High-speed vision with a 3D-stacked SPAD image sensor

Citation for published version:

Digital Object Identifier (DOI):
10.1117/12.2586883

Link:
Link to publication record in Heriot-Watt Research Portal

Document Version:
Publisher's PDF, also known as Version of record

Published in:
Advanced Photon Counting Techniques XV

Publisher Rights Statement:
Copyright 2021 Society of PhotoOptical Instrumentation Engineers (SPIE). One print or electronic copy may be made for personal use only. Systematic reproduction and distribution, duplication of any material in this publication for a fee or for commercial purposes, and modification of the contents of the publication are prohibited.

Istvan Gyongy, Germán Mora Martín, Alex Turpin, Alice Ruget, Abderrahim Halimi, Robert Henderson, and Jonathan Leach "High-speed vision with a 3D-stacked SPAD image sensor", Proc. SPIE 11721, Advanced Photon Counting Techniques XV, 1172105 (12 April 2021); https://doi.org/10.1117/12.2586883

General rights
Copyright for the publications made accessible via Heriot-Watt Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
Heriot-Watt University has made every reasonable effort to ensure that the content in Heriot-Watt Research Portal complies with UK legislation. If you believe that the public display of this file breaches copyright please contact open.access@hw.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
High-speed vision with a 3D-stacked SPAD image sensor

Gyongy, Istvan, Mora Martín, Germán, Turpin, Alex, Ruget, Alice, Halimi, Abderrahim, et al.


Event: SPIE Defense + Commercial Sensing, 2021, Online Only
High-speed vision with a 3D-stacked SPAD image sensor

Istvan Gyongy*, Germán Mora-Martín, Alex Turpin, Alice Ruget, Abderrahim Halimi, Robert Henderson, Jonathan Leach

*a School of Engineering, Institute for Integrated Micro and Nano Systems, The University of Edinburgh, Edinburgh EH9 3FF, UK

*b School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh EH14 4AS, UK

*c School of Computing Science, University of Glasgow, Glasgow G12 8QQ, UK

ABSTRACT

3D sensing devices are becoming increasingly prevalent in robotics, self-driving cars, human-computer interfaces, as well as consumer electronics. Recent years have seen single-photon avalanche diodes (SPADs) emerging as one of the key technologies underlying 3D time-of-flight sensors, with the capability to capture accurate 3D depth maps in a range of environmental conditions, and with low computational overhead. In particular, direct ToF SPADs (dToF), which measure the return time of back-scattered laser pulses, form the backbone of many automotive LIDAR systems. We here consider an advanced direct ToF SPAD imager with a 3D-stacked structure, integrating significant photon processing. The device generates photon timing histograms in-pixel, resulting in a maximum throughput of 100’s of giga photons per second. This advance enables 3D frames to be captured at rates in excess of 1000 frames per second, even under high ambient light levels. By exploiting the re-configurable nature of the sensor, higher resolution intensity (photon counting) data may be obtained in alternate frames, and depth upscaled accordingly. We present a compact SPAD camera based on the sensor, enabling high-speed object detection and classification in both indoor and outdoor environments. The results suggest a significant potential in applications requiring fast situational awareness.

Keywords: SPAD image sensor, 3D time-of-flight imaging, LIDAR, 3D-stacked image sensor, TCSPC, Object detection, High-speed imaging

1. INTRODUCTION

Time-of-Flight (ToF) or LIDAR sensors are being increasingly adopted in autonomous systems and robotics, forming a key component of vision systems tasked with recognising and responding to obstacles as quickly as possible. Therefore, having a high frame rate coupled with low latency in the imaging becomes important. ToF sensors are also found in smartphones, were being able to track and interpret the 3D environment in real-time is again desirable, for example to ensure a seamless experience in augmented reality applications.

In a ToF image sensor, a transmitter emits some form of illumination onto the scene of interest, and the back-scattered light is then detected by a receiver. By estimating the time taken for the signal to return, the corresponding distance can be deduced based on the speed of light, and a 3D image built up.

Amongst the multitude of detector options for implementing ToF, single-photon avalanche diodes (SPADs) offer advantages from the perspective of high-speed imaging. SPADs can detect individual photons of light with picosecond timing resolution, whilst operating close to the quantum limit. There is thus the potential of extracting accurate depth estimates from just the handful of signal photons that are collected over the short exposure times associated with high-speed imaging. Furthermore, as SPADs can be implemented in standard CMOS processes, arrays can be built with integrated processing for an all-digital 3D imaging system that is essentially free of any electronic noise. Combining a 2D array of SPADs, with pulsed, wide beam illumination enables fast, parallelised 3D image acquisition.

