HARDVS: Revisiting Human Activity Recognition with Dynamic Vision Sensors

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]School of Computer Science and Technology, Anhui University, Hefei 230601, 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.
wangxiaocvpr@foxmail.com, {17398389386, zeyiabc}@163.com, zhiminbao@tencent.com, {linzhu, yhtian}@pku.edu.cn, guoqi.li@ia.ac.cn, wangyw@pcl.ac.cn

Abstract

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. Both the dataset and source code will be released at https://github.com/Event-AHU/HARDVS.

Introduction

With the rapid development of the smart city, recognizing human behavior (i.e., Human Activity Recognition, HAR) accurately and efficiently is becoming an extremely urgent task. Most researchers develop the HAR algorithms (Kong and Fu 2018; Ahmad et al. 2021) based on RGB cameras which are widely deployed and easy to collect the data. With the help of large-scale benchmark datasets and deep learning, HAR in regular scenarios has been studied to some extent. However, the storage, transmission, and analysis of surveillance video set limit the demands in the practical systems due to the usage of RGB sensors. More in detail, the standard RGB cameras have a limited frame rate (e.g., 30 FPS) which makes it hard to capture the fast-moving objects and is easily influenced by motion blur. The low dynamic range (60 dB) makes the RGB sensors work poorly in low illumination. It also suffers from the high redundancy between nearby frames which needs more storage and energy consumption. Privacy protection also greatly limits its development. Therefore, a natural question is *do we have to recognize human activities using the RGB sensors?*

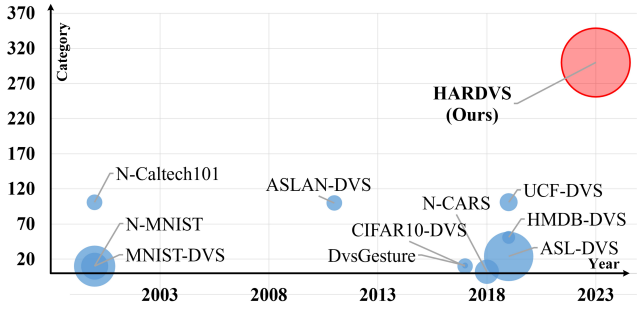


Figure 1: Comparison between existing datasets and our proposed HARDVS dataset for event based video classification.

Recently, the biologically inspired sensors (called event cameras), such as DAVIS (Brandli et al. 2014), CeleX (Chen and Guo 2019), ATIS (Posch, Matolin, and Wohlgenannt 2010), and PROPHESEE [1], drawing more and more researcher's attention. Different from RGB cameras which record light in a synchronous way (i.e., the video frame), the event cameras output events (or spikes) asynchronously which corresponds to the illumination variation. In another word, each pixel of event cameras independently records a binary value only when the light changes exceed a threshold. Events for the increase and decrease of illumination are called ON and OFF events respectively. Due to the unique sampling mechanism, the asynchronous events are spatially sparse but temporally dense. It is less affected by motion blur, therefore, is suitable for capturing fast-moving human

*Corresponding author: Bo Jiang

[1]https://www.prophesee.ai

actions, such as the magician's fast-moving palm, and movement recognition of sports players. It has a higher dynamic range (120 dB) and lower latency, which enables it to work well even in low illumination compared with standard RGB cameras. In addition, the storage and energy consumption are also significantly reduced (Gallego et al. 2020; Wang et al. 2021a; Li et al. 2022; Zhu et al. 2022, 2021). Event streams highlight the contour information and protect personal privacy to a large extent. According to the aforementioned observation and thinking, we are inspired to address human activity recognition in the wild by using event cameras.

Although there are already several benchmark datasets proposed for classification tasks (Bi et al. 2020; Amir et al. 2017; Li et al. 2017; Serrano-Gotarredona and Linares-Barranco 2015; Kuehne et al. 2011; Soomro, Zamir, and Shah 2012; Kliper-Gross, Hassner, and Wolf 2011; Planamente et al. 2021; Cannici et al. 2021), however, most of them are simulated/synthetic datasets that are transformed from RGB videos with the simulator. Some researchers attain the event data by recording the screen while displaying RGB videos. Obviously, these datasets are hard to reflect the features of event cameras in real-world scenarios, especially fast-motion and low-light scenarios. ASL-DVS is proposed by Bi et al. (Bi et al. 2020) which is consisted of 100800 samples but can only be used for hand gesture recognition with 24 classes. DvsGesture (Amir et al. 2017) is also limited by its scale and categories in the deep learning era. In addition, some datasets have become saturated in performance, for example, Wang et al. (Wang et al. 2019a) already achieved 97.08% on the DvsGesture (Amir et al. 2017) dataset. Therefore, the research community still has insistent demands for a large-scale HAR benchmark dataset recorded in the wild.

