TIME-E2V: Overcoming limitations of E2VID

Mira Adra
GTD International
mira.adra@gtd.eu

Jean-Luc Dugelay
Eurecom
jld@eurecom.fr

Abstract

In the field of action recognition, event cameras have marked a breakthrough by capturing motion dynamics beyond the capability of traditional cameras, thanks to their high temporal sensitivity. However, the asynchronous and sparse nature of event data challenges their use with traditional convolutional neural networks (CNNs). The E2VID model offers a solution by transforming event data into continuous video frames, enabling the use of standard CNNs for event-based data analysis. However, it struggles with accurately capturing motion speed variations and pauses, limiting its effectiveness in scenarios where temporal dynamics are crucial. In response, we introduce TIME-E2V, which integrates spatial embeddings from E2VID with LSTM-derived temporal embeddings from frame timestamps. This combination is processed by a modified 3D convolutional network (C3D), leveraging its inherent strengths in video analysis. Our proposed approach not only overcomes E2VID's challenges but also delivers competitive performance across a wide range of dynamic scenes with the leading action recognition networks for event cameras, including those based on Spiking Neural Networks.

1. Introduction

A new era in action recognition was brought about by the emergence of event cameras, which provided unmatched temporal resolution and spatial dynamics. Unlike traditional cameras, these neuromorphic sensors excel in capturing movement dynamics due to their high temporal and spatial dimensions, making them particularly suited for applications requiring detailed motion analysis and low latency. Their event-based data representation, triggered by changes in pixel brightness, not only ensures efficient data processing but also a heightened level of security, a critical advantage for battery-powered devices or high security applications.

In the past, researchers have focused on developing neural network architectures that are suitable for the sparse and asynchronous nature of event camera data. Early endeavours favoured Graph CNNs [16], due to their proficiency in maintaining the spatiotemporal integrity of such data. However, they faced challenges in scalability and computational efficiency when processing the high-dimensional event stream. This motivated researchers to investigate alternative techniques including 2D frame reconstructions, to bridge the gap and allow the integration of event data with conventional computer vision algorithms. Despite various attempts at these reconstructions—ranging from histograms to time surfaces— they often result in a loss of spatial or temporal resolution and their deviation from conventional image representations complicates their adoption in pre-existing computer vision frameworks, which are predominantly designed for conventional video or image data. The search for a more direct and efficient processing method led to the exploration of Spiking Neural Networks (SNNs) [13] which mimic the operational principle of the human brain, processing data in a discrete manner through spikes, which aligns perfectly with the nature of the event-based data generated by neuromorphic cameras and allows their direct processing as spike tensors. A recent breakthrough came with the proposal of a transformer-based architecture for action recognition [3], setting new benchmarks in accuracy within this domain. However, it has a significant trade-off between model performance and complexity.

A particularly noteworthy advancement is the development of the E2VID model [12]. It aims to bridge the gap in event-based research by converting the event stream into grayscale video sequences with high temporal resolution. This approach facilitates the application of state-of-the-art video processing algorithms, such as the 3D convolutional (C3D) architecture, to event data, thus leveraging the advanced capabilities of these models for action recognition tasks. The E2VID ability to generate high-quality, temporally consistent video frames from event data while preserving the intrinsic advantages of event cameras has not only positioned it as a superior alternative to SNNs, but also enabled it to achieve competitive performance. However, the model's limitations in accurately representing action speeds

and capturing pauses in motion point to the need for further innovation.

In this paper, we primarily aim to highlight the impact of the E2VID model in overcoming the limited research done in the event domain by finding a representation that allows us to leverage the traditional video-based convolutional networks for action recognition. What we do differently from the original paper is that we also venture into an unexplored downstream of E2VID in action recognition. In that context, we aim to call attention to a major limitation in the model —its compromised ability to depict action speed— and propose a novel dual-channel architecture designed to overcome this issue. In summary, our work offers these key contributions:

- We verify the efficiency of the E2VID model, but also identify its inherent limitations, particularly its failure to accurately capture action speeds within reconstructed videos.
- We introduce an innovative architecture that integrates the C3D action recognition model with an additional branch for temporal embeddings, demonstrably surpassing the performance of SNNs dedicated to event data processing.
- Through comprehensive experimentation, we define the optimal event sampling parameters and ensure our model's robustness against the grayscale nature of the reconstructed frames, thereby paving the way for more accurate and versatile action recognition applications.

