High-speed, super-resolution 3D imaging using a SPAD dToF sensor

Germán Mora-Martín [1], Jonathan Leach [2], Robert K. Henderson [1], Istvan Gyongy [1]

[1]The University of Edinburgh, Institute for Integrated Micro and Nano Systems, Edinburgh, U.K.
[2]Heriot-Watt University, Institute of Photonics and Quantum Sciences, Edinburgh, U.K.
Istvan.Gyongy@ed.ac.uk Tel: +44 131 651 7054

***Abstract*—**High-speed 3D time-of-flight (ToF) imaging has the potential to offer improved situational awareness in robotics and automotive applications as well as assisting photogrammetry-based high-speed scientific imaging such as material testing. This paper uses a CMOS SPAD dToF sensor for depth and intensity imaging at up to 10kFPS. Depth maps are upscaled from a resolution of 64×32 to 256×128 using a recently proposed video super-resolution techniques tailored to SPADs. We also present preliminary results from the application of the sensor to human activity recognition (HAR).

I. INTRODUCTION

The use of SPAD-based 3D depth sensors has become widespread in the last few years, with the sensors finding applications in smartphones, robotics, and even home appliances [1]. SPADs have also become a key technology in LIDAR for autonomous systems [2]. By integrating SPAD arrays with processing logic, solid-state, all-digital receivers can be implemented that provide accurate depth maps even in high ambient conditions. However, array sizes tend to be limited, leading to a relatively low angular resolution when imaging in a flash modality. Instead of using flood illumination, some SPAD modules project a dot array (using a diffractive optical element [3]) which increases the SNR in the spots and thus the range but again results in sparse spatial sampling. There is therefore an interest in using post-processing to improve the lateral resolution of depth maps, as well as to provide scene interpretation, especially for long-range targets subject to significant pixelation.

II. SENSOR ARCHITECTURE

We used a high-speed SPAD dToF sensor in our study [4], capable of running at frame rates in the 10 kFPS range (>100 kFPS for on-chip depth computation). The sensor, implemented in STMicroelectronics' 40nm technology, comprises 64×32 pixels, each pixel consisting of a 4×4 array of SPADs and processing logic. A time-gated, multi-event histogramming TDC is integrated into each pixel, generating an 8-bin histogram with a resolution down to ~250 ps [4]. The time gate functionality enables the histogram to be shifted in time to extend the range of the sensor. Three main mechanisms are available for setting the time gate positions of individual pixels: (1) internal control that automatically tracks the signal peak via in-pixel background estimation and peak detection, (2) internal control that continually cycles across up to 128-time gate positions, and (3) external control, potentially based on guidance from an additional sensor [5] (such as a stereo vision system). On-chip (column parallel) depth computation and selective readout options are available to provide further data compression. In addition to time-resolved imaging, the sensor offers a 128×128 photon counting (intensity) imaging modality. Figure 1 shows a portable camera setup consisting of a custom PCB (housing the SPAD and an FPGA module), a compact 850 nm VCSEL illumination module, and a laser range finder for reference depth measurements. The camera is connected to a laptop which controls and powers the camera; a Matlab software interface provides real-time visualisation of the captured data. For post-processing, we used a desktop computer (HP EliteDesk 800 G5 TWR) with an RTX2070 GPU.

III. HIGH-SPEED, SUPER-RESOLVED DATA

Figure 2 depicts examples captured at 200 FPS in the tracking modality (mode 1 above) of the sensor. The scene is of two people, one running and the other waving, in an open space with objects scattered around. The figure shows depth data obtained by applying centre-of-mass peak extraction on the histogram frame (panel a) as well as the upscaled version (from 64×32 to 256×128) of this data following neural network super-resolution processing (b) [6], which is seen to lead to an improvement in the profiles of the people. Unlike commonly used, intensity-guided approaches [7], the upscaling is based entirely on the depth data here. Processing speeds above 30 FPS are achieved. Figure 3 shows data acquired indoors at 10 kFPS of a balloon being burst. Three sequences are given: intensity frames (3a), depth frames (3b), and super-resolved depth data (3c). The rupture of the balloon is captured in high temporal detail, demonstrating the potential of SPAD cameras in specialised high-speed imaging applications [8], especially where high sensitivity is required, as offered by state-of-the-art SPADs [9]