In this paper, we propose a new large-scale benchmark dataset, termed HARDVS, to address the problem on the lack of real event data. Specifically, our proposed HARDVS dataset contains more than 100K video clips recorded with a DAVIS346 camera, each of them lasting for about 5-10 seconds. It contains 300 categories of human activities in daily life, such as *drinking, riding a bike, sitting down*, and *washing hands*. The following factors are taken into account to make our data more diverse, including *multi-views*, *illuminations*, *motion speed*, *dynamic background*, *occlusion*, *flashing light*, and *photographic distance*. To the best of our knowledge, our proposed HARDVS is the first real, large-scale, and challenging benchmark dataset for human activity recognition in the wild. A comparison between existing recognition datasets and our HARDVS is illustrated in Fig. 1.

Based on our newly proposed HARDVS dataset, we construct a novel event-based human action recognition framework, termed ESTF (Event-based Spatial-Temporal Transformer). As shown in Fig. 3, the ESTF transforms the event streams into spatial and temporal token sequences and learns the dual features by employing SpatialFormer (SF) and TemporalFormer (TF) respectively. Further, we propose a FusionFormer to realize the message passing between the spatial and temporal features. The aggregated representation is added with features of dual branches as the input for the subsequent learning blocks, respectively. The outputs will be concatenated and fed into two MLP layers for the final action prediction.

To sum up, the contributions of this paper can be concluded as the following three aspects:

- We propose a new large-scale neuromorphic dataset for human activity recognition, termed HARDVS. It contains more than 100K samples with 300 classes, and fully reflects the challenging factors in the real world.
- We propose a novel Event-based Spatial-Temporal Transformer (ESTF) approach for human action recognition by exploiting spatial and temporal feature learning and fusing them with Transformer networks. Extensive experiments on multiple event-based classification datasets fully demonstrate the effectiveness of our proposed ESTF approach.
- We re-train and report the performances of multiple popular HAR algorithms on our HARDVS dataset, which provide extensive baselines for future works to compare on the HARDVS dataset.

Related Works

HAR using Event Sensors Compared with RGB cameras, few researchers focus on event camera-based HAR. Arnon et al. (Amir et al. 2017) propose the first gesture recognition system based on TrueNorth neurosynaptic processor. Xavier et al. (Clady et al. 2017) propose an event-based luminance-free feature for local corner detection and global gesture recognition. Chen et al. (Chen et al. 2021) propose a hand gesture recognition system based on DVS and also design a wearable glove with a high-frequency active LED marker that fully exploits its properties. A retinomorphic event-driven representation (EDR) is proposed by Chen et al. (Chen et al. 2019), which can realize three important functions of the biological retina, i.e., the logarithmic transformation, ON/OFF pathways, and integration of multiple timescales. Graph neural networks (GNN) and SNNs are also exploited for event-based recognition. Specifically, Wang et al. (Wang et al. 2021b) adopt GNNs and CNNs for gait recognition. Xing et al. design a spiking convolutional recurrent neural network (SCRNN) architecture for event-based sequence (Xing, Di Caterina, and Soraghan 2020). According to our observations, these works are evaluated only on simple HAR datasets or simulated datasets. It is necessary and urgent to introduce a large-scale HAR dataset for current evaluation.

Event Benchmark Datasets for HAR As shown in Table 1, most of the existing event camera-based datasets for recognition are artificial datasets. Usually, the researchers display the RGB HAR datasets on a large screen and record the activity with neuromorphic sensors. For example, the N-Caltech101 (Orchard et al. 2015) and N-MNIST (Orchard et al. 2015) are recorded with an ATIS camera which contains 101 and 10 classes, respectively. Bi et al. (Bi et al. 2020) also transform popular HAR datasets into simulated event flow, including HMDB-DVS (Bi et al. 2020; Kuehne et al. 2011), UCF-DVS (Bi et al. 2020; Soomro, Zamir, and Shah 2012), and ASLAN-DVS (Kliper-Gross, Hassner, and Wolf 2011), which further expands the number of

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

Table 1: Comparison of event datasets for human activity recognition. MVW, MILL, MMO, 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.

datasets available for HAR. However, these simulated event datasets hardly reflect the advantages of event cameras, such as low light, fast motion, etc. There are three realistic event datasets for classification, i.e., the DvsGesture (Amir et al. 2017), N-CARS (Sironi et al. 2018), and ASL-DVS (Bi et al. 2020), but these benchmarks are limited by their scale, categories, and scenes. Specifically, these datasets contain 11, 2, and 24 classes only, and also rarely take challenging factors like multi-view, motion, and glitter into consideration. Compared with existing datasets, our proposed HARDVS dataset is large-scale (100K samples) and category-wide (300 classes) for deep neural networks. Our sequences are recorded in the wild and fully reflect the features of the aforementioned attributes. We believe our proposed benchmark dataset will greatly promote the development of event-based HAR.

HARDVS Benchmark Dataset

Highlights 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:** As we all know, large-scale datasets play a very important role in the deep learning era. In this work, we collect more than 100k DVS event sequences to meet the needs for large-scale training and evaluation of HAR. **2). Wide varieties:** Thousands of human activities can exist in the real world, but existing DVS-based HAR datasets only contain limited categories. Therefore, it is hard to fully reflect the classification and recognition ability of HAR algorithms. Our newly proposed HARDVS contains 300 classes which are several times larger than other DVS datasets. **3). Different capture distances:** The HARDVS dataset is collected under various distances, i.e., 1-2, 3-4, and more than 5 meters. **4). Long-term:** Most of the existing DVS-based HAR datasets are microsecond-level, in contrast, each video in our HARDVS dataset lasts for about 5 seconds. **5). Dual-modality:** The DAVIS346 camera can output both RGB frames and event flow, therefore, our dataset can also be used for HAR by fusing video frames and events. In this work, we focus on HAR with DVS only, but the RGB frames will also be released to support the research on dual-modality fusing based HAR.

