Event Vision in Egocentric Human Action Recognition*

Francisco J. Moreno-Rodríguez[1], V. Javier Traver[2[0000−0002−1596−8466]], Francisco Barranco[3[0000−0002−3721−0170]], Mariella Dimiccoli[4[0000−0002−2669−400X]], and Filiberto Pla[2[0000−0003−0054−3489]]

[1] Universitat Jaume I, Castelló, Spain
[2] Institute of New Imaging Technologies, Universitat Jaume I, Castelló, Spain
[vtraver,pla]@uji.es
[3] Dept. of Comp. Architecture and Technology, CITIC, University of Granada, Spain
fbarranco@ugr.es
[4] Institut de Robòtica i Informàtica Industrial (CSIC-UPC)
mdimiccoli@iri.upc.edu

Abstract. This paper lies at the intersection of three research areas: human action recognition, egocentric vision, and visual event-based sensors. The main goal is the comparison of egocentric action recognition performance under either of two visual sources: conventional images, or event-based visual data. In this work, the events, as triggered by asynchronous event sensors or their simulation, are spatio-temporally aggregated into event frames (a grid-like representation). This allows to use exactly the same neural model for both visual sources, thus easing a fair comparison. Specifically, a hybrid neural architecture combining a convolutional neural network and a recurrent network is used. It is empirically found that this general architecture works for both, conventional gray-level frames, and event frames. This finding is relevant because it reveals that no modification or adaptation is strictly required to deal with event data for egocentric action classification. Interestingly, action recognition is found to perform better with event frames, suggesting that these data provide discriminative information that aids the neural model to learn good features.

Keywords: Egocentric view · action recognition · event vision

1 Introduction

In contrast to the more widespread third-person vision (3PV), first-person (egocentric) vision (1PV) provides unique insights of the scene as observed directly from the privileged point of view of the camera wearer, as well as their activities

* Work supported by project UJI-B2018-44 from *Pla de promoció de la investigació de la Universitat Jaume I*, Spain, the research network RED2018-102511-T, from the Spanish *Ministerio de Ciencia, Innovación y Universidades*, and by National Grant PID2019-109434RA-I00/ SRA (State Research Agency /10.13039/501100011033).

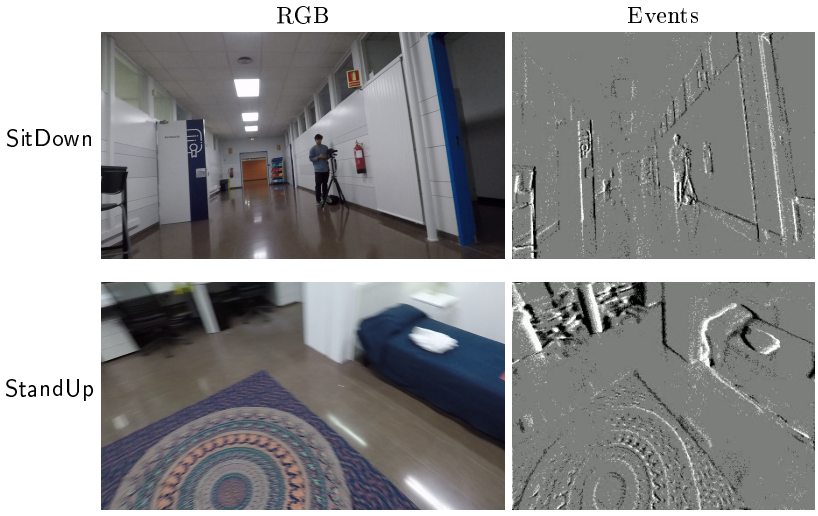

Fig. 1. Examples of RGB image (left) and simulated event frame (right) during a SitDown (up) and StandUp (down) sequences from the dataset used. The event frames have different size as a requirement of the video-to-event simulator used. The gray-level display of the event frames represents the aggregated polarity of the events as per Eq. (1).

