Classification of Human Activity by Event-based Vision Sensors using Echo State Networks

Rohan Saini
Kyushu Institute of Technology, 2-4 Hibikino, Wakamatsu-ku, Kitakyushu, 808-0196, Japan

Aryan Rakheja
Kyushu Institute of Technology, 2-4 Hibikino, Wakamatsu-ku, Kitakyushu, 808-0196, Japan

Ryuta Toyoda
Kyushu Institute of Technology, 2-4 Hibikino, Wakamatsu-ku, Kitakyushu, 808-0196, Japan

Yuichiro Tanaka
Kyushu Institute of Technology, 2-4 Hibikino, Wakamatsu-ku, Kitakyushu, 808-0196, Japan

Hakaru Tamukoh
Kyushu Institute of Technology, 2-4 Hibikino, Wakamatsu-ku, Kitakyushu, 808-0196, Japan
Email: saini-rohan413@mail.kyutech.jp, rakheja-aryan859@mail.kyutech.jp, toyoda.ryuta785@mail.kyutech.jp, tanaka-yuichiro@brain.kyutech.ac.jp, tamukoh@brain.kyutech.ac.jp

Abstract

We propose a system for human activity recognition using an event-based vision sensor (EVS) with echo state networks (ESNs). Conventional cameras are susceptible to motion blur and require computationally intensive methods, whereas EVS provides no motion blur and low latency. Our research aims to enable accurate recognition of human activities by using energy-efficient methods. Therefore, we adopt ESNs, which require low computational costs, for the classifier. Additionally, we use feature extraction algorithms such as optical flow and histogram of gradients to improve accuracy. We used an EVS activity recognition dataset created by us containing six human activities and 600 videos. The results showed that our hybrid approach outperformed several techniques. We achieved 89% accuracy when trained with ridge regression.

Keywords: Echo state networks (ESNs), Event-based vision sensor (EVS), Human action recognition

1. Introduction

Human activity recognition is a vital task in today's modern world and is used in many industries. Primarily, traditional cameras serve to accomplish this task, yet they present numerous drawbacks such as motion blur, increased latency, and reduced dynamic ranges. Convolutional neural networks (CNNs) are widely used for this purpose, but they have drawbacks like high power consumption and difficulty in handling temporal data.

To solve these problems, we propose a human activity recognition system with an event-based vision sensors (EVS) [1] and reservoir computing (RC) models, which complement our motive by efficiently processing temporal data to improve accuracy. EVS enables high-speed, low-power object recognition, transforming robotics and neuromorphic research. We used echo state networks (ESNs) as RC implementations along with two feature extraction methods namely optical flow [2] and histogram of oriented gradients (HOG) [3].

2. Related Works

This section provides recent related works utilizing EVS and RC to overcome the problems related to high computational cost, power consumption, and limitations in real-time performance. We also discuss the differences between our proposed method and the works in this section.

The study "Human Activity Recognition with Event-Based Dynamic Vision" [4], used an EVS for action recognition but relied primarily on recurrent neural networks, which can still be computationally intensive.

Another notable work, "Event-based timestamp image encoding network for human action recognition and anticipation" [5], proposed a multi-sensor fusion approach using RC to improve recognition accuracy. This approach displayed high computational cost because of heavy reliability on CNN and long short-term memory (LSTM).

3. Proposed Method

This study integrates ESNs with EVS for human action recognition. Figure 1 shows an example of videos recorded by EVS. The blue parts indicate negative events that are below the negative event threshold and the red parts are positive events above the positive threshold. The threshold and reference voltage are constant values. Before feeding the EVS videos to the ESNs, we applied two feature extraction methods to the videos: optical flow and (HOG).

© The 2025 International Conference on Artificial Life and Robotics (ICAROB2025), Feb.13-16, J:COM HorutoHall, Oita, Japan

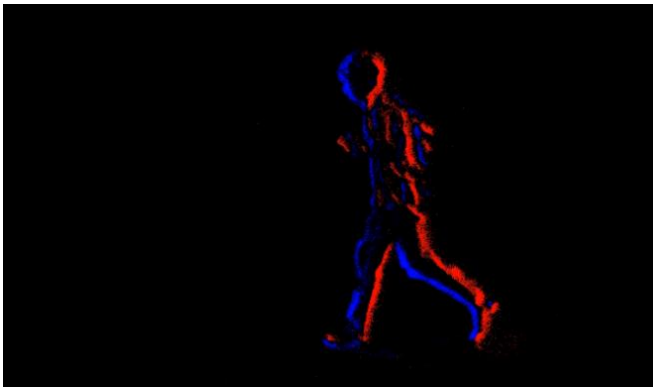

Fig. 1. Running person recorded by EVS

Optical Flow

Optical flow is a computer vision technique that quantifies the motion of objects by tracking the displacement of pixels between consecutive frames. The brightness change equation is:

$$I_x u + I_y v + I_t = 0 \quad (1)$$

Here, $\boldsymbol{I_x}$ and $\boldsymbol{I_y}$ are spatial gradients, and $\boldsymbol{I_t}$ is the temporal intensity change. *u* and *v* denote the horizontal and vertical flow components.

