Learning Bottleneck Transformer for Event Image-Voxel Feature Fusion based Classification

Chengguo Yuan[1] , Yu Jin[1] , Zongzhen Wu[1] , Fanting Wei[1] , Yangzirui Wang[1] , Lan Chen[1] , and Xiao Wang[1]

[1] Anhui University, Hefei City, 230601, Anhui Province, China

https://github.com/Event-AHU

Scan for Source Code !!!

Introduction

Recognizing target objects using an event-based camera draws more and more attention in recent years. Existing works usually represent the event streams into point-cloud, voxel, image, etc, and learn the feature representations using various deep neural networks. Their final results may be limited by the following factors: monotonous modal expressions and the design of the network structure. To address the aforementioned challenges, this paper proposes a novel dual-stream framework for event representation, extraction, and fusion. This framework simultaneously models two common representations: event images and event voxels. By utilizing Transformer and Structured Graph Neural Network (GNN) architectures, spatial information and three-dimensional stereo information can be learned separately. Additionally, a bottleneck Transformer is introduced to facilitate the fusion of the dual-stream information. Extensive experiments demonstrate that our proposed framework achieves state-of-the-art performance on two widely used event-based classification datasets.

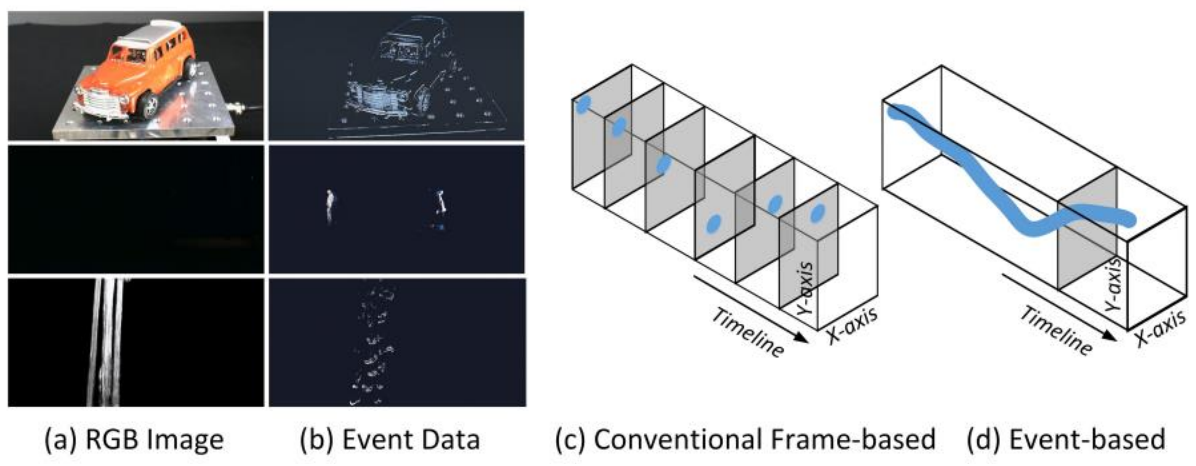


Fig. 1: Comparison of the frame- and event-based (a, b) shows representative samples in regular scenarios, low illumination, and fast motion. (c, d) illustrates the different types of raw data representation of frame- and event-based camera

Method

Given an input event stream consisting of hundreds of thousands of events, our approach involves several steps to enhance the representation. **Initially**, we employ event frame stacking and voxel construction techniques to generate event frame and voxel representations, respectively. **Subsequently**, we utilize two intermediate representations, namely event frame and voxel graph, to capture the spatio-temporal relationships within the event stream. **What's more**, to further improve the feature descriptors for event frame and graph based event representation, we propose a novel dual branch learning network. **Finally**, we combine these representations to create a comprehensive representation for event data, enabling effective recognition.

