High-speed object detection using SPAD sensors

Proceedings Volume 11693, Photonic Instrumentation Engineering VIII; 116930L (2021)
https://doi.org/10.1117/12.2577545

PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

High-speed object detection using SPAD sensors

Mora Martín, Germán, Turpin, Alex, Ruget, Alice, Halimi, Abderrahim, Henderson, Robert, et al.

Germán Mora Martín, Alex Turpin, Alice Ruget, Abderrahim Halimi, Robert Henderson, Jonathan Leach, Istvan Gyongy, "High-speed object detection using SPAD sensors," Proc. SPIE 11693, Photonic Instrumentation Engineering VIII, 116930L (5 March 2021); doi: 10.1117/12.2577545

SPIE.

Event: SPIE OPTO, 2021, Online Only

High-speed object detection using SPAD sensors

Germán Mora-Martín*[a], Alex Turpin[b], Alice Ruget[c], Abderrahim Halimi[c], Robert Henderson[a], Jonathan Leach[c], Istvan Gyongy*[a]
[a] School of Engineering, Institute for Integrated Micro and Nano Systems, The University of Edinburgh, Edinburgh EH9 3FF, UK
[b] School of Computing Science, University of Glasgow, Glasgow G12 8QQ, UK
[c] School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh EH14 4AS, UK
* Corresponding author: german.mora@ed.ac.uk; istvan.gyongy@ed.ac.uk

ABSTRACT

3D-imaging is used in a wide range of applications such as robotics, computer interfaces, autonomous driving or even capturing the flight of birds. Current systems are often based on stereoscopy or structured light approaches, which impose limitations on standoff distance (range), and require textures in the scene or accurate projection patterns. Furthermore, there may be significant computational requirements for the generation of 3D maps.

This work considers a system based on the alternative approach of time-of-flight. A state-of-the art single-photon avalanche diode (SPAD) image sensor is used in combination with pulsed, flood-type illumination. The sensor generates photon timing histograms in pixel, achieving a photon throughput of 100's of Gigaphotons per second. This in turn enables the capture of 3D maps at frame rates >1kFPS, even in high ambient conditions and with minimal latency.

We present initial results on processing data frames from the sensor (in the form of 64×32, 16-bin timing histograms, and 256×128 photon counts) using convolutional neural networks, with the view to localize and classify objects in the field of view, with low latency. In tests involving three different hand signs, with data frames acquired with millisecond exposures, a classification accuracy of >90% is obtained, with histogram-based classification consistently outperforming intensity-based processing, despite the former's relatively low lateral resolution. The total, GPU-assisted, processing time for detecting and classifying a sign is under 25 ms.

We believe these results are relevant to robotics or self-driving cars, where fast perception, exceeding human reaction times is often desired.

Keywords: 3D-imaging, Time-of-flight, SPAD, LiDAR, Object Detection, Neural Networks.

1. INTRODUCTION

3D-imaging has a growing range of applications like face recognition in smartphones[1], augmented reality[2], autonomous driving[3], robotics and ballistics[4]. Depth information from a scene can be captured in many ways such as stereoscopy or structured light[5], but these often require significant computational requirements along with detailed projection patterns. An alternative approach to retrieve depth is by using time-of-flight (ToF)[6], which uses a modulated or pulsed light source (typically a LED or a laser diode) emitting at IR wavelengths. The light illuminates the scene and the sensor measures the time for the back-scattered light to return. ToF cameras typically attain speeds of 10-60 frames per second (FPS). There are two types of such cameras: indirect (iToF) and direct (dToF). An iToF sensor measures the time dependent intensity information to deduce the delay between emitted and received signal rather than the ToF itself[7]. Photo-demodulator or photodiode pixels with multiple storage nodes are used, whose activation is synchronised with the illumination signal. This technology, as found for example the Microsoft Azure Kinect[8] and Lucid Helios2[9] cameras, can provide high lateral resolution (pixel counts of 640×480 or more) and millimeter accuracy. However, the range is usually limited to a few meters, due to an inherent tradeoff between range and accuracy[10]. Furthermore, multipath issues[11] can arise in certain scenes, leading to inaccurate depth estimates which may confuse computer vision systems. Ambient illumination (in outdoor use) can also impact the accuracy.

Photonic Instrumentation Engineering VIII, edited by Yakov Soskind, Lynda E. Busse, Proc. of SPIE
Vol. 11693, 116930L · © 2021 SPIE · CCC code: 0277-786X/21/$21 · doi: 10.1117/12.2577545

Direct ToF on the other hand, uses a sub-nanosecond electronic stopwatch to time the echo from a pulsed light source. These devices are typically implemented using high sensitivity avalanche photo-diode (APD) or single-photon avalanche diode (SPAD) sensors. Highly accurate depth measurements over hundreds of meters and beyond are possible, even in daylight conditions, leading to dToF's adoption in automotive LiDAR systems[12-13]. Traditionally, dToF systems used a single point sensor, or an array of point sensors, and necessitated optical scanning to cover a FoV, thus limiting the rate of acquisition of 3D data. Implementing dToF in large array format sensors was difficult due to the challenges in integrating suitable timing electronics into the sensor. However, more recently, SPAD-based dToF devices in line[14], and image sensor format[15-16] have been developed, with the premise of high-speed, high accuracy 3D imaging when combined with blade-scanned, and flood illumination sources, respectively.