2. Related Works

The emergence of event cameras—which stand out for their great spatiotemporal resolution and energy efficiency—has set off a paradigm shift in the analysis of motion dynamics. Their behaviour is said to be motion-centric as they only capture the dynamics in a scene thus removing background redundancy, saving energy, and economising on the data size as pixels are triggered only when there is change in brightness level. When compared to standard imaging devices, these neuromorphic sensors [9] offer a greater dynamic range and responsiveness. However, the distinctive nature of their data they produce poses compatibility challenges with traditional networks designed for the visible domain and thus necessitates novel processing approaches to fully realise its potential.

Early attempts to convert event data into a format compatible with standard video processing techniques primarily employed one of these two representations: Time surfaces [8] which suffered from temporal aliasing as the sampling rate is too low to capture the rapid changes in motion and histogram of events image [6] which was limited by the inability to convey the exact temporal order of events. More recent approaches like the one demonstrated by Zhu et al. [17], proceeds to aggregate events in a three-dimensional space-time volume for unsupervised learning of optical flow from event data. Despite preserving some temporal information, they often fail to reflect the continuous flow of events due to their fixed bin sizes.

A more promising path toward enhanced event data processing was to design a network that is compatible with the asynchronous nature of event data rather than trying to integrate event data with traditional architectures. In light of that, Tavanaei et al. [13] proposed the Spiking Neural Networks built upon the Integrate and Fire neuron which allows them by nature to process information as temporal spikes instead of numeric values. This spike-based computational model not only has the potential to process spatiotemporal data but also allows us to feed the event data directly into the network as Spike Tensors without pre-processing. While promising for object detection [2] and facial expression recognition [1] tasks, their application to action recognition, which involves complex temporal patterns, remains challenging. These tasks require advanced temporal encoding and decoding mechanisms—areas where SNN research is still developing.

Acknowledging the limitations of SNNs, Rebecq et al. [12] revisit the event-to-image reconstruction approach and try to bridge the gap between event data and traditional CNNs as efficiently as possible. They proposed a transformative solution to event-based video reconstruction, utilizing a neural network to reconstruct video frames from event data. This model successfully preserved the spatial and temporal integrity of the scenes, setting a new benchmark for the field. However, it faces challenges with processing rapid motion and motion pauses. One effective solution for preserving temporal embeddings in traditional CNN models was the one adopted by Wang et al. [15], which is the introduction of Recurrent Neural Networks (RNN), particularly the Long short-term memory (LSTM) network, as a temporal feature extractor after the C3D network. This approach is what inspired our model architecture, however, we modify the placement of the LSTM in our model according to our goal. Our TIME-E2V model is developed in response to these findings, aiming to overcome the specific limitations identified in methods like E2VID. To our knowledge, this is the first demonstration of downstream applications focused on temporal analysis for action recognition using event data.

3. Limitations of the E2VID model

Even though the E2VID model represents a significant step forward in event-to-video reconstruction and demonstrates the feasibility of preserving temporal and spatial details within reconstructed videos, a thorough evaluation of its performance points to a significant drawback in its ability to preserve time, which manifests in two notable cases.