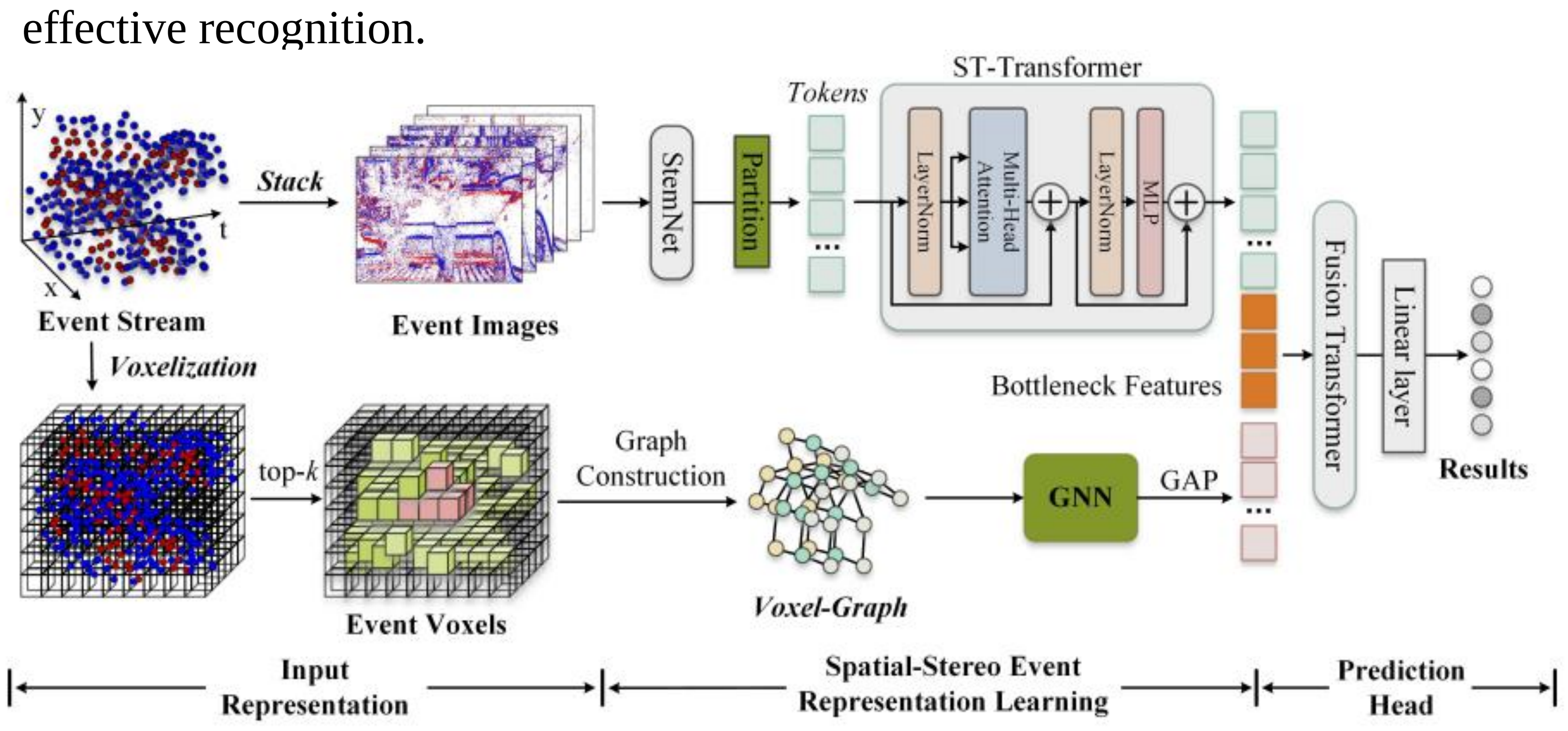


Fig. 2: An overview of our proposed Image-Voxel Feature Learning framework for event-based recognition.

Some details about the network architecture are listing below.

Input Representation. To reduce the number of events, we employ some down-sampling techniques. We first transform the asynchronous event flow into the synchronous event images by stacking the events in a time interval based on the exposure time. We also employ voxelization to obtain voxel representation.

Graph Neural Networks for Event Voxel Encoding. We construct a geometric neighboring graph for voxel event data. To be specific, each node v_i represents a voxel o_i , the edge e_{ij} exists between node v_i and v_j , if the Euclidean distance between their 3D coordinates is less than a threshold R. We adapt Gaussian Mixture Model(GMM), convolution to learn the effective representations for voxel graph.

Spatial-temporal Transformer for Event Frame Encoding. We extract initial CNN features and embed event frames through StemNet. After obtaining the initial features, we designed an ST-Transformer module to further achieve a better representation of spatio-temporal information.

Bottleneck Transformer. In order to achieve the interaction between Event Images and Event Voxels information representations and learn a unified spatio-temporal context data representation. We also designed the Fusion Transformer module and introduced the Bottleneck mechanism.

Experiment

In this work, we utilized two datasets, namely N-MNIST , and ASL-DVS , to evaluate our proposed model. The tables show the results of our methods on different datasets and the comparison with Other SOTA algorithms.

