n-HAR: A Neuromorphic Event-Based Human Activity Recognition System Using Memory Surfaces

Bibrat Ranjan Pradhan, Yeshwanth Bethi, Sathyaprakash Narayanan, Anirban Chakraborty, Chetan Singh Thakur
Indian Institute of Science, Bangalore
Karnataka, India
{bibratp, yeshwanthb, sathyaprakas, anirban, csthakur}@iisc.ac.in

***Abstract*—In recent years, a new generation of low-power, neuromorphic, event-based vision sensors has been gaining popularity for their very low latency and data sparsity. Though the conventional frame-based cameras have advanced in a lot of ways, they suffer from data redundancy and temporal latency. The bio-inspired artificial retinas eliminate the data redundancy by capturing only the change in illumination at each pixel and asynchronously communicating in binary spikes. In this work, we propose a system to achieve the task of human activity recognition based on the event-based camera data. We show that such tasks, which generally need high frame rate sensors for accurate predictions, can be achieved by adapting existing computer vision techniques to the spiking domain. We used event memory surfaces to make the sparse event data compatible with deep convolutional neural networks (CNNs). We leverage upon the recent advances in deep convolutional networks based video analysis and adapt such frameworks onto the neuromorphic domain. We also provide the community with a new dataset consisting of five categories of human activities captured in real world without any simulations. We achieved an accuracy of 94.3% using event memory surfaces on our activity recognition dataset.**

I. Introduction

Neuromorphic event-based camera systems are bio-inspired vision sensors that output spikes representing the pixel-level illumination changes instead of standard intensity frames. A major drawback of frame-based camera systems is that they sample the information at fixed intervals, without taking into account the dynamics of the scene. This might lead to both acquiring redundant data or missing important data between the fixed intervals. The information about underlying scene dynamics, which is useful for tasks like activity recognition, can be lost when frame-based camera systems are used. The event-based systems offer significant advantage over standard cameras in terms of high dynamic range, no motion blur, and latency in the order of microseconds. Event-based imaging systems are, therefore, very good at capturing the dynamic content of a scene. Because of these advantages, there have been many recent developments in computer vision algorithms to solve problems of optical flow estimation [1], gesture recognition [2], unsupervised feature extraction and learning [3] [4], motion analysis [5], and tracking [6] in event domain. Event-based camera systems are also becoming popular in areas that need high frame rates. In several applications, vision-based human activity recognition tasks require high frame rate video input to ensure minimum motion blur. Also, the data generated from frame-based camera systems is redundant when there is no motion or slow motion occurring in the scene. This makes event-based cameras a better fit for the task, as they avoid data redundancy by recording only the *changes in illumination* rather than the illumination measure of the scene. Moreover, by precisely timing the changes in each pixel, event-based cameras inherently encode the motion information of the scene, and are useful in extracting optical flow and other motion-based features easily. Event-based data can also act as a good substitute for computationally expensive optical flow features, which might aid in realtime activity recognition. Unlike frame-based cameras, event-driven sensors provide data in the Address Event Representation (AER) format [7]. This is significantly different from the way the information is encoded in RGB frames and hence the existing computer vision algorithms cannot be directly applied to analyze this data. The conventional computer vision algorithms have to be adapted to the event domain data, which is sparse and asynchronous. In this work, we use memory surfaces of the event data to adapt deep neural network models to achieve the task of human activity recognition in neuromorphic data.

II. Methodology

A. Neuromorphic Sensors

In this work we collected the dataset (described in Section III.A) with a camera that belongs to a novel class of imaging devices known as silicon retinas. They are a neuromorphic approach to visual sensory transduction and seek to replicate the robustness, efficiency, and low power consumption of biological vision systems. There are two popular camera systems that are implemented on this model, namely, Asynchronous Time-based Image Sensor (ATIS) [8] and Dynamic Vision Sensor (DVS) [9]. The conventional cameras have a fixed-length and global exposure time by which all the pixels would be sampled at a regular time interval, resulting in copious amounts of redundant data. To combat this, each pixel of the ATIS camera has an independent and asynchronous pipeline, by which it responds to the environment only when