Automotive LiDAR has been a strong driver of research into high-level interpretation of depth. Object classification has originally been done with machine learning techniques (such as support vector machines (SVMs), k-nearest neighbours or gradient boosting trees) where objects are trained by its size, shape, intensity, etc[17]. More recently, deep learning techniques have emerged as a better approach for classification tasks by using convolutional neural networks (CNNs), where features are automatically extracted from the input[18]. Intensity or RGB images have been preferred to perform such tasks due to the high image resolution with well-known neural networks such as ImageNet, AlexNet or YOLO[19]. LiDAR data has also been tested in networks such as PointNet or RangeNet but they are still not accurate enough due to the sparsity of data over distance[20-21].

The present paper considers 3D imaging over a short range (<10m), and constitutes an initial investigation into whether accurate interpretation of 3D depth data may be possible, despite the relatively low lateral resolution typically provided by dToF (in a non-scanning or "flash" operation). In particular, we consider a recently developed, reconfigurable SPAD sensor, capable of high speeds of operation.

2. SPAD SENSOR

This work uses a state-of-the-art SPAD array sensor with dToF architecture for high-speed 3D-imaging[22]. The total resolution of the array is 256×256 pixels, containing 64×64 macropixels with 4×4 SPADs each. This SPAD has the capability to operate in two main modes: intensity or photon count mode at the full resolution of the SPAD and time correlated single photon counting (TCSPC) mode, at 64×64. The latter mode provides a 16-bin photon timing histogram per 4×4 SPAD macropixel with a minimum temporal resolution of 500 ps per bin (the temporal bin width being adjustable). Multi-event histogramming being used, ensure robustness to ambient illumination[23]. Half of the rows remain unused to double the capturing speed of the sensor, thus having final resolutions of 256×128 for intensity mode and 64×32 for TCSPC mode at speed up to ≈1000 FPS[23].

A firmware has been developed to operate the sensor in a hybrid mode while still attaining high speeds. In the hybrid mode, intensity and TCSPC frames are captured in alternate frames. One of the advantages of this mode is that it enables the intensity guided upscaling of depth maps, as it has been demonstrated in recent publications[24-25].

The histogram to depth conversion is done thanks to an Opal Kelly XEM7310 FPGA integration module interfacing the SPAD data along with a Matlab script. Depth values are estimated by using a centre-of-mass calculation from each histogram following the equation:

$$d = \frac{\sum_{t=\max(d_{max}-t_l,1)}^{\min(d_{max}+t_r,16)} t \max(0,h_t-b)}{\sum_{t=\max(d_{max}-t_l,1)}^{\min(d_{max}+t_r,16)} \max(0,h_t-b)} \quad (1)$$

Where h_t is the histogram bin at a given macropixel (from 1 to 16), d_{max} is the index of the bin with the maximum count, b is the median of the bins (measure of ambient level) and t_l, t_r are parameters with a value corresponding to the width of the histogram peak, typically t_l, $t_r = 2$. In other words, the depth is calculated by using a centroid around the highest peak, providing sharp edges in depth maps. Alternatively, taking a centroid of all the bins (after compensating for the ambient level) would resemble more to the estimation of depth of an iToF camera, providing blurrier edges on its depth map. Figure 1 shows examples of a intensity and depth frame and the histogram corresponding to two different pixels of the latter.