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:* We collect different views of the same behavior to mimic practical applications, including front-, side-, horizontal-, top-down-, and bottom-up-views. *(b). Multi-illumination:* High dynamic range is one of the most important features of DVS sensors, therefore, we collect the videos under scenarios with strong-, middle-, and low-light. Note that, 60% of each category are videos with low-light. Our dataset also contains many videos with *glitter*, because we find that the DVS sensor is sensitive to flashing lights, especially at night. *(c). Multi-motion:* We also highlight the features of DVS sensors by recording many actions with various motion speeds, such as slow-, moderate-, and high-speed. *(d). Dynamic background:* As it is relatively easy to recognize actions without background objects, i.e., stationary DVS camera, we also collect many actions with a dynamic moving camera to make our dataset challenging enough. *(e). Occlusion:* In the real world, human actions can be occluded commonly and this challenge is also considered in our dataset.

Data Collection and Statistic Analysis The HARDVS dataset is collected with a DAVIS346 camera whose resolution is 346×260. There is a total of five persons involved in the data collection stage. From a statistical perspective, our dataset contains a total of $107,646$ video sequences and 300 classes of common human activities. We split $60\%, 10\%,$

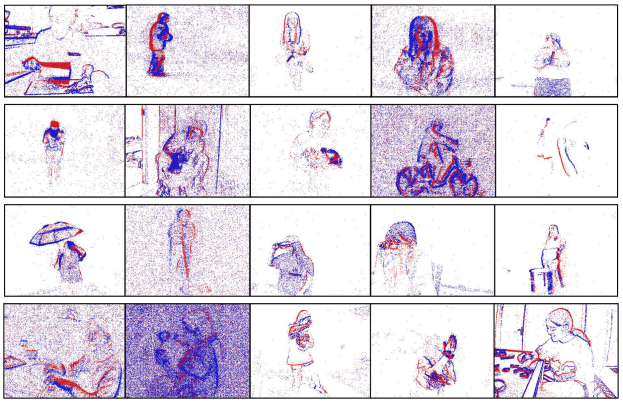


Figure 2: Illustration of some representative samples of our proposed HARDVS dataset.

and 30% of each category for training, validating, and testing, respectively. Totally, the number of videos in the training, validating, and testing subset is 64526|10734|32386, respectively. With the aforementioned characteristics, we believe our HARDVS dataset will be a better evaluation platform for the neuromorphic classification problem, especially for the human activity recognition task.

Methodology

Overview In this section, we devise a new Event-based Spatial-Temporal Transformer (ESTF) approach for event-stream data learning. As shown in Fig. 3, the proposed ESTF architecture contains three main learning modules, i.e., i) Initial Spatial and Temporal Embedding, ii) Spatial and Temporal Enhancement Learning, and iii) Spatial-Temporal Fusion Transformer. Specifically, 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 (FusionFormer) block is designed to integrate the spatial and temporal cues together for the final feature representation. The details of these modules are introduced below.

Initial Spatial and Temporal Embedding Different from frame-based sensors which capture a global image at each time, the event cameras asynchronously capture the intensity variations in the log-scale. That is, each pixel outputs a discrete event (or spike) independently when the visual change exceeds a pre-defined threshold. Usually, we use a 4-tuple $\{x, y, t, p\}$ to represent the discrete event of a pixel captured with DVS, where x, y are spatial coordinates, t is the timestamp, and $p \in \{1, -1\}$ is the polarity of brightness variation. Following previous works (Wang et al. 2019b; Zhu and Yuan 2018; Fang et al. 2021; Yao et al. 2021), 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. Let $\mathcal{E} = \{E_1, E_2 \cdots E_T\} \in \mathbb{R}^{H \times W \times T}$ be the collection of the sampled input event frames. Following existing work (Tran et al. 2015), we set $T = 8$ in our experiments. For each event frame E_t, we adopt StemNet (the ResNet-18 (He et al. 2016) is selected in our experiments) to extract an initial CNN feature descriptor for it and denote $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2 \cdots \mathcal{X}_T\} \in \mathbb{R}^{T \times h \times w \times c}$ as the collection of T event frames. Based on it, we respectively extract spatial and temporal embeddings. To be specific, for the temporal branch, we adopt a convolution layer to reduce the feature size to obtain $\mathcal{X}^t \in \mathbb{R}^{T \times \frac{h}{2} \times \frac{w}{2} \times c'}$ and reshape it to the matrix form as $X^t \in \mathbb{R}^{T \times d}$, 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 $\mathcal{X}$ to $\mathcal{X}^s \in \mathbb{R}^{T \times \frac{h}{2} \times \frac{w}{2} \times d}$. Then, we conduct the merging/summation operation on the time dimension and reshape it to the matrix form $X^s \in \mathbb{R}^{N \times d}$ where $N = \frac{hw}{4}$. Hence, both spatial and temporal embeddings have the same d-dim feature descriptors.