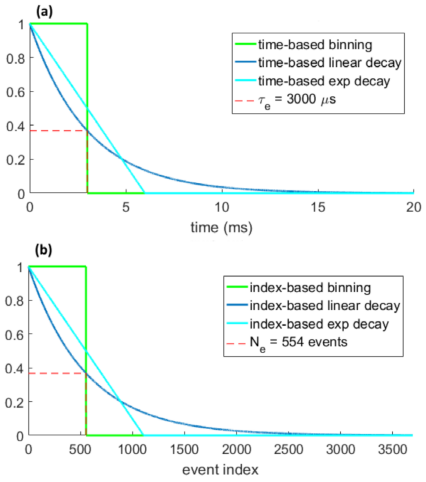

Fig. 1. Plots of the six methods for generating time and index surfaces. Panel (a) Shows the three time-based kernels over time. Panel (b) shows the value of the event-based kernel as a function of event index. Figure courtesy: Afshar *et al.* [11].

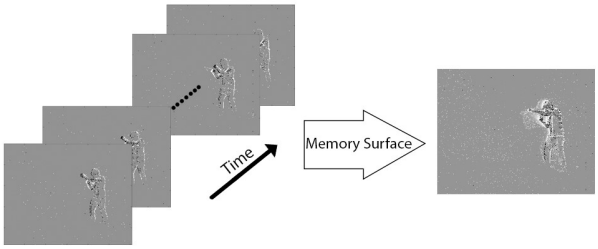

Fig. 2. The sequence on the left shows a series of events captured at each pixel along time and the frame on the right shows the time-based event memory surface calculated from events.

a change occurs in the log intensity scale of illumination of that corresponding pixel. This independent pixel setup of the camera allows it to have individual exposure time for each pixel, hence reprimanding the need for global exposure time [10]. In addition, the asynchronous nature permits a frame-rate-independent output with high temporal resolution and dynamic range, by which it eliminates motion blurs.

B. Event Data Quantization

As mentioned earlier, event-based sensors generate data that is sparse and asynchronous in the AER format. However, almost all current computer vision algorithms heavily rely on densely quantized frames. To adapt these algorithms to the event domain, the asynchronous event data has to be quantized into data structures compatible with the current computer vision techniques. Among such quantization methods, two such prominent techniques are:

Time Quantization: The easiest way to convert the event data into a series of two-dimensional (2D) frames is by sampling all the spikes at each pixel at regular time intervals and binning all the spikes that occurred between the time intervals. This way of quantization would produce frames similar to the normal frames generated by generic camera systems. But, quantizing the events in this way will sacrifice the important advantage of having information about the dynamics of the scene to a finer and accurate temporal resolution.

Memory Surfaces : The disadvantage of time-based quantization can be overcome by using time/memory surfaces. Lagorce *et al.* [12] introduced using layers of time-decaying event surfaces and feature-based clustering for hierarchical learning. Afshar *et al.* [11] proposed a variation of time surfaces, called Memory surfaces, using three different types of time-based kernels (Binning, Linear, and Exponential) that can be used across both time and event index. Memory surfaces capture the information of the time at which the events occurred in between the sampling intervals (see Fig. 1). This eliminates the disadvantage of the time quantization technique. Event memory surfaces were generated for the dataset collected and used as input data structures to the deep learning models in our experiment. The estimated memory surface for one such example event sequence can be seen in Fig. 2.

C. Human Activity Recognition

The recent promising performances achieved by the state-of-the-art deep learning models in the field of visual activity analysis motivated us to explore how similar frameworks can be built for the task of human activity recognition for event data. The analysis of activity is about understanding the motion pattern and in both the event-based and frame-based data activity recognition tasks, the objective is to decode the most relevant motion embedded in the scene and employ machine learning to perform the analytics. The current deep learning models are good at this analysis with frame-based data. In this work, we want to test the hypothesis that models suitable for event-based data can evolve from the models used for classical activity analysis on RGB videos. This motivates us to treat the problem of event data analysis as a domain adaptation problem, and try to adapt deep learning for conventional video analytics to the neuromorphic data domain.

Two-stream architecture [13] is a popular framework where two independent 2D convolutional neural networks (CNNs) pre-trained on ImageNet dataset [14] are used for activity recognition. One network is trained on one RGB frame per video-clip, and the second is trained on dense optical flow maps. As an extension of the above model, a long-term recurrent convolutional network (LRCN) [15] model was used where the architecture consists of time-distributed 2D CNN models, followed by the use of time-pooling layers to capture the temporal information. Several variants of the Two Stream architecture [16][13] currently hold the state of the art performance on popular datasets, e.g., UCF-101[17](98%) and HMDB-51[18](80.2%). Among other architectures, Conv3D [19] and the time-distributed version of Conv3D with subsequent pooling of these features [20] also capture rich temporal correlation of spatial features.