and behaviour. Therefore, it is not surprising that gesture, action, or activity recognition, which have been widely explored in third-person contexts, have also been investigated over the last decade for the egocentric case. Although significant progress has been achieved, both 1PV and 3PV action recognition approaches have been dominated by the use of regular visual sensors. Recently, however, an alternative sensing paradigm, event-based neuromorphic visual sensors, or event cameras, have been receiving increasing attention. Unlike conventional cameras, these bio-inspired event-based sensors deliver a source of sparse and asynchronous flow of events corresponding to luminance changes detected with a very precise timing. Despite the potential benefits of event cameras in egocentric computational vision, action recognition with event-based visual data on egocentric visual streams has not been addressed. Actually, to the best of our knowledge, only one work exists (Section 2). The work reported in this paper combines these three research topics (1PV, action recognition, and event-based vision), and explores whether event-based representations can contribute to egocentric action recognition.

Please, note that the use of the term *event* throughout this paper refers to the low-level concept associated to brightness changes as delivered by event cameras. In particular, these events should not be confused to the higher-level events representing either temporally contiguous images identified as a unit, or to semantic concepts (such as "door open", "keys dropped") detected from a sequence of images.

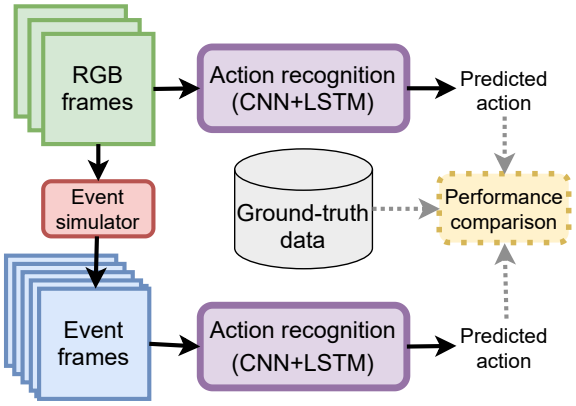

Fig. 2. Overview of the work. Note how event frames can be produced at a higher frame rate than the original RGB frames by using high-quality video interpolators. Exactly the same network (Fig. 3) is used in both visual sources (Fig. 1), but each is trained on their respective data.

Overview and contributions. We propose the use of event frames [12,24] and a hybrid neural architecture (convolutional and recurrent neural networks) by leveraging authors' recent related work on egocentric gesture recognition using conventional (*not* event-based) eyewear cameras [18,17]. Specifically, the main contribution of this work is to compare conventional gray-level frames with event-based data (Fig. 1) for egocentric action recognition using exactly the same neural model (Fig. 2).

2 Related work

Taking into account the problem and our particular approach, a brief overview of related work on the following areas is considered: motion estimation, visual events, egocentric vision, and action recognition.

(Ego)motion estimation. Estimation of head motion [28] is generally relevant in the context of 1PV. More generally, homography estimation methods based on deep learning have been proposed recently [9,20,31] and have therefore the potential to properly characterise egocentric-related movements.

Event-based visual data. Event cameras [12] offer a very distinct alternative to how conventional cameras operate, by delivering visual changes as a series of sparse and asynchronous events with a time resolution of a few microseconds. Each of such events typically encodes the sign of the brightness change and its spatial information. The main benefits of event cameras is a low-power consumption, high-temporal resolution, and low-latency event stream. Understandably, these cameras or their simulations have mostly been used for low-level visual tasks such as tracking [3] or vehicle control [29], which most straightforwardly can exploit these properties. Generally speaking, tasks requir-

ing ultra fast visual processing are favoured by smart compression and high temporal resolution of event cameras, for instance to overcome the undesirable motion blur in traditional cameras. However, despite its potential advantages, the use of these event-based visual data for other visual tasks has been limited so far. Only recently, events have been studied for 3D pose estimation [25] and gesture recognition [2,30] of hands. Action recognition from 3PV datasets has also been explored [15,14] with event data, which reinforces the rising interest of this sensing paradigm.