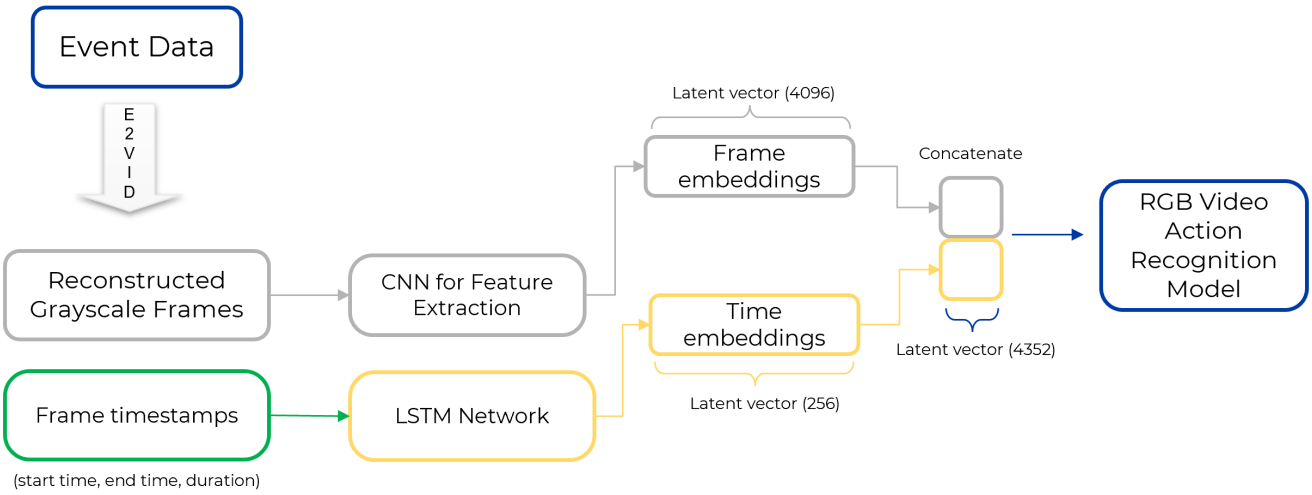

Figure 1. Illustration of the proposed model.

First, it is unable to accurately represent the speed of actions, which is crucial for distinguishing between similar activities, such as running versus walking. Second, this limitation is further highlighted by the model's struggle to capture pauses in motion when employing a fixed event sampling rate for frame generation, as a pause can be considered an extremely slow action that the model fails to record accurately.

In order to provide a clearer understanding of the E2VID model's capabilities, we explore the two fundamental approaches it uses for frame reconstruction: Fixed Time Duration and Fixed Number of Events. The latter results in an asynchronous frame sequence, producing a higher number of frames during periods of intense motion and fewer when there is little movement. On the other hand, the Fixed Time Duration method produces frames within set time windows, regardless of whether or not there is action in the frames. Accordingly, a thorough comparative study was performed to determine which method most effectively captures the temporal nuances essential for accurate action recognition. This study will be detailed in the subsequent 'Experimental Results' section, following an introduction to the datasets used for benchmarking our model. Notably, we have verified that the most optimal approach for our application is the Fixed Number of Event approach with a sampling rate of 0.25 events per pixel per event window.Therefore, we have deliberately chosen this sampling technique for all subsequent analysis and methodology in the paper.

Consequently, after obtaining the reconstructed video frames using the Fixed Number of Events sampling approach, we generate a video at a rate of 30 frames per second[1]. Upon comparison with its visible counterpart, we realise that the reconstructed video from event frames appears to be in slow motion where an action of 7 seconds in the visible domain is represented as 17 seconds in the reconstructed domain. This proves that the E2VID does not conserve speed. However, on the bright side, this sampling approach generates a higher number of event frames than that with fixed time duration which means it preserves more details about the pattern of the motion.

In another experiment, we recorded a video of a person walking, stopping for a few seconds, then proceeding to walk again. Using the temporal approach, the stop in motion is translated into consecutive frames where the person stays in the same position indicating that he stopped. However, for the second approach, the frames show a continuous motion as the stop did not generate enough events for the model to accumulate in the event window for frame generation. Consequently, our analysis proves that the E2VID model has a significant shortcoming: it struggles with temporal representation, which affects its ability to correctly capture motion pauses and effectively represent real-time speed in reconstructed movies. These efforts underscore the ongoing need for models that can accurately represent the speed and subtlety of movements, a critical aspect for applications in action recognition. And, this is the challenge we aimed to overcome with our proposed model.