Spatial and Temporal Enhancement Learning Based on the above initial spatial embeddings $X^s \in \mathbb{R}^{N \times d}$ and temporal embeddings $X^t \in \mathbb{R}^{T \times d}$, we then devise our Spatial and Temporal Enhancement Learning (STEL) module to further enrich their representations. 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. A residual connection is also employed, as shown in Fig. 3. To be specific, given spatial embeddings $X^s \in \mathbb{R}^{N \times d}$, we first incorporate the position encoding (Dosovitskiy et al. 2020) to obtain $\bar{X}^s \in \mathbb{R}^{N \times d}$ which represents N number of the input tokens with d-dim feature descriptor. Then, the outputs of SF block are summarized as follows,

$$Y^s = LN(\bar{X}^s + MSA(LN(\bar{X}^s))) \quad (1)$$

$$\widetilde{X}^s = Y^s + MLP(Y^s) \quad (2)$$

In contrast to input $\bar{X}^s$, the output $\widetilde{X}^s$ provides the spatial-aware enhanced representations by employing the MSA mechanism to model the spatial relationships of different event patches. Similarly, given $\bar{X}^t \in \mathbb{R}^{T \times d}$ representing T temporal tokens with position encoding, the outputs of TF block are summarized as follows,

$$Y^t = LN(\bar{X}^t + MSA(LN(\bar{X}^t))) \quad (3)$$

$$\widetilde{X}^t = Y^t + MLP(Y^t) \quad (4)$$

Compared with the input $\bar{X}^t$, the outputs $\widetilde{X}^t \in \mathbb{R}^{T \times d}$ provide a temporal-context enhanced representations for T number of frame tokens thanks to the MSA mechanism to model the dependencies of different event frames.

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 (FusionF) module. To be specific, let $\widetilde{X}^s$ and $\widetilde{X}^t$ denote the outputs of previous SF and TF blocks respectively. We first collect the N spatial and T temporal tokens together and feed them to a unified Transformer block

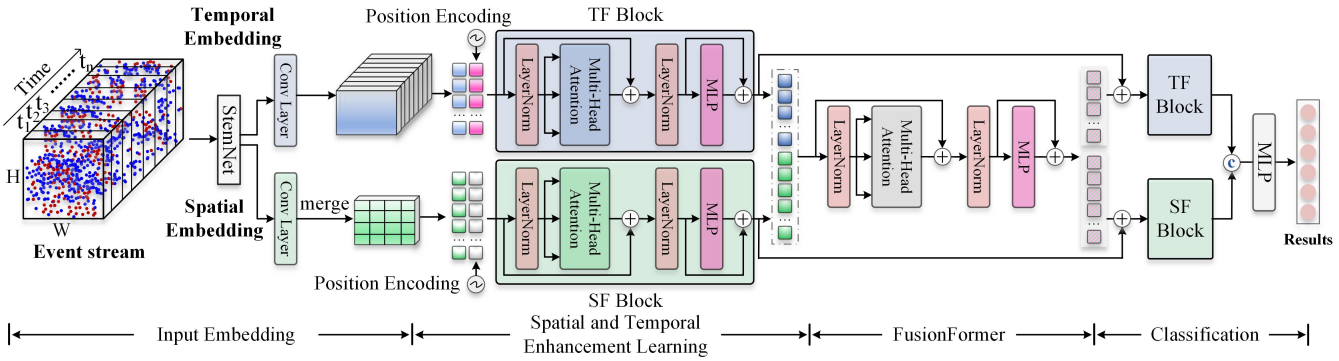


Figure 3: An overview of our proposed ESTF framework for event-based human action recognition. It transforms the event streams into spatial and temporal tokens and learns the dual features using multi-head self-attention layers. Further, a FusionFormer is proposed to realize message passing between the spatial and temporal features. The aggregated features are added with dual features as the input for subsequent TF and SF blocks, respectively. The outputs will be concatenated and fed into MLP layers for action prediction.

which includes multi-head self-attention (MSA) and MLP submodule, i.e.,

$$Z = [\widetilde{X}^t, \widetilde{X}^s] \in \mathbb{R}^{(T+N)\times c} \tag{5}$$

$$Y = LN(Z + MSA(LN(Z))) \tag{6}$$

$$\widetilde{Z} = Z + Y + MLP(Y) \tag{7}$$

Afterword, we split $\widetilde{Z}$ into $\{\widetilde{Z}^s, \widetilde{Z}^t\}$ where $\widetilde{Z}^s \in \mathbb{R}^{N\times d}$ and $\widetilde{Z}^t \in \mathbb{R}^{T\times d}$ and further employ the above SF (Eqs.(1,2)) and TF (Eqs.(3,4)) block to respectively enhance their representations as follows,

$$F^s = SF(\widetilde{Z}^s), F^t = TF(\widetilde{Z}^t) \tag{8}$$

Finally, we concatenate both F^s and F^t together and reshape the concatenated features to the vector form. After that, we utilize a two-layer MLP to output the final class label prediction, as shown in Fig. 3.