Egocentric vision and action recognition. Egocentric visual streams can be useful for a variety of tasks, including visual life-logging [4], hand analysis [6], activity recognition [23,21], social interaction analysis [1,11], video summarisation [8,27], etc. To the best of our knowledge, only conventional cameras have been used for action recognition in the context of egocentric vision, an exception being a very recent work [22], which introduces N-EPIC-Kitchens dataset, an event-based version of EPIC-Kitchens [7], a well-known dataset of egocentric videos captured with conventional wearable cameras. Interestingly, the findings of this parallel work align with ours (as discussed in Section 4.3) in that action recognition performance with event-based visual data is on par or better than with conventional imaging, even though different datasets and approaches have been used in our respective works.

3 Methodology

How events are defined and event frames generated are described first (Section 3.1. Then, the neural network architecture proposed for action recognition and training details are provided (Section 3.2).

3.1 Event data generation

Event cameras [12] trigger a flow of events $\mathbf{e}_k, k \in \{1, 2, \ldots\}$, each represented by a tuple $\mathbf{e}_k = (\mathbf{x}_k, t_k, p_k)$, where $\mathbf{x}_k = (x_k, y_k)$ represents the spatial coordinates where a brightness change higher than a threshold C has been detected at time t_k since the last event at that same pixel location, and $p_k \in \{-1, +1\}$ is the polarity (sign) of the change.

The dataset used in our work (Section 4.1) consists of conventional videos only. Therefore, visual events were simulated from the regular video frames. Event camera simulators are widely used in the literature [24,12,22] since they render realistic events as compared to actual event cameras, and facilitate developing and benchmarking algorithms under controlled conditions prior to the use of specific event cameras. In our case, event simulation turns out to be very useful for the comparison between conventional frame-based and event-based methods for action recognition. Our choice was the simulator v2e [13], a Python tool for realistic event synthesis.

Since events are triggered asynchronously and correspond to spatio-temporal sparse data, there are several possibilities when it comes to representing and processing them. To process events one-by-one, spiking neural networks are possibly

the straightforward choice [5], but these networks are more recent and far less known than conventional neural networks. Alternatively, we adopt a common approach [15,14,22] of aggregating these events into a grid representation. Events can essentially be aggregated by count (every n events) or by time (at regular time intervals). We aggregate events $\mathbf{e}_k = (\mathbf{x}_k, t_k, p_k)$ into 2D event frames $\mathbf{E}_t$ at regular time steps $t \in \{T, 2T, \ldots\}$ taking into account their polarity p_k [24],

$$\mathbf{E}_t(\mathbf{x}) = \sum_k p_k \cdot \frac{t_k - (t-T)}{T}, \quad \forall\ k \text{ such that } \mathbf{x}_k = \mathbf{x}, \text{and } t - T < t_k \leq t, \quad (1)$$

where $T = 5$ ms in our case, which is 8 times faster than the original frame rate (25 fps) due to video interpolation performed by the event simulator using high-end algorithms such as *Super SloMo* [16]. Examples of such event frames are given in Fig. 1.

Although these grid-like representations do not process events natively as they are triggered, and therefore a lag is introduced and some temporal resolution can be lost, grouping events helps exploit their spatio-temporal consistency, and allows existing and widely proven conventional neural models (such as convolutional neural networks) to be used straightforwardly. Also, for object classification and visual odometry tasks, architectures based on the grid representation have been reported to outperform specialised algorithms designed for event cameras [24]. For the purposes of our work this is also an ideal choice because it enables easier and fairer comparison between RGB and event-based visual data with a common learning model.

3.2 Action recognition

A motion-based recognition approach is proposed that combines a convolutional neural network (CNN) and a recurrent network (Fig. 3), as in our recent work on egocentric gesture recognition with regular eyewear cameras [17,18]. The idea is to estimate head-induced motion with frame-to-frame homography estimation (i.e. an estimate of the global camera motion), and then use these estimates as input to a long short-term memory (LSTM) so that action-relevant visual dependencies more distant in time are captured by the overall network. For homography estimation, the chosen CNN model [31] features two favourable properties: it is unsupervised, and it is designed to be robust against independently moving objects not following the global camera motion.

