MoveEnet: Online High-Frequency Human Pose Estimation with an Event Camera

Gaurvi Goyal, Franco Di Pietro, Nicolo Carissimi, Arren Glover, Chiara Bartolozzi
Event-Driven Perception for Robotics
Istituto Italiano di Tecnologia, Italy
{gaurvi.goyal, franco.dipietro, nicolo.carissimi, arren.glover, chiara.bartolozzi}@iit.it

Abstract

Human Pose Estimation (HPE) is crucial as a building block for tasks that are based on the accurate understanding of human position, pose and movements. Therefore, accuracy and efficiency in this block echo throughout a system, making it important to find efficient methods, that run at fast rates for online applications. The state of the art for mainstream sensors has made considerable advances, but event camera based HPE is still in its infancy. Event cameras boast high rates of data capture in a compact data structure, with advantages like high dynamic range and low power consumption. In this work, we present a system for a high frequency estimation of 2D, single-person Human Pose with event cameras. We provide an online system, that can be paired directly with an event camera to obtain high accuracy in real time. For quantitative results, we present our results on two large scale datasets, DHP19 and event-Human 3.6m. The system is robust to variance in the resolution of the camera and can run at up to 100Hz and an accuracy 89%.

1. Introduction

Human Pose Estimation (HPE) is often defined as the localisation of a fixed number of body joints in a human agent and is widely considered a crucial building block in a variety of human-centric tasks [29, 30, 36, 41]. In robotics, IoT or smart home applications that need to assess the environment and human agents, each person's pose estimation is the preliminary stage that feeds action recognition, posture, emotion and intent estimation pipelines [8, 9]. To this end, HPE must be accurate and fast, while taking up minimal computation power. Due to this instrumental role, HPE has been investigated strongly in the recent years, with various sensory modalities, like digital cameras, marker based systems like motion capture and RGBD sensors [8, 28, 31, 37]

On the other hand, the popularity of neuromorphic event cameras has increased in the recent years, as they deliver information in a more compact data format, thus, the processing can be faster, with lower computational load, freeing it up for other components in a larger system. This has boosted interest in formulation of algorithms and systems that leverage such advantages for low-latency tracking, motion estimation, gesture recognition, etc. (see [12]). However, the field is quite recent and there are a plethora of visual tasks that need further exploration to move towards full event-driven pipelines for vision, including HPE. Specific applications can benefit substantially from a high-frequency, marker-less system that event cameras can uniquely provide, like tracking fast motions in sports applications.

Figure 1. Sample result of MoveEnet from the event-Human 3.6m dataset superimposed on EROS representation

In this work, we exploit the recent advancements in Deep Learning based HPE to create a pipeline that combines the advantages of event cameras with the performance of Artificial Neural Networks (ANNs). We take the approach of using a smart image-like representation, EROS (described in Sec. 3.1) of the asynchronous event stream from event cameras, that supports the re-use of existing and robust ANNs. This representation is chosen to overcome two major roadblocks in developing event-based HPE. On one hand, it allows for a low cost and fast conversion of available large-scale image-based HPE datasets into an EROS-like representation, leveraging their diversity for effective

pre-training of an ANN, before fine-tuning on event camera datasets. On the other hand, EROS keeps persistent activity in face of the *blind-spot problem*, whereby any body part that is stationary becomes invisible in data representations with fixed number of events or temporal window.

As architecture for HPE, we traded off accuracy and efficiency in the available state of the art, choosing MoveNet [1] for its high frequency inference.

In this paper, we propose a Human Pose Estimation system, MoveEnet, that can take events as input from a camera and estimate 2D pose of the human agent in the scene. The final system can be attached to any event camera, regardless of resolution. We demonstrate the results on 2 large scale benchmark dataset, DHP19 [3] which has been acquired by event cameras directly and Human 3.6m [20], a popular video based benchmark converted to events by us. The resulting pose is can run at up to 100Hz frequency on a GPU enabled machine, and up to 20 Hz on a CPU-only machine with 89% accuracy. Code for running the model and conversion of datasets is made available to the community[1].

2. Related Works

We considered works published in peer reviewed venues as state of the art and for comparative analysis.

Frame-based Human Pose Estimation Many works have been published in Human Pose Estimation in the last few years [5,11,21] in the RGB domain, varying from skeleton pose, localising few predefined poses [5] to volumetric pose represented by a mesh over the full body volume [25]. A few surveys covering the area in detail are already available [2, 8, 24, 38], even as more works are published every year. Most recent works rely on ANNs, either following a bottom-up approach or a top-down, holistic approach. One of the seminal works in the area, OpenPose [4] follows the bottom-up approach, first detecting limbs in the view and then grouping them for each human agent. Though many more works have been published since, OpenPose still remains widely used due to ease of use and high accuracy and offers a reasonable baseline method.

Bottom-up approaches are widely used and are accurate on images but when moving to videos, there is no direct way to optimise the system, even though consecutive frames are generally relatively similar. This is possible in a top-down approach [29, 36], where single person pose estimation is run for each person detected by a person detector running on the raw input. For a video, the detection can be done less often than the pose estimation, being computationally lighter. One recent high frequency system available is MoveNet, though no paper has been published associated with it at the time of writing [1].