Loss Function Our proposed ESTF framework can be optimized in an end-to-end way. The standard cross-entropy loss function is adopted to measure the distance between our model prediction and ground truth as,

$$Loss = -\frac{1}{B}\sum_{b=1}^{B}\sum_{n=1}^{N} Y_{bn} log P_{bn} \tag{9}$$

where B denotes the batch size and N denotes the number of event classes. Y and P represent the ground truth and predicted class labels of the event sample, respectively.

Experiments

Dataset and Evaluation Metrics In this work, three datasets are adopted for the evaluation of our proposed model, including **N-Caltech101** (Orchard et al. 2015), **ASL-DVS** (Bi et al. 2020), and our newly proposed **HARDVS**. More details about these datasets can be found in Table 1. The widely used **top-1** and **top-5 accuracy** are adopted as evaluation metrics.

RG-CNNs	VMV-GCN	EV-VGCNN	EST
0.657	**0.778**	0.748	0.753
M-LSTM	**AMAE**	**HATS**	**Ours**
0.738	0.694	0.642	**0.832**

Table 2: Results on N-Caltech101 (Orchard et al. 2015).

Comparison with SOTA Algorithms

Results on N-Caltech101 (Orchard et al. 2015). As shown in Table 2, our proposed method achieves 0.832 on the top-1 accuracy metric which is significantly better than the compared models by a large margin. For example, the VMV-GCN achieves 0.778 on this benchmark dataset which ranks second place, meanwhile, our model outperforms it by up to 5.4%. The M-LSTM is an adaptive event representation learning model which obtained 0.738 only on this dataset. EV-VGCNN is a graph neural network based model which obtains 0.748 and is also worse than ours. These experimental results fully demonstrate the effectiveness of our proposed spatial-temporal feature learning for event-based pattern recognition.

Results on ASL-DVS (Bi et al. 2020). As shown in Table 3, the performance on this dataset is already close to saturation and most of the compared models achieve more than 0.95+ on the top-1 accuracy. Note that, the VMV-GCN (Xie et al. 2022) achieves 0.989 on this benchmark dataset which ranks the second place. Thanks to our proposed spatial-temporal feature learning and fusion modules, we set new state-of-the-art performance on this dataset, i.e., 0.999 on the top-1 accuracy. Therefore, we can conclude that our method almost completely solves the simple gesture recognition problem defined in the ASL-DVS.

Results on HARDVS. From the experimental results reported in the ASL-DVS (Bi et al. 2020) and N-Caltech101 (Orchard et al. 2015), we can find that existing event-based recognition datasets are almost saturated. The newly proposed HARDVS dataset can bridge this gap

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

Table 3: Results on the ASL-DVS (Bi et al. 2020) dataset.

Algorithm	Event		MAC	Parameter
ResNet18	49.20	56.09	17.2G	11.7M
C3D	50.52	56.14	0.2G	147.2M
R2Plus1D	49.06	56.43	40.7G	63.5M
TSM	52.63	60.56	0.7G	24.3M
ACTION-Net	46.85	56.19	34.7G	27.9M
TAM	50.41	57.99	33.1G	25.6M
V-SwinTrans	51.91	59.11	17.5G	27.8M
TimeSformer	50.77	58.70	107.3G	121.2M
SlowFast	50.63	57.77	0.7G	33.6M
ESTF (Ours)	51.22	57.53	17.6G	46.1M

Table 4: Results on the newly proposed HARDVS dataset.

and further boost the development of event-based human action recognition. As shown in Table 4, we re-train and test multiple state-of-the-art models for future works to compare on the HARDVS benchmark dataset, including C3D (Tran et al. 2015), R2Plus1D (Tran et al. 2018), TSM (Song et al. 2019), ACTION-Net (Wang, She, and Smolic 2021), TAM (Liu et al. 2021b), Video-SwinTrans (Liu et al. 2021a), TimeSformer (Bertasius, Wang, and Torresani 2021), SlowFast (Feichtenhofer et al. 2019). It is easy to find that these popular and strong recognition models still perform poorly on our newly proposed HARDVS dataset. To be specific, the R2Plus1D (Tran et al. 2018), ACTION-Net (Wang, She, and Smolic 2021), and SlowFast (Feichtenhofer et al. 2019) only achieve $49.06|56.43$, $46.85|56.19$, and $50.63|57.77$ on the top-1 and top-5 accuracy respectively. The recently proposed TAM (Liu et al. 2021b) (ICCV-2021), Video-SwinTrans (Liu et al. 2021a) (CVPR-2022), TimeSformer (Bertasius, Wang, and Torresani 2021) (ICML 2021) also obtain $50.41|57.99$, $51.91|59.11$, and $50.77|58.70$ on the two metrics respectively. Compared with these models, our proposed spatial-temporal feature learning and fusion modules perform comparable to or even better than these SOTA models, i.e., $51.22|57.53$. It is because our proposed spatial and temporal feature enhancement module works well in capturing motion cues in the event streams. Overall, our proposed model is effective for event-based human action recognition tasks and may be a good baseline for future works to compare.

