HARDVS: Revisiting Human Activity Recognition with Dynamic Vision Sensors

Anhui University

EVENT-AHU RESEARCH

https://github.com/Event-AHU

Xiao Wang[1],, Zongzhen Wu[1], Bo Jiang[1*], Zhimin Bao[2], Lin Zhu[3], Guoqi Li[4,5], Yaowei Wang[5], Yonghong Tian[5,6,7]

[1]Anhui University, Hefei City, 230601, Anhui Province, China,
[2]Tencent, [3]Beijing Institute of Technology, [4]University of Chinese Academy of Sciences,
[5]Peng Cheng Laboratory, China, [6]National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University, China, [7]School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University, China.

Scan for Source Code !!!

HARDVS

Introduction

The main streams of human activity recognition (HAR) algorithms are developed based on RGB cameras which usually suffer from illumination, fast motion, privacy preservation, and large energy consumption. Meanwhile, the biologically inspired event cameras attracted great interest due to their unique features, such as high dynamic range, dense temporal but sparse spatial resolution, low latency, low power, etc. As it is a newly arising sensor, even there is no realistic large-scale dataset for HAR. Considering its great practical value, in this paper, we propose a large-scale benchmark dataset to bridge this gap, termed HARDVS, which contains 300 categories and more than 100K event sequences. We evaluate and report the performance of multiple popular HAR algorithms, which provide extensive baselines for future works to compare. More importantly, we propose a novel spatial-temporal feature learning and fusion framework, termed ESTF, for event stream based human activity recognition. It first projects the event streams into spatial and temporal embeddings using StemNet, then, encodes and fuses the dual-view representations using Transformer networks. Finally, the dual features are concatenated and fed into a classification head for activity prediction. Extensive experiments on multiple datasets fully validated the effectiveness of our model.

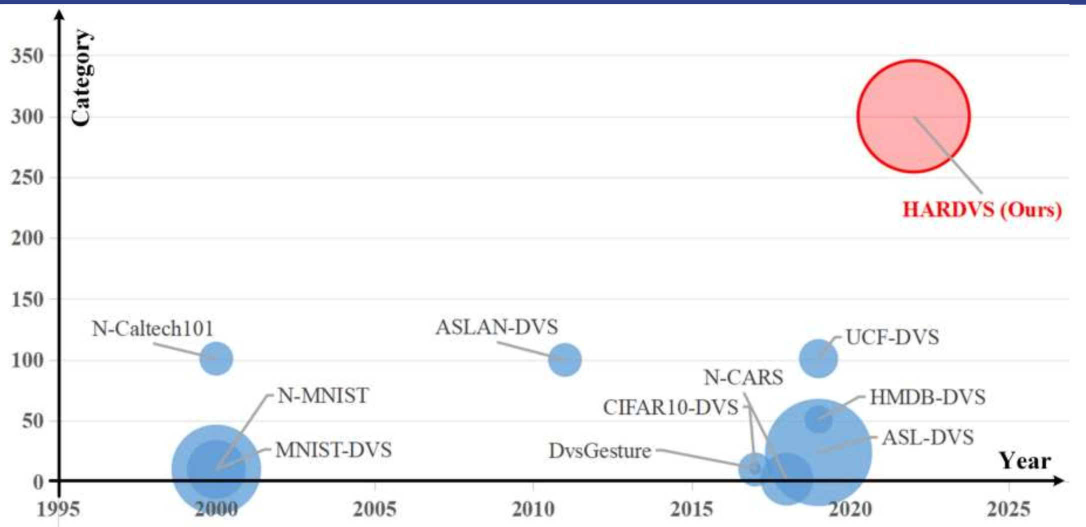

Figure 1: Comparison between existing datasets and our proposed HARDVS dataset for event based video classification.

HARDVS Benchmark Dataset

We aim to provide a good platform for the training and evaluation of DVS-based human activity recognition. When constructing the HARDVS benchmark dataset, the following attributes/highlights are considered: 1). Large-scale. 2). Wide varieties. 3). Different capture distances. 4). Longterm. 5). Dual-modality.
Our dataset considers multiple challenging factors which may influence the results of HAR with the DVS sensor. The detailed introductions can be found below: (a). Multi-view. (b). Multi-illumination. (c). Multi-motion. (d). Dynamic background. (e). Occlusion.

Table 1: Comparison of event datasets for human activity recognition. M-VW, M-ILL, M-MO, DYB, OCC, and DR denotes multi-view, multi-illumination, multi-motion, dynamic background, occlusion, and duration of the action, respectively. Note that we only report these attributes of realistic DVS datasets for HAR.