A number of algorithms have been proposed for the fast processing of SPAD ToF data; they typically rely on high resolution photon time stamps, as provided by conventional SPAD ToF architectures, and therefore implementation for practical, high ambient scenes may challenging due to photon pile-up issues.

*istvan.gyongy@ed.ac.uk
2. DIRECT TOF WITH SPAD

In direct ToF (dToF) image sensors based on SPADs, short (typically nanosecond long) laser pulses are emitted to illuminate the scene. Every time a laser pulse is sent, this starts an electronic timer, and any back-scattered photons from the target that are subsequently detected by the SPAD here are then timed. From these time-stamps, a photon timing histogram is constructed, and depth is estimated by extracting the temporal position of the peak of the histogram. By having an array of pixels with this functionality (or scanning the light source across the scene), a 3D image can then be composed.

dToF devices based on SPADs are often referred to as single-photon LIDAR sensors, and the histogram generation is typically carried out off-chip, with every pixel reporting a single time stamp (corresponding to the first detected photon) per output frame. The frame of time stamps, which is usually acquired over multiple laser cycles, is then transferred on to a separate processing unit, where histogram accumulation and peak extraction take place. However, such an approach proves problematic in high ambient conditions (or for high signal returns, say, from retro-reflected targets), as it leads to pile-up distortion: an exponential decay in the histogram due to the dominance of early photons, which can mask or distort the signal peak. The effect on the resulting ToF depth map can be drastic, as illustrated in Figure 1. Even when the conditions for pile-up distortion are not reached, the high-rates at which time-stamps are generated can lead to data bottlenecks and loss of information, thereby rendering the approach unsuitable for high-speed imaging in practical conditions. There is thus a need for on-chip processing to reduce data rates, whilst preserving the timing information of signal photons. A number of hardware-based solutions have duly been proposed, to combat the problem of photon timing and readout pile-up.

One of the main methods is coincidence detection\(^8\), which requires multiple photon detections within a certain time window, for a photon event to be registered. The underlying assumption is that signal photons (from the laser pulse) will arrive in a burst, so the method filters out solitary photons, which are taken to be background or ambient photons. For the filtering to work effectively, a per-pixel adaptive mechanism is required for the coincidence parameters (the photon threshold and length of time window). This in turn requires continuous monitoring of the per-pixel background and signal levels and can make the implementation of the scheme challenging.

A second, somewhat related method uses an electronic time gate\(^9\) that can be swept in time, and tries to “home in” on time region where signal photons arrive, and ignore any photons arriving outside this time window.

A third method (which can be potential combined with coincidence detection or time gating), is to implement on-chip histogramming that is capable of registering multi-events per laser cycle\(^10\). The concept is to accommodate as many of the detected photons as possible to prevent pile-up and maximise information on depth. As the histogram is created on-chip (potentially even in-pixel), this results in significant data compression so read out data bottleneck is avoided as well.
3. SPAD CAMERA

3.1 Sensor characteristics

With the requirement for on-chip processing in mind, the SPAD chip depicted in Figure 2 was developed\(^\text{10}\). The chip is designed in 3D stacked technology and is composed of two tiers: a top tier with the SPAD detector array, and a bottom tier of photon processing. The detectors and digital logic therefore do not compete for area, and so substantial processing can be integrated into the pixels, without compromising the overall photosensitive area or fill factor of the array. There are a total of 256\(\times\)256 detectors, grouped into 64\(\times\)64 macropixels. Each macropixel is thus composed of a group of 4\(\times\)4 SPADs, with a processing unit underneath. Readout is over 8, 100MHz output lines, giving a maximum frame rate of 760FPS for whole array or over 1kFPS for \(\frac{1}{2}\) array as used here.

The chip is highly reconfigurable and supports a number of modes of operation. Here two of the main modes are considered: in-pixel histogramming at the macropixel resolution, and photon counting (intensity) imaging at the SPAD resolution. Figure 2 shows example data from the sensor from the histogramming mode. Each macropixel provides a 16-bin (14-bit/bin) photon timing histogram, from which depth can be extracted, with low computational overhead, using a centre-of-mass calculation across the peak histogram bin and its neighbours. Performing this calculation across the whole array leads to the depth map depicted in the figure.

The width of the histogram bins is programmable, with a minimum size of 500ps, but provided that the laser pulse is spread over multiple bins (and a sufficient number of photons are collected within the histogram), sub-bin precision is obtained in the depth estimates. Thus, in practice, sub-cm precision can be readily achieved. As the histogram frames are generated at a rate of >1KFPS, 3D depth images can obtained at the same rate. Figure 3 gives an example frame from a high-speed sequence, capturing the moment that a balloon bursts. The diagram shows the histograms produced by three neighbouring macropixels, observing the edge of the tear in the balloon. It is noted that the top macropixel is seeing a signal return from the inside of the balloon, whilst the bottom macropixel is registering a return from the outside of the balloon (as the peak is earlier in time). The macropixel in the middle is seeing two peaks: it is detecting both the inside and the outside of the balloon.