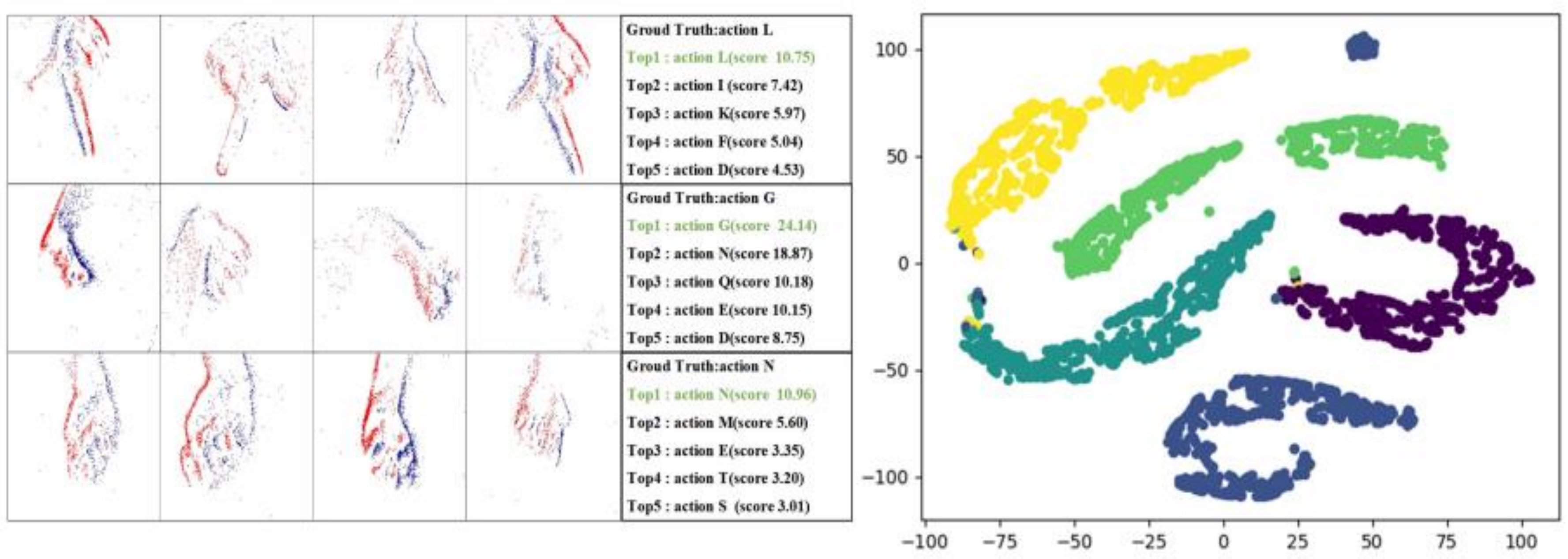
Table 1: Results on the ASL-DVS [15] dataset.

EST [35]	**AMAE** [36]	**M-LSTM** [37]	**MVF-Net** [38]	**EventNet** [39]
0.979	0.984	0.980	0.971	0.833
RG-CNNs [15]	**EV-VGCNN** [40]	**VMV-GCN** [41]	**EV-Gait-3DGraph** [9]	**Ours**
0.901	0.983	0.989	0.738	0.996

Table 2: Results on the N-MNIST [34] dataset.

EST [35]	**M-LSTM** [37]	**MVF-Net** [38]	**Gabor-SNN** [42]	**EvS-S** [27]
99.0	98.6	98.1	83.7	99.1
HATS [42]	**EventNet** [39]	**RG-CNNs** [15]	**EV-VGCNN** [40]	**Ours**
99.1	75.2	99.0	99.4	98.9

This picture exhibits the visualization of top-5 recognition results and feature distribution on the ASLDVS dataset.



To help researchers better understand the method we proposed, we conduct comprehensive experiments of component analysis on the DVS128-Gait-Day dataset and ASL-DVS dataset to check their influence on the overall model.

Table 3: Ablation study on DVS128-Gait-Day dataset [33].

Index	Component	Results
1	Event image only	95.2
2	Event voxel only	98.0
3	Event Image + Voxel	98.7

Table 4: Ablation study on ASL-DVS [15].

Index	Component	Results
1	w/o Bottleneck Feature	98.5
2	w/o FusionFormer	98.3

Conclusion

Our paper introduces a novel dual-stream framework for event representation, extraction, and fusion. The proposed framework simultaneously models two common representations: event images and event voxels. By leveraging Transformer and Structured Graph Neural Network (GNN) architectures, spatial information and three dimensional stereo information can be learned separately. Moreover, the introduction of a bottleneck Transformer facilitates the fusion of the dual-stream information. These findings highlight the effectiveness of the dual-stream framework in addressing the limitations of existing approaches and improving the recognition accuracy in event-based object recognition tasks.

Reference

1. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "**Imagenet classification with deep convolutional neural networks.**" NIPS-2012.
2. K. He, X. Zhang, S. Ren, and J. Sun, "**Deep residual learning for image recognition**," CVPR 2016.
3. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "**Attention is all you need**," NIPS-2017.