3D Convolutional Spiking Neural Network for Human Action Recognition Using Modulating STDP With Global Error Feedback

1st Thoshara Nawarathne
Department of Electrical and Computer Engineering
University of Calgary
Calgary, Canada
thosharamalathinawar@ucalgary.ca

2nd Henry Leung
Department of Electrical and Computer Engineering
University of Calgary
Calgary, Canada
leungh@ucalgary.ca

2024 IEEE International Systems Conference (SysCon) | 979-8-3503-5880-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/SysCon61195.2024.10553446

***Abstract*—Video action recognition using 3D Convolutional Neural Networks (CNN) become an increasingly popular strategy in past years with the evolution of machine learning and computer vision. However, the higher memory and computation capacity requirement of these networks leads to the use of low-power memory-saving neural networks to perform video action recognition tasks efficiently. Spike-based information processing and computation of bio-inspired spiking convolutional neural networks perform an essential role when comes to energy efficient memory saving computation for video classification and action recognition tasks which allow on-chip real-time processing. This paper proposes a novel 3D Convolutional Spiking Neural Network (CSNN) architecture with modulating STDP supervised learning via global error feedback for human action recognition in video data. The proposed model includes two 3D convolutional layers, followed by two spiking neuron layers, modeled using Leaky Integrate and Fire (LIF) neurons for feature extraction from video data. Using the modulating STDP learning rule with global error feedback, this model can successfully recognize human actions from video data allowing online parallel computations. The proposed network experimented on two datasets: one 3D image dataset - synthesized 3D MNIST and one video dataset - UCF 101 human action recognition dataset and achieved 71.6% and 63.7% recognition accuracy.**

***Index Terms*—Asynchronous video, 3D convolutional spiking neural networks, action recognition, UCF 101 dataset**

I. Introduction

The advancement of artificial intelligence(AI) and computer vision causes the exponential growth of asynchronous video-based applications over the past years. Object and activity recognition in videos plays a crucial role in asynchronous video processing and related applications. Currently, asynchronous video-related applications influence the quality of life by a significant degree. Since, the Covid-19 pandemic remote working becomes popular which laid a strong infrastructure for asynchronous video communication [1] to keep the workers connected and involved irrespective of their nationalities, time region. Asynchronous Video Interviews (AVI) [2] are considered one of the most efficient types of job interviews. Automatic recognition of personality characteristics [3] in AVI is an open research area to analyze human-machine interaction and personality assessment which are key factors in hiring. Asynchronous videos are also popular in social media networks because most consumer applications such as Youtube, Tiktok, and Snapchat are solely dependent on them. Hence, asynchronous video analysis becoming an important research area in computer vision and AI. The state-of-art research area for asynchronous video analysis is artificial neural networks [3], [4]. However, there is a bottleneck for this approach since these algorithms require a high computation and memory capacity. On the other hand, Spiking Neural Networks (SNN), the third generation of neural networks has been proven theoretically that their spiking information transmission directs to high computational performance with low cost and low memory requirements [5]. However, SNN-based applications are currently rare in the picture because the training of the SNN models is considerably more challenging than the ANN due to the non-differentiable nature of the spike activation. Because of this non-differentiability, the back-propagation [6] utilized in supervised learning cannot be adapted to the SNN, and there are no specific globally accepted training algorithms to apply for the SNN. As a result, there are a limited number of studies that use SNN networks for the video analysis tasks such as object/action recognition and classification, and their performances need to be greatly improved.

This paper focuses on a novel video action detection architecture utilizing a 3D Convolutional Spiking Neural Network (CSNN) with modulating STDP [7] learning rule via global error feedback [8]. This network has access to both spatial and temporal features in video data and permits multi-layer supervised learning on-chip online fashion. The proposed architecture was compared with the conventional 3D CNN architecture and two spiking network models: unsupervised 3D CSNN with SVM classifier [9] and STS - ResNet [10] to evaluate the performance of the network. Two datasets are used here: A synthesized 3D MNIST dataset and UCF 101 dataset [11] with selected classes. The former was used for image classification while the second one was used for human action recognition in video data. Our experiment show that the proposed algorithm achieves relatively higher performance in terms of recognition accuracy. However, the proposed work's

spiking nature will lead to higher computational capacity with low memory requirements. The learning algorithm facilitates real-time on-chip learning with global error feedback which allows parallelization to lead to fast computation and suitable to apply on neuromorphic devices which are the veins of edge computing. The remainder of the paper is presented as follows: section 2 briefly explains related studies in the literature, and section 3 presents the proposed network architecture of this study. Section 4 presents the experiment results and section 5 concludes the study.

