# Spiking Neural Network Enhanced Hand Gesture Recognition Using Low-Cost Single-Photon Avalanche Diode Array

Zhenya Zang, Xingda Li, Xi Chen, David Day Uei Li

## Abstract

We present a compact spiking convolutional neural network (SCNN) and spiking multilayer perceptron (SMLP) to recognize ten types of gestures in dark and high-ambient light environment, using a 9.6 USD single-photon avalanche diode (SPAD) array. In our hand gesture recognition (HGR) system, photon intensity data from the sensor is leveraged to train and test the network. A vanilla convolutional neural network (CNN) is also implemented to compare the performance of the SCNN with the same network topologies and training strategies. The SCNN is trained from scratch instead of being converted from the CNN. The test of the networks is conducted in both dark and ambient light (AL)-corrupted environments. The result indicates that SCNN achieves comparable accuracy (90.8%) to CNN (92.9%) and exhibits much lower computational complexity with only 8 timesteps. The code and dataset are available at

<https://github.com/zzy666666zzy/TinyLiDAR_NET_SNN>.

## Introduction

Spiking neural network (SNN), known as a typical implementation of neuromorphic computing, is an emerging algorithm for a wide range of applications in vision [1][2][3][4], due to its high computational energy efficiency and accuracy. Hardware communities also leverage the high efficiency to design neuromorphic chips, such as Tianjic [5], TrueNorth [6], Liohi [7], etc. SNN usually processes data from event-based cameras that only record moving objectives and generate event-sequences, represented by spike streams. Differentiating with conventional artificial neural network (ANN) using multiply-accumulate (MAC) as basic operators, SNN includes accumulation and dot-product, consuming less hardware and performing higher computational efficiency. our hand gesture recognition. SNN has been adopted by dynamic vision sensors (DVS) for hand gesture recognition (HGR) [8][2].

Due to the small spatial resolution (4×4 or 8×8) of existing ST’s SPAD array products, it is arduous to recognize multiple stationary gestures.

In this work, while maintaining the accuracy, we aim to pursue a low-cost gesture recognition, in terms of product price and computing. The contribution of this work is three-fold:

1. We generate and opensource a dataset of gestures captured by a low-cost SPAD array. With the datasets, we train SCNN and SMLP from scratch instead of using conversion to guarantee the accuracy and computing.
2. We evaluate SCNN’s and SMLP’s accuracy of classification in inference phase with and without ambient light (AL).
3. We discuss the computational complexity of SCNN and SMLP and demonstrate the efficacy of hardware implementation.

## Related Work

### ST’s SPAD Using Machine Learning

The series of low-cost, portable sensor from ST become increasingly popular in the research of computer vision and pattern recognitions, with the synergy of machine learning. A sensor embedding a VSCEL module and a SPAD array, VL53L1X, was used to classify five types of objects by using the on-chip histogram method to capture reflected photon from the objectives [9]. The same sensor was employed to capture low-resolution depth images and generate high-resolution and skeleton images using DL [10]. Also, five VL53L1X were mounted on a small drone to realize obstacle avoiding and maze-solving [11]. ST also developed its integrated software tool to realize dynamic gesture recognition using time-of-flight technology [12]. The sensor was also successfully integrated with a costume RISC-V processor on a drone to measure its motion and distance from the ground, thereby assisting detecting human poses using DL [13].

### SPAD Sensors Emitting and Processing Spikes

A monolithic chip [14] integrating a SPAD array a SNN processor was designed to process spike trains encoded from 2-D intensity, 3-D depth, and dim-vision imaging. A monolithic chip [15] embedding an analog SNN processor was proposed to process raw photon events from the SPAD sensor on the same chip.

## Data Acquisition and Processing

### Sensor Configuration

### The SPAD sensor we are using is VL53L8CH, which is mounted on an ST’s evaluation board, NUCLEO-F401RE. The key feature of the SPAD sensor that we use is that it supports 15 fps for 8×8 spatial resolution. Due to the limited memory of the processor, the number of time bins of in-pixel histograms should be smaller than 18, with the time resolution (bin width in the histogram) of ca. 123.3 ps. The integration time is 5ms. Ranging frequency is 60 Hz. All the mentioned parameters were configured using the CPU on the evaluation board via its firmware. The reason we used intensity (photon counts) rather than depth images is because the pixel-wise depth data is reconstructed by ST’s fitting algorithm that slightly introduces errors to images. The compiled code of configuration will be transferred to the SPAD sensor through an I2C interface. As the sensor outputs photon counts, depth data, and histograms concurrently, we further configured the firmware making the sensor only output photon counts. A Python script receives the number of photo counts through an UART interface.

A group of images of different colors

Description automatically generated

Figure 1. Images in the first row are normalized photon counts images. Ones in the second row are encoded images using Poisson encoder.