Ablation Study

Component Analysis As shown in Table 5, three main modules are analyzed on the N-Caltech101 and HARDVS datasets, including SpatialFormer (SF), TemporalFormer (TF), and FusionFormer. For the N-Caltech101, we can find that our baseline method ResNet18 (He et al. 2016) achieves 72.14 on the top-1 accuracy metric. When introducing the TemporalFormer (TF) into the recognition framework, the overall performance can be significantly improved by $+9.4$, and achieves 81.54. When the SpatialFormer (SF) is adopted for long-range global feature relation mining, the recognition results can be enhanced to 80.47, and the improvement is up to $+8.33$. When both modules are all utilized for joint spatial-temporal feature learning, a better result can be obtained, i.e., 82.89. If the FusionFormer is adopted to achieve interactive feature learning and information propagation between the spatial and temporal Transformer branches, the best results can be achieved, i.e., 83.17 on the top-1 accuracy. Similar conclusions can be found from the experimental results of HARDVS dataset. Based on the experimental analysis for Table 5 and Table 2, we can draw the conclusion that our proposed modules all contribute to final recognition results.

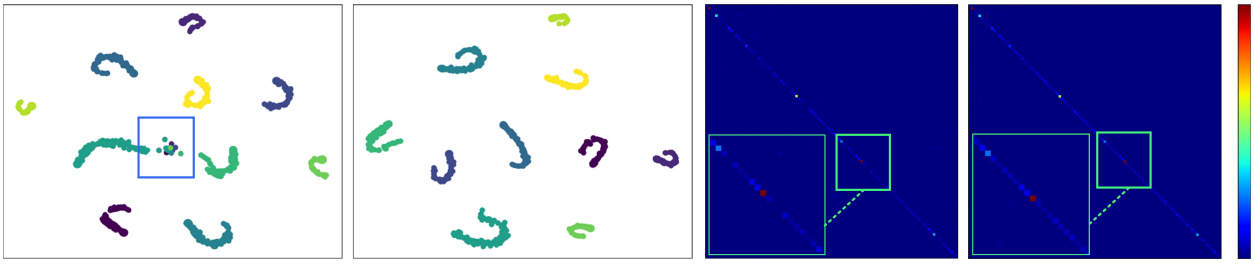
ResNet	TF	SF	FF	N-Caltech101	HARDVS
✓				72.14	49.20
✓	✓			81.54	49.65
✓		✓		80.47	50.81
✓	✓	✓		82.89	51.06
✓	✓	✓	✓	83.17	51.22

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

Analysis on Number of Input Frames In this paper, we transform the event streams into an image-like representation for classification. In our experiments, 8 frames are adopted for the evaluation of our model. Actually, various event frames can be obtained with different intervals of the time windows. In this part, we test our model with 4, 6, 8, 10, 12, and 16 frames on the N-Caltech101 dataset and report the results in Fig. 5. It is easy to find that the mean accuracy is 73.67, 75.94, 77.37, 73.49, 75.11, and 73.03, correspondingly, and the highest mean accuracy can be obtained when 8 frames are adopted. For the decrease in accuracy when the frames are larger than 8, we think this may be caused by the fact that the event streams are partitioned into more frames and each frame will be more sparse. Therefore, this will lead to sparse edge information which is very important for recognition.

Analysis on Split Patches of Spatial Data In this paper, the spatial features are partitioned into non-overlapped patches. We test multiple scales in this subsection, including 8×8, 6×6, and 4×4. As illustrated in Fig. 5 (right), the best performance can be obtained when 4×4 is adopted, i.e., 83.17, 94.20, and 77.37 on the top-1, top-5, and mean accuracy respectively.

Analysis on Layers of Transformer Layers As we all know, the self-attention or Transformer layers can be stacked multiple times for more accurate recognition, as validated in many works. In this experiment, we also test different Transformer layers to check their influence on our model. As shown in Fig. 5 (middle), four different settings are tested, i.e., 1, 2, 3, and 4 layers, and the corresponding mean accuracy is 77.37, 75.20, 76.05, and 74.64. We can find that higher recognition results can be obtained when the Transformer is set as 1 to 3 layers. Maybe a larger dataset is



(a) ResNet (HARDVS) (b) Ours (HARDVS) (c) ResNet (N-Caltech101) (d) Ours (N-Caltech101)

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). Best viewed by zooming in.

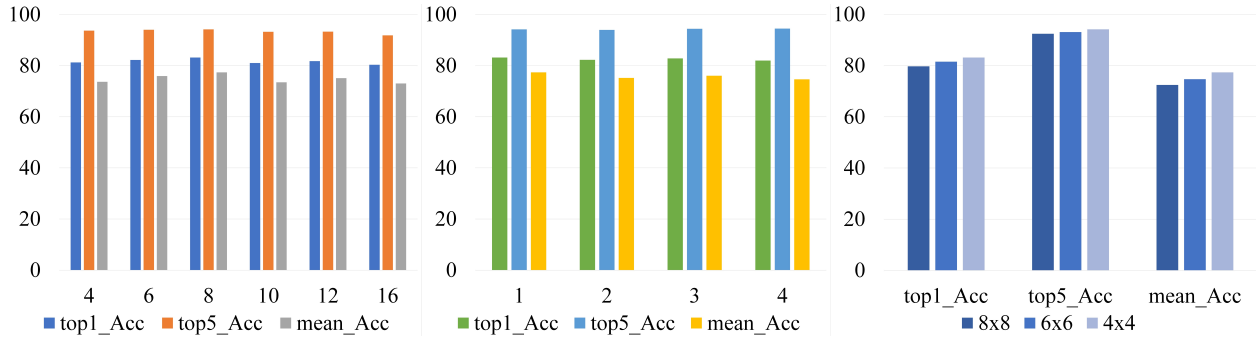


Figure 5: Results of different (left) input frames; (middle) transformer layers; (right) patch sizes on the HARDVS dataset.

needed to train deeper Transformer layers.