Note that this exact model is used in both RGB (gray) frames and event frames (inputs I_t and I_{t-1} in Fig. 3). In both cases, input values are normalized to the the range $[0, 1]$. We use the homography-estimation CNN as pretrained by their authors on sequences exhibiting motions different to those occurring in some of the actions in our experimental dataset. More importantly, only conventional images were used for training, *not* event frames. Although our purpose here is not obtaining accurate motion estimates, since only discriminative motion features are required, a natural question is whether using such CNN with event frames as input is a sensible choice. Therefore, as a sanity check, we briefly

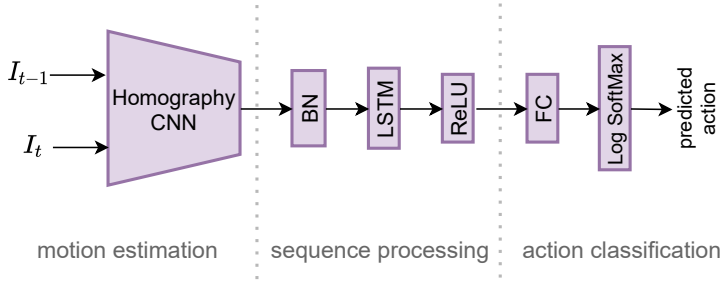

Fig. 3. Neural model for action recognition. The input is a pair of consecutive frames and the output is the predicted action. BN=Batch normalisation, FC=Fully Connected (dense) layer.

tested whether motion can be reasonably estimated with this off-the-shelf CNN even with event frames as input. It turned out that, compared to ground-truth motion, estimates were different with RGB and event frames, but both were reasonable. Therefore, the same pre-trained CNN was used for RGB and Events data, and only the LSTM and the fully connected (FC) layers are data-specific and supervisedly trained on the respective sequences. We leave as future work to train from scratch or fine-tune the homography CNN with data sequences for our specific task and visual formats.

As stated above, the event simulator produces event frames at higher temporal resolution than the available RGB videos. For the sake of fairer comparison, however, we temporally subsampled the event frames to match frame rates of the RGB and Events sequences.

Shorter and overlapped fragments (clips) from the training videos were used for training, thus having a larger training set. Classification is performed frame-wise, which is harder than video-wise classification because not the full video is observed, but this scenario lends itself to online and early action recognition. For video-level classification, majority voting is applied. The network was trained for 100 epochs, using a learning rate of 10^{-4}. The hidden size of the LSTM was set to 128.

4 Experimental work

The dataset used is first introduced (Section 4.1). Next, the experiments and the results are described (Section 4.2) and discussed (Section 4.3).

4.1 Dataset

A subset from *First2Third-Pose* dataset [10] has been selected for the purpose of this work. Although events could be generated from *Charades-Ego*, an egocentric

Table 1. Actions from *First2Third-Pose* used in our work. Number of videos and mean and standard deviation of video lengths (in frames) for each action are given.

Action	Description	Main motion	Videos	
			No.	Length
StandUp	Standing up from the ground	Vertical positive	58	88 ± 22
SitDown	Sitting on ground	Vertical negative	61	94 ± 28
Rot	Rotating	Full-body, horizontal	71	194 ± 70
Turn	Head turning	Head, horizontal	59	294 ± 67

action dataset [26], *First2Third-Pose*, was chosen as a part of a research collaboration[5], and the dataset N-EPIC-Kitchens [22], with available events simulated from egocentric videos of human activity, is very recent, and was not available by the time this work was developed [19], and it has not actually been released yet. *First2Third-Pose* features synchronised pairs of first- and third-view videos of 14 people performing 40 activities. Although mainly intended for 3D pose estimation, this dataset can also be useful for activity (action) recognition. Here we focus on the egocentric view and select four representative actions. The rationale for this choice was to have both dissimilar and similar actions, so that the proposed approach can be assessed in easier and harder scenarios. Therefore, two groups of two actions each were chosen (Table 1), with a total of 249 videos.

