



Human activity recognition using a single-photon direct time-of-flight sensor

GERMÁN MORA-MARTÍN,^{1,*}  STIRLING SCHOLES,² ROBERT K. HENDERSON,¹  JONATHAN LEACH,²  AND ISTVAN GYONGY¹ 

¹*School of Engineering, Institute for Integrated Micro and Nano Systems, The University of Edinburgh, Edinburgh EH9 3FF, UK*

²*School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh EH14 4AS, UK*

**german.mora@ed.ac.uk*

Abstract: Single-Photon Avalanche Diode (SPAD) direct Time-of-Flight (dToF) sensors provide depth imaging over long distances, enabling the detection of objects even in the absence of contrast in colour or texture. However, distant objects are represented by just a few pixels and are subject to noise from solar interference, limiting the applicability of existing computer vision techniques for high-level scene interpretation. We present a new SPAD-based vision system for human activity recognition, based on convolutional and recurrent neural networks, which is trained entirely on synthetic data. In tests using real data from a 64×32 pixel SPAD, captured over a distance of 40 m, the scheme successfully overcomes the limited transverse resolution (in which human limbs are approximately one pixel across), achieving an average accuracy of 89% in distinguishing between seven different activities. The approach analyses continuous streams of video-rate depth data at a maximal rate of 66 FPS when executed on a GPU, making it well-suited for real-time applications such as surveillance or situational awareness in autonomous systems.

© 2024 Optica Publishing Group under the terms of the [Optica Open Access Publishing Agreement](#)

1. Introduction

Human activity recognition (HAR) has become increasingly significant in computer vision due to its applications in video surveillance, health care services, human-computer interaction, and autonomous driving [1–4]. In the latter, activity recognition presents advantages in situational awareness over object detection by providing a higher level interpretation of the environment [5], enabling autonomous cars to anticipate and respond to hazards more effectively.

Initial approaches for HAR involved encoding motion-related information in single RGB images [6] or computing changes in direction, speed, and shape of a space-time volume (STV) [7]. However, these approaches are inflexible in accommodating variations in the activities, such as different view points or occlusion. More sophisticated algorithms based on local descriptors using histogram of oriented gradients [8], histogram of optical flows [9], or support vector machines (SVM) [10] lead to improved HAR performance. Nevertheless, these methods tend to be computationally expensive and incompatible with real-time processing. The arrival of deep learning methods has impacted the field of HAR, with significant performance improvements when using spatio-temporal networks such as 3D convolutional neural networks (CNN) [11,12]. Alternatively, recurrent neural networks (RNN) are typically better suited to handle sequential data and, in particular, convolutional long-short-term memory (ConvLSTM) layers have been key in further improving the performance of HAR [13–15].

HAR based on depth data has become popular as it guarantees good contrast between people and the background (even when camouflage is used [16]), as well as preserving privacy. It typically calculates 3D-skeletons via optical flow or dynamic images (in combination with RGB information) prior to activity recognition [17–20]. In [21], an RNN based on ConvLSTM layers is used to perform HAR from high-resolution, indoor, short-range depth data with an average accuracy of 75% on the NTU RGB+D dataset.

In this paper, we extend depth-based HAR to outdoor, longer range scenarios to support the target application of autonomous systems. Existing techniques are not suitable in this domain as they rely on indirect ToF sensors, which have a limited range outdoors. Furthermore, they assume that people are captured with a high pixel resolution, which is not the case when imaging over longer distances (unless the field-of-view (FoV) is limited). We therefore adopt direct Time-of-Flight (dToF) imaging, based on a Single-Photon Avalanche Diode (SPAD) sensor, which estimates depth by illuminating the scene with a pulsed light source, and measuring the time of arrival of backscattered photons [22]. These sensors are well-suited for long-range LIDAR as they can obtain precise depth estimates even from low photon returns [23–26]. Recent SPAD dToF cameras have been developed in image sensor format and, when combined with flood illumination, enable high-speed 3D imaging without any optical scanning [27,28].

Even with a suitable depth sensor identified, challenges remain in HAR over long distances due to the low transverse resolution as well as noise from solar interference [29]. Therefore, a robust approach is required that overcomes the effects of pixelation (with objects represented by just a few pixels), noise, and can operate in real-time. Although higher resolution LIDAR systems are available commercially (these typically use mechanical or MEMS scanning [30,31]), the use of a SPAD sensor here is motivated by two reasons: (1) the availability of accurate simulation models for the generation of synthetic data for neural network training, and (2) accessibility to raw single-photon histogram data, providing control over depth map generation (e.g. choosing from multiple surface returns). Nevertheless, the approach presented here is expected to be adaptable to other LIDAR systems. Indeed, even high resolution LIDAR (e.g. 0.1° as in [30]) suffers from significant pixelation when capturing long-range targets (a person at 200 m only being 1-2 pixels across).

In this paper, we avoid the intermediate steps of pose estimation or depth upscaling [32–34] so as to minimise latency. Instead, a CNN is trained for people and object segmentation based on depth (at the native resolution of the dToF camera), and a convLSTM network, inspired by [21], is adapted for HAR. The approach assumes a SPAD dToF sensor with 64×32 pixel resolution and a FoV such that human limbs are approximately one pixel across. A method similar to the one presented in [34] is used for the generation of a synthetic training and test dataset. The performance of the method is subsequently analysed for both the synthetic dataset, and real data captured from a state-of-the-art SPAD dToF sensor [35].