Model Parameters and Running Efficiency The storage space occupied by our checkpoint is 377.34 MB and the number of parameters is 46.71 M. The MAC is 17.62 G tested using toolkit *ptflops* [2]. Our model spends 25 ms for each video (8 frames used) in our proposed HARDVS dataset on a server with GPU RTX-3090.

Visualization

As shown in Fig. 4 (a, b), we select 10 classes of actions defined in the HARDVS dataset and visualize the features by projecting them into 2D plane using *tSNE* toolkit [3]. It is easy to find that partial data samples are not discriminated well using the baseline ResNet18, such as the regions highlighted in blue bounding box. In contrast, our proposed ESTF model achieves a better feature representation learning and more categories are classified well. For the confusion matrix on N-Caltech101 dataset, as shown in Fig. 4 (c, d), we can find that our proposed ESTF achieves significant improvement compared with our baseline ResNet18. All in all, we can draw the conclusion that our proposed spatial-temporal feature learning method works well for event-based action recognition.

[2]https://pypi.org/project/ptflops/

[3]https://github.com/mxl1990/tsne-pytorch

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. These videos reflect various views, illuminations, motions, dynamic backgrounds, occlusion, etc. More than 10 popular and recent classification models are evaluated for future works to conduct comparisons. 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 ALS-DVS datasets. We hope the proposed dataset and baseline approach will boost the further development of event camera based human action recognition.

Acknowledgements

This work is supported by National Natural Science Foundation of China (Grants 62102205, 62236009, U22A20103, 62302041, 62332002, 62027804, 62088102), National Science Foundation for Distinguished Young Scholars (62325603), Anhui Provincial Key Research and Development Program under Grant 2022i01020014, Natural Science Foundation of Anhui Province under Grant 2108085Y23, China National Postdoctoral Program for Innovative Talents under contract No. BX20230469. Multi-source Cross-

platform Video Analysis and Understanding for Intelligent Perception in Smart City U20B2052, Peng Cheng Laboratory Research Project No. PCL2023A08. The authors acknowledge the High-performance Computing Platform of Anhui University for providing computing resources.

References

Ahmad, T.; Jin, L.; Zhang, X.; Lin, L.; and Tang, G. 2021. Graph Convolutional Neural Network for Action Recognition: A Comprehensive Survey. *IEEE Transactions on Artificial Intelligence*.

Amir, A.; Taba, B.; Berg, D.; Melano, T.; McKinstry, J.; Di Nolfo, C.; Nayak, T.; Andreopoulos, A.; Garreau, G.; Mendoza, M.; et al. 2017. A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7243–7252.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is space-time attention all you need for video understanding? In *ICML*, volume 2, 4.

Bi, Y.; Chadha, A.; Abbas, A.; Bourtsoulatze, E.; and Andreopoulos, Y. 2020. Graph-based spatio-temporal feature learning for neuromorphic vision sensing. *IEEE Transactions on Image Processing*, 29: 9084–9098.

Brandli, C.; Berner, R.; Yang, M.; Liu, S.-C.; and Delbruck, T. 2014. A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10): 2333–2341.

Cannici, M.; Plizzari, C.; Planamente, M.; Ciccone, M.; Bottino, A.; Caputo, B.; and Matteucci, M. 2021. N-rod: A neuromorphic dataset for synthetic-to-real domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1342–1347.

Chen, G.; Xu, Z.; Li, Z.; Tang, H.; Qu, S.; Ren, K.; and Knoll, A. 2021. A novel illumination-robust hand gesture recognition system with event-based neuromorphic vision sensor. *IEEE Transactions on Automation Science and Engineering*, 18(2): 508–520.

Chen, H.; Liu, W.; Goel, R.; Lua, R. C.; Mittal, S.; Huang, Y.; Veeraraghavan, A.; and Patel, A. B. 2019. Fast retinomorphic event-driven representations for video gameplay and action recognition. *IEEE Transactions on Computational Imaging*, 6: 276–290.

Chen, S.; and Guo, M. 2019. Live demonstration: CeleX-V: a 1M pixel multi-mode event-based sensor. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1682–1683. IEEE.

Clady, X.; Maro, J.-M.; Barré, S.; and Benosman, R. B. 2017. A motion-based feature for event-based pattern recognition. *Frontiers in neuroscience*, 10: 594.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Fang, W.; Yu, Z.; Chen, Y.; Huang, T.; Masquelier, T.; and Tian, Y. 2021. Deep residual learning in spiking neural networks. *NeurIPS*.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6202–6211.