Videos of different actions tend to have significantly different lengths (last column in Table 1). In particular, StandUp and SitDown are shorter (about 1–6 seconds), which last about 9-20 seconds. These differences pose a practical challenge in terms of which video clip length to use for training. Long clips can be more informative, but fewer of them can be extracted from those original videos which are shorter. With shorter clips, more training instances can be sampled from short videos, but each clip characterises worse the action sequences, and therefore they will be harder to model. Overall, clips of size $S = 35$ frames with an overlap of $O = 25$ frames were found to be a good compromise. Clips with these (S, O) values will be referred to as 'short' clips. The dataset was split into 80%-20% training-validation sets in terms of individual clips. Additionally, since the same action is performed by different subjects and in different places (Fig. 4), specific subjects-based and location-based splits will be used in some tests, as detailed below.

4.2 Results

To better understand the performance of the recognition system, different conditions are separately considered.

Two-way classifications. A simple binary classification between two dissimilar actions, Rot and SitDown, was first tested, and Events outperformed RGB (Table 2, Test 1). Arguably, Events deals better with this unbalanced class case.

[5] Red Española de Aprendizaje Automático y Visión Artificial para el Análisis de Personas y la Percepción Robótica (ReAViPeRo), https://www.init.uji.es/reavipero.

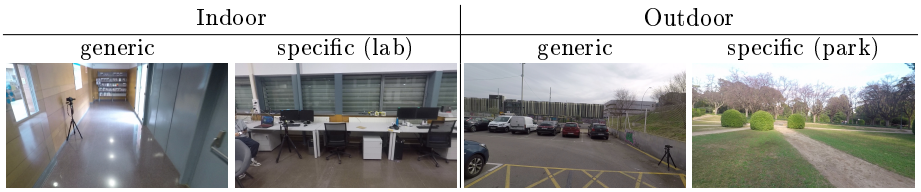

Fig. 4. Different places where actions in *First2Third-Pose* were recorded.

Table 2. Recognition results, accuracy (%) at video level.

Split	Test	RGB	Events
Clips (80%-20%)	1. Two-way, dissimilar, short [Fig. 5]	75.4	**86.7**
	2. Two-way, similar, short	**75.8**	65.48
	3. Two-way, similar, long	70.6	**81.8**
	4. Three-way, short	68.3	**75.8**
	5. Four-way, short [Fig. 6(top)]	52.7	**56.9**
	6. Four-way, long [Fig. 6(below)]	55.5	**64.5**
Places	7. Four-way, long	58.4	**67.4**
Subjects	8. Four-way, long	53.6	**74.6**

An example comparing RGB and Events at frame-level classification (Fig. 5) illustrates that with most frames correctly labelled, Events can correctly classify the sequence at video-level, whereas RGB cannot. For a harder scenario, two similar actions (Rot and Turn) are considered. Results (Table 2, Test 2) reveal that in this case RGB outperforms Events. One possible explanation is that these actions include faster motions, which might violate the small-baseline transformations assumption of the homography network. Under these circumstances, RGB might fare better due to the more similar nature to the images used for training the homography CNN. Interestingly, if the clip length and overlap are enlarged to $S = 84$ and $O = 80$, to account for the longer videos in these actions, performance (Table 2, Test 3) drops by about 5 percentage points in RGB, and increases by 15 points in Events. This highlights the impact that the clip length has.

Three- and four-way classification. When considering Rot, SitDown and Turn, results (Table 2, Test 4) are understandably worse for both RGB and Events. However, Events outperforms RGB. Finally, when considering the four actions, performance drops (Table 2, Test 5), but longer snippets help (Table 2, Test 6), particularly when using Events. The confusion matrices (Fig. 6) indicate that misclassifications mostly happen between the two similar-motion groups of actions (SitDown vs StandUp and Rot vs Turn). They also illustrate that, generally, better performance is obtained with long clips, both with RGB and Events. The overall trend of better recognition with Events over RGB can also be observed.