II. Related Works

There are a reasonable number of studies focusing on object/action recognition in video data with different algorithms. Among them, the CNN and convolutional [12]–[17] Long Short-Term Memory (LSTM) networks [18], [19] are the most common methods in the literature. 2D CNN which is a widely used method for image classification [13], [15], [17] is also used for object/action recognition tasks. However, the recognition performance quite lags behind the 3D CNN architecture [13]. In recent years, the 3D CNN model is widely used for video data because of its capability in analyzing spatial-temporal features effectively [20]. Furthermore, it has the state-of-art action recognition performance [19], [20]. Even though these methods achieve higher levels of recognition accuracy they are computationally intensive and require an excess of memory and time for computation. The brain-inspired SNNs are considered the green neural networks that require low power and memory capability with effective spiking nature [21]. Recent studies more focused on building video classifiers/action recognition algorithms using the novel SNN architecture.

Convolutional SNN for video analysis, classification, and object/action recognition widely has been considere for the emergence of event-based vision cameras. Recently, 3D CSNN networks come into the picture for video action recognition type tasks [9], [10]. Most of these studies focus on learning using Spiking Timing Dependent Plasticity (STDP) [22], Back Propagation Through Time (BPTT), surrogate gradient with backpropagation, or use of ANN-SNN conversion for training because of the non-differentiable nature of the spike activation. The STDP learning rule is more biologically plausible and suitable for training shallow unsupervised networks, but the deep supervised learning objectives cannot be fulfilled with the STDP learning rule. The Reference [9] considers unsupervised learning using STDP followed by a Support Vector Machine (SVM) classifier for training 2D and 3D CSNN. The surrogate gradient networks use gradient approximation for the spiking function for the error feedback. In [23], surrogate gradient-based direct training is used for event-based spatiotemporal dynamic data recognition. However, these approximations deviate from the biological neuron perspective. Some studies consider training the network using ANN and convert the weights to a SNN model [24], [25]. While the performance of these types of networks are improved, they omit the temporal coding features of SNN topology. BPTT is widely used for training recurrent neural networks. The same approach can be used to train SNN by unrolling the network through each time step and getting the gradient spatially using gradient approximation to obtain the overall backpropagation at the end. The classification/recognition accuracy of these networks is satisfactory, with a trafeoff of enhance memory requirement cause by, unrolling the network. Reeference [10] focuses on spatial temporal feature extraction using convolutional residual network (STS Resnet) trained by hybrid BPTT and ANN-SNN conversion algorithm. A supervised learning algorithm is proposed in [7] to train multilayer SNN by employing spatial error backpropagation with the local STDP learning rule allowing online on-chip learning. Most of the works considered above are trained by propagating error layer by layer. In [8] uses global random feedback with direct error propagation from output layers is applied to hidden layers for effective training in multilayer SNN.

This work considers building a 3D CSNN architecture for video action recognition using the novel supervised learning rule that integrates modulating STDP [7] with and modifying it using the global error feedback.

III. Methodology

We propose a 3D Convolutional Spiking Neural Network (3D CSNN) architecture as depicted in Figure 1. It is equipped with a pre-processing section of video data: frame extraction and spike encoding, a spiking feature extractor, and a classifier.