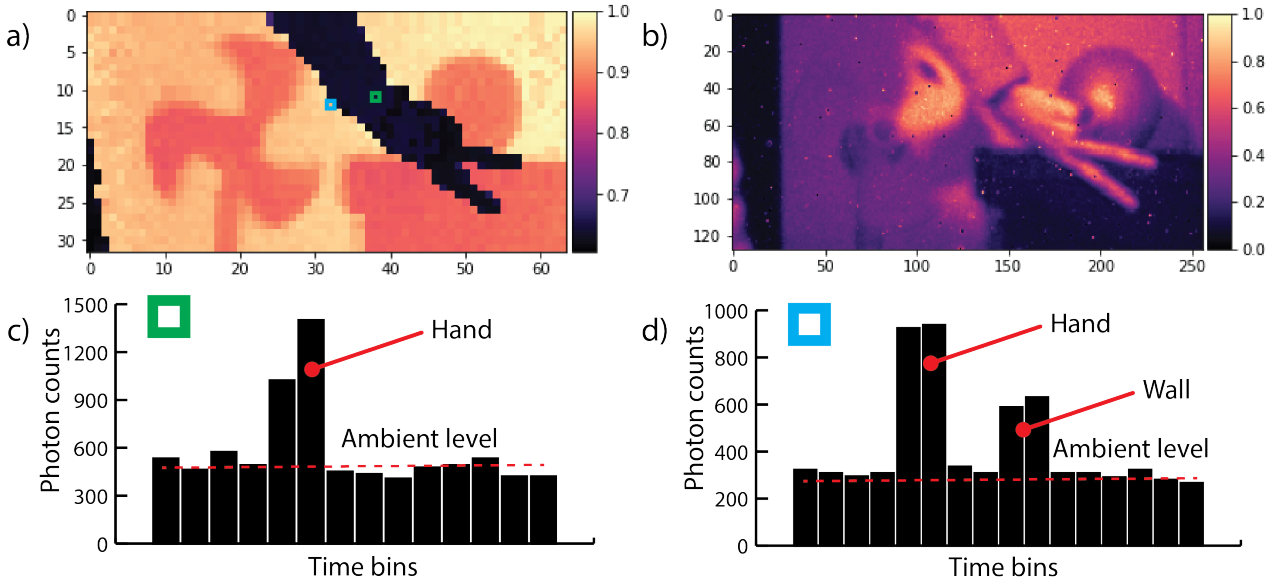

Figure 1. a) Normalized depth map (64×32), 1 corresponding to 4m. b) Normalized intensity map (256×128). c) 16-bin histogram corresponding to the green pixel from a), revealing a peak at the distance of the hand. d) 16-bin histogram corresponding to the blue pixel from a), revealing a double peak from hand and the wall (edge).

3. NEURAL NETWORK FOR OBJECT DETECTION

The main goal of this paper is to test the capabilities of the sensor to recognize certain objects while running at low lateral resolution and high speeds using histogram data. Recently, convolutional neural networks (CNN) have proved to be an excellent tool for object detection in images. In particular, the U-net network has shown in the past good results for segmentation of images, giving a pixel-wise description of every class present in the scene[26].

In this paper, different adapted versions of U-net are used to fit the input of each class of data: depth (64×32), intensity (256×128), histogram (64×32×16) and a combination of intensity and depth (I+D), where the depth is up-scaled using nearest neighbours interpolation and then stacked together on the third axis (256×128×2). Table 3 (Appendix) gives a layer-by-layer description of the neural network. The output of this network provides a binary mask: 0 denoting background and 1 denoting the pixel contains an object of interest. The result is cropped and fed into a simpler neural network which predicts the class of the input. The input is cropped differently for each type of data: 28×28 for depth images, 112×112 for intensity images and 28×28×16 for histograms. For I+D, we input 3 types of data: depth, intensity and both stacked. The probabilities for each class and type are summed up and the highest one remains as the final output. Combining depth and intensity can enhance object detection since they can complement each other, especially under low light levels and short ranges. Figure 2 shows a detailed diagram of the whole process and a description of the classification neural network used for each data type.