Generalisation ability. Finally, we evaluate how the system generalises to different places and subjects. For places, the training split includes three cases

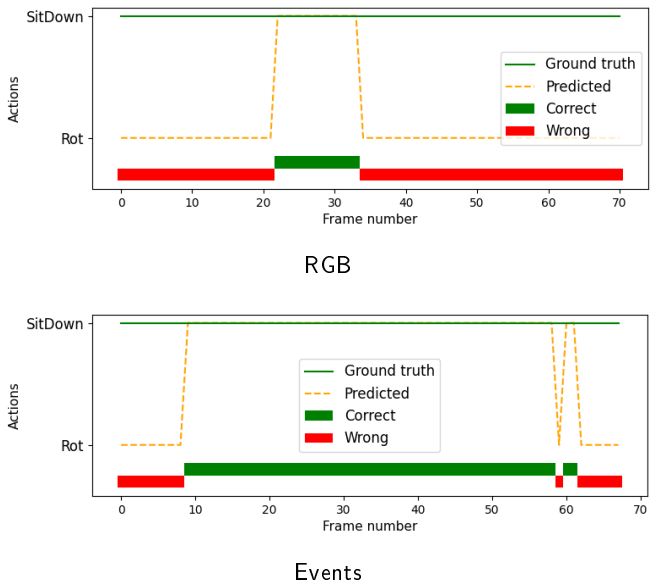

Fig. 5. Frame-level classification for a SitDown sequence (Fig. 1a) for RGB and Events. In RGB, most frames are mislabelled and therefore the action is misclassified at video level.

('outdoor', 'lab', 'park') and we evaluate on 'indoor'; this split represents roughly a 80%-20% split. While 'lab' is also an indoors place, its sequences contain more objects and texture that might facilitate homography estimates, whereas 'indoor' places are usually in texture-less corridor areas with many lighting artefact, or exterior parts of a building, with windows and light issues. Again, Events outperforms RGB (Table 2, Test 7), one possible reason being that light reflections might misguide RGB, whereas no (false) events are produced by these action-irrelevant data. Regarding the subject-split better results are also obtained with Events (Table 2, Test 8). As for why the performance of Events in this case (74.6%) is higher than in any other 4-way test considered, a tentative explanation is that by including all repetitions of the same subject in the training split, the model captures a useful variability that might possibly be not present with other train-val splits.

4.3 Interpretation and discussion

As these results illustrate, better recognition performance is obtained with event visual data than with regular images. A likely explanation is that the event representation acts as feature selection and encodes motion-like information, thus helping the network to more directly focus on relevant and discriminative data. Related to this hypothesis, we plan to compare the performance of event-based data with edge-like or motion-like data precomputed on conventional frames.

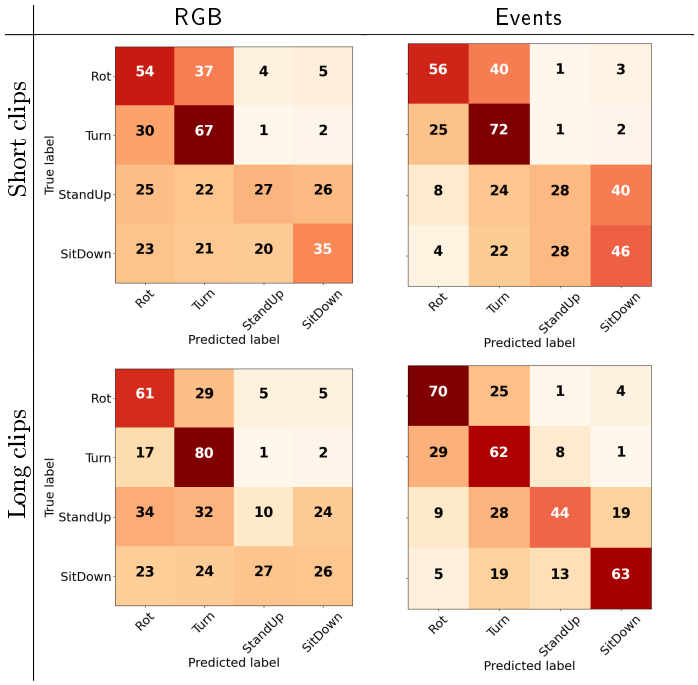

Fig. 6. Confusion matrices (values in %) for 4-way video-level classification with RGB (left) and Events (right) for short (top) and long (bottom) clips.