Event-based Human Pose Estimation In event-based vision, there are very few models addressing Human Pose Estimation. EventCap [39] employs an hybrid event camera, that creates both an asynchronous event stream as well as low frequency grayscale intensity images, to obtain a high frequency 3D volumetric pose. The accuracy of the pose estimation relies on the optimization step between the two frames, that can be executed only once the next grayscale image is available, making it a strictly offline system. LiftMono-HPE [32] is an ANN based system that estimates a 3D skeleton pose from a monocular event camera. It uses integrated images obtained by accumulating a fixed number of events, from which it estimates the pose on 3 orthogonal planes, followed by triangulation. The system uses the torso length of the person to estimate depth as a prior, and thus is challenging to use in real world scenario and runs at about 2Hz online.

EventHPE [43] requires both events and grayscale images created by an event camera. It first uses the grayscale image to estimate a starting pose. Simultaneously the events and grayscale image are fed to an ANN to estimate the optical flow and the silhouette of the person. Another ANN combines these to initial pose to estimate the change in the pose. Majority of latest event cameras do not provide these grayscale images, but this method cannot work without them, making it less interesting for online testing. EventPointPose [7] estimates 2D human pose with low latency, by converting the events to a 3D point cloud which is fed to a point cloud based ANNs to estimate the pose. Training labels are generated with different modalities, but the preferred uses the mid point of the moving window, adding latency (and hence inaccuracy) to the system, if considered for an online application. The baseline method for DHP19 [3] uses input from two cameras to triangulate 3D HPE.

To the author's best knowledge, there are not any available system that run directly on the camera's event stream. In this work, we present a system that not only runs online but is also light-weight and updates high frequency and high accuracy human pose estimates from the event stream.

Representations A single event sourced from an event camera is not sufficiently informative to make complex estimates such as in HPE. A number of events must be accumulated, either by the algorithm itself [33], or in a preprocessing layer that creates a spatial or spatio-temporal representation. Such a representation provides context to each event, and thus can be used to provide a more informative input to the algorithm [12].

[1] https://github.com/event-driven-robotics/hpe-core

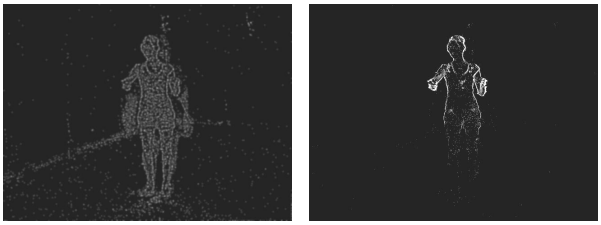
Events represent incremental change in light, and such can be integrated to produce absolute intensity images, however camera bias and noise typically results in poor quality representations. Error can be decayed over time to produce reasonable imagery over short periods of time [34], which we call polarity-integrated images (PIM) in this work. [19] represented events in a binary 2D image-like matrix that indicated event presence. The Surface of Active Events [27] encodes the time of the latest event for every pixel location. Hierarchy of Time Surfaces (HOTS) [22] calculated features based on neighbourhoods of spatio-temporal patterns, which are assigned to the pixel locations with the help of indices. Histogram of Averaged Time-Surfaces (HATS) [35] applied filtering in the space-time window to calculate the pixel values, minimising noise. 3D voxel-like structures [42] can represent events with a third, temporal, dimension. The timestamp and polarity can be encoded in many different ways [12].

An important characteristic of an event representation is the selection of which (or how many) events should be encoded at any point in time. Simple metrics such as a moving time window or a constant number of events can be highly effective under certain application constraints. However, anything that is not moving in the scene in this window is invisible to the camera (blind spots). A moving camera creates a large, but highly variable, number of events, which also create an issue with this type of selection methods, as the computation time strongly depends on the number of events and can lead to large system latency.

Thus some other representations are designed for speed invariance and persistence of features across time as they accumulate the events. The Speed Invariant Time-surface (SITS) [26] accumulates all past events and produces a linear scaling based on event order, ignoring the value of the timestamp. Task dependent tuning has moved away from the temporal domain, and placed on local region of interest size, which is less sensitive and requires less modification. Other representations that embody this property are Threshold Ordinal Surface (TOS) [15] and Exponentially Reduced Ordinal Surface (EROS) [13], which (among other differences) use a non-linear and exponential decay, respectively.

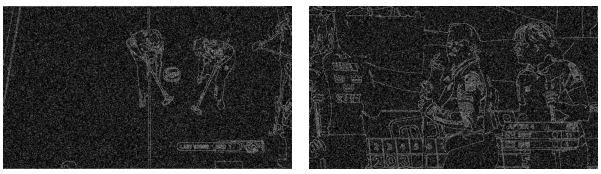
3. Methodology

Our approach is to leverage powerful, but efficient-by-design, frame-based ANN architectures for HPE. The core of this work is the use of a light-weight image-like event representation, that solves the disappearance of static body parts (blind-spots) and allows for pre-training on widely available frame-based datasets with high accuracy ground truth followed by fine tuning with native event-camera datasets.