The lateral resolution of depth frames can be computationally increased to the SPAD resolution by operating the sensor in a hybrid modality\(^\text{11}\), whereby depth frames are captured together with intensity frames in a time-interleaved way. Depth upscaling is then performed, guided by the intensity data\(^\text{12}\). Alternatively, rather than upscaling depth using the sensor’s own intensity data, it is also possible to do so using data from a separate high-speed sensor. Figure 4 shows selected example frames from a sequence, where the SPAD was operated in depth mode only, and its output fused with intensity (RGB) data from a high-speed camcorder.
Figure 3. Depth frame capturing the moment of a balloon bursting (right), with the outputs of three neighbouring macropixels (left), observing the edge of the tear in the balloon, being shown.

Figure 4. Selected frames from a ~1kFPS sequence capturing a ball thrown into a plate with milk. The depth frames from the SPAD are upsampled based on intensity data from a high-speed camcorder (Sony FDR-AX700) and the intensity (RGB) data then overlaid.
3.2 Experiments with portable camera

To capture data in a variety of environments, a portable camera system was built around the sensor, as depicted in Figure 5. The system is configured to have a 10m unambiguous range, and features an integrated 850nm laser source, producing 10ns laser pulses with 2W of peak optical power, and 6MHz repetition rate.

With the benefit of the portable system, a range of datasets were captured both indoors and outdoors, to investigate the detection of objects based on the output of the sensor. Figure 6, reproduced from reference 13, illustrates some of the initial data that was collected indoors. The objective with this particular data was to evaluate how accurately the system could discriminate between different hand gestures, namely rock, paper and scissors. As shown in the figure, three different environmental conditions were considered: low ambient with a plain background, low ambient with background objects (including a rotating fan), and the same background but with high ambient level. To produce this, a lamp was used to illuminate the scene (with no ambient filter in front of the SPAD), resulting in a signal to background ratio <1 in the histograms. The figure shows example depth and intensity frames in each of the three scenarios, obtained with a 5ms exposure time (equivalent to 200FPS). A U-net, a type of convolutional neural network, was then trained to localise and classify the images, the bottom two rows in the figure showing the output of the neural net for the example image frames (in terms of bounding boxes for the localisation, and the results of the classification).

It was found that the highest accuracy in the classification – typically well over 90% – was obtained when the neural net is provided with raw histogram frames, rather than depth data-derived from the histograms, or intensity frames. Even when a combination of depth data and intensity data is presented to the neural net (so two frames and twice amount of total data points), there is no performance advantage over a single histogram frame. Whilst the superior performance of histogram-based processing over intensity is surprising (given the low lateral resolution of histogram frames), the advantage over depth-based processing is perhaps to be expected. Histograms provide a richer data set than depth alone, for example edges of objects show multiple peaks (as in Figure 3), and are thus captured with greater fidelity than a simple depth map (which just takes one of the peaks or the average of the peaks).

The neural network processing on histogram frames currently takes around 25ms/frame (with assistance of a GPU), but there are plans to speed up the processing. Furthermore, the dataset is currently being expanded with a bigger variety of scenes, involving multiple objects and different types of motion.

![Data capture with the portable camera system based on the SPAD](image-url)
4. CONCLUSIONS

We discussed the benefits of SPADs in 3D-ToF imaging, together with the importance of on-chip data processing or compression in achieving high frame rates, especially under high ambient light levels. In particular, we highlighted the advantages of multi-event histogramming, in the context of a recent SPAD sensor that exploits 3D-stacking technology to generate photon timing histograms in-pixel.

Thanks to the resulting data compression, the device can generate depth maps at rates exceeding 1kFPS, and with sub-cm precision. The sensor can be operated in a hybrid ToF and intensity imaging modality to upscale depth in x,y guided by the intensity data, or data from a separate high-speed RGB sensor can be used for this purpose.

We discussed preliminary results, suggesting accurate object detection performance, when processing the output of the sensor with a trained convolutional neural network, especially when the processing is based on raw histogram data.

ACKNOWLEDGEMENTS

This research was supported by EPSRC through grants EP/M01326X/1 and EP/S001638/1. AH would like to acknowledge the support of the UK’s Royal Academy of Engineering (RF/201718/17128) and DSTL Dasa project DSTLX1000147844. The authors are grateful to STMicroelectronics and the ENIAC-POLIS project for chip fabrication.

REFERENCES