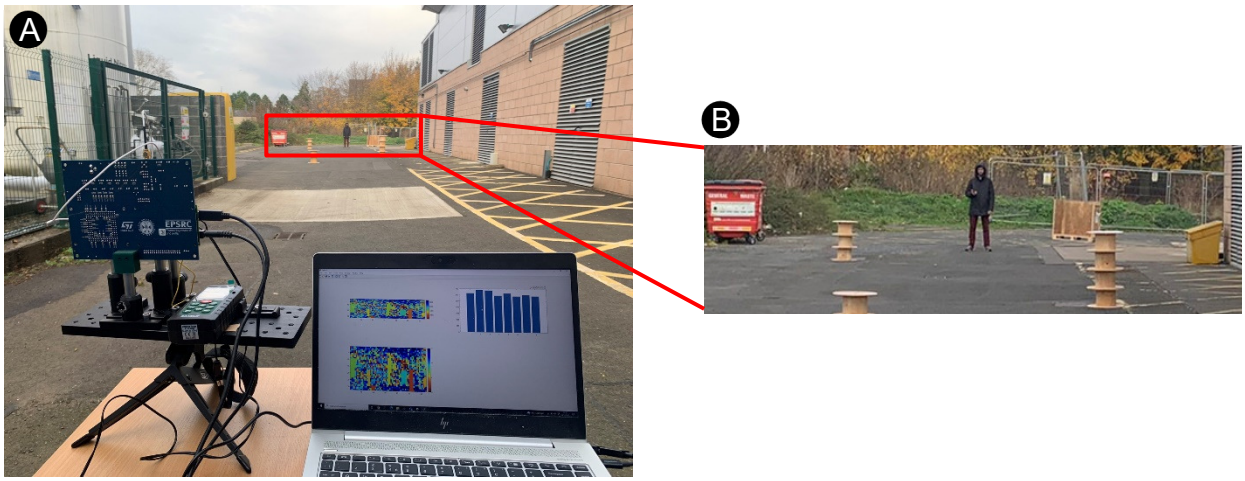


Figure 1. a) Picture of the camera setup. A 25 mm/f1.4 objective is used in front of the sensor, giving a 20×5 degree field-of-view (FOV), together with a 10 nm ambient filter. Illumination is provided by a compact 850 nm VCSEL source with 10 ns pulse width and 60 W peak optical power triggered at 1.2 MHz. b) The FOV of the camera.

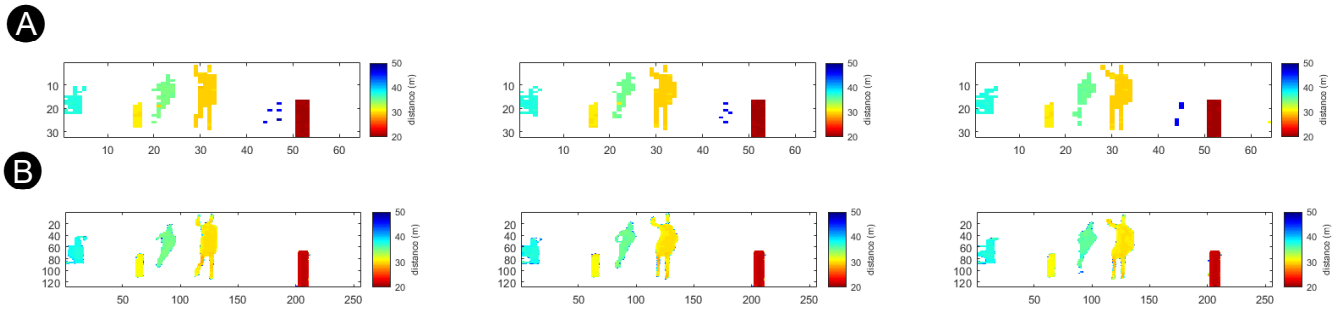
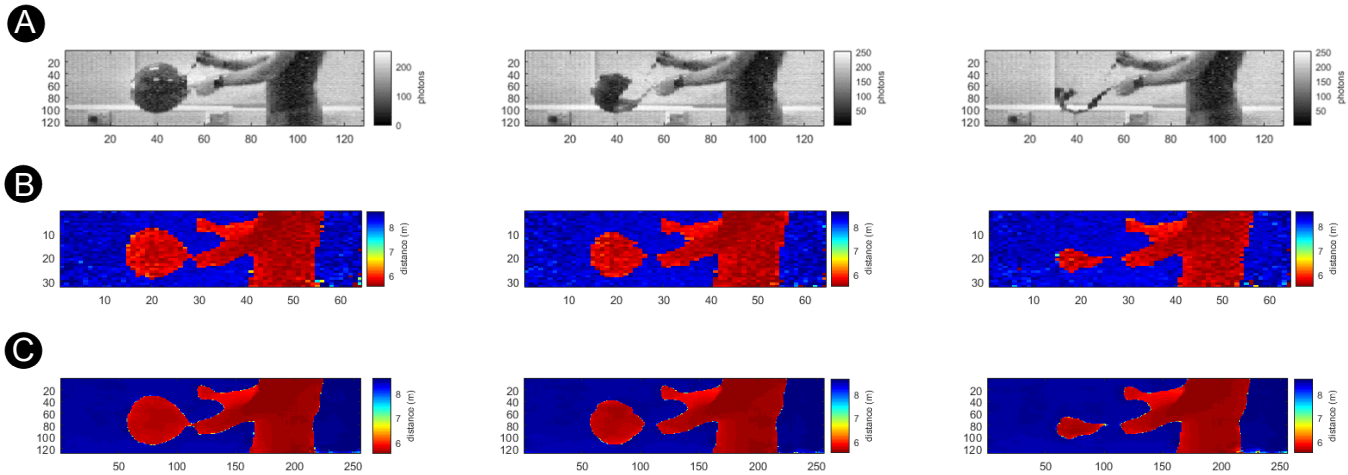


