly efficiently parallel operations, with the entire algorithm completing in less than 100 μs. Combined with the reduced data needing to be transferred off-sensor (4 bytes per frame), allows for this entire process to be performed at extreme frame-rates of over 10,000 Hz. The typical limiting factor becomes the strength of illumination, determining the maximum rate at which clear images of the eye are obtainable.

Many commercial gaze trackers utilize Near-infrared (NIR) illumination and imaging to acquire a reliable image of the subject's eye, where the pupil is one of the darkest observed features in the image. In head mounted setups using NIR LEDs, the reflections of these LEDs themselves upon the eye are often utilized for gaze estimation, typically by comparing these reflection locations to that of the pupil. The locations of these reflections can also be performed entirely upon the SCAMP PPA, thresholding the image to create a mask of the brightest regions as shown in Figure 1, and then extracting the bounding box of each sequentially.

III. Demonstration Setup

A chin rest is used to keep the user's head in a roughly fixed position in front of a display screen, and a SCAMP-5 system is accordingly positioned to obtain clear images of the user's eye such as shown in Figure 1. A simple calibration is then performed by the user fixating on a sequence of targets on a display screen directly in front of them. After-which the user's estimated gaze vector is projected to and displayed on the screen as a gaze "cursor" allowing them to interact with some simple example programs such as shown in Figure 4.

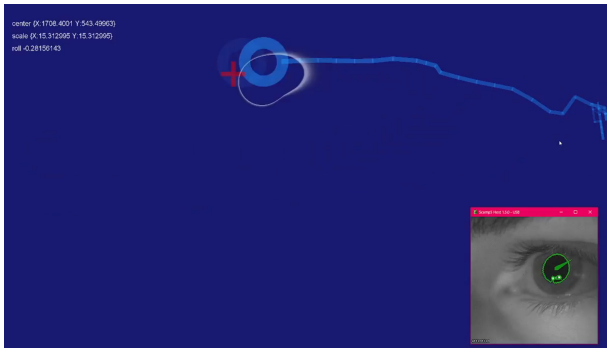

Fig. 4. A program displaying the users gaze estimated by our PPA setup (blue circle) along with that of a commercial desk mounted eye tracker (Tobii Eye Tracker 5, white contour), as the user looks to chase targets (red cross).

This estimated gaze is based purely upon the bounding box of the user's pupil, output by the PPA sensor; however please note in this demonstration, our goal is to show the potential advantages of using such a PPA sensor rather than a specific gaze tracking methodology. One can envision future PPA sensors being mounted within devices such as virtual reality headsets, enabling high rate, ultra low latency gaze tracking at minimal power and computational cost.

References

[1] A. S. Kaplanyan, A. Sochenov, T. Leimkühler, M. Okunev, T. Goodall, and G. Rufo, "Deepfovea: neural reconstruction for foveated rendering and video compression using learned statistics of natural videos," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 6, pp. 1–13, 2019.

[2] L. Bose, P. Dudek, J. Chen, S. J. Carey, and W. W. Mayol-Cuevas, "Fully embedding fast convolutional networks on pixel processor arrays," in *European Conference on Computer Vision*, 2020, pp. 488–503.

[3] L. Bose, J. Chen, S. J. Carey, P. Dudek, and W. Mayol-Cuevas, "Visual odometry for pixel processor arrays," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4604–4612.

[4] S. J. Carey, A. Lopich, D. R. Barr, B. Wang, and P. Dudek, "A 100,000 fps vision sensor with embedded 535GOPS/W 256×256 SIMD processor array," in *2013 Symp. on VLSI Circuits*. IEEE, 2013, pp. C182–C183.

[5] A. N. Angelopoulos, J. N. Martel, A. P. Kohli, J. Conradt, and G. Wetzstein, "Event based, near-eye gaze tracking beyond 10,000 hz," *IEEE Transactions on Visualization and Computer Graphics (Proc. VR)*, 2021.