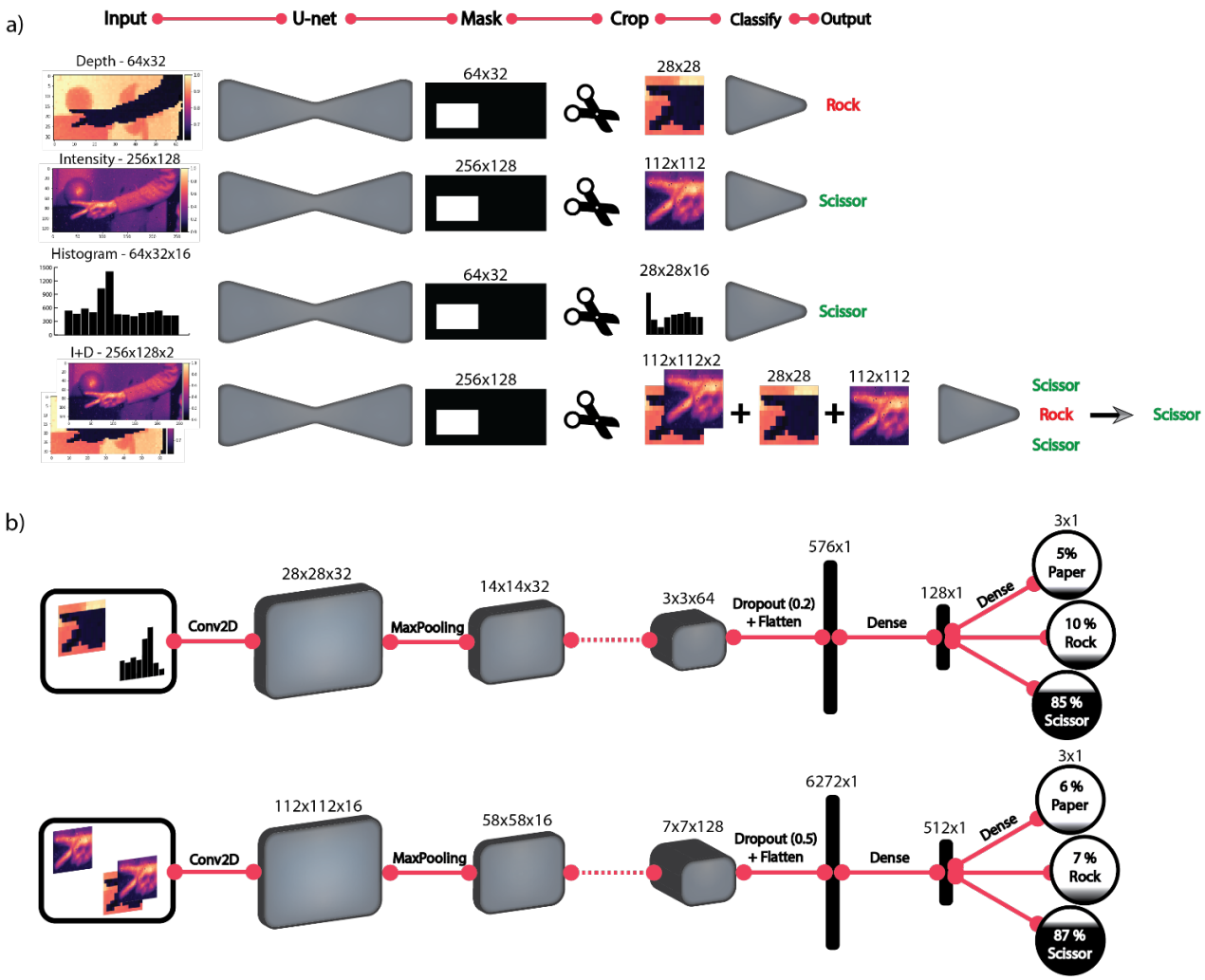

Figure 2. a) Schematic for different data types of the neural network process to reach an output of localization and classification. b) Classification neural network layer diagram (top - depth and histograms and below - intensity and I+D).

4. CAMERA SET-UP

The SPAD camera (Quantic 4×4 sensor) triggers an IR laser source (850 nm) which illuminates the scene of interest. The laser (Iberoptics Time-of-flight Illuminator) has a maximum output power of 2W distributed uniformly over a FoV of 20º. It emits 10 ns pulses at a repetition rate of 6 MHz, best suited for short range (up to 10 m) applications. The laser has cylindrical shape (19 mm diameter and 28 mm long) which is ideal for building a compact set-up. A 6 mm focal length C-mount lens (MVL6WA) is placed next to the laser source to match the illumination cone and capture the returning photons. The lens can optionally insert an ambient filter (Thorlabs FL850-10) to further supress the undesired effect of ambient photons. All the components are enclosed into a 3D-printed box (180×180×88 mm) to maximise compactness and portability of the set-up. The box also includes a slot to place a portable battery to power the SPAD camera and laser and a hole compatible with tripod mounts, providing versatility of imaging outdoors.

The SPAD camera is also connected to a laptop via a USB3.0 connection to display the output in real-time, like explained in Section 2. Figure 3 shows a graphical representation of the actual SPAD camera box used in this work.

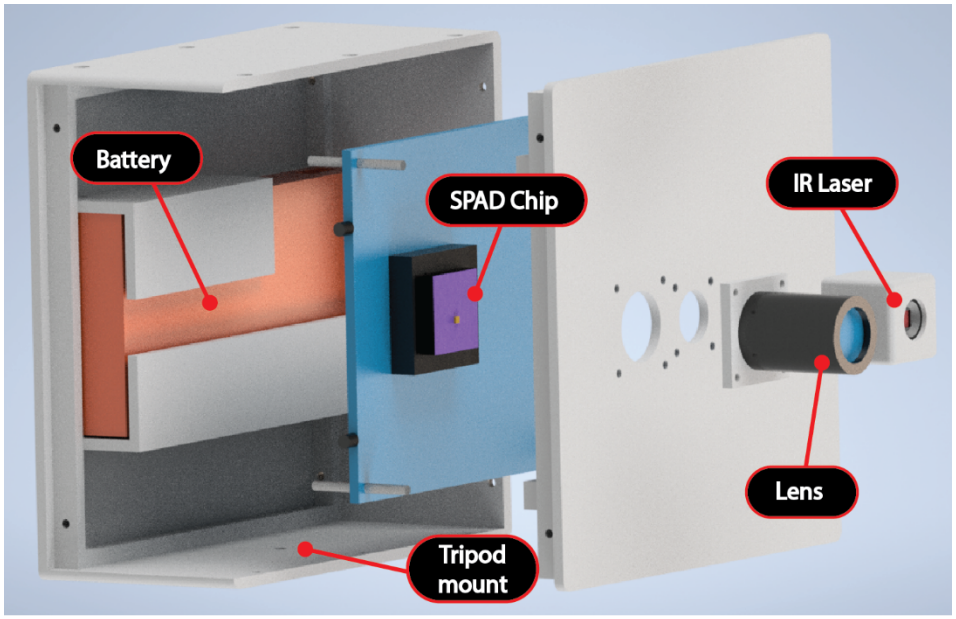

Figure 3. Rendered image of the SPAD camera. The camera contains a lens, an IR laser, a SPAD chip in a board and a battery and tripod mount for portability.