2. Data preparation

The open-source Airsim platform is used together with Unreal Engine like in [34] to generate a synthetic dataset for the purposes of training and testing the segmentation and HAR networks [36,37]. By using synthetic data, we can create large and diverse datasets much faster than it would be possible through physical experiments. However, we need to ensure that the generated data accurately represents the characteristics of real world data. In this work, RGB, segmentation (background corresponding to 0, people to 1 and objects to 2), and depth maps are generated with a lateral resolution of 512×128 and a horizontal FoV of 20° in Unreal, and used to create simulated depth frames, with a resolution of 64×32 , which mimic the SPAD sensor [35] used for collecting real data. Each pixel in the sensor is composed of 4×4 SPAD detectors, and the overall aspect ratio of the array is 4:1 due to the processing units next to every column of pixels. The high-resolution depth map and grayscale version of the RGB data (representing surface reflectivity) are used to synthesise SPAD data with 8-bin temporal photon histograms, using a multi-event, in-pixel histogramming TDC architecture with adjustable bin size. In the model, the signal photon rate is proportional to the surface reflectivity extracted from the grayscale frame and inversely proportional to the square of the distance to the surface [38]. On the other hand, the background photon rate is proportional to the surface reflectivity and the ambient level, with both parameters being randomised to create scenes with different signal-to-background

ratio (SBR) levels. The SBR is defined as the ratio of signal photons and background photons across all bins in the histogram, and averaged over all SPAD pixels. To account for Poisson noise in the signal and background photon counts, the photon timing histograms computed for each pixel are randomised according to Poisson statistics. Finally, depth information is extracted from histograms using centre-of-mass peak extraction [39]. To define the signal-to-noise ratio (SNR) in the dToF SPAD frames, we use Eq. (1) from [40], which gives the depth precision as a function of signal and background photon levels, amongst other parameters [41]

$$\delta = \frac{\sigma}{\sqrt{N_{sig}}} \sqrt{1 + \frac{1}{12} \left(\frac{a}{\sigma} \right)^2 + 4\sqrt{\pi} \left(\frac{\sigma}{a} \right) \frac{b}{N_{sig}}}, \quad (1)$$

where σ is the standard deviation of the IRF, N_{sig} is the expected total number of photons in the dominant signal peak of the histogram, a is the histogram bin width and b is the expected number of background photons in each histogram bin. In the case the error in the depth estimate is background dominated (as it is typical for flash LIDAR configuration), the standard deviation in the depth estimate is proportional to $\frac{\sqrt{b}}{N_{sig}}$. We therefore define the SNR for a given pixel as $\frac{N_{sig}}{\sqrt{b}}$, which is a measure of how well defined the signal peak is (as N_{sig} gives the size of the peak while $\sqrt{b}$ corresponds to the standard deviation of the noise floor). The overall SNR in a given frame is calculated as the average SNR across all SPAD pixels.

To obtain ground truth segmentation data which accurately matches the profile of objects in the synthetic SPAD frames, the high-resolution segmentation map provided by Unreal Engine is downsampled by passing it through the SPAD model (note that basic nearest neighbours downsampling results in disparities in the profiles). For this downsampling process, we use the

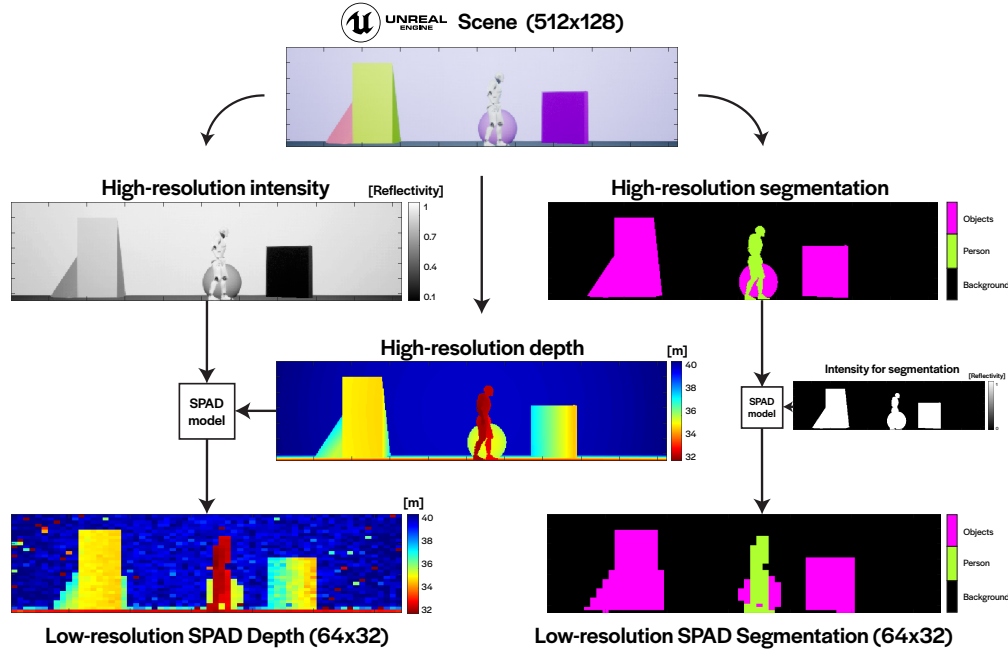


Fig. 1. Synthetic data generation workflow diagram. RGB (converted to its grayscale equivalent), depth, and segmentation frames are captured from Unreal Engine using Airsim. The SPAD optical model is used to obtain low-resolution SPAD depth and segmentation data to train and test the scene segmentation and HAR networks.

high-resolution segmentation as a pseudo-depth input and a binarised version of the segmentation as a pseudo-intensity input (background set to 0 and objects and people to 1). Figure 1 shows a diagram summarising the use of Unreal data to obtain synthetic SPAD dToF data through the optical model.