Dataset	Year	Sensors	Scale	Class	Resolution	Real	M-VW	M-ILL	M-MO	DYB	OCC	DR
ASLAN-DVS	2011	DAVIS240c	3, 697	432	240 × 180	✗	-	-	-	-	-	-
MNIST-DVS	2013	DAVIS128	30, 000	10	128 × 128	✗	-	-	-	-	-	-
N-Caltech101	2015	ATIS	8, 709	101	302 × 245	✗	-	-	-	-	-	-
N-MNIST	2015	ATIS	70, 000	10	28 × 28	✗	-	-	-	-	-	-
CIFAR10-DVS	2017	DAVIS128	10, 000	10	128 × 128	✗	-	-	-	-	-	-
HMDB-DVS	2019	DAVIS240c	6, 766	51	240 × 180	✗	-	-	-	-	-	-
UCF-DVS	2019	DAVIS240c	13, 320	101	240 × 180	✗	-	-	-	-	-	-
N-ImageNet	2021	Samsung-Gen3	1, 781, 167	1000	480 × 640	✗	-	-	-	-	-	-
ES-ImageNet	2021	-	1, 306, 916	1000	224 × 224	✗	-	-	-	-	-	-
N-EPIC-Kitchens	2022	-	10, 000	-	-	✗	-	-	-	-	-	-
N-ROD	2022	-	41,877	51	640 × 480	✗	-	-	-	-	-	-
DvsGesture	2017	DAVIS128	1, 342	11	128 × 128	✓	✗	✓	✗	✗	✗	-
N-CARS	2018	ATIS	24, 029	2	304 × 240	✓	✓	✗	✓	✗	✓	-
ASL-DVS	2019	DAVIS240	100, 800	24	240 × 180	✓	✗	✗	✗	✗	✗	0.1s
PAF	2019	DAVIS346	450	10	346 × 260	✓	✗	✗	✗	✗	✗	5s
DailyAction	2021	DAVIS346	1, 440	12	346 × 260	✓	✓	✓	✗	✗	✗	5s
HARDVS (Ours)	2023	DAVIS346	107, 646	300	346 × 260	✓	✓	✓	✓	✓	✓	5s

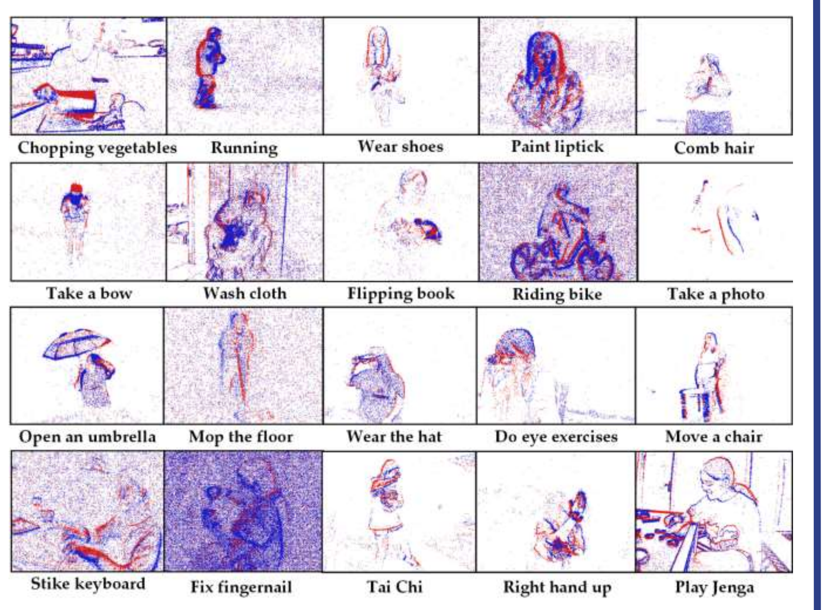

Figure 2: Illustration of some representative samples of our proposed HARDVS dataset.

Method

Given the input event-stream data, we first extract the initial spatial and temporal embeddings respectively. Then, a Spatial and Temporal Feature Enhancement Learning module is devised to further enrich the event-stream data representations by deeply capturing both spatial correlation and temporal dependence of event stream. Finally, an effective Fusion Transformer block is designed to integrate the spatial and temporal cues together for the final feature representation.
Some details about the network architecture are listing below.