5. RESULTS DISCUSSION

We have tested the object detection performance for different data types that can be obtained from the SPAD camera: depth, intensity, histogram and I+D. For simplicity, three different hand signs (paper, rock, and scissor) are selected as classes for the experiment. An iToF-like approach of extracting depth as explained in Section 2 was also tested, but was finally discarded from the results discussion since it provides no significant difference with respect to the dToF approach.

Even though the camera can reach capturing speeds above 1000 FPS, due to limitation in power of the light source, the data was captured at around 200 FPS with temporal resolution of 4 ns per bin in order to maintain accurate estimations of depth. The data was recorded for short ranges using three different datasets: the first one includes the hand and a wall without ambient light, the second one remains in the dark but includes objects in the background (a football and a rotating fan) and the third one presents both objects and ambient light. An additional dataset is analysed, which is a combination of the previous three together. From each dataset, 1000 frames of each gesture are captured, giving a total of 3000 per dataset and a grand total of 9000 frames to give as much variety as possible. These frames are split into three sets: a training set (70% of the data), a validation set (15% of the data), which gives an unbiased evaluation of the model to fit on the training set and the test set (15% of the data), which determines the performance of the neural network. An example of a test frame in different modes before and after being run by the neural network is shown in Figure 4.

To accelerate the training and processing of the neural network, a RTX2070 GPU is used, attaining processing speeds of >40 FPS for intensity, depth and histograms and >30 FPS for I+D. The neural network training is run 10 times to capture the random variability in the classification accuracy, since it is typical to have slightly different performances in the test every time the network is trained. Table 1 summarizes the average accuracy and standard deviation for each dataset and each mode.

The results show high accuracies for all modes in all datasets. This suggests that the neural network has learnt successfully to extract key features in order to output correctly the position and class of unseen data. The unusually high accuracies are an indicator that test images are similar to the training dataset, thus explaining the present accuracies reaching values close to 100%. Therefore, the discussion is focused on the relative accuracies between each mode rather than the absolute quantities attained. To prove the model is not completely overfitted and the results are meaningful, the average accuracies for test datasets which have been trained in a different dataset are tested (e.g. testing object/light dataset under no object / no light trained model). Table 2 summarizes these values and shows lower values as expected, but high enough (>80% at least in all cases) to show that the models are learning and not overfitting. Intensity images suffer from a higher loss as their appearance changes considerably depending on whether ambient light is present.

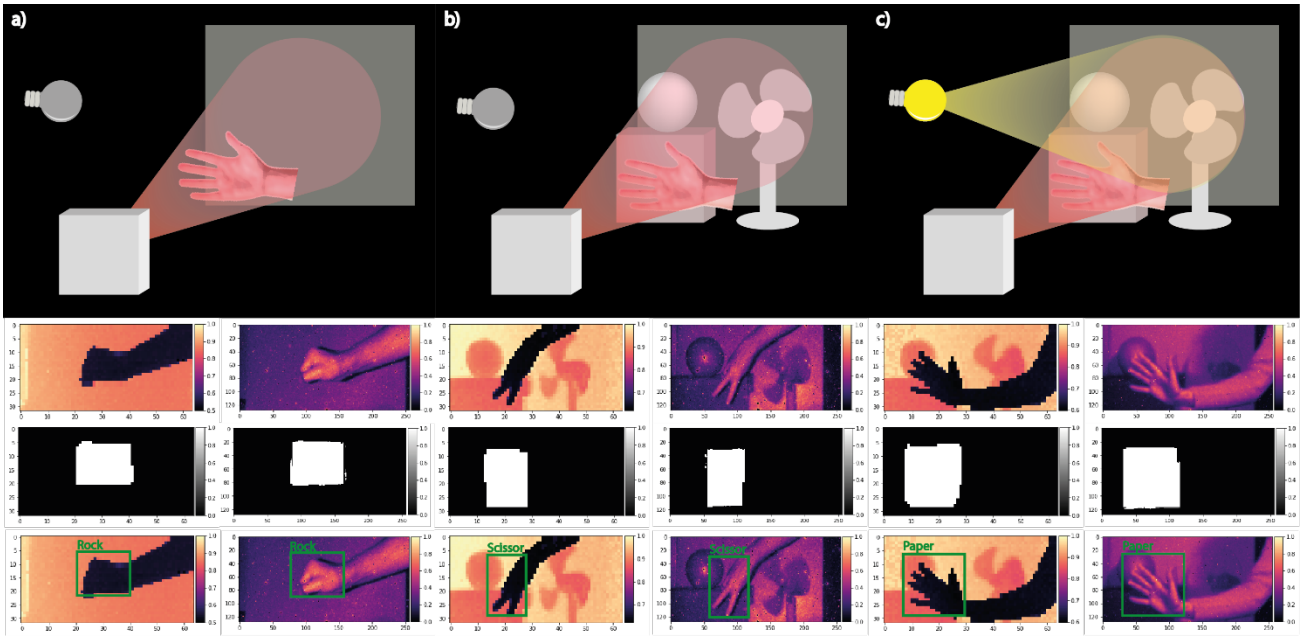

Figure 4. In a column, from top to bottom: representation of the dataset, input image (depth – left and intensity - right), output localization network (binary mask) and final output determining the class of the sign. a) Dataset without background objects neither ambient light. b) Dataset with objects and without ambient light. c) Dataset with objects and ambient light.