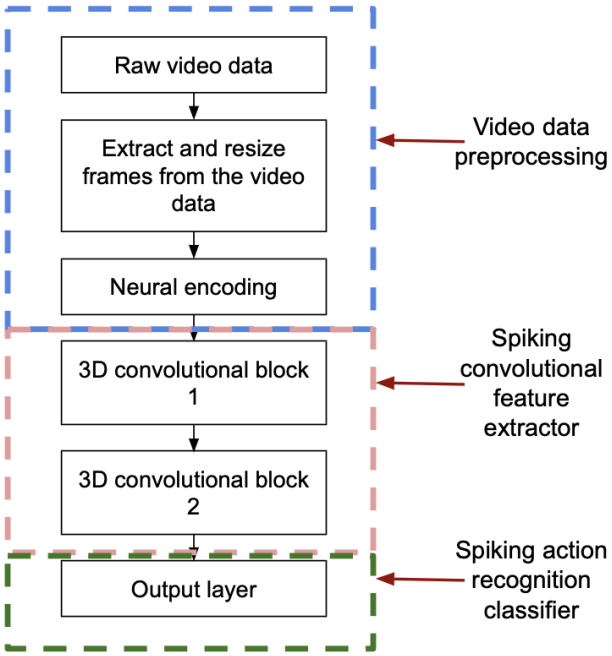

Fig. 1. Overall schematic of the proposed network.

A. Video data preprocessing

The video data is converted into sequences of frames before applying it to the network. Frames are series of image data

of a video; when there is an action in the video the variation of pixel intensity can be observed from one frame to another frame. In this study, an equal number of frames was extracted for each video clip and they were then resized. Since the information transfer and processing in SNN is achieved via spikes/train of spikes, there is a need to convert these real-valued data to spikes before feeding them into the proposed architecture. Different encoding methods can be applied to obtain the spike trains depending on the data type. However, Poisson-distributed spike encoding is employed here because of its simplicity. For each frame, this spike encoding method was applied to generate spike trains where the probability of spike generation is proportional to the intensity level of the given pixel. Then these encoded spike trains were then fed to the 3D CSNN model.

B. Neuron Model

The proposed network consists of 2 spiking convolutional blocks, each block with a 3D convolutional layer followed by a spiking neuron layer. The basic building block of the spiking neuron layer is the neuron model. The Leaky-Integrate-and-Fire (LIF) model [26] is used here. This LIF model is much simple and can closely mimic the biological neuron. Discrete-time neuron dynamics of the LIF model can be expressed by

$$v_j(t) = v_j(t-1)e^{-\delta t/\tau} + \sum_i w_{ij} H(v_i(t) - \theta) \quad (1)$$

$$v_j(t^+) \rightarrow v_r \quad (2)$$

$v_j(t)$ corresponds to the leaky membrane potential voltage of j^{th} neuron. The sigma function accumulates the spike inputs to the j^{th} neuron from the i^{th} neuron. Here H represents the heavy side step function. At each time step $v_j(t)$ increases or decreases by acquiring input spikes multiplying by w_{ij} which is the weight or the learnable parameter of the network called synaptic efficacy, if the $v_j(t)$ exceeds the threshold voltage θ, an spike emits with magnitude 1. After emitting spikes, the voltage was set to the resting potential v_r. as in Eq.(2). If there are no input spikes to the j^{th} neuron the membrane potential decays exponentially with the time constant τ at discrete time steps δt. For this analysis θ was set to $10mV$, τ = $20mS$ with δt = $1mS$.

C. Spiking Feature Extractor and Classifier

The proposed 3D CSNN architecture can be used to analyze the positional variations of action in video data. In the 3D CSNN layer, the kernel can move in three directions (x,y,z). For the proposed architecture x and y are referred to the width and height which utilized for extracting spatial features while the z-axis represents the temporal depth: the number of frames here, which is used for temporal feature extraction. For this study for both convolutional layers equipped the 64 filters a kernel size of $3 * 3 * 3$ and a stride of 1 for each direction. The operation from the input layer to the end of the 1^{st} convolutional block is depicted in Figure 2.