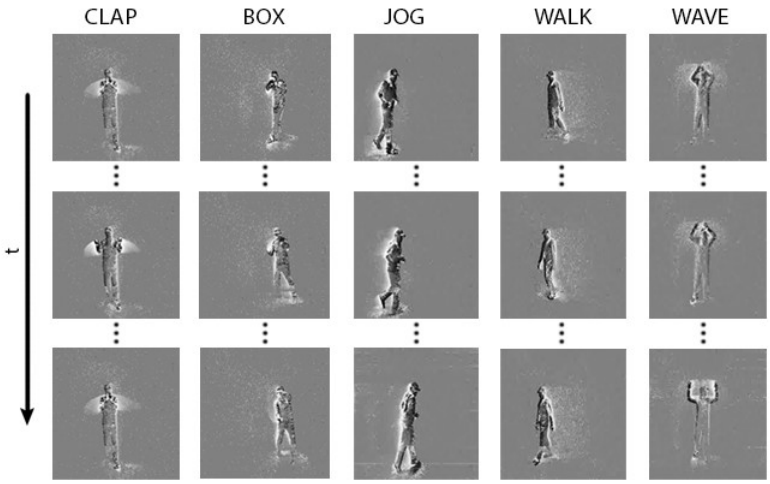

Fig. 3. Samples from each activity category in the dataset converted to event memory surfaces.

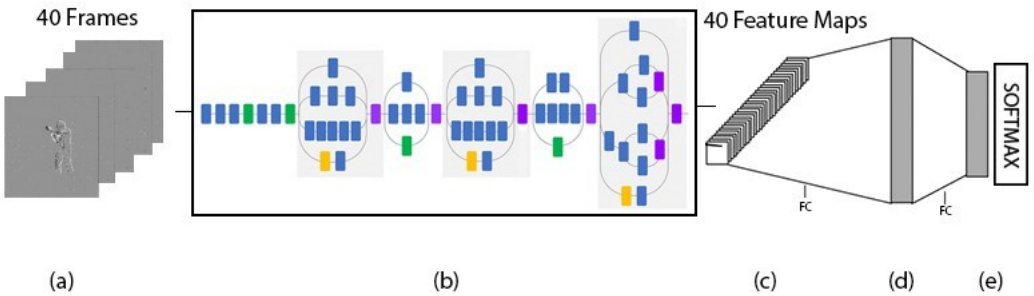

Fig. 4. n-HAR Architecture: (a): 40 Input memory surface frames to the model, (b): Inception Layers, (c): 40 Feature representations from 40 input frames, (d): Flatten and Fully connected with 512 units and (e): Fully connected layer with 512 units and Softmax output.

D. Architecture

We explored several popular architectures such as the 3D CNN towards the human activity recognition task on the event data. 3D convolutional layers extract features across volumes of input frames and learn temporal correlation between frames in the window, and fully connected layers with a final softmax classifier were used at the output for the classification. However, through our experiments we observed that an Inception V3 model [21], when pre-trained on Imagenet [14] dataset, acts as the best performing base network that can further be adapted to build the proposed Neuromorphic Human Activity Recognition (n-HAR) framework for activity recognition on event data.

n-HAR Model Architecture : Each input to the model is a set of 40 uniformly sampled frames from a video. We extract rich spatial features from each frame in the input by passing them through a InceptionV3 [21] network that was pre-trained on the ImageNet [14] dataset. The features are the output of the deepest average pooling layer in the Inception V3 model. We flattened these features before feeding them to a two layered perceptron (fully connected) network, each having 512 nodes. Only the parameters of the fully connected network were kept trainable. Dropout of 0.5 on the fully connected layers was used for regularization and a softmax classifier at the output was used for classification. Figure 4 shows a schematic of the n-HAR model architecture.