Table 1. Average processing speed, average classification accuracy and standard deviation of the accuracy for every dataset.

	Processing speed (FPS)	No Object / No light (NONL)		Object / No Light (ONL)		Object / Light (OL)		Total	
		Average	St. Dev	Average	St. Dev	Average	St. Dev	Average	St. Dev
Intensity (256×128)	41	98.59	0.22	98.59	0.57	98.50	0.73	98.78	0.31
Depth (64×32)	**45**	98.63	0.26	98.70	0.28	98.39	0.36	98.59	0.17
Histogram (64×32×16)	41	**99.44**	0.38	**99.10**	0.41	**99.26**	0.21	**99.26**	0.16
I+D (256×128×2)	32	99.02	0.23	**99.13**	0.49	**99.47**	0.20	**99.31**	0.28

Depth accuracies across all the datasets follow the expected trend: the accuracy of the neural network reduces marginally when ambient light is present in the data, which adds noise onto depth maps. The intensity mode displays similar values to those from the depth neural network, showing a slightly better performance on the overall dataset. This should be expected since this mode offers higher resolution images, which should help the network to discern better between classes. Nevertheless, the histogram and the I+D data present the most interesting findings provided in this section. The accuracies were submitted to a t-student test, determining that the average accuracies of histogram and I+D data were statistically superior from the intensity and depth ones. When comparing histogram and I+D data, the test concluded that there were no significant differences between their averages. Table 4 (Appendix) summarizes the t_{stat} and t_{crit} parameters between each data type. If $-t_{crit} < t_{stat} < t_{crit}$, the averages are not significantly different and vice versa.

Table 2. Average classification accuracy and standard deviation for every test dataset trained on a different dataset.

	Test – OL Trained - NONL		**Test – NONL Trained - OL**		**Test – OL Trained - ONL**		**Test – ONL Trained - OL**	
	Average	**St. Dev**	**Average**	**St. Dev**	**Average**	**St. Dev**	**Average**	**St. Dev**
Intensity (256×128)	82.88	0.24	81.27	0.42	81.56	0.24	89.57	0.30
Depth (64×32)	86.05	0.28	86.23	0.24	**95.62**	0.20	90.91	0.19
Histogram (64×32×16)	**88.47**	0.35	**91.67**	0.39	**95.57**	0.25	**93.58**	0.25
I+D (256×128×2)	86.32	0.21	87.22	0.46	**94.89**	0.27	92.01	0.23

In the I+D data type, we provide intensity and depth information on the same pixel, which helps the neural network to reach higher accuracies than the constituent data types processed individually. In other words, combining both types of data counters any weaknesses the individual networks might have (such as ambient light with depth) to create a more robust approach. For histograms, we observe similar values to the I+D data, but the added advantage is that it only uses a single, as opposed to two, data frames from the sensor (so the data acquisition time is halved), and processing is 28% faster. These results suggest that histograms represent a rich source of information that can be exploited, despite possessing low lateral resolution, for fast and accurate classification. This may be explained by the fact that a single pixel's histogram encompasses information about the depth, the surface reflectivity, the ambient level and the presence of edges between two objects, represented by a double peak.

6. CONCLUSIONS

This work presents the application of a dToF SPAD sensor to high-speed object detection. A neural network has been trained and tested for different forms of data from the sensor (depth, intensity, histogram and a combination of intensity and depth) and under different scenarios to compare the classification accuracy of the hand signs rock, paper and scissor. Processing speeds at video rates are obtained but we aim to increase these values by making modifications to the neural network and processing code while maintaining high accuracies.

The highest level of accuracy is typically achieved for histogram and intensity plus depth (I+D) data. This is despite the histogram data having four times lower lateral resolution than the intensity data, and half the overall data size of the I+D data, 64x32x16 vs 256x128x2 (the I+D data relying on two separate data frames from the SPAD sensor, obtained consecutively).

This points to the histogram data having potential in high-speed high-level scene interpretation, despite the modest lateral resolution and low number of temporal bins (16).

We observe that this preliminary study should continue by capturing more datasets involving a larger range of objects to classify, multiple objects to detect in the same scene (with occluded objects too) and more variability in depth to provide a more robust evidence for the advantages of photon timing histograms for object detection versus depth or intensity.