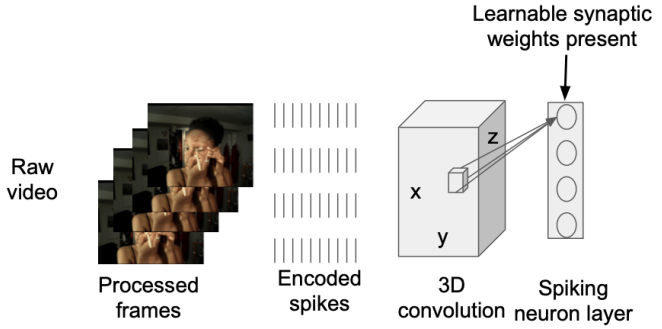

Fig. 2. Spiking convolutional operation.

As shown in the figure the generated spike trains convolve with the 3D convolutional filters, and the features are stored in the spiking neuron layer(act as synapses) as the synaptic weight. This layer acts as the input layer for the second convolutional block. This block has a similar function as in the previous block. The output from this block is connected to the final action recognition layer with spiking neurons. To achieve classification supervised training rule is applied with modulating STDP algorithm.

D. Training Algorithm

We propose a new learning algorithm for SNN based on our previous work [7]. The learning rule in [7] assigns temporal error using the local STDP learning rule which is unsupervised. To allow supervised learning, the spatial error was considered by implementing backpropagation at a given time step, hence, this rule is called modulated STDP. The local temporal error assignment allows updating the weights in an online manner. Therefore, the proposed network can train in real-time and is capable to build in neuromorphic hardware devices. However, instead of propagating error layer by layer for multilayer SNN as in [7] a global random feedback [8] is proposed here to allow direct strengthening of synaptic weights in all layers as shown in Figure 3.

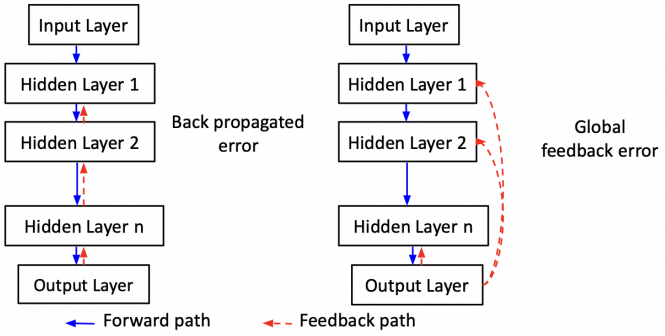

Fig. 3. Layer by layer error propagation vs global feedback error propagation.

The training of this network can be summarized in Eqs. (3)-(6). To update the synaptic weights of the network each neuron in the model keep an eligibility trace (λ) which is given by

$$\lambda_i(t) = \begin{cases} 1 & if v_i(t) > \theta \\ \lambda_i(t-1).e^{-\delta t/\tau_\lambda} & Otherwise \end{cases} \quad (3)$$

Where the τ_λ is the time constant which determnines the decay of the eligibility trace. Considering local pre (i) and post (j) synaptic neurons the synaptic weights are updated using the STDP learning rule as

$$\Delta W_{ij}(t) = \begin{cases} +\lambda_i(t) & if\ v_i(t) > \theta \\ -\lambda_j(t) & if\ v_j(t) > \theta \end{cases} \quad (4)$$

If there is a difference between the target and actual output at the output layer, a modulated signal $(M_{out}(t))$ will be created as shown in Eq.(5) it needs which need to forward back to the hidden layers to adjust synaptic weights for supervised training.

$$M_{out}(t) = target\ spikes - actual\ spikes \quad (5)$$

This backward pass of error signal is performed parallelly from the output layer to all the hidden layers by assigning random weights for the feedback paths. Let $M_i(t)$ be the globally feedback error signal from the output layer to the i^{th} neuron of the considered hidden layer, the synaptic weight update can be represented

$$\Delta W_{ij}(t) = \eta \lambda_i(t) M_j(t) \quad (6)$$

Where η is the learning rate. At a given time step thenetwork can update its synaptic weights in real-time allowing it to fabricate on the neuromorphic hardware.