3. Scene segmentation and human activity recognition networks

As in [42], the structure of U-net [43] is adapted here for the fast localisation of people and general objects. The input of the network is a 64×32 depth image normalised between 0 and 1, and the output has three channels ($64 \times 32 \times 3$), corresponding to detection of background (0), people (1), and objects (2). Without the objects class, the network is less robust since it tends to predict objects as people rather than background in examples where there is limited depth contrast between features.

For HAR, the network used in this paper is adapted from the stateless network in [21] because of its simplicity, speed, and high accuracy on high-resolution depth sequences. Note that [21] excludes a segmentation step since human features are predominant, whereas here they represent just a few pixels and could be occluded by other elements in the scene. Additionally, our approach is designed to analyse HAR from continuous streams of depth data containing multiple people. The adapted network contains fewer layers with modified parameters to suit the smaller resolution of the data here (e.g. smaller convolution kernels and strides are used). The network comprises of two branches, each featuring multiple recurrent blocks designed to extract features from video data. Both branches are then added together and go through a decision block that includes convolutional and pooling layers. Finally, the output class is extracted through a softmax activation function. A layer by layer description of the network is available in the Supplemental material. The network is trained in stateless mode, which fixes the amount of input frames to be processed per sequence. In this work, input sequences are resampled to a total of 32 frames to provide a suitable balance between performance and inference time. While a stateful training might seem more suitable in HAR to cope with variable lengths, it is important to note that stateful networks are much more complex and unstable to train [21] (they can easily diverge in the training process). Additionally, its use would be impractical in real time scenarios since there is no prior knowledge of the start and end time of actions.

The workflow of the whole scheme is the following: first, a $64 \times 32 \times 32$ depth sequence is passed through the segmentation network to localise people in each frame and to assign corresponding bounding boxes. The detection confidence for each pixel in the segmentation map has to be $>80\%$ to be considered as a valid detection. Next, the depth sequence is cropped spatially around each person in frames of 32×32 pixels. Note that spatially cropping the data is not a necessary step for HAR, but in doing so, the processing speed of the network is increased due to the reduced input size (which removes redundant data). The cropped sequence is also filtered, by assigning a fixed value of 1 to pixels that correspond to either background or objects, to limit their influence on the inference. Finally, the cropped sequence is analysed by the HAR network, which outputs an activity from the following set: remaining idle, walking, running, crouching down, standing up, waving, or jumping. The training dataset is randomly shuffled to prevent biasing the weights from the network to specific scenarios. Figure 2 shows a diagram summarising the steps involved in this approach to perform HAR.

When there are multiple people detected in a scene, it is crucial to assign the same bounding box label consistently to each person. To avoid the accidental swapping of bounding box labels, the prior and current bounding box locations of each person are compared and labelled appropriately. This prevents sudden jumps in the input sequence to the HAR network, therefore reducing the number of misclassifications. In the case that a person is not detected in a given frame, its prior bounding box is assigned (unless the person is going out of frame). When two people are crossing, considered here as the lateral distance between the centres of their bounding boxes