Histogram of Oriented Gradients (HOG)

The HOG computes gradient, which is the rate and direction of change in intensity for each pixel in an image, capturing texture and edge information.

$$G_x(r,c) = I(r,c+1) - I(r,c-1) \quad (2)$$
$$G_y(r,c) = I(r-1,c) - I(r+1,c) \quad (3)$$

In Equations 2 and 3, G_x and G_y are the gradients in x and y (horizontal and vertical) directions, r and c refer to the rows and columns, and *I* signifies the pixel intensity.

Echo State Networks (ESNs)

Figure 2 shows the complete architecture of the model, and how the feature input array is given to an ESN. ESNs use a reservoir of recurrently connected neurons to nonlinearly convert inputs into temporal patterns in a high-dimensional space. Weights in the hidden layer are fixed and not trainable. The sparsely hidden layer usually has less than 10% connectivity to make various patterns in the reservoir. Only output layer weights are trained so that the network generates outputs by utilizing the patterns in the reservoir and therefore, its training cost is low compared to normal neural networks. The reservoir state *x*(*t*) evolves as:

$$x(t) = f(W_{in}\, u(t) + W_{res}\, x(t-1)) \quad (3)$$

Here W_{in} and W_{res} are the weight matrices of the input to reservoir layers and in the reservoir layer, respectively. $u(t)$ is input, and f is a non-linear function that was tanh.

For the input to the network, we use either the optical flow feature or the HOG feature from an EVS video. For the readout of the network, we use either a linear model or the

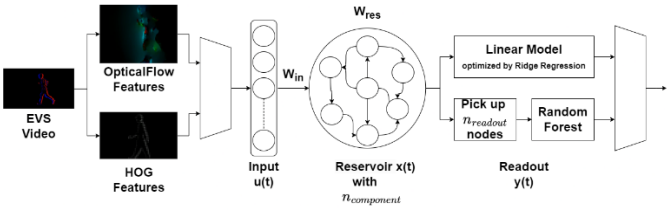

Fig. 2. Task flow for the ESN-based architecture

random forest. Therefore, the proposed framework includes four variants in total.

For the variant with the linear model, the output of the network *y*(*t*) is computed as follows:

$$y(t) = W_{out} x(t) \quad (4)$$

where W_{out} is the weight matrix between the reservoir and readout layers. The W_{out} is optimized by the ridge regression [6].

For the variant with random forest, the output of the network y can be expressed as:

$$y = f_{RF}(f_{ESN}(U, \theta_{ESN}), \theta_{RF}) \quad (5)$$

where f_{RF} is the random forest prediction, f_{ESN} is the ESN transformation function, U is the input signal, θ_{ESN} is $\{n_{readout}, n_{components}$, weight scaling$\}$, θ_{RF} is $\{n_{estimators}$, max depth$\}$. $n_{readout}$ is the number of units in the readout layer of the ESN, $n_{components}$ represents the number of internal units in the reservoir ESN, weight scaling is the adjustment of the strength of the connections between the input layer and the reservoir layer, $n_{estimators}$ represents the number of decision trees in the random forest, and max depth describes the maximum depth of each decision tree in the random forest.

4. Results and Discussion

We used a dataset consisting of 600 EVS videos (with 80% allocated for training and 20% for testing) and labeled four action classes.
Class 1: Represents the action of Walking.
Class 2: Represents the action of Jogging.
Class 3: Represents the action of Running.
Class 4: Represents the action of Boxing.
We used 500 nodes for the reservoir layer with 25% connectivity. The spectral radius of the reservoir weight connections was set as 0.59 to satisfy the echo state property, which is characteristic to ensure the reproducibility of the reservoir. We adopted a grid search to find the best parameters for the random forest. The maximum depth of the random forest (max depth) was 10 and the maximum number of $n_{estimators}$ was 75.

Figure 3 shows the graph for training accuracy with respect to the number of training videos. It depicts the training comparison between all variants we used. Table 1 shows the comparison of performance between the techniques we used.

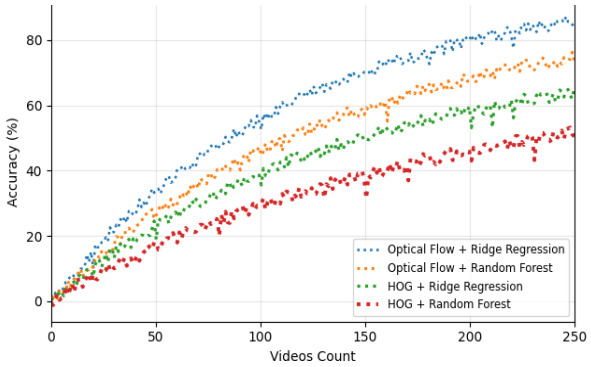

Fig. 3. Training Accuracy Graph

Table 1. Comparison of performances among variants

Feature Set	Training Method	Accuracy (%)
Optical Flow	Ridge Regression	89
Optical Flow	Random Forest	80
HOG	Ridge Regression	65
HOG	Random Forest	58