### Neural Network Details

The training datasets include 5,100 images of 10 kinds of gestures, including 5,000 gesture images and 1,100 images without gestures. With the same ratio, the test datasets have 1,100 images, with 100 images for each gesture and 100 images for no-gesture condition. We implemented the networks using PyTorch. SpikingJelley [16] was imported to training the SCNN and SMLP. As spiking neurons generate Dirac delta-alike spikes that is non-differentiable during back-propagation, we used Sigmoid surrogate function to calculate surrogate gradient, where is a parameter controlling the slope of the function. We utilized Bicubic interpolation [17] from OpenCV to enhance the spatial resolution from 8×8 to 25×25 yet try to keep high fidelity. During the training phase, normalized images () in each batch were encoded as Poisson events with *T* (configure to 8 here) timesteps, using Poisson encoder[16] with time-step index

Encoded images are shown in Fig. Notably, the input to spiking networks is not the encoded images but spikes at each timestep. Architectures of SCNN and SMLP are depicted in Fig. Due to the simplicity of topologies, three networks were trained using Intel (R) Core i5 CPU @ 3.1 GHz.

We used early stopping strategy with 20 epoch patients to avoid over-fitting. Adam is the optimizer. The learning rate is 10­-3. Cross-entropy is the loss function.

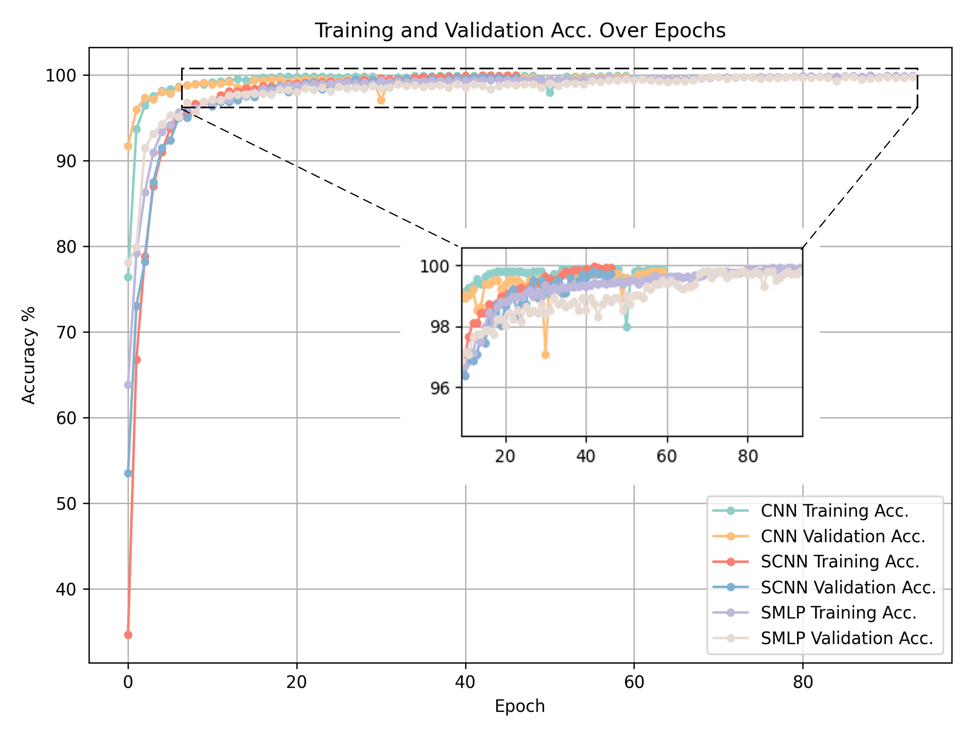


Figure . Training and validation accuracy over epochs. CNN, SCNN, and SMLP terminate training at 58, 48, and 93 epochs, respectively.

Two types of SNNs, SCNN and SMLP were constructed to investigate their differentiation in term of performance. In SCNN, batch normalization follows each convolutional layer to accelerate the training convergence. Integrate-and-Fire (LF) neuro was adopted as the activation function. Besides, SMLP includes two fully connected layer followed by drop-off modules to alleviate overfitting. Details of the architectures are depicted in Fig. We also designed a vanilla CNN with the same training strategies and network topology to compare the training performance versus the SNNs. As shown in Fig., the vanilla CNN shows faster convergence than SCNN and SMLP. And the final accuracy of three networks is nearly same.

## Evaluation on Test Datasets

We collected two sets of test datasets with and without AL to validate the robustness of our SNN. The AL is a 60-Watt LED bulb that illuminate the environment and is placed ca. 1 meter away from the sensor. When AL is applied on the sensor, background photons from AL randomly spread over all the time bin in histograms all pixels. We used confusion matrix to evaluate the accuracy of classification of three networks. As shown in Fig, all models testing clear datasets yield higher accuracy than the noisy dataset corrupted by the AL. In Fig. () SCNN and SMLP perform comparable accuracy to CNN. SCNN exhibits high accuracy yet slightly lower accuracy than CNN.