This will provide insights into what actually and intrinsically aids the recognition task.

The results are somehow surprising and particularly interesting because a generic learning model has been used, without any adaptation or specific design that takes into account the idiosyncrasies of event visual data. Furthermore, the CNN used as part of the architecture for homography estimation was completely "event agnostic", since it was trained on regular image sequences only.

The benefits in recognition performance using events observed in our work are in agreement with a recent study using different dataset and architectures [22]. Additionally, these authors explore the role of motion information from event frames by using existing neural architectures to extract temporal information. They conclude the significantly positive effect brought by this temporal information. Very interestingly, this finding aligns with the motion-like information captured by our network at both, frame-to-frame level (through the homography CNN) and inter-frame level (through the LSTM).

5 Conclusions

Although event vision, based on neuromorphic event cameras, have witnessed increased research attention over the last few years, mostly from the robotics and

computer vision communities, its use in egocentric action recognition has been very limited. For egocentric videos, this work compares the action recognition performance of regular gray-level frames to that of a grid-based event representation. Importantly, exactly the same neural network architecture is used in both cases, without any ad hoc architecture that more specifically accounts for the nature of event-based data. Experimental results reveal that action recognition using events outperforms that of using gray-level frames, thus encouraging further work on event vision for egocentric action recognition tasks. Future research work includes using more actions and other datasets. Another interesting direction is exploring other architectures such as the spiking neural networks, and compare them with the current model.

References

1. Alletto, S., Serra, G., Calderara, S., Cucchiara, R.: Understanding social relationships in egocentric vision. Pattern Recognition **48**(12), 4082–4096 (2015)
2. Amir, A., et al.: A low power, fully event-based gesture recognition system. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 7388–7397 (2017)
3. Barranco, F., Fermuller, C., Ros, E.: Real-time clustering and multi-target tracking using event-based sensors. In: IEEE Intl. Conf. on Intelligent Robots and Systems (IROS). pp. 5764–5769 (2018)
4. Bolaños, M., Dimiccoli, M., Radeva, P.: Toward storytelling from visual lifelogging: An overview. IEEE Transactions on Human-Machine Systems **47**(1), 77–90 (2017)
5. Cordone, L., Miramond, B., Ferrante, S.: Learning from event cameras with sparse spiking convolutional neural networks. In: Intl. Joint Conf. on Neural Networks (IJCNN) (2021)
6. Cruz, S., Chan, A.: Is that my hand? An egocentric dataset for hand disambiguation **89**, 131–143 (2019)
7. Damen, D., et al.: The EPIC-KITCHENS dataset: Collection, challenges and baselines. IEEE Trans. on Pattern Analysis & Machine Intelligence **43**(11), 4125–4141 (2021)
8. del Molino, A.G., Tan, C., Lim, J., Tan, A.: Summarization of egocentric videos: A comprehensive survey. IEEE Transactions on Human-Machine Systems **47**(1), 65–76 (Feb 2017)
9. DeTone, D., Malisiewicz, T., Rabinovich, A.: Deep image homography estimation. CoRR **abs/1606.03798** (2016), https://arxiv.org/abs/1606.03798
10. Dhamanaskar, A., Dimiccoli, M., Corona, E., Pumarola, A., Moreno-Noguer, F.: Enhancing egocentric 3D pose estimation with third person views. CoRR **abs/2201.02017** (2022), https://arxiv.org/abs/2201.02017
11. Felicioni, S., Dimiccoli, M.: Interaction-gcn: A graph convolutional network based framework for social interaction recognition in egocentric videos. In: IEEE Intl. Conf. on Image Processing (ICIP). pp. 2348–2352 (2021)
12. Gallego, G., et al.: Event-based vision: A survey. IEEE Trans. on Pattern Analysis & Machine Intelligence **44**, 154–180 (Jan 2022)
13. Hu, Y., Liu, S.C., Delbruck, T.: v2e: From video frames to realistic DVS events. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 1312–1321 (2021)