III. Experiment and Inference

A. Dataset

Few simulated datasets [22] already exist in event domain for action/activity recognition. The major drawback of these simulated datasets is that they are not close to real-world data because of the limitations of their setup. These were created by pointing event-based sensors to a monitor projecting the subjects of interest and trying to minimize the artifacts caused by the display frequency. However, this can never replicate the real-world scenarios as they are limited by the frame rate intrinsic to the videos projected on the monitor. To avoid this, we collected a dataset of real-world activities using an ATIS camera mounted on a tripod. Thirty subjects with diversity in height and gender were part of the dataset collection. Each person performed five categories of activities and then event memory surfaces of each clip were generated and quantized at a rate of 30 FPS. The dataset is imbalanced, but still has sufficient number of videos per class; Boxing: 475, Clapping: 435, Jogging: 860, Walking: 791, Waving: 530. n-HAR dataset has in total of 3091 videos whereas popular datasets like

KTH [23] having similar classes, has only 599 videos. This dataset will be released to both the neruromorphic and the vision community to facilitate further research and development. Figure 3 shows different samples across the categories of a subject from the dataset. The full dataset can be accessed from *http://neuronics.dese.iisc.ac.in/research/research-highlights/n-har/*

B. Training

The model was trained using Keras [24] with Tensorflow [25] as backend. The training and testing have been split with a ratio of 3:1. Each input to both models consists of 40 frames of an activity. The Adam optimizer [26] was used to minimize the cross-entropy loss for multi-class classification. A learning rate of 1e-5 with a decay factor of 1e-6 was used. Early stopping with a patience of 5 epochs was imposed on the training.

C. Results

Our proposed system achieved an average of **94.3%** accuracy across all the classes in the testing split. This high value of accuracy can be partially credited to the use of pre-trained ImageNet [14] weights in our architecture. No Data Augmentation was done during training or testing as we used static feature extraction technique before trainable fully connected layers, hence reducing the number of model parameters. Figure 5 shows the confusion matrix of the predicted labels for our model. It's almost diagonal form may be because we have sufficient number of data points per class (Above 400) for proper class separation required for classification. Analysis of the confusion matrix shows that some of jogging data-points are classified as walking. This can be because some of the jogging subjects were jogging slower than usual, as this dataset was captured with subjects moving in a small indoor space. Some minor confusion is also observed between boxing and clapping as both these actions involve subjects moving their hands in an outstretched position in front of them. Similarly, some overlap is seen between boxing and jogging, as the subjects were asked to move around the limited space while boxing. Table I shows the precision, recall, and F1-score of each class for the model. Approximately **0.9** to **1** precision and recall scores per class signifies that the class imbalance in the n-HAR dataset does not affect the performance of the model.

TABLE I
PRECISION, RECALL AND F1-SCORE OF EACH CLASS FOR N-HAR MODEL

Class	Precision	Recall	F1-score
Boxing	0.89	0.91	0.9
Clapping	0.9	0.88	0.89
Jogging	0.91	1	0.95
Walking	0.99	0.9	0.95
Waving	1	0.99	0.99

The aim of this experiment was to assert that activity recognition is a very natural computer vision challenge that can, in principle, be solved in neuromorphic videos as well. The results indicate that pre-trained weights like ImageNet [14], learnt from normal RGB frames, can be used as basic building blocks that are further adapted in a framework designed to achieve the task of human activity recognition in event-based frames generated using event memory surfaces.

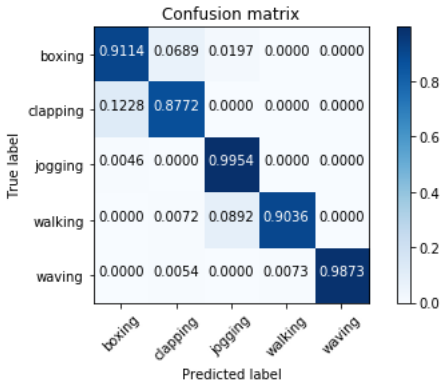

Fig. 5. Confusion Matrix showing how often the labels predicted by n-HAR model match the actual ground truth labels for different activity categories.

IV. Conclusion

In this work, we show that existing frame-based activity recognition techniques can be adapted to event data from neuromorphic vision sensors. We use event memory surfaces to allow the sparse and asynchronous data in the event domain to become compatible with the deep convolution neural network architectures. To the best of our knowledge, this is the first attempt of human activity recognition on a real event-based dataset, as opposed to synthetic datasets [22] that use techniques like capturing videos played on monitors to simulate existing datasets into event datasets. We also provide a real-world event-based dataset for human activity recognition comprising of five categories. In our future work, we aim to explore how more complex activities are represented in the event domain and to provide video analytics solution that successfully recognizes such activities. To design deep CNN architectures potentially utilizing the sparsity property of the event data to extract feature representations at a much lower computational complexity would be our long term goal.