Figure 2. Selected frames from data captured in tracking mode at 200 FPS with 8 ns bin size and 16 time gate positions (giving 81.6m of unambiguous range) a) depth maps obtained by centre-of-mass processing of histogram frames b) corresponding super-resolution depth maps. Only pixels which are detecting surfaces in the 20-50 m range are plotted in panel (a).



64×32 depth data derived from histogram frames c) the data in panel b upscaled to 128×256 after super-resolution processing. Note that the photon counting, and depth sequences were captured separately.

IV. HUMAN ACTIVITY RECOGNITION

Human activity recognition (HAR) has gained importance in computer vision due to its applications in video surveillance, health care services, human-computer interaction, and autonomous driving [10]. Using just depth information for HAR has become a popular research topic due to the preservation of privacy and fast speed compared with other methods (e.g., using 3D skeletons) [11]. Furthermore, depth-based detection has the potential to work even when there is no colour contrast between the person and the background (for example when camouflage is used [12]). However, one of the key challenges is to overcome the low transverse resolution of depth maps when imaging from a distance.

Recurrent neural networks (RNNs) are a very useful type of network for sequential data. In particular, convolutional long-short-term memory (Conv-LSTM) layers are key to learning spatio-temporal features from data [13]. In [11], an RNN based on conv-LSTM layers is used to perform HAR on high-resolution, indoor, and short-range depth data. In this work, we use a similar network to perform HAR on data from a SPAD dToF sensor involving longer-range, outdoor sequences, where SNR is typically lower, and objects can become heavily pixelated.

The method is designed to perform HAR on sequences of any length. First, a 64×32 depth sequence is captured and passed through a U-net-like network to localise people [14]. Next, the depth sequence is cropped spatially around each person in frames of 16×32 pixels and then resized to 32×32 pixels. Finally, the cropped sequence is analysed by the RNN network and outputs an activity from the following set: remaining idle, walking, running, crouching down, standing up, waving, or jumping. Figure 4 shows a diagram summarising the steps involved in this approach to perform HAR.

Unreal Engine [15] was used to generate a large and diverse training dataset, shared for both human localisation and HAR networks. Ground truth information for a variety of sensors can be extracted from virtual environments. In this work, depth, intensity, and segmentation frames of size 512×128 are recorded and used as inputs in an optical model to simulate data from a SPAD dToF sensor. To match the sensor architecture described here, the model assumes 4×4 SPAD macropixels, a pixel resolution of 64×32 (with an aspect ratio of 4:1), and in-pixel histogramming.

Figure 4 shows the confusion matrix of all activities considered here, indicating % of samples predicted in a given class in the test dataset (data unseen by the model). The overall accuracy for the test dataset is 91.5%. Activities corresponding to standing up, walking, running, jumping, and waving are detected with a recall higher than 90% while crouching down, and remaining idle has lower recall values (though a precision of >93%). False positives can occur due to similarities between two actions (such as crouching down slightly before jumping) or failure to localise the person accurately (in some cases due to distracting features in the background). The network is able to perform HAR from a sequence with a latency of 150 ms.