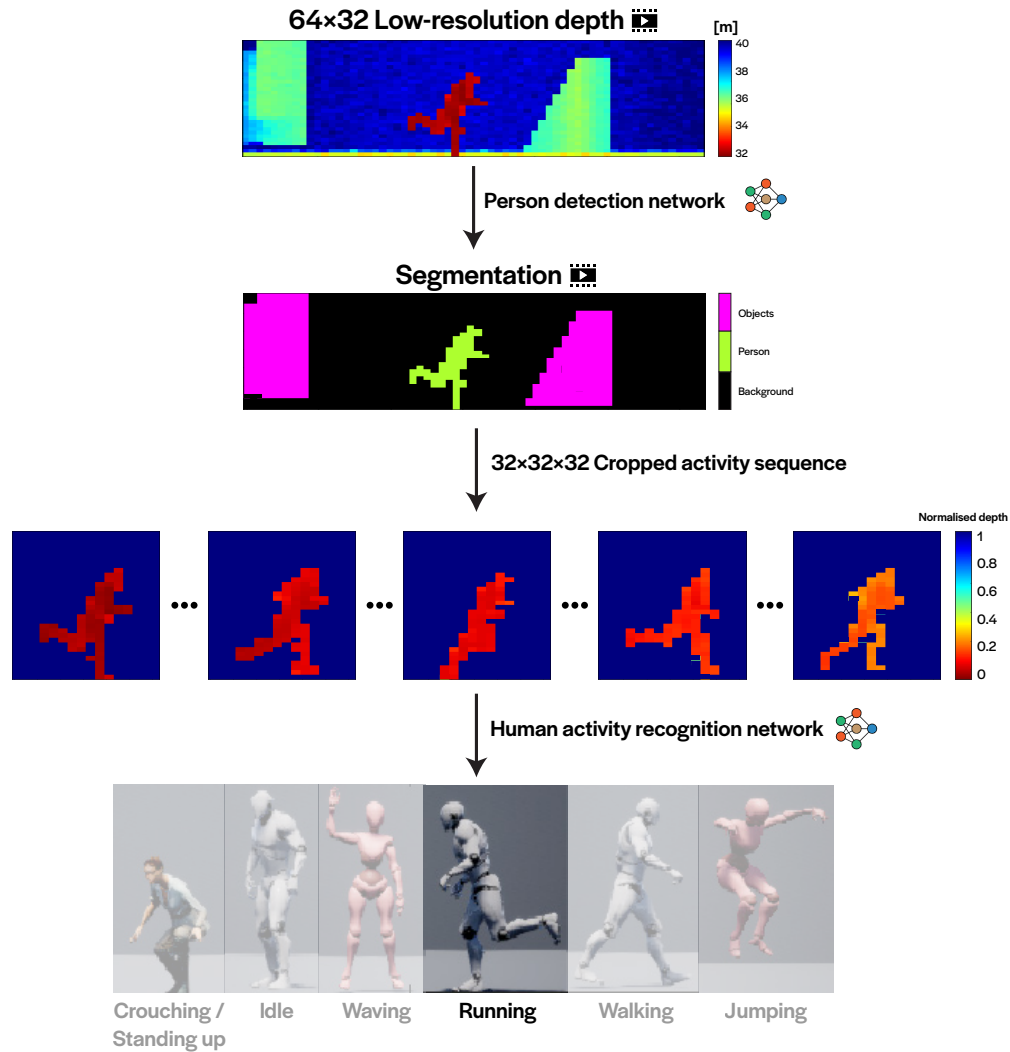


Fig. 2. Human activity recognition workflow diagram. A low-resolution depth sequence is captured and segmented via a scene segmentation network. Based on the localisation of the person, the sequence is cropped and filtered by assigning a fixed value of 1 to pixels that correspond to background and objects. The cropped sequence (32 frames) is passed through a second network evaluating the activity performed (e.g. running).

being below 5 pixels, the assignment of labels can be prone to errors. To prevent this, the average depth position of each person is calculated before and after two people cross each other, and bounding boxes are labelled accordingly (e.g. the person who is in front as two people cross each other, remains in front immediately after crossing). Note that there might be some scenarios involving complex trajectories as people cross each other leading to mislabelled bounding boxes, but this was not observed for the cases studied here.

The neural networks are implemented in Tensorflow using Keras [44]. The training stage is performed using the Adam optimiser [45] in a desktop computer (HP EliteDesk 800 G5 TWR) with the assistance of a RTX2070 GPU to accelerate the task. For the localisation network, the performance is tracked at every step by the F-score and the loss to be minimised is the focal

Tversky loss with a maximum of 50 epochs and batch size of 32. This loss is used to compensate for class imbalance, as the background is often more represented than other classes. Early stopping is introduced to avoid overfitting the model, whereby the training process is terminated after 5 epochs and the weights corresponding to the minimum loss are recorded. The learning rate of the model is set to $1e-3$. For the HAR network, the performance is tracked at every step by the categorical accuracy and the loss to be minimised is the categorical crossentropy, with a maximum of 50 epochs and batch size of 8. As before, early stopping after 5 epochs is imposed but in this case the learning rate is set to $5e-5$.

4. Results

4.1. HAR on synthetic data

The performance of the segmentation network is evaluated for a model trained with 80,000 examples. These examples include a diverse set of randomised depth data with one or two people (with random orientations), featuring multiple objects at different locations and depths (ranging from 30 to 40 metres). The scenes also feature randomised reflectivity and background photon levels (SBR ranging from 0.15-0.7 and SNR from 2-10). The validation dataset corresponds to 10% of the training set, providing an unbiased evaluation of the neural network. The test dataset consists of 2,100 sequences containing 32 frames each. Using the intersection over union (IoU) with threshold 0.5 as a metric to assess the performance of the network, the percentage of correctly localised samples is 96%. Figure 3 shows a pair of examples where the network fails to provide a correct segmentation of people. In the first case, the detection confidence for the person is below 80% whereas in the second case, an object with dimensions similar to a person has been mislabelled.

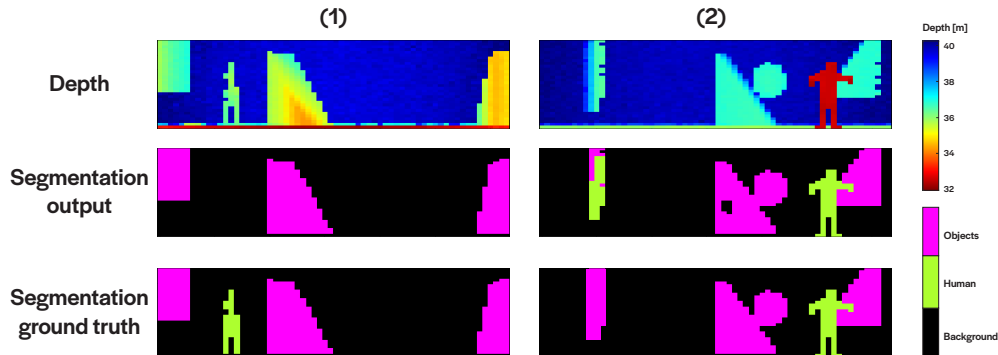


Fig. 3. Examples of poor segmentation from the synthetic test dataset, showing depth information, segmentation output, and segmentation ground truth. (1) example where the detection confidence for the person is below 80% and as a consequence, the person is not detected. (2) example of misclassification where an object has been detected as a person.