V. Acknowledgment

Research facilities for this work were supported by the Pratiksha trust grant PratikshaYI/2017-8512 and Broadcom by conducting the Neurocom 2018 workshop which facilitated the collection of the dataset.

References

[1] R. Benosman, S.-H. Ieng, C. Clercq, C. Bartolozzi, and M. Srinivasan, "Asynchronous frameless event-based optical flow," *Neural Networks*, vol. 27, pp. 32 – 37, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0893608011002930

[2] J. H. Lee, T. Delbruck, M. Pfeiffer, P. K. Park, C.-W. Shin, H. Ryu, and B. C. Kang, "Real-time gesture interface based on event-driven processing from stereo silicon retinas," *IEEE transactions on neural networks and learning systems*, vol. 25, no. 12, pp. 2250–2263, 2014.

[3] X. Lagorce, S.-H. Ieng, X. Clady, M. Pfeiffer, and R. B. Benosman, "Spatiotemporal features for asynchronous event-based data," *Frontiers in neuroscience*, vol. 9, p. 46, 2015.

[4] M. Giulioni, F. Corradi, V. Dante, and P. Del Giudice, "Real time unsupervised learning of visual stimuli in neuromorphic vlsi systems," *Scientific reports*, vol. 5, p. 14730, 2015.

[5] S. Litzenberger and A. Sabo, "Can silicon retina sensors be used for optical motion analysis in sports?" *Procedia Engineering*, vol. 34, pp. 748–753, 2012.

[6] X. Lagorce, C. Meyer, S.-H. Ieng, D. Filliat, and R. Benosman, "Asynchronous event-based multikernel algorithm for high-speed visual features tracking," *IEEE transactions on neural networks and learning systems*, vol. 26, no. 8, pp. 1710–1720, 2015.

[7] K. A. Boahen, "Point-to-point connectivity between neuromorphic chips using address events," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 47, no. 5, pp. 416–434, 2000.

[8] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128 x 128 120 db 15us latency asynchronous temporal contrast vision sensor," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, Feb 2008.

[9] C. Posch, D. Matolin, and R. Wohlgenannt, "A qvga 143 db dynamic range frame-free pwm image sensor with lossless pixel-level video compression and time-domain cds," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 1, pp. 259–275, 2011.

[10] G. Cohen, S. Afshar, B. Morreale, T. Bessell, A. Wabnitz, M. Rutten, and A. van Schaik, "Event-based sensing for space situational awareness," *The Journal of the Astronautical Sciences*, pp. 1–17, 2017.

[11] S. Afshar, G. Cohen, T. J. Hamilton, J. Tapson, and A. van Schaik, "Investigation of event-based memory surfaces for high-speed tracking, unsupervised feature extraction and object recognition," *arXiv preprint arXiv:1603.04223*, 2016.

[12] X. Lagorce, G. Orchard, F. Galluppi, B. E. Shi, and R. B. Benosman, "Hots: a hierarchy of event-based time-surfaces for pattern recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 7, pp. 1346–1359, 2017.

[13] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.

[14] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, vol. 00, June 2018, pp. 248–255. [Online]. Available: doi.ieeecomputersociety.org/10.1109/CVPR.2009.5206848

[15] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[16] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[17] 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.

[18] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European conference on computer vision*. Springer, 2016, pp. 20–36.

[19] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[20] X. Wang, L. Gao, J. Song, and H. Shen, "Beyond frame-level cnn: saliency-aware 3-d cnn with lstm for video action recognition," *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 510–514, 2017.

[21] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[22] Y. Hu, H. Liu, M. Pfeiffer, and T. Delbruck, "Dvs benchmark datasets for object tracking, action recognition, and object recognition," *Frontiers in neuroscience*, vol. 10, p. 405, 2016.

[23] I. Laptev, B. Caputo *et al.*, "Recognizing human actions: a local svm approach," in *Pattern Recognition, International Conference on(ICPR)*. IEEE, 2004, pp. 32–36.

[24] F. Chollet *et al.*, "Keras: The python deep learning library," *Astrophysics Source Code Library*, 2018.

[25] H.-Y. Chen *et al.*, "Tensorflow–a system for large-scale machine learning," in *OSDI*, vol. 16, 2016, pp. 265–283.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.