Initial Spatial and Temporal Embedding. we first transform the asynchronous event flows into the synchronous event images by stacking the events in a time interval based on the exposure time. we adopt StemNet(ResNet-18) to extract an initial CNN feature descriptor for it and denote $\chi \in \mathbb{R}^{H\times W\times T\times c}$.For the temporal branch, we adopt a convolution layer to reduce the feature size to obtain $X^{\mathrm{T}} \in \mathbb{R}^{T\times d}$, where $d=\frac{h}{2}\times\frac{w}{2}\times c'$. For the spatial branch, we first adopt a convolution layer to resize the features. Then, we conduct the merging operation on the time dimension and reshape it to the matrix form $X^{\mathrm{S}} \in \mathbb{R}^{N\times d}$ where N=$\frac{hw}{4}$.

Spatial and Temporal Enhancement Learning(STEL)**.** The proposed STEL module involves two blocks, i.e., Spatial Transformer (SF) block, and Temporal Transformer (TF) block, which respectively capture the spatial correlations and temporal dependences of event data to learn context enriched representations. The SF block includes multi-head self-attention (MSA) and MLP module with a LayerNorm (LN) used between two modules.

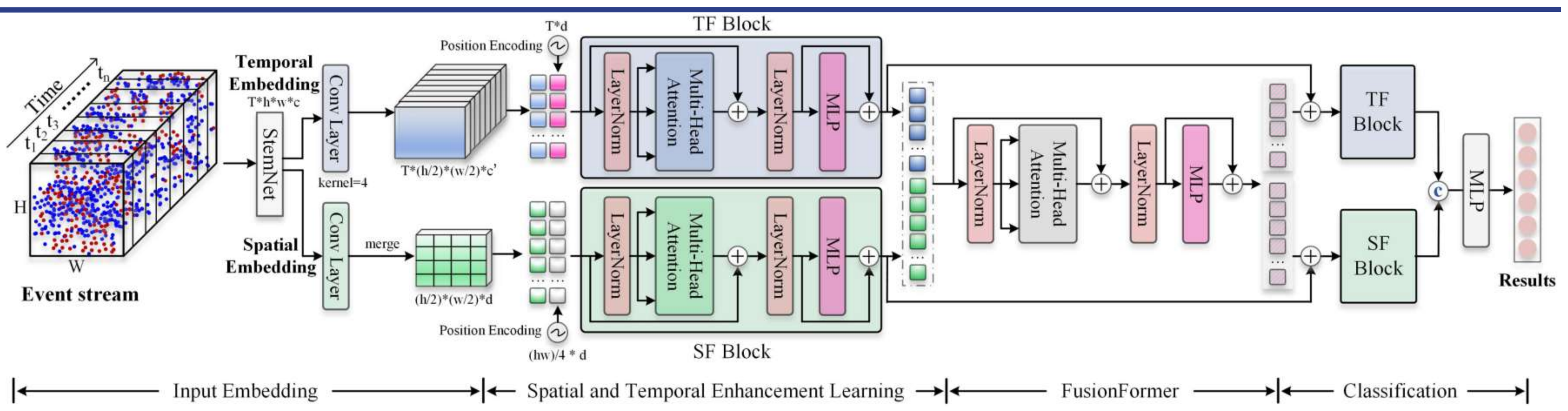

Figure 3: An overview of our proposed ESTF framework for event-based human action recognition.

Fusion Transformer. In order to conduct the interaction between the above ST and TF blocks and learn a unified spatio-temporal contextual data representations, we also design a Fusion Transformer module. We first collect the N spatial and T temporal tokens together and feed them to a unified Transformer block which includes multi-head self-attention (MSA) and MLP submodule.

Experiment

In this work, we utilized three datasets, namely N-Caltech101, ASL-DVS, and HARDVS to evaluate our proposed model.

Table 2: Results on N-Caltech101.

EventNet	Gabor-SNN	RG-CNNs	VMV-GCN	EV-VGCNN	EST
0.425	0.196	0.657	0.778	0.748	0.753
ResNet-50	**MVF-Net**	**M-LSTM**	**AMAE**	**HATS**	**Ours**
0.637	0.687	0.738	0.694	0.642	0.832

Table 3: Results on ASL-DVS.