ACKNOWLEDGEMENTS

This work was supported by EPSRC through grants EP/M01326X/1 and EP/S001638/1. Also it is supported by the UK Royal Academy of Engineering Research Fellowship Scheme (Project RF/201718/17128) and DSTL Dasa project DSTLX1000147844). The authors are grateful to STMicroelectronics and the ENIAC-POLIS project for chip fabrication.

REFERENCES

[1] Gang Pan, Shi Han, Zhaohui Wu and Yueming Wang, "3D Face Recognition using Mapped Depth Images," IEEE Computer Science 175–175 (2006).
[2] Steger, C., Ulrich, M., and Wiedemann, C., [Machine Vision Algorithms and Applications], Wiley-VCH, 425-428 (2018).
[3] Chen, X., Ma, H., Wan, J., Li, B. and Xia, T., "Multi-view 3D object detection network for autonomous driving," Proc. 30th IEEE Conference on Computer Vision and Pattern Recognition, 6526-6534 (2017).
[4] Jing, C., Potgieter, J., Noble, F. and Wang, R., "A comparison and analysis of RGB-D cameras' depth performance for robotic application," 24th International Conference on Mechatronics and Machine Vision in Practice, 1-6 (2017).
[5] Giancola, S., Valenti, M., and Sala, R., [A survey on 3D cameras: Metrological comparison of time-of-flight, structured-light and active stereoscopy technologies], Springer, 12-20 (2018).
[6] Horaud, R., Hansard, M., Evangelidis, G. and Ménier, C., "An overview of depth cameras and range scanners based on time-of-flight technologies," Machine Vision and Applications 27(7), 1005-1020 (2016).
[7] David, R., Allard, B., Branca, X., and Joubert, C., "Study and Design of an Integrated CMOS Laser Diode Driver for an iToF-based 3D Image Sensor," CIPS 2020; International Conference on Integrated Power Electronics System, 1-6 (2020).
[8] Microsoft, "Azure Kinect DK", 2020, https://azure.microsoft.com/en-gb/services/kinect-dk/, (28 January 2021).
[9] Lucid Vision Labs, "Lucid Helios2", 2020, https://thinklucid.com/helios-time-of-flight-tof-camera/, (28 January 2021).
[10] Pancheri, L., Stoppa, D., Gonzo, L., and Dalla Betta, G.F., "A CMOS range camera based on single photon avalanche diodes," Technisches Messen, 74(2), 57-62 (2007).
[11] Marco, J., Hernandez, Q., Muñoz, A., Dong, Y., Jarabo, A., Kim, M.H., Tong, X. and Gutierrez, D., "DeepToF: Off-the-shelf real-time correction of multipath interference in time-of-flight imaging," ACM Transactions on Graphics, 36(6) (2017).
[12] Afshar, S., Hamilton, T. J., Davis, L., Van Schaik, A., and Delic, D., "Event-based processing of single photon avalanche diode sensors," IEEE Sensors Journal, 20(14), 7677-7691 (2020).
[13] A. Wallace, A. Halimi, G.S. Buller, "Full Waveform LiDAR for Adverse Weather Conditions," IEEE Trans. vehicular Technology, Vol. 69, Issue 7 (2020).
[14] Niclass, C., Soga, M., Matsubara, H. and Kato, S., "A 100-m range 10-Frame/s 340×96-Pixel Time-of-Flight Depth Sensor in 0.18- m CMOS," IEEE Journal of Solid-State Circuits, 48(2), 559-572 (2013).
[15] Henderson, R.K, Johnston, N., Rocca, F., Chen, H., Li, D.D., Hungerford, G., Hirsch, R., McLoskey, D., Yip, P. and Birsch, D.J.S., "A 192×128 Time Correlated SPAD Image Sensor in 40nm CMOS Technology", IEEE Journal of Solid-State Circuits (2019).
[16] Zhang, C., Lindner, S., Antolovi, I.M., Pavia, J.M. and Wolf, M., "A 30-frames/s, 252×144 SPAD Flash LiDAR With Integrated Histogramming," IEEE Journal of Solid-State Circuits, 54(4), 1-15 (2019).
[17] Li, Y. and Ibanez-Guzman, J., "LiDAR for Autonomous Driving," IEEE Signal Processing Magazine, 37(4), 50-61 (2020).
[18] Albawi, S., Mohammed, T.A., and Al-Zawi, S., "Understanding of a convolutional neural network," Proc. International Conference on Engineering and Technology, ICET, 1-6 (2018).
[19] Redmon, J. and Farhadi, A., "YOLOv3: An incremental improvement," ArXiv (2018).
[20] Chen, X., Gigu, P. and Behley, J. (n.d.)., "SuMa++: Efficient LiDAR-based Semantic SLAM," Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS) (2019).
[21] Aßmann, A., Stewart, B., and Wallace, A. M., "Deep Learning for LiDAR Waveforms with Multiple Returns," IEEE 2020 28th European Signal Processing Conference (EUSIPCO) (pp. 1571-1575), (2021).