Gallego, G.; Delbrück, T.; Orchard, G.; Bartolozzi, C.; Taba, B.; Censi, A.; Leutenegger, S.; Davison, A. J.; Conradt, J.; Daniilidis, K.; et al. 2020. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1): 154–180.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Kliper-Gross, O.; Hassner, T.; and Wolf, L. 2011. The action similarity labeling challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3): 615–621.

Kong, Y.; and Fu, Y. 2018. Human action recognition and prediction: A survey. *arXiv preprint arXiv:1806.11230*.

Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; and Serre, T. 2011. HMDB: a large video database for human motion recognition. In *2011 International conference on computer vision*, 2556–2563. IEEE.

Li, H.; Liu, H.; Ji, X.; Li, G.; and Shi, L. 2017. Cifar10-dvs: an event-stream dataset for object classification. *Frontiers in neuroscience*, 11: 309.

Li, J.; Wang, X.; Zhu, L.; Li, J.; Huang, T.; and Tian, Y. 2022. Retinomorphic Object Detection in Asynchronous Visual Streams. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, February 22-March 1, 2022*, 1332–1340. AAAI Press.

Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; and Hu, H. 2021a. Video swin transformer. *arXiv preprint arXiv:2106.13230*.

Liu, Z.; Wang, L.; Wu, W.; Qian, C.; and Lu, T. 2021b. Tam: Temporal adaptive module for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13708–13718.

Orchard, G.; Jayawant, A.; Cohen, G. K.; and Thakor, N. 2015. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in neuroscience*, 9: 437.

Planamente, M.; Plizzari, C.; Cannici, M.; Ciccone, M.; Strada, F.; Bottino, A.; Matteucci, M.; and Caputo, B. 2021. Da4event: towards bridging the sim-to-real gap for event cameras using domain adaptation. *IEEE Robotics and Automation Letters*, 6(4): 6616–6623.

Posch, C.; Matolin, D.; and Wohlgenannt, R. 2010. 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*, 46(1): 259–275.

Serrano-Gotarredona, T.; and Linares-Barranco, B. 2015. Poker-DVS and MNIST-DVS. Their history, how they were made, and other details. *Frontiers in neuroscience*, 9: 481.

Sironi, A.; Brambilla, M.; Bourdis, N.; Lagorce, X.; and Benosman, R. 2018. HATS: Histograms of averaged time surfaces for robust event-based object classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1731–1740.

Song, X.; Lan, C.; Zeng, W.; Xing, J.; Sun, X.; and Yang, J. 2019. Temporal–spatial mapping for action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(3): 748–759.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

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

Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; and Paluri, M. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 6450–6459.

Wang, Q.; Zhang, Y.; Yuan, J.; and Lu, Y. 2019a. Space-time event clouds for gesture recognition: From RGB cameras to event cameras. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1826–1835. IEEE.

Wang, X.; Li, J.; Zhu, L.; Zhang, Z.; Chen, Z.; Li, X.; Wang, Y.; Tian, Y.; and Wu, F. 2021a. VisEvent: Reliable Object Tracking via Collaboration of Frame and Event Flows. *arXiv preprint arXiv:2108.05015*.

Wang, Y.; Du, B.; Shen, Y.; Wu, K.; Zhao, G.; Sun, J.; and Wen, H. 2019b. EV-gait: Event-based robust gait recognition using dynamic vision sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6358–6367.

Wang, Y.; Zhang, X.; Shen, Y.; Du, B.; Zhao, G.; Lizhen, L. C. C.; and Wen, H. 2021b. Event-stream representation for human gaits identification using deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wang, Z.; She, Q.; and Smolic, A. 2021. ACTION-Net: Multipath Excitation for Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13214–13223.

Xie, B.; Deng, Y.; Shao, Z.; Liu, H.; and Li, Y. 2022. VMV-GCN: Volumetric Multi-View Based Graph CNN for Event Stream Classification. *IEEE Robotics and Automation Letters*, 7(2): 1976–1983.

Xing, Y.; Di Caterina, G.; and Soraghan, J. 2020. A new spiking convolutional recurrent neural network (SCRNN) with applications to event-based hand gesture recognition. *Frontiers in Neuroscience*, 14: 1143.

Yao, M.; Gao, H.; Zhao, G.; Wang, D.; Lin, Y.; Yang, Z.; and Li, G. 2021. Temporal-wise Attention Spiking Neural Networks for Event Streams Classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10221–10230.

Zhu, A. Z.; and Yuan, L. 2018. EV-FlowNet: Self-Supervised Optical Flow Estimation for Event-based Cameras. In *Robotics: Science and Systems*.

Zhu, L.; Li, J.; Wang, X.; Huang, T.; and Tian, Y. 2021. NeuSpike-Net: High Speed Video Reconstruction via Bio-inspired Neuromorphic Cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2400–2409.

Zhu, L.; Wang, X.; Chang, Y.; Li, J.; Huang, T.; and Tian, Y. 2022. Event-based Video Reconstruction via Potential-assisted Spiking Neural Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3594–3604.