E. Performace Comparison

The proposed architecture was compared by implementing the non-spiking 3D CNN model and the spiking 3D CNN model trained using STDP learning rule with SVM classifier developed in [9]. Initially synthesized 3D MNIST dataset was utilized as the toy dataset to evaluate the performance of the network. Then, a video dataset was used to analyze the model. For the video action recognition scenario, feature extraction of the conventional CNN is equipped with the frame generation and resizing as preprocessing. The spiking network with SVM classifier follows every preprocessing step mentioned previously(section 3.1). The conventional CNN model uses two 3D Convolutional layers follows by two max-pooling layers, while the STDP-based CSNN model the same network architecture with unsupervised learning and SVM classification. Moreover, the STS feature extractor [10] is used to compare the performance of the proposed architecture for action recognition in video data.

IV. Results

The proposed network experimented on two datasets: the synthesized 3D MNIST dataset and UCF 101 - Human Action Recognition Dataset [11]. For comparison, a conventional 3D CNN network, a spiking 3D CNN network with the unsupervised STDP learning rule and SVM classifier, and the STS Res-net were considered here. The synthesized 3D MNIST dataset includes two categories training and testing with a 6:1 ratio while the UCF 101 dataset is split into three parts for training, validation, and testing with 80:10:10 ratios. The proposed network was experimented for 5 times each with 50 epochs using different sets of training, validation, and testing data.

A. Synthesized 3D MNIST Dataset - Classification

This dataset was created by replicating the dimensions of the original 2D MNIST handwritten digits dataset: 60,000 training samples and 10,000 testing samples categorized into 10 classes. Each synthesized image sample has a dimension of $28 * 28 * 28$. These data samples are encoded into spike trains using the Poisson-distribute spike encoder and fed to the 3D CSNN architecture. The learning rate of the network was set to 0.01, and the cross entropy loss function was selected as the loss function with the Adam optimizer. The training and validation curves of this experiment are shown in Figure 4.

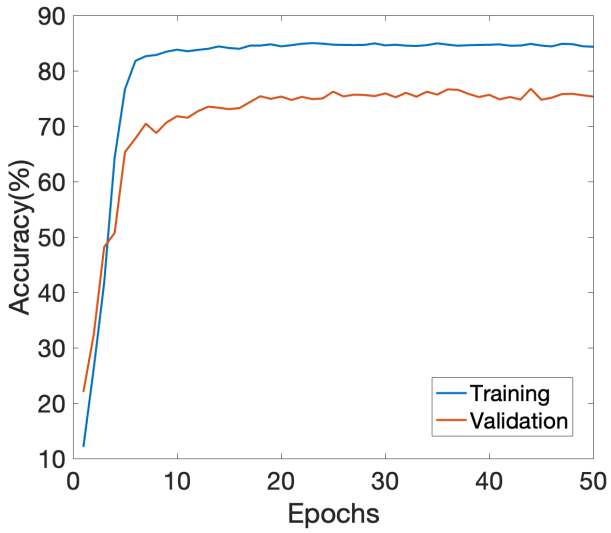

Fig. 4. Training and validation accuracy curves for the synthesized MNIST dataset.

The training and testing accuracies for the proposed network, the conventional 3D CNN network, and the 3D CSNN network with the unsupervised STDP learning rule with the SVM classifier are tabulated in Table 1. Accordingly, even though the conventional 3D CNN model has the best performance, the proposed network performs well compared to the other spiking networks for this classification task.

TABLE I
Classification Accuracy for Synthesized 3D MNIST Dataset

Network Architecture	Training Accuracy	Testing Accuracy
Proposed	84.2	71.6
Conventional 3D CNN	99.2	77.1
STDP rule with SVM	82.1	67.4

B. UCF 101 Dataset - Human Action Recognition

The UCF 101 dataset contains realistic human action videos from youtube categorized under 101 classes. This analysis was performed with ten randomly selected classes from the dataset, each with 100 video clips. It was observed that the video clips have different time lengths: frames and a minimum of 28 frames. Therefore, from each sample video 20 frames were extracted to feed to the proposed network. For each frame, the Poisson distributed spike encoding was applied to obtain the spike trains. The proposed network was trained by setting the learning rate η to 0.01 with cross-entropy loss and the Adam optimizer. The training and validation curves for the proposed network are depicted in Figure 5. The action recognition accuracy of this network with other implemented networks is tabulated in Table 2. While the proposed network is outperformed by the conventional 3D CNN network. It has the best performance among the spiking neural networks in terms of action recognition accuracy.