4. Methodology

4.1. Model Architecture

To address the challenge of the reconstructed event frames in accurately capturing temporal dynamics, we propose the TIME-E2V, an integrated model that combines spatial and temporal information to enhance action recognition accuracy. Figure 1 presents a schematic overview of our framework. At the base of our approach is the E2VID model, which takes as input the non-uniform event data

[1]Visual materials related to this paper will be made available upon acceptance.

stream and generates their corresponding grayscale video frames. These sequences are then sampled at a granularity of 30 frames per second, and passed through a feature extraction network, which consists of several layers of 3D convolutions and pooling layers, producing spatial embeddings that capture the scene's visual details. Simultaneously, we extract temporal embeddings by recording the start and end timestamp and calculating the duration between successive frames and saving them as tuples. This temporal data is processed through an LSTM network, which is a type of RNN, specifically designed to handle sequential data and can effectively capture long-term dependencies, making it ideal for modeling the temporal patterns inherent in the event data captured over time.

The resulting spatial and temporal embeddings are then concatenated along the feature dimension to form a unified spatiotemporal representation. This comprehensive embedding feeds into our action recognition model, which serves as the backbone of our method and is based on the C3D model introduced by Tran et al. [14] for video action recognition , which has been adapted to incorporate temporal data alongside the conventional frame input directly from raw video frames without requiring pre-segmented or handcrafted features. By integrating these temporal cues with spatial features, the model gains an enhanced ability to distinguish between actions, especially those that are closely related or that unfold over varying timescales.

This process not only preserves the spatial precision of the scenes captured by event cameras but also ensures compatibility with conventional video analysis techniques. By combining the strengths of E2VID with LSTM and a modified C3D model, our system sets a new standard for accuracy and efficiency in interpreting event camera data, marking a pivotal step forward in the domain of neuromorphic computing in the domain of motion dynamics.

4.2. Datasets

In order to verify the results of our model we rely on some benchmark datasets for action recognition, revisiting them with a novel perspective aligned with our current objectives. These include the Gait3 Dataset [4], adapted specifically for this study, alongside the TUM Action Recognition [11] and the DailyAction-DVS [10] datasets, each chosen for their relevance and contribution to the domain of action recognition.

Gait3 dataset: This dataset was originally recorded for the purpose of gait recognition and was modified for evaluating the model's performance on speed-characterized actions. Initially containing 3 classes : walking, running, and walking with backpack, we split this dataset into two parts: Gait3 Speed dataset with the classes walking and running to assess the model's ability to differentiate between the same action done in two speeds. And Gait3 Backpack, which aims at evaluating the details of the action (whether or not the person has a backpack while walking). Both datasets have 18 minutes of data across 69 people in both visible and event domain pairs.

Despite the dataset's scope being limited to just two classes per dataset, it is highly relevant for this specific application. With data pairs in both the visible and event domains, it serves as an ideal testbed to specifically highlight the E2VID model's challenges in discerning motion speed and showcase how our model overcomes these limitations.

DailyAction-DVS: This dataset consists of a diverse range of 12 types of motions, from everyday tasks like walking to challenging moves like lifting and bending, captured under various circumstances and across 15 people. The footage, which has an average duration of 12 minutes per action, provides a strong basis for evaluating the effectiveness of our methodology in everyday scenarios.

ActionTUM: This dataset consists of 291 recordings of ten distinct activities performed by 15 persons, captured using a DAVIS camera from 3 different viewpoints. For every activity, there are about 2.5 minutes of footage. The TUM dataset, despite its smaller size, is essential for assessing the model's performance in a scenario with a constrained dataset since it provides information about the model's flexibility and generalization abilities.