The performance of HAR is evaluated for a model trained with 7,600 depth sequences, 10% of which are reserved for the validation dataset. The test dataset is equivalent to the segmentation network. Figure 4(a) 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) and Table 1 shows the results in terms of recall, precision and F-score. Precision measures the ratio between true positive predictions and total predicted positive observations. Recall, or sensitivity, measures the ratio between true positive predictions and all observations in a class. Finally, the F-score is the harmonic mean of the precision and recall. The overall weighted average and simple average (with all classes having the same weight) of the accuracy are both 92% and the network is able

to perform HAR from the raw data sequence with a maximal rate of 66 FPS. All activities are detected with a recall higher than 90% with the exception of walking (though a precision of 98% is attained for that action). We note that the network is highly sensitive to the training data (exhibits high variance) due to the similarity between some actions, which results in different inference performance when training the data repeatedly over the same dataset. As a consequence, the weights of the network can easily change to favor a specific class, resulting in an asymmetric confusion matrix. However, we note that the average performance over all categories tends to be consistent. False positives occur mostly due to similarities between two actions (such as running and walking) or, less frequently, due to a failure to localise the person accurately (in some cases due to distracting features in the background as seen in Fig. 3). [Visualization 1](#) shows example sequences of each activity considered in this paper being segmented and classified correctly. The video includes RGB data, segmentation output and depth combined with HAR results.

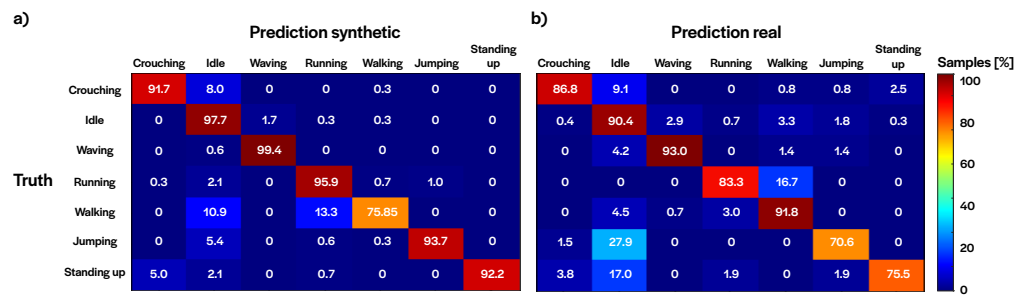


Fig. 4. Confusion matrix of activities representing % of samples predicted in each class for a) synthetic dataset and b) real dataset. Example: 4.2% of waving data is confused with remaining idle.

Table 1. Performance parameters (precision, recall, F1-score) and number of samples for synthetic test dataset (left) and real test dataset (right).

Activity	Synthetic				Real			
	Precision [%]	Recall [%]	F-score [%]	# Samples	Precision [%]	Recall [%]	F-score [%]	# Samples
Crouching	95	92	93	302	96	91	93	121
Idle	77	98	86	300	82	90	86	272
Waving	98	99	99	312	92	93	93	142
Running	86	96	91	294	91	83	87	180
Walking	98	76	86	294	90	92	91	401
Jumping	99	94	96	226	84	71	77	68
Standing up	100	92	96	282	90	81	85	53

Synthetic activity sequences involving two people crossing each other have been also captured and tested. Although the observed HAR performance is promising, a detailed assessment has yet not been carried out as the primary objective was to evaluate whether people can be properly tracked for segmentation purposes. Figure 5 depicts example frames of a depth sequence of two people jumping and running, being properly localised and classified. This illustrates the effectiveness of the method in maintaining the correct bounding box for each person even when they cross each other.

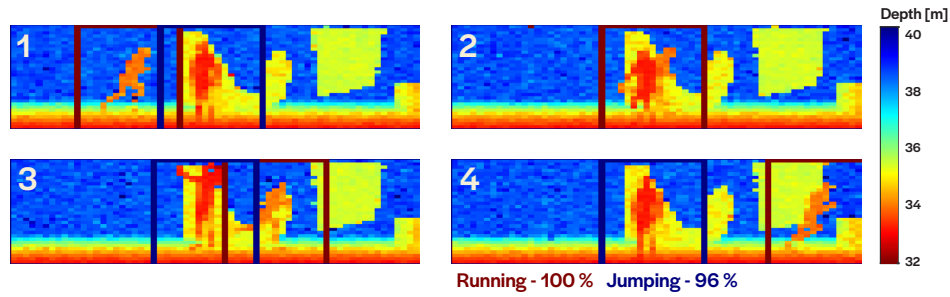


Fig. 5. Example of a synthetic sequence where two people cross each other. The bounding boxes are preserved for each person throughout the sequence and the network is able to predict correctly their activities. On average, the sequence has a SBR of 0.13, SNR of 2.52 and 49 average signal photons.

4.2. HAR on experimental data

Activity sequences were captured outdoors using a 64×32 SPAD dToF sensor at 50 FPS to generate a test dataset from real data [35]. The SPAD sensor features on-chip, multi-event histogramming which enables high-speed 3D imaging without the need of summing binary frames of photon time stamps. The near infrared laser source (850 nm), triggered from the SPAD, is spread over the FoV of the sensor ($20 \times 5^\circ$) using cylindrical lenses. The laser has an optical peak power of 60 W and emits 10 ns pulses with a repetition rate of 1.2 MHz. A 25 mm/f1.4 objective lens (Thorlabs MVL25M23) is used in front of the SPAD, together with a 10 nm bandpass ambient filter (Thorlabs FL850-10). Figure 6 shows an image of the portable camera system to capture real data. Figure 7 compares a sequence of a person walking captured by a real dToF sensor (Fig. 7(a)) with synthetic SPAD dToF data (Fig. 7(b)). The visual similarity between the two sequences justifies the use of synthetic SPAD data for training.