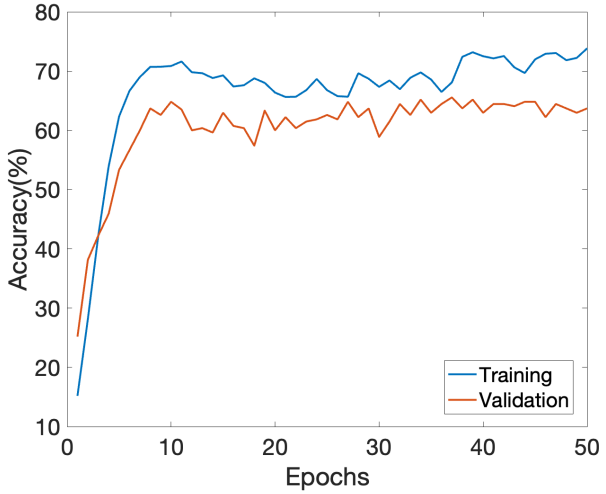

Fig. 5. Training and validation accuracy curves for the UCF 101 dataset with selected classes.

TABLE II
CLASSIFICATION ACCURACY FOR UCF 101 DATASET WITH SELECTED CLASSES

Network Architecture	Training Accuracy	Testing Accuracy
Proposed	73.8	63.7
Conventional 3D CNN	79.2	71.1
STDP rule with SVM	69.1	57.9
STS		42.4

V. CONCLUSIONS

Low-powered and real-time action recognition in asynchronous videos has several advantages. This work proposes a 3D convolutional SNN learned with a novel modulating STDP rule via global error feedback for action recognition in video data. The spiking nature of the network leads to low-power computing, while the learning algorithm has the ability to train on-chip-online manner. The proposed architecture was assessed with the conventional 3D CNN architecture, a spiking CNN architecture with the biologically plausible STDP learning rule with the SVM classifier, and the STS feature extractor. While the conventional 3D CNN deep learning architecture outperforms all SNN apporaches in terms of accuracy, the proposed method has the best classification/recognition accuracy compared to the other spiking network architectures considered in this study. It is noted that hyperparameters were selected manually here. The performance of the network can be improved with automated hyperparameter tuning and varying the depth with more convolutional blocks.

REFERENCES

[1] Lowenthal, Patrick R., Jered Borup, Richard Edward West and Leanna M. Archambault. “Thinking beyond Zoom: Using Asynchronous Video to Maintain Connection and Engagement during the COVID-19 Pandemic.” The Journal of Technology and Teacher Education 28 (2020): 383-391.

[2] Patrick D. Dunlop, Djurre Holtrop, and Serena Wee. 2022. ”How asynchronous video interviews are used in practice: A study of an Australian-based AVI vendor.” International Journal on Selection and Assigment 30, 3 (Jan 2022), 448–455.

[3] H. -Y. Suen, K. -E. Hung and C. -L. Lin, ”TensorFlow-Based Automatic Personality Recognition Used in Asynchronous Video Interviews,” in IEEE Access, vol. 7, pp. 61018-61023, 2019, doi: 10.1109/ACCESS.2019.2902863.

[4] L. Hemamou, G. Felhi, J. -C. Martin and C. Clavel, ”Slices of Attention in Asynchronous Video Job Interviews,” 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), Cambridge, UK, 2019, pp. 1-7, doi: 10.1109/ACII.2019.8925439.

[5] SW. Maass. 1997. Networks of spiking neurons: The third generation of neural network models. Neural Networks 10, 9 (1997), 1659–1671.

[6] Barry J. Wythoff. 1993. Backpropagation neural networks: A tutorial. Chemometrics and Intelligent Laboratory Systems 18, 2 (1993), 115–155.