EST	AMAE	M-LSTM	MVF-Net	ResNet-50
0.979	0.984	0.980	0.971	0.886
EventNet	**RG-CNNs**	**EV-VGCNN**	**VMV-GCN**	**Ours**
0.833	0.901	0.983	0.989	0.999

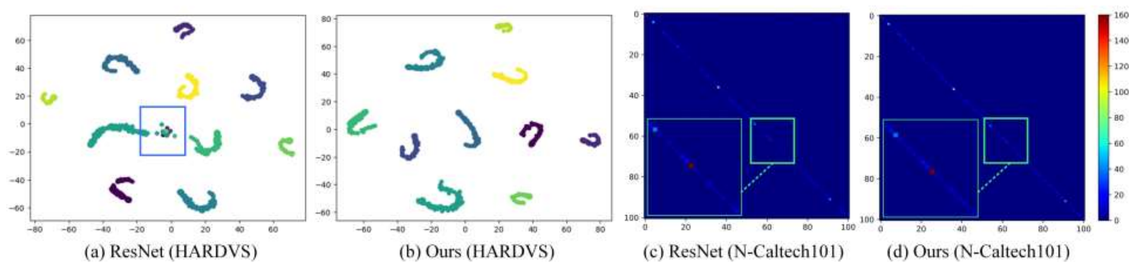

Figure 4: Visualization of feature distribution of our baseline and newly proposed ESTF on HARDVS dataset (a, b) and confusion matrix of baseline ResNet and our model on N-Caltech101 dataset (c, d).

Table 4: Results on HARDVS.

Algorithm	Publish	Backbone	Event	MAC	Param.
ResNet18	CVPR-2016	ResNet18	49.20 | 56.09	17.2G	11.7M
C3D	ICCV-2015	CNN	50.52 | 56.14	0.2G	147.2M
R2Plus1D	CVPR-2018	ResNet-34	49.06 | 56.43	40.7G	63.5M
TSM	ICCV-2019	ResNet-50	52.63 | 60.56	0.7G	24.3M
ACTION-Net	CVPR-2021	ResNet-50	46.85 | 56.19	34.7G	27.9M
TAM	ICCV-2021	ResNet-50	50.41 | 57.99	33.1G	25.6M
V-SwinTrans	CVPR-2022	Swin Transformer	51.91 | 59.11	17.5G	27.8M
TimeSformer	ICML-2021	ViT	50.77 | 58.70	107.3G	121.2M
SlowFast	ICCV-2019	ResNet-50	50.63 | 57.77	0.7G	33.6M
ESTF (Ours)	-	ResNet18	51.22 | 57.53	17.6G	46.1M

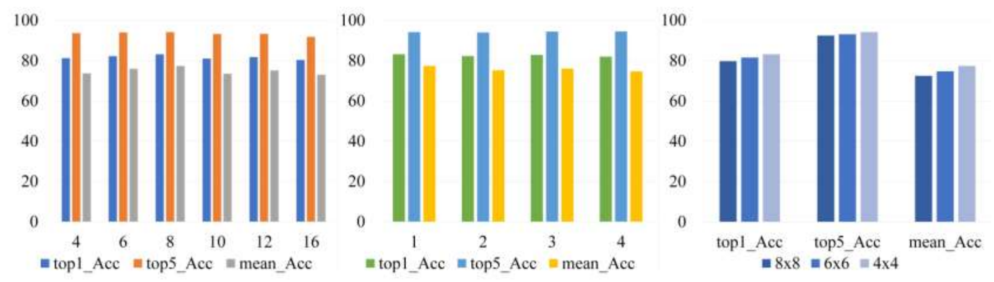

Figure 5: Results of different (left) input frames; (middle) transformer layers; (right) patch sizes on the HARDVS dataset.

Table 5: Component Analysis on the N-Caltech101 and HARDVS Dataset.

No.	ResNet	TF	SF	FF	N-Caltech101	HARDVS
1	✓				72.14	49.20
2	✓	✓			81.54	49.65
3	✓		✓		80.47	50.81
4	✓	✓	✓		82.89	51.06
5	✓	✓	✓	✓	83.17	51.22

Conclusion

In this paper, we propose a large-scale benchmark dataset for event-based human action recognition, termed HARDVS. It contains 300 categories of human activities and more than 100K event sequences captured from DAVIS346 camera. In addition, we also propose a novel Event-based Spatial-Temporal Transformer (short for ESTF) that conducts spatial-temporal enhanced learning and fusion for accurate action recognition. Extensive experiments on multiple benchmark datasets validated the effectiveness of our proposed framework. It sets the new SOTA performances on N-Caltech101 and ALSDVS datasets. We hope the proposed dataset and baseline approach will boost the further development of event camera based human action recognition.