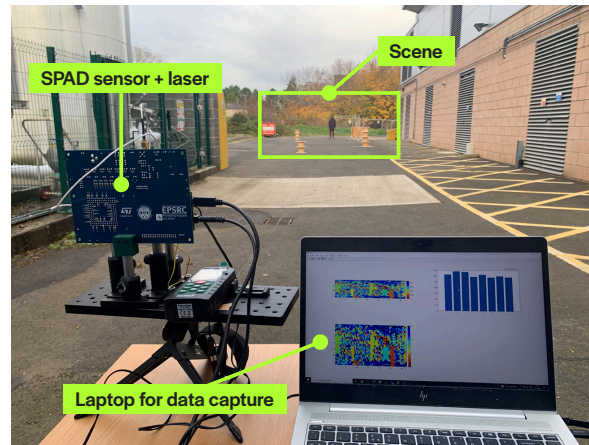


Fig. 6. Picture of the portable camera system used in this paper to capture real SPAD dToF data. The system includes a SPAD dToF sensor, an FPGA integration module (Opal Kelly XEM7310), a 850 nm laser source, a 25 mm/f1.4 objective lens with a 10 nm bandpass filter, and a laptop for data capture.

The performance of HAR is evaluated for a test dataset containing multiple sequences with a total of 1,237 blocks of 32 frames. Each sequence is temporally cropped so that it is a multiple of

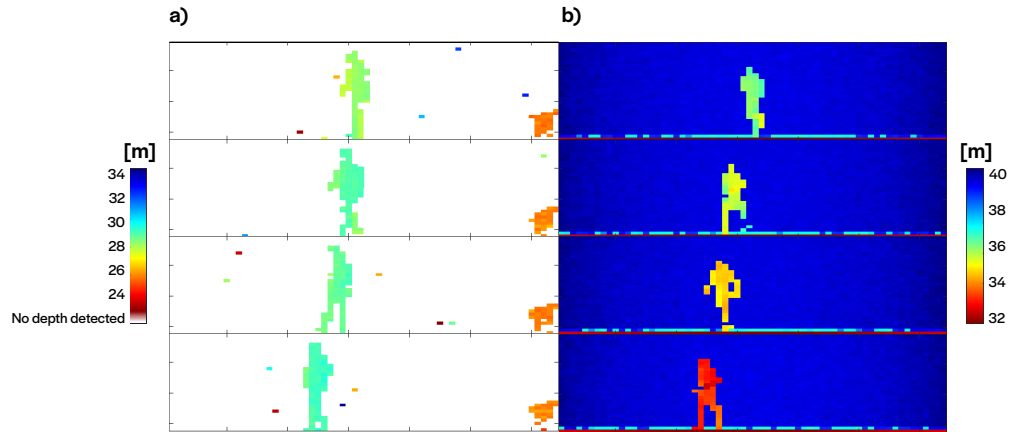


Fig. 7. Comparison of selected frames from a sequence of a person walking captured with a) real SPAD camera (50 FPS, mean SBR 0.17, mean SNR 11, and 716 average signal photons for the person in the first frame) and b) synthetic SPAD camera (model-generated data, mean SBR 0.19, mean SNR 2.82, and 42 average signal photons).

32 frames. Each block is created from 64 frames overlapped 50% with the previous block (e.g. block 1 consists of frames 1-64, and block 2 corresponding to frames 32-96), which are then resampled to 32 frames. This is done to guarantee that all activities, regardless of their different duration, fit within a single block captured at 50 FPS. Frames are processed at a maximum rate of 66 FPS, exceeding the acquisition speed, which demonstrates the scheme's potential for real-time operation. Figure 4(b) shows the confusion matrix of all activities captured with the real SPAD dToF and Table 1 shows results in terms of recall, precision and F-score. The overall weighted average accuracy is 89% and its simple average accuracy is 87%, which is comparable to the synthetic dataset. A small portion of the misclassifications come from poor segmentation. Another source of misclassification comes from the division of data in blocks, which can lead to sequences that include a transition between two actions, making the ground truth ambiguous. This ambiguity significantly impacts the apparent performance for some actions, most notably jumping, which only has 70.6% recall. The action gets confused with idle in 27.9% of cases, with the majority representing transitions between the two actions. Despite the elevated number of false positives, we note that jumping is nevertheless detected accurately in at least one of the blocks for every instance of the action (the action being typically spread over 2-3 blocks). A similar explanation applies for the confusion between standing up and crouching down with idle. Another common confusion comes from walking and running, which have subtle differences between them. We note that, while the synthetic dataset that the network is trained on features a range of walking and running speeds, the speed of motion is constant in any given sequence (i.e. there is no acceleration), whilst in the real dataset the speed varies throughout a sequence, which makes it harder to distinguish the two activities. This may explain the performance disparity between the real and synthetic datasets in the running and walking categories. The small level of false positives between crouching, jumping and standing up, are due to the similarities between parts of these actions. We note that the low sample size in some categories (especially jumping and standing up with approximately 60 samples each) means that the elements in the confusion matrix will be subject to a level of statistical variance. However, the combined accuracy figure of 89% gives a good indication of the general performance of the approach.