4.3. Training Implementation

For the E2VID model, we use the original codebase [2] and generate frames with a fixed number of 0.25 events per pixel per event window. The total number of events per frame can then be calculated as the product of the height, width, and the number of events per pixel. The LSTM component of our network architecture is made up of 2 layers with 256 hidden layers each and configured with a bidirectional structure. To prevent overfitting, we also apply a dropout of 0.4 between the 2 LSTM layers.The resulting spatial embeddings (4096 dimensions) are concatenated with the temporal embeddings (256 dimensions) from the LSTM network to form a spatiotemporal embedding of size 4352 dimensions.

For the C3D model adaptation, which integrates temporal with spatial data, we initialize with pre-trained weights from Sports-1M dataset [7] and train for 100 epochs starting at a learning rate of 0.01 that reduces by a factor of 0.1 every 20 epochs. The C3D model is composed of 8 convolutional layers, 5 max-pooling layers, and 2 fully connected layers, followed by a softmax output layer. We also use the SGD optimizer that employs a dual learning rate strategy: base learning rate for most parameters and a tenfold increase for LSTM and the final layer parameters. Additionally, for comparative analysis, we train a Spiking-Element-

[2] https://github.com/uzh-rpg/rpg_e2vid

Wise ResNet model (SEW-ResNet) following the exact architecture proposed in [5] and implemented with SpikingJelly to enhance deep learning in SNNs by using residual connections.

5. Experimental results

5.1. Comparison of E2VID Sampling Methods

Building on the previously described frame reconstruction techniques of the E2VID model, we conducted a comprehensive series of tests to determine the optimal sampling technique for enhancing action recognition accuracy. Accordingly, we explored various sampling rates, including different fixed time intervals (15, 30, 60 fps) and fixed numbers of events per pixel (0.15, 0.25, 0.35). After generating the video frames from the event stream under these conditions, we assessed the model's accuracy on the two subdivisions of the Gait3 dataset thereby ensuring a thorough evaluation by encompassing a wide spectrum of spatiotemporal dynamics.

Sampling Rate	Gait3 Speed	Gait3 Backpack
15 fps	82.14%	92.85%
30 fps	77.14%	88.57%
60 fps	51.78%	76.78%

Table 1. Comparison of Model Accuracy Across Different Frame Rates.

Sampling Rate	Gait3 Speed	Gait3 Backpack
n = 0.15	85.71%	91.07%
n = 0.25	91.07%	98.21%
n = 0.35	96.42%	100%

Table 2. Comparison of Model Accuracy Across Different Number of Events per pixel.

As demonstrated in Table 1, for fixed time intervals, a lower frame rate yielded higher accuracy, as each frame covers a longer duration, capturing more events. In contrast, Table 2's outcomes reveal a complex correlation between sampling rate and model accuracy when considering fixed numbers of events per pixel. Event though a sample rate of 0.35 yields the best accuracy, 100% for Gait3 Backpack in particular, it also represents an extreme, at which any increases may result in declining gains. Although slightly less accurate, a rate of 0.25 could in fact provide a more balanced approach especially if we view this situation as setting a threshold rather than direct correlation. It maintains a good trade-off between detail and frame volume by maintaining a high level of detail without going over the point where there are too few frames.

In Table 3, we selected the best-performing sampling rate for each technique, as determined from Tables 1 and 2, and compared their performance. The Fixed Number of Events method notably outperformed Fixed Time Duration, yielding an accuracy of 91.07% for the Gait3 Speed category and a 98.21% for the Gait3 Backpack category. This suggests that the dynamic approach of sampling frames based on event density provides a more accurate representation of the actions and effectively captures both rapid and detailed movements. Accordingly, we will proceed with the Fixed Number of Events method for our model evaluation to best demonstrate its performance.

Sampling Method	Gait3 Speed	Gait3 Backpack
Fixed Time duration	82.14%	92.85%
Fixed Number of Events	91.07%	98.21%

Table 3. Comparison of Model Accuracy Across Sampling Methods.

5.2. Results on DailyAction and ActionTUM

For evaluation, we train our proposed TIME-E2V on two benchmark datasets: DailyactionDVS and ActionTUM, to verify its performance and how it would generalize across a diverse range of actions. Additionally, for comparative analysis, we train the SEW-ResNet model and a variant of our model without the temporal component, referred to as E2V, to verify the impact of incorporating temporal information on recognition performance.