14. Huang, C.: Event-based timestamp image encoding network for human action recognition and anticipation. In: Intl. Joint Conf. on Neural Networks (IJCNN) (2021)
15. Innocenti, S.U., Becattini, F., Pernici, F., Del Bimbo, A.: Temporal binary representation for event-based action recognition. In: Intl. Conf. on Pattern Recognition (ICPR) (2021)
16. Jiang, H., et al.: Super SloMo: High quality estimation of multiple intermediate frames for video interpolation. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 9000–9008 (2018)
17. Marina-Miranda, J.: Head and Eye Egocentric Gesture Recognition for Human-Robot Interaction using Eyewear Cameras. Master's thesis, Universitat Jaume I, Castellón, Spain (Jul 2021)
18. Marina-Miranda, J., Traver, V.J.: Head and eye egocentric gesture recognition for human-robot interaction using eyewear cameras (Submitted Sept 2021), under review. Preprint at http://arxiv.org/abs/2201.11500
19. Moreno-Rodríguez, F.J.: Visual Event-based Egocentric Human Action Recognition. Master's thesis, Universitat Jaume I, Castellón, Spain (Jul 2021)
20. Nguyen, T., Chen, S.W., Shivakumar, S.S., Taylor, C.J., Kumar, V.: Unsupervised deep homography: A fast and robust homography estimation model. IEEE Robotics and Automation Letters **3**(3), 2346–2353 (2018)
21. Núñez-Marcos, A., Azkune, G., Arganda-Carreras, I.: Egocentric vision-based action recognition: A survey. Neurocomputing **472**, 175–197 (2022)
22. Plizzari, C., Planamente, M., Goletto, G., Cannici, M., Gusso, E., Matteucci, M., Caputo, B.: E^2(go)motion: Motion augmented event stream for egocentric action recognition. CoRR **abs/2112.03596** (2021), https://arxiv.org/abs/2112.03596
23. Possas, R., Caceres, S.P., Ramos, F.: Egocentric activity recognition on a budget. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 5967–5976 (2018)
24. Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D.: Events-To-Video: Bringing modern computer vision to event cameras. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 3852–3861 (2019)
25. Rudnev, V., Golyanik, V., Wang, J., Seidel, H.P., Mueller, F., Elgharib, M., Theobalt, C.: EventHands: Real-time neural 3D hand pose estimation from an event stream. In: Intl. Conf. on Computer Vision (ICCV) (2021)
26. Sigurdsson, G.A., Gupta, A.K., Schmid, C., Farhadi, A., Karteek, A.: Actor and observer: Joint modeling of first and third-person videos. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) pp. 7396–7404 (2018)
27. Traver, V.J., Damen, D.: Egocentric video summarisation via purpose-oriented frame scoring and selection. Expert Systems with Applications **189** (2022)
28. Tsutsui, S., Bambach, S., Crandall, D., Yu, C.: Estimating head motion from egocentric vision. In: ACM Intl. Conf. on Multimodal Interaction. pp. 342–346 (2018)
29. Vitale, A., Renner, A., Nauer, C., Scaramuzza, D., Sandamirskaya, Y.: Event-driven vision and control for UAVs on a neuromorphic chip. In: IEEE Intl. Conf. on Robotics and Automation (ICRA). pp. 103–109 (2021)
30. Xing, Y., Di Caterina, G., Soraghan, J.: A new spiking convolutional recurrent neural network (SCRNN) with applications to event-based hand gesture recognition. Frontiers in Neuroscience **14** (2020)
31. Zhang, J., et al.: Content-aware unsupervised deep homography estimation. In: European Conf. on Computer Vision (ECCV). pp. 653–669. Springer (2020)