Visualization 2 shows a sequence captured with a SPAD dToF (at 50 FPS and a range >30 metres) of a person performing all activities considered in this paper. The person is correctly detected throughout the sequence with the activities also being correctly estimated.

[Visualization 3](#) shows another example captured with the same SPAD of two people crossing each other multiple times. The people are shown to be successfully tracked with bounding boxes, while their activities are correctly predicted.

5. Conclusions

We have presented a system for outdoor, human activity recognition based on SPAD dToF data. The development of the system involved the generation of a synthetic dataset with people performing a total of seven different actions, which was processed into realistic SPAD dToF frames with randomised signal and noise levels. The synthetic data was in turn used to develop a pair of networks: a CNN for person detection followed by an RNN for activity classification. Test results based on synthetic, as well as a real dataset, captured at >30 metres and at different SNR conditions (ranging from 2-10), indicate high levels of HAR accuracy, with values of 92% and 89%, respectively, demonstrating that the scheme successfully overcomes pixelation effects and depth noise from solar interference. The approach is designed to analyse continuous streams of depth data, and detect people and classify their activities, achieving a maximal processing rate of 66 FPS when running on a GPU.

Preliminary results indicate that the method can be extended to multiple people in the FoV, with people correctly tracked when crossing each other. Future work will focus on gathering and analysing more real SPAD dToF data involving multiple people. It will also consider concatenating intensity information to the depth input and evaluate its potential benefits in activity classification. Gaining further insight into the limitations of HAR (e.g., the smallest person size in terms of pixels that the network can still reliably recognise) is also of interest.

Whilst the scheme is currently demonstrated at a range of 30-40 metres, we expect that it can be readily extended to longer ranges by adopting a suitably powerful laser source (and adjusting the receiver optics to preserve the size of people in the FoV). In this study we used a laser with an optical peak power of 60 W, compared to the >1 kW used by typical automotive flash LIDAR [46].

The method could be particularly well suited for autonomous systems seeking to obtain high-level information on their environment with low latency. By understanding the activity of people in the FoV, systems are more likely to predict how situations may evolve and avoid potential accidents. The method would also be useful in security and surveillance applications where the identity of the person is preferred to be kept private.

Funding. Defence Science and Technology Laboratory (DSTLX1000147352, DSTLX1000147844).

Acknowledgements. The authors are grateful to STMicroelectronics for chip fabrication. Portions of this work were presented at the International Image Sensor Workshop in 2023 [47].

Disclosures. The authors declare no conflicts of interest.

Data availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

Supplemental document. See [Supplement 1](#) for supporting content.

References

1. S. Vishwakarma and A. Agrawal, "A survey on activity recognition and behavior understanding in video surveillance," *The Visual Computer* **29**(10), 983–1009 (2013).
2. X. Zhou, W. Liang, and K. I.-K. Wang, "Deep-learning-enhanced human activity recognition for internet of healthcare things," *IEEE Internet of Things Journal* **7**(7), 6429–6438 (2020).
3. L. Chen, N. Ma, and P. Wang, "Survey of pedestrian action recognition techniques for autonomous driving," *Tsinghua Sci. Technol.* **25**(4), 458–470 (2020).
4. A. Del Bimbo, R. Cucchiara, S. Sclarof, *et al.*, "Pattern Recognition," *ICPR International Workshops and Challenges: Virtual Event*, January 10–15, 2021, Proceedings, Part III, vol. 12663 (Springer Nature, 2021).
5. A. Gupta, A. Anpalagan, L. Guan, *et al.*, "Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues," *Array* **10**, 100057 (2021).