Activity sequences were captured using the SPAD dToF sensor at 50 FPS to generate a test dataset from real data. Figure 6 compares a sequence of a person walking captured by the real dToF sensor (Fig. 6a) with corresponding synthetic SPAD data (Fig. 6b). The visual similarity between the two sequences appears to justify the use of synthetic SPAD data for training. Indeed, preliminary results suggest similar performance on the real dataset to the results on the synthetic dataset, HAR predicting activities such as running, walking, and standing up with high sensitivity, whilst crouching down and remaining idle have reduced recall values (but high precision).

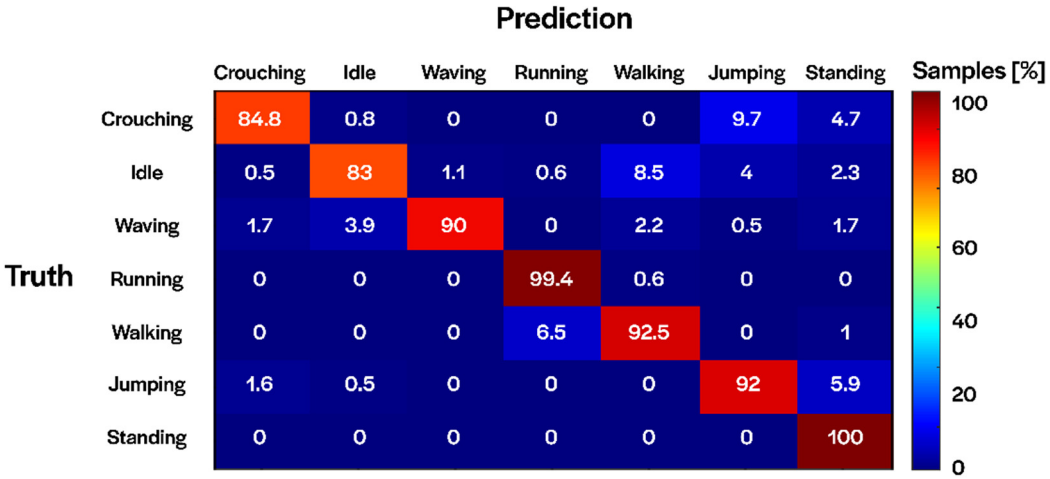


Figure 4: Confusion matrix of activities representing % of samples predicted in each class. Example: 1.6% of jumping data is confused with crouching down. Each class has approximately 200 samples.

V. CONCLUSIONS

We demonstrated the application of a dToF SPAD sensor in high-speed imaging and showcased the use of deep learning models, trained on synthetic SPAD data, to overcome the limited transverse resolution and provide upscaled depth maps or human activity recognition (HAR). Future work will attempt to improve segmentation and extend the method to multiple people within the field of view.

Acknowledgments— This research was supported by EPSRC via grants EP/M01326X/1, EP/S001638/1 and DSTL Dasa project DSTLX1000147844. The authors are grateful to STMicroelectronics for chip fabrication.

REFERENCES

[1] https://www.st.com/content/st_com/en/about/media-center/press-item.html/t4210.html *Last visited 06/04/2023*
[2] Li et al., *Federated learning: Challenges, Methods, and Future Directions,* IEEE Signal Process. Mag, 37(4), 2020
[3] *Breaking Down iPad Pro 11's LiDAR Scanner*, EE Times. Available online: https://www.eetimes.com/breaking-down-ipad-pro-11s-lidar-scanner
[4] Gyongy et al., *A direct time-of-flight image sensor with in-pixel surface detection and dynamic vision,* JSQTE 2023
[5] Taneski et al., *Guided Flash Lidar: A Laser Power Efficient Approach for Long-Range Lidar,* IISW 2023
[6] Mora Martín et al., *Video super-resolution for single-photon LIDAR,* Optics Express 31, 7060-7012 2023
[7] Ferstl et al., *Image guided depth upsampling using anisotropic total generalized variation,* ICCV 2013
[8] Etoh et al., *Needs, requirements, and new proposals for ultra-high-speed video cameras in Japan*, ICHSPP 1994
[9] Shimada et al., *A SPAD depth sensor robust against ambient light: the importance of pixel scaling and demonstration of a 2.5 µm pixel with 21.8% PDE at 940 nm,* IEDM 2022
[10] Song et al., *Pattern recognition, ICPR international workshops and challenges,* ICPR 2021
[11] Sánchez-Caballero et al., *Real-time human action recognition using raw depth video-based recurrent neural networks,* Multimedia Tools and Applications 2022
[12] Tachella et al., *Real-time 3D reconstruction from single-photon LIDAR data using a plug-and-play point cloud denoisers,* Nature Communications 2019
[13] Xingjian et al., *Convolutional LSTM network: a machine learning approach for precipitation nowcasting,* Advances in neural information 2015
[14] Ronneberger et al., *U-Net: convolutional networks for biomedical image segmentation,* MICCAI 2015
[15] Epic Games, *Unreal Engine*