5. Conclusion

We proposed an EVS-ESN-based architecture with optical flow feature extraction demonstrating a promising accuracy of 89% with ridge regression and an 80% accuracy with random forest. This shows the potential of ESN for these use cases. For the future scope, to further improve this study, we could use more sophisticated model architectures like graph neural networks (GNNs) [7]. It could be because optical flow features represent motion and spatial dynamics by encoding them into a graph. GNNs could learn more spatial-temporal features than ESNs. Further, optical flow may not always align with Euclidean representation of data, which refers to a space or a data structure that does not follow the traditional Euclidean geometry. Non-Euclidean data has many features or dimensions, which cannot be represented in a traditional Euclidean space. GNNs excel in handling non-euclidean data.

Acknowledgments

This work received support from JSPS KAKENHI Grant Numbers 23H03468, as well as from JST
ALCA-Next Grant Number JPMJAN23F3. This work was partially supported by Sony Semiconductor Solutions Corporation.

References

1. Sony Semiconductor Solutions Corporation, “Semiconductor Solutions for Electric Vehicles,” 2024. [Online]. Available: https://www.sony-semicon.com/en/technology/industry/evs.html
2. S. S. Beauchemin and J. L. Barron, “The computation of optical flow,” *ACM computing surveys (CSUR)*, vol. 27, no. 3, pp. 433–466, 1995.
3. 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)*, 2005, pp. 886–893 vol. 1. doi: 10.1109/CVPR.2005.177.
4. P. Pansuriya, N. Chokshi, D. Patel, and S. Vahora, “Human activity recognition with event-based dynamic vision sensor using deep recurrent neural network,” *International Journal of Advanced Science and Technology*, vol. 29, no. 4, pp. 9084–9091, 2020.
5. C. Huang, “Event-based timestamp image encoding network for human action recognition and anticipation,” in *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–9.
6. G. C. McDonald, “Ridge regression,” *Wiley Interdiscip Rev Comput Stat*, vol. 1, no. 1, pp. 93–100, 2009.
7. F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, “The Graph Neural Network Model,” *IEEE Trans Neural Netw*, vol. 20, no. 1, pp. 61–80, 2009, doi: 10.1109/TNN.2008.2005605.

Authors Introduction

Mr. Rohan Saini

He received his Bachelor’s degree in Engineering in 2024 from the Faculty of Computer Science and Engineering, Dronacharya College of Engineering in India. He is currently a master's student at Kyushu Institute of Technology, Japan.

Mr. Aryan Rakheja

He received his Bachelor’s degree in Engineering in 2024 from the Faculty of Computer Science and Engineering, Dronacharya College of Engineering in India. He is currently a master's student at Kyushu Institute of Technology, Japan.

Mr. Ryuta Toyoda

He received the B.Eng. degree from Kyushu Institute of Technology, Japan, in 2023. He is a master's degree student at the Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology. His research interests include developing processing circuits for EVS.

Dr. Tanaka Yuichiro

He received the B.E., M.E., and Ph.D. degrees from the Kyushu Institute of Technology, in 2016, 2018, and 2021, respectively. He was a Research Fellow with the Japan Society for the Promotion of Science (JSPS), from 2019 to 2021. He was an Assistant Professor at the Research Center for Neuromorphic AI Hardware, Kyushu Institute of Technology, from 2021 to 2024. He has been an Associate Professor at the Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology since 2024. His research interests include neural networks, digital hardware implementation, and home service robots. He is a member of IEICE and JNNS.

Prof. Hakaru Tamukoh

He received his B.Eng. degree from Miyazaki University, Japan in 2001. He received his M.Eng. and Ph.D. degrees from Kyushu Institute of Technology, Japan in 2003 and 2006, respectively. He was a postdoctoral research fellow of the 21st century center of excellent program at Kyushu Institute of Technology, from April 2006 to September 2007. He was an assistant professor at Tokyo University of Agriculture and Technology, from October 2007 to January 2013. He joined the Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology, Japan as an associate professor in February 2013. He has been a professor at the Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology, Japan, since April 2021. His research interests include hardware/software complex systems, digital hardware design, brain-inspired AI, neural networks, soft computing, and home service robots. He was the author of papers that won the Best Paper Award at IEEE/INNS IJCNN 2019, the Best Live Demonstration Award at IEEE ISCAS 2019, and the Best Paper Award at ICONIP 2013. He won the world championship of robot competition five times at RoboCup @Home 2017, 2018, 2024 and World Robot Challenge 2018, 2020. He received the METI Minister's Award for Excellence in 2018 and 2021, and the RSJ Special Award from the Robotics Society of Japan in 2018. He is a member of IEEE, IEICE, SOFT, JNNS, JSAI, and RSJ.