[7] D. G. Peterson, T. Nawarathne, and H. Leung. 2023. Modulating STDP With Back-Propagated Error Signals to Train SNNs for Audio Classification. IEEE Transactions on Emerging Topics in Computational Intelligence 7 (2023), 89–101.

[8] Zhao D, Zeng Y, Zhang T, Shi M, and Zhao F. 2020. GLSNN: A Multi-Layer Spiking Neural Network Based on Global Feedback Alignment and Local STDP Plasticity. Front Comput Neuroscience (2020).

[9] M. El-Assal, P. Tirilly, and I. M. Bilasco. 2022. 2D versus 3D Convolutional Spiking Neural Networks Trained with Unsupervised STDP for Human Action Recognition. 2022 International Joint Conference on Neu- ral Networks (IJCNN) (2022), 1–8.

[10] Ali Samadzadeh, Fatemeh Sadat Tabatabaei Far, Ali Javadi, Ahmad Nickabadi, and Morteza Haghir Chehreghani. 2022. Convolutional Spiking Neural Networks for Spatio-Temporal Feature Extraction. (2022).

[11] Soomro Khurram, Zamir Amir, and Shah Mubarak. 2012. A Dataset of 101 Human Actions Classes From Videos in The Wild. (2012).

[12] S. Ji, W. Xu, M. Yang and K. Yu, ”3D Convolutional Neural Networks for Human Action Recognition,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 221-231, Jan. 2013, doi: 10.1109/TPAMI.2012.59.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems (2012).

[14] Sykora P., Kamencay P., Hudec R., and Vrskova R. 2022. Human Activity Classification Using the 3DCNN Architecture. Applied Sciences 12, 2 (2022).

[15] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations (ICLR) (2015).

[16] Stan W. Smith. 2021. S. N. Boualia and N. E. B. Amara. In 3D CNN for Human Action Recognition, Vol. 18. Signals & Devices (SSD), Monastir, Tunisia, 276–282.

[17] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. International Conference on Computer Vision and Pattern Recognition (CVPR), 2015 (2015).
[18] Orozco Carlos, Xamena Eduardo, Buemi Maria, and Berlles Julio. 2020. Human Action Recognition in Videos using a Robust CNN LSTM Approach. Ciencia y Tecnología (2020), 21–34.
[19] Mahshid Majd and Reza Safabakhsh. 2020. Correlational Convolutional LSTM for human action recognition. Neurocomputing 396 (2020), 224–229.
[20] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features With 3D Convolutional Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).
[21] Yongqiang Cao, Yang Chen, and Deepak Khosla. 2015. Spiking deep convolutional neural networks for energy efficient object recognition. International Journal of Computer Vision 113 (2015), 54–66.
[22] Caporale Natalia and Dan Yang. 2008. Spike Timing–Dependent Plasticity: A Hebbian Learning Rule. Annual Review of Neuroscience 31, 1 (2008), 25–46.
[23] G.C. Qiao, N. Ning, Y. Zuo, S.G. Hu, Q. Yu, and Y. Liu. 2021. Direct training of hardware-friendly weight binarized spiking neural network with surrogate gradient learning towards spatio-temporal event-based dynamic data recognition. Neurocomputing 457 (2021), 203–213.
[24] Peter U Diehl, Daniel Neil, Jonathan Binas, Matthew Cook, Shih-Chii Liu, and Michael Pfeiffer. 2015. Fast-classifying, high- accuracy spiking deep networks through weight and threshold balancing. International Joint Conference on Neural Networks (IJCNN) (2015), 1–8.
[25] Evangelos Stromatias, Miguel Soto, Teresa Serrano-Gotarredona, and Bernabe Linares-Barranco. 2017. An event- driven classifier for spiking neural networks fed with synthetic or dynamic vision sensor data. Frontiers in neuroscience (2017).
[26] W. Gerstner, W. M. Kistler, R. Naud, and L. Paninski. 2014. From Single Neurons to Networks and Models of Cognition. International Conference on ComputerCambridge University Press (2014).