A group of blue squares

Description automatically generated

Figure 2. Confusion matrix of CNN, SCNN, and SMLP with and without AL, where accuracy in each figure indicates the performance of classification from models.

We also evaluate the computational workload regarding training and inference. As shown in Fig., SCNN and SMLP consume longer time in training and similar inference (1,100 images in test datasets) time to CNN. However, the sparsity of spiking network shows a great potential of hardware implementation.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Model Size (MB) | Parameters (MB) | Training time (s) | Inference time (ms) | #Operations |
| CNN | 0.34 | 0.042 |  | 0.043 | 2.14×105 |
| SCNN | 0.34 | 0.042 | 939.21 | 0.41 | 1.50×108 |
| SMLP | 1.02 | 0.12 | 376.22 | 0.064 | 2.12×106 |

## Conclusion

Reference

[1] S. Afshar, T. J. Hamilton, L. Davis, A. Van Schaik, and D. Delic, ‘Event-Based Processing of Single Photon Avalanche Diode Sensors’, *IEEE Sens. J.*, vol. 20, no. 14, pp. 7677–7691, 2020, doi: 10.1109/JSEN.2020.2979761.

[2] S. Singh, A. Sarma, S. Lu, A. Sengupta, V. Narayanan, and C. R. Das, ‘Gesture-SNN: Co-optimizing accuracy, latency and energy of SNNs for neuromorphic vision sensors’, in *2021 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, 2021, pp. 1–6. doi: 10.1109/ISLPED52811.2021.9502506.

[3] X. She and S. Mukhopadhyay, ‘Speed: Spiking neural network with event-driven unsupervised learning and near-real-time inference for event-based vision’, *IEEE Sens. J.*, vol. 21, no. 18, pp. 20578–20588, 2021.

[4] Y. Lin and E. Charbon, ‘Spiking Neural Networks for Active Time-Resolved SPAD Imaging’, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 8147–8156.

[5] J. Pei *et al.*, ‘Towards artificial general intelligence with hybrid Tianjic chip architecture’, *Nature*, vol. 572, no. 7767, pp. 106–111, 2019.

[6] F. Akopyan *et al.*, ‘Truenorth: Design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip’, *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 34, no. 10, pp. 1537–1557, 2015.

[7] M. Davies *et al.*, ‘Loihi: A neuromorphic manycore processor with on-chip learning’, *Ieee Micro*, vol. 38, no. 1, pp. 82–99, 2018.

[8] A. M. George, D. Banerjee, S. Dey, A. Mukherjee, and P. Balamurali, ‘A reservoir-based convolutional spiking neural network for gesture recognition from dvs input’, in *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2020, pp. 1–9.

[9] C. Callenberg, Z. Shi, F. Heide, and M. B. Hullin, ‘Low-cost SPAD sensing for non-line-of-sight tracking, material classification and depth imaging’, *ACM Trans. Graph. TOG*, vol. 40, no. 4, pp. 1–12, 2021.

[10] A. Ruget *et al.*, ‘Pixels2Pose: Super-resolution time-of-flight imaging for 3D pose estimation’, *Sci. Adv.*, vol. 8, no. 48, p. eade0123, 2022.

[11] S. Pikalov, E. Azaria, S. Sonnenberg, B. Ben-Moshe, and A. Azaria, ‘Vision-Less Sensing for Autonomous Micro-Drones’, *Sensors*, vol. 21, no. 16, p. 5293, 2021.

[12] STMicroelectronics, ‘Turnkey gesture recognition solution based on VL53L5CX, VL53L7CX and VL53L8CX multizone Time-of-Flight ranging sensors’. STMicroelectronics, 2023. [Online]. Available: https://www.st.com/en/embedded-software/stsw-img035.html#documentation

[13] D. Palossi *et al.*, ‘Fully onboard ai-powered human-drone pose estimation on ultralow-power autonomous flying nano-uavs’, *IEEE Internet Things J.*, vol. 9, no. 3, pp. 1913–1929, 2021.

[14] X. Yang *et al.*, ‘A Bio-Inspired Spiking Vision Chip Based on SPAD Imaging and Direct Spike Computing for Versatile Edge Vision’, *IEEE J. Solid-State Circuits*, 2023.

[15] M. S. A. Shawkat, M. M. Adnan, R. D. Febbo, J. J. Murray, and G. S. Rose, ‘A Single Chip SPAD Based Vision Sensing System With Integrated Memristive Spiking Neuromorphic Processing’, *IEEE Access*, vol. 11, pp. 19441–19457, 2023.

[16] W. Fang *et al.*, ‘SpikingJelly: An open-source machine learning infrastructure platform for spike-based intelligence’, *Sci. Adv.*, vol. 9, no. 40, p. eadi1480, 2023.

[17] ‘OpenCV: Geometric Image Transformations’. [Online]. Available: https://docs.opencv.org/4.x/da/d6e/tutorial\_py\_geometric\_transformations.html