As shown in Table 4, our model exhibits a pronounced enhancement in performance against SEW-ResNet, with an impressive gain of nearly 6% in accuracy on the ActionTUM dataset, achieving 97.26%. Moreover, the E2V model trained on reconstructed frames, noted a still superior accuracy over SNNs—95.85% on Daily Action and 93.15% on ActionTUM, reinforcing the point that leveraging the strength of traditional C3D models outperforms SNNs trained with event data for action recognition. Furthemore, upon comparison of the two variants of our model, we deduce that integration of temporal embeddings enhances our own model's accuracy by 0.69% for Daily Action and a more significantly 4.11% for ActionTUM.

Model	DailyAction-DVS	ActionTUM
SEW-ResNet	94.69%	89.69%
E2V	95.85%	93.15%
TIME-E2V	96.54%	97.26%

Table 4. Performance comparison of different models on action recognition tasks.

Based on this significant discrepancy in improvement, we notice that the ActionTUM dataset, despite its lower baseline accuracy, may have benefit more from the temporal resolution that our embeddings provided. This could be due to the fact that it includes actions like kicking, throw-

ing, and turning around that are more dependent on speed and precise timing for accurate recognition. This significant gain highlights the impact of our temporal embeddings in enhancing the model's ability to recognize the speed and dynamic of actions and achieve cutting-edge results, making it superior to both SNNs and its own version in the absence of these embeddings.

5.3. Results on Gait3 Dataset

For the Gait3 Dataset, we employed a different strategy to directly address and highlight the E2VID model's difficulty in accurately capturing motion speed. We conducted our tests across the two distinct Gait3 Speed and Gait3 Backpack datasets. By comparing our model's performance on these tasks with that of traditional RGB frames, we aim to demonstrate the unique advantages that reconstructed event frames provide in action recognition scenarios, effectively showcasing our TIME-E2V model's ability to tackle the E2VID model's challenges.

Model	Modality	Gait3 Speed	Gait3 Backpack
C3D [14]	RGB	95.83%	98.07%
E2V	Event	91.07%	98.21%
TIME-E2V	Event	94.64%	98.21%

Table 5. Action recognition analysis for Gait3 datasets.

For the 'Walking versus Running' task, Table 5 shows that the introduction of time embeddings to event frames significantly increased the model's accuracy by 3.57%, from 91.07% to 94.64%, almost matching the 95.83% accuracy achieved with the original RGB videos. This demonstrates that our model has effectively learned to differentiate speed variations in motion. However, for Gait3 Backpack, the accuracy remained consistent at 98.21% with the inclusion of time embeddings, matching the accuracy of the RGB model. The absence of change indicates that time embeddings have a negligible impact on actions that are not speed-dependent but rely on recognizing spatial details, such as the presence of a backpack.

In both cases, the accuracy of event frames with time embeddings meets or exceeds the performance of the visible domain model. This integration enables the system to capture the subtle variations of motion speed with a level of precision that was previously unattainable with event data alone, underscoring the efficacy of our approach in overcoming the limitations of the E2VID model and highlighting the potential of event-based data in action recognition tasks.

6. Ablations

6.1. Effect of Grayscale on the model performance

A critical consideration for our study is the impact of color on model accuracy, given that the E2VID model exclusively produces grayscale images. This raises an important question about the potential advantage color information may offer in visible spectrum data, potentially skewing accuracy comparisons favorably towards it over event-based data. To address this, we convert the Gait3 dataset to grayscale, enabling a more equitable comparison of model performance across both domains. For our analysis, the dataset is partitioned into training, validation, and testing segments with a distribution of 60%, 20%, and 20%, respectively, facilitating a thorough evaluation of this effect.

Dataset	RGB Videos	Grayscale Videos
Gait3 Backpack	98.07%	98.07%
Gait3 Speed	93.75%	93.75%

Table 6. Action recognition accuracies for C3D on Gait3 dataset.