6. A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Machine Intell.* **23**(3), 257–267 (2001).
7. A. Yilmaz and M. Shah, "Actions sketch: A novel action representation," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1 (IEEE, 2005), pp. 984–989.
8. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1 (Ieee, 2005), pp. 886–893.
9. N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Computer Vision—ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7–13, 2006. Proceedings, Part II 9*, (Springer, 2006), pp. 428–441.
10. D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Computer vision and image understanding* **115**(2), 224–241 (2011).
11. S. Ji, W. Xu, M. Yang, *et al.*, "IEEE Trans. Pattern Anal. Mach. Intell.," *IEEE transactions on pattern analysis and machine intelligence* **35**(1), 221–231 (2013).
12. D. Tran, L. Bourdev, R. Fergus, *et al.*, "Learning spatiotemporal features with 3D convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, (2015), pp. 4489–4497.
13. X. Shi, Z. Chen, H. Wang, *et al.*, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems* **28**, 1 (2015).
14. S. K. Yadav, K. Tiwari, H. M. Pandey, *et al.*, "Skeleton-based human activity recognition using convLSTM and guided feature learning," *Soft Computing* pp. 1–14 (2022).
15. Z. Zhang, Z. Lv, C. Gan, *et al.*, "Human action recognition using convolutional LSTM and fully-connected LSTM with different attentions," *Neurocomputing* **410**, 304–316 (2020).
16. J. Tachella, Y. Altmann, and N. Mellado, "Real-time 3D reconstruction from single-photon LIDAR data using plug-and-play point cloud denoisers," *Nat. Commun.* **10**(1), 4984 (2019).
17. S. K. Yadav, K. Tiwari, H. M. Pandey, *et al.*, "Skeleton-based human activity recognition using convLSTM and guided feature learning," *Soft Computing* pp. 1–14 (2022).
18. C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2016), pp. 1933–1941.
19. S. Kamal and A. Jalal, "A hybrid feature extraction approach for human detection, tracking and activity recognition using depth sensors," *Arab J Sci Eng* **41**(3), 1043–1051 (2016).
20. M. K. Sain, R. H. Laskar, J. Singha, *et al.*, "Hybrid deep learning model-based human action recognition in indoor environment," *Robotica* **41**(12), 3788–3817 (2023).
21. A. Sánchez-Caballero, D. Fuentes-Jiménez, and C. Losada-Gutiérrez, "Real-time human action recognition using raw depth video-based recurrent neural networks," *Multimed Tools Appl* **82**(11), 16213–16235 (2023).
22. R. Horaud, M. Hansard, G. Evangelidis, *et al.*, "An overview of depth cameras and range scanners based on time-of-flight technologies," *Machine Vision and Applications* **27**(7), 1005–1020 (2016).
23. O. Kumagai, J. Ohmachi, M. Matsumura, *et al.*, "7.3 a 189×600 back-illuminated stacked SPAD direct time-of-flight depth sensor for automotive LIDAR systems," in *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 64 (2021), pp. 110–112.
24. J. Rapp, J. Tachella, and Y. Altmann, "Advances in single-photon LIDAR for autonomous vehicles: Working principles, challenges, and recent advances," *IEEE Signal Process. Mag.* **37**(4), 62–71 (2020).
25. J. Peng, Z. Xiong, H. Tan, *et al.*, "Boosting photon-efficient image reconstruction with a unified deep neural network," *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 1–18 (2022).
26. Y. Hong, Y. Li, and C. Dai, "Image-free target identification using a single-point single-photon LiDAR," *Opt. Express* **31**(19), 30390–30401 (2023).
27. R. K. Henderson, N. Johnston, and F. M. Della Rocca, "A 192 × 128 time correlated SPAD image sensor in 40-nm CMOS technology," *IEEE J. Solid-State Circuits* **54**(7), 1907–1916 (2019).
28. M. Laurenzis, "Single photon range, intensity and photon flux imaging with kilohertz frame rate and high dynamic range," *Opt. Express* **27**(26), 38391–38403 (2019).
29. S. Scholes, A. Ruget, G. Mora-Martín, *et al.*, "Dronesense: The identification, segmentation, and orientation detection of drones via neural networks," *IEEE Access* **10**, 38154–38164 (2022).
30. V. L. Ouster, Puck hi-res LiDAR sensor, (2016) [retrieved 21 March 2024], <https://velodynelidar.com/products/puck-hi-res/>.
31. Titan m1, (2022) [retrieved 21 March 2024], <https://www.neuvition.com/products/titan-m1.html>.
32. A. Ruget, M. Tyler, and G. Mora Martín, "Pixels2pose: Super-resolution time-of-flight imaging for 3d pose estimation," *Sci. Adv.* **8**(48), eade0123 (2022).
33. D. B. Lindell, M. O'Toole, and G. Wetzstein, "Single-photon 3d imaging with deep sensor fusion," *ACM Trans. Graph.* **37**(4), 1–12 (2018).
34. G. Mora-Martín, S. Scholes, and A. Ruget, "Video super-resolution for single-photon LIDAR," *Opt. Express* **31**(5), 7060–7072 (2023).
35. I. Gyongy, A. T. Erdogan, N. A. Dutton, *et al.*, "A direct time-of-flight image sensor with in-pixel surface detection and dynamic vision," *IEEE Journal of Selected Topics in Quantum Electronics* pp.1–12 (2023).
36. Games Epic, Unreal engine, (2019).

37. S. Shah, D. Dey, C. Lovett, *et al.*, “AirSim: High-fidelity visual and physical simulation for autonomous vehicles,” in *Field and Service Robotics*, M. Hutter and R. Siegwart, eds. (Springer International Publishing, 2018), pp. 621–635.
38. S. Scholes, G. Mora-Martín, and F. Zhu, “Fundamental limits to depth imaging with single-photon detector array sensors,” *Sci. Rep.* **13**(1), 176 (2023).
39. I. Gyongy, S. W. Hutchings, and A. Halimi, “High-speed 3D sensing via hybrid-mode imaging and guided upsampling,” *Optica* **7**(10), 1253–1260 (2020).
40. I. Gyongy, N. A. Dutton, and R. K. Henderson, “Direct time-of-flight single-photon imaging,” *IEEE Trans. Electron Devices* **69**(6), 2794–2805 (2022).
41. L. J. Koerner, “Models of direct time-of-flight sensor precision that enable optimal design and dynamic configuration,” *IEEE Trans. Instrum. Meas.* **70**, 1–9 (2021).
42. G. Mora-Martín, A. Turpin, and A. Ruget, “High-speed object detection with a single-photon time-of-flight image sensor,” *Opt. Express* **29**(21), 33184–33196 (2021).
43. O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” 2015.
44. F. Chollet, “Keras,” (2015).
45. D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, 2017.
46. M. E. Warren, “Automotive LIDAR technology,” in *2019 Symposium on VLSI Circuits*, (2019), pp. C254–C255.
47. G. Mora-Martín, J. Leach, R. K. Henderson, *et al.*, “High-speed, super-resolution 3D imaging using a SPAD dToF sensor,” in *Proc. Int. Image Sensor Workshop*, (2023).