[22] Hutchings, S.W., Johnston, N., Gyongy, I., Al Abbas, T., Dutton, N.A.W., Tyler, M., Chan, S., Leach, J. and Henderson, R.K., "A Reconfigurable 3-D-Stacked SPAD Imager with In-Pixel Histogramming for Flash LIDAR or High-Speed Time-of-Flight Imaging," IEEE Journal of Solid-State Circuits, 54(11), 2947-2956 (2019).
[23] Gyongy, I., Hutchings, S.W., Halimi, A., Tyler, M., Chan, S., Zhu, F., McLaughlin, S., Henderson, R.K. and Leach, J., "High-speed 3D sensing via hybrid-mode imaging and guided upsampling," Optica, 7(10), 1253 (2020).
[24] Ruget, A., McLaughlin, S., Henderson, R.K., Gyongy, I., Halimi, A. and Leach, J., "Robust super-resolution depth imaging via a multi-feature fusion deep network", Optics Express, 1-16 (2021).
[25] Turpin, A., Musarra, G., Kapitany, V., Tonolini, F., Lyons, A., Starshynov, I., Villa, F., Conca, E., Fioranelli, F., Murray-Smith, R., and Faccio, D., "Spatial images from temporal data," ArXiv, 1-12 (2019).
[26] Ronneberger, O., Fischer, P. and Brox, T., "U-net: Convolutional networks for biomedical image segmentation," Springer International Publishing, 234-241 (2015).

APPENDIX

Table 3. Layer-by-layer description of the adapted U-net network. The input layer size is adapted to each data type.

Layer number	Layer name	Output Shape	Activation type	Connected to layer number
1	Input	(X,Y)	-	-
2	Conv2D (3,3)	(X,Y,32)	ReLu	1
3	Conv2D (3,3)	(X,Y,32)	ReLu	2
4	MaxPooling2D (2,2)	(X/2,Y/2,32)	-	3
5	Conv2D (3,3)	(X/2,Y/2,64)	ReLu	4
6	Conv2D (3,3)	(X/2,Y/2,64)	ReLu	5
7	MaxPooling2D (2,2)	(X/4,Y/4,64)	-	6
8	Conv2D (3,3)	(X/4,Y/4,128)	ReLu	7
9	Conv2D (3,3)	(X/4,Y/4,128)	ReLu	8
10	MaxPooling2D (2,2)	(X/8,Y/8,128)	-	9
11	Conv2D (3,3)	(X/8,Y/8,256)	ReLu	10
12	Conv2D (3,3)	(X/8,Y/8,256)	ReLu	11
13	MaxPooling2D (2,2)	(X/16,Y/16,256)	-	12
14	Conv2D (3,3)	(X/16,Y/16,512)	ReLu	13
15	Conv2D (3,3)	(X/16,Y/16,512)	ReLu	14
16	UpSampling2D (2,2)	(X/8,Y/8,512)	-	15
17	Concatenation	(X/8,Y/8,768)	-	12 & 16
18	Conv2D (3,3)	(X/8,Y/8,256)	ReLu	17
19	Conv2D (3,3)	(X/8,Y/8,256)	ReLu	18
20	UpSampling2D (2,2)	(X/4,Y/4,256)	-	19
21	Concatenation	(X/4,Y/4,384)	-	9 & 20
22	Conv2D (3,3)	(X/4,Y/4,128)	ReLu	21

23	Conv2D (3,3)	(X/4,Y/4,128)	ReLu	22
24	UpSampling2D (2,2)	(X/2,Y/2,128)	-	23
25	Concatenation	(X/2,Y/2,192)	-	6 & 24
26	Conv2D (3,3)	(X/2,Y/2,64)	ReLu	25
27	Conv2D (3,3)	(X/2,Y/2,64)	ReLu	26
28	UpSampling2D (2,2)	(X,Y,64)	-	27
29	Concatenation	(X,Y,96)	-	3 & 28
30	Conv2D (3,3)	(X,Y,32)	ReLu	29
31	Conv2D (3,3)	(X,Y,32)	ReLu	30
32	Conv2D (1,1)	(X,Y)	Sigmoid	31

Table 4. T-student parameters t_{crit}, t_{stat} and results (average accuracies statistically equal or unequal) between all data types.

	t_{crit}	t_{stat}	**Result**	t_{crit}	t_{stat}	**Result**	t_{crit}	t_{stat}	**Result**	t_{crit}	t_{stat}	**Result**
	Intensity			**Depth**			**Histogram**			**I+D**		
Intensity	-	-	-	2.10	-1.63	**Equal**	2.14	4.18	**Unequal**	2.10	3.88	**Unequal**
Depth	2.10	-1.63	**Equal**	-	-	**-**	2.10	-8.54	**Unequal**	2.10	6.67	**Unequal**
Histogram	2.14	4.18	**Unequal**	2.10	-8.54	**Unequal**	-	-	**-**	2.10	-0.50	**Equal**
I+D	2.10	3.88	**Unequal**	2.10	6.67	**Unequal**	2.10	-0.50	**Equal**	-	-	**-**