The experimental results in Table 6 reveal that the accuracy remains consistent between the original visible spectrum data and its grayscale-converted counterpart. This consistency underscores that the absence of color information does not disadvantage the grayscale event frames produced by the E2VID model. Consequently, we can confidently assert that our comparison between visible and grayscale event data is fair, ensuring that any observed differences in model performance are attributable to factors other than color information.

7. Conclusions

In this work, we introduced TIME-E2V, a novel framework designed to enhance action recognition capabilities by integrating temporal embeddings with event data, thereby overcoming the limitations of the E2VID model. We used state-of-the-art SNNs and the E2V model as baselines for comparison. As a future perspective, we plan to integrate the recent advancements in video transformers into our baseline, as impressive results have been reported on the DVS Gesture dataset. When evaluated on benchmark datasets, our TIME-E2V model significantly outperformed its variant, achieving nearly a 4.11% increase on the Action-TUM dataset. This approach not only addresses the E2VID model's limitations but also demonstrates how event-based vision systems can be made more adaptable and efficient, expanding their applicability in dynamic and complex visual applications.

8. Acknowledgements

This research is a part of the HEIMDALL project, funded by the BPI as part of the AAP I-Démo.

References

[1] S. Barchid, B. Allaert, A. Aissaoui, J. Mennesson, and C. Djeraba. Spiking-fer: Spiking neural network for facial expression recognition with event cameras. https://doi.org/10.48550/arxiv.2304.10211, 2023. arXiv:2304.10211.

[2] L. Cordone, B. Miramond, and P. Thierion. Object detection with spiking neural networks on automotive event data. In *2022 International Joint Conference on Neural Networks (IJCNN)*, 2022.

[3] T. De Blegiers, I. R. Dave, A. Yousaf, and M. Shah. Eventtransact: A video transformer-based framework for event-camera based action recognition. https://doi.org/10.48550/arxiv.2308.13711, 2023. arXiv:2308.13711.

[4] M. J. Eddine and J. Dugelay. Gait3: An event-based, visible and thermal database for gait recognition. In *2022 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2022.

[5] W. Fang, Z. Yu, Y. Chen, T. Huang, T. Masquelier, and Y. Tian. Deep residual learning in spiking neural networks. https://hal.science/hal-03482280, 2021. HAL: hal-03482280.

[6] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, 2022.

[7] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[8] 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*, 39(7):1346–1359, 2017.

[9] P. Lichtsteiner, C. Posch, and T. Delbrück. A 128×128 120 db 15 μs latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43(2):566–576, 2008.

[10] Q. Liu, X. Dong, H. Tang, M. De, and G. Pan. Event-based action recognition using motion information and spiking neural networks. In *Proceedings of Thirtieth International Joint Conference on Artificial Intelligence*, 2021.

[11] S. Miao, G. Chen, X. Ning, Y. Zi, K. Ren, Z. Bing, and A. Knoll. Neuromorphic vision datasets for pedestrian detection, action recognition, and fall detection. *Frontiers in Neurorobotics*, 13, 2019.

[12] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. https://arxiv.org/pdf/1904.08298v1, 2019. arXiv:1904.08298v1.

[13] A. Tavanaei, M. Ghodrati, S. R. Kheradpisheh, T. Masquelier, and A. S. Maida. Deep learning in spiking neural networks. *Neural Networks*, 111:47–63, 2019.

[14] 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 (ICCV)*, 2015. arXiv: https://doi.org/10.1109/iccv.2015.510.

[15] X. Wang, Z. Miao, R. Zhang, and S. Hao. I3d-lstm: A new model for human action recognition. *IOP Conference Series: Materials Science and Engineering*, 569(3):032035, 2019.

[16] Y. Wang et al. Event-stream representation for human gaits identification using deep neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1–1, 2021.

[17] A. Z. Zhu, L. Yuan, K. Chaney, and K. Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. https://doi.org/10.1109/cvpr.2019.00108, 2019. CVPR 2019.