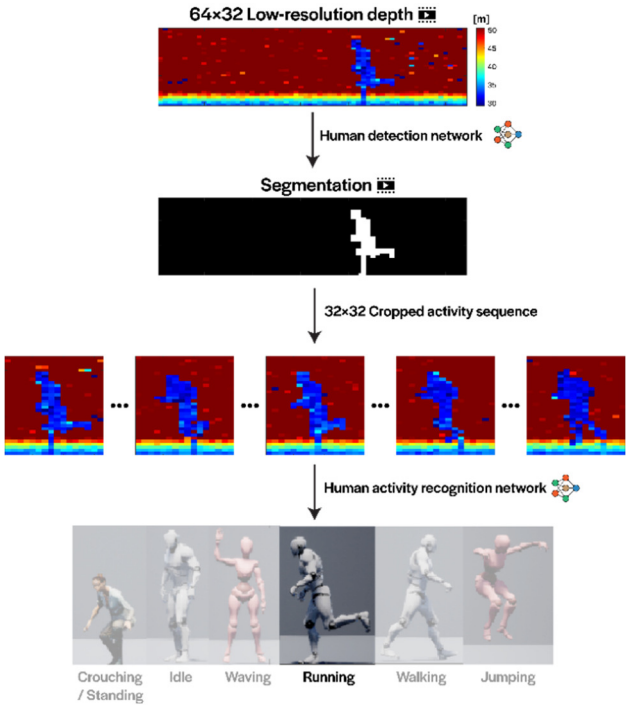


Figure 5. Human activity recognition workflow diagram. A low-resolution depth sequence is captured and segmented via a human segmentation network (similar to U-net). Based on the localisation of the human, the sequence is cropped accordingly. The cropped sequence is passed through a second network evaluating the activity performed (e.g. running).

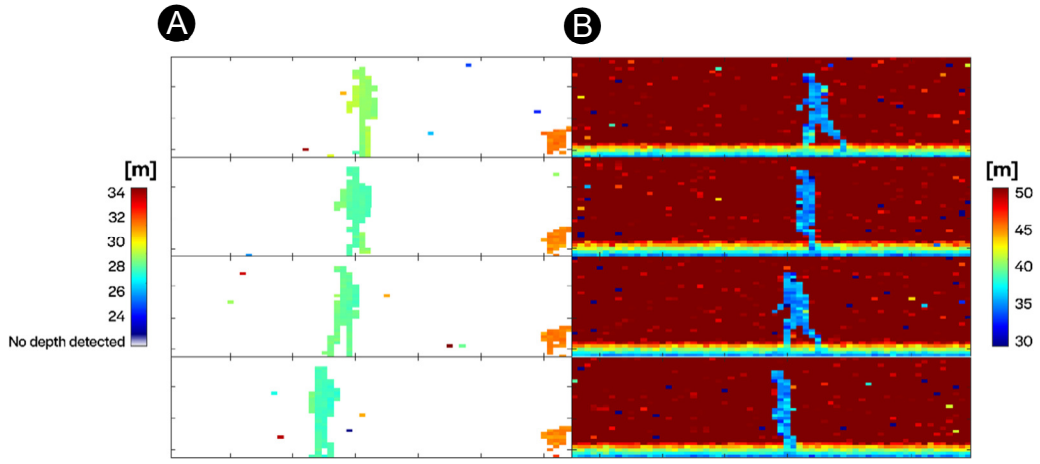


Figure 6. Comparison of selected frames from a sequence of a person walking captured with a) a real SPAD camera (50 FPS, mean signal-to-background photons ratio (SBR) of 0.17 and 716 average signal photons for the person in the first frame) b) a virtual SPAD camera (model-generated data, mean SBR 0.19 and 42 average photons).