(a) EROS (b) FixedCount

Figure 2. Samples of EROS and fixedCount of the same moment for a sample from event-Human 3.6m dataset. The legs are not visible in the fixedCount representation as the legs are not moving and, hence, not generating events. The hand-like shape to the right of the leg is an artifact created by the EROS representation, as a trade-off for the persistence it offers.



(a) (b)

Figure 3. Samples of the EROS-like representation from images from the MPII [2] dataset.

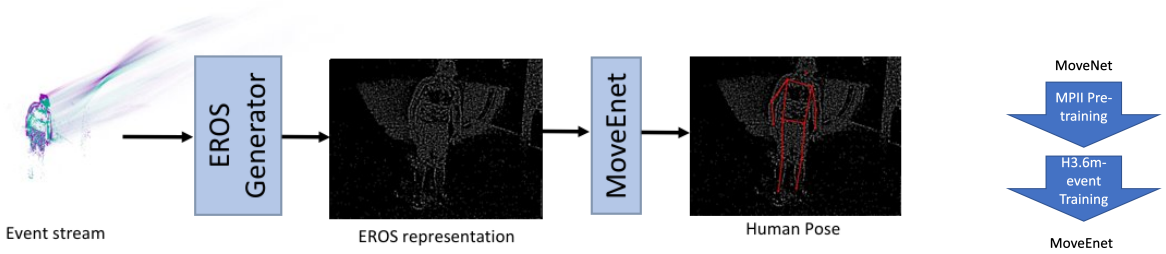
3.1. Event-based Input to the Network

The first challenge for event-based HPE comes as events are a very different data representation compared to images. To bootstrap from traditional HPE domain, the events and images must be converted into a common representation.

Commonly used approaches, such as the accumulation of the most recent events defined by a temporal window [14] or with a fixed number of events (fixed count) [3, 32] are simple to implement. However, the representation formed is far from an image taken by a traditional camera and suffers from artificially introduced motion blur, that is inherently absent in the event-stream, or lack the persistence needed to overcome the *blind spots*, illustrated in Fig. 2, whereby the representation of static body parts fades away when a person moves a set of limbs while the others remain static.

We use the EROS [13] to both mitigate *blind spots* and artificial motion blur, and also bridge the gap from events to RGB for training purposes. Learning efficiency is boosted when, for any given output, the input is consistent; therefore it is important that the event-based representation is consistent independent of the dynamics of the scene. EROS has the following beneficial properties:

- Speed invariance: a consistent representation should be produced if the person is moving quickly or slowly;
- Local-region update: two limbs should be represented identically even if one is moving quickly, while the



(a) (b)

Figure 4. Methodological pipelines. (a) indicates the overall MoveEnet system at inference. Events are converted to the EROS representation. This is sent through the Deep network to get the Pose. The MoveEnet model indicated in this figure is trained by steps shown in (b): pre-trained on the EROS-like MPII-based image dataset, followed by finetuning on event-Human 3.6m.

other is stationary;

- Persistence: if the person stops moving the representation is not modified and the person remains in the representation;
- Edge features: the surface resembles an edge map making it compatible with edge-extraction of RGB images;

The EROS is not perfect in these aspects and artefacts accumulate on the surface depending on the quality of the input events and the type of scene being observed. The data-driven, network trained on the EROS offers some mitigation of noise and artefacts in the representation.

Finally, EROS is a light-weight representation (around 10 million events per second) and remains asynchronous (it can be queried with millisecond resolution), supporting the targeted goal of high-frequency, on-line HPE.

3.2. Training Paradigm

Training is performed in two-stages: pre-training from RGB image-based datasets and refinement on datasets of event-streams, as shown in Fig. 4b.

3.2.1 Pre-training using RGB Images

Pre-training is now considered to be standard practice, due to its contribution to generalization and robustness in Deep Learning models, given the variety and diversity incorporated in the pre-training datasets, in terms of factors like viewpoints, lighting conditions, environments etc. [16, 40]. In frame-based vision, pre-training was made possible by the availability of large scale image and video datasets. In contrast, large scale datasets in event vision are few and far between [3].

An important aspect of this work was to unlock RGB datasets for training neural networks for event cameras. We bridge the gap between events and frames by processing each data source to arrive at a common, compatible data structure.

EROS generates an edge-like output, with other event-based artefacts. To convert RGB images to a similar representation, we perform a canny edge detection on the image to find points of image gradients and add salt-and-pepper noise, as in [6]. Samples of EROS-like images obtained from RGB datasets are shown in Fig. 3

3.2.2 Refinement from Event-streams

Pre-training on EROS-like images does not capture all the nuances of the artefacts in real EROS representations. The model is therefore refined on the EROS produced from event data.

The ground-truth in HPE datasets is created using motion capture systems with millisecond resolution. The EROS is speed invariant, and can be queried asynchronously at the precise millisecond to synchronise with the ground-truth, resulting in highly accurate association. Representations that accumulate events (i.e. temporal window or number of events), instead have a non-discrete temporal period that is represented. Ambiguity in exact joint position arises if the joint moves many pixel during this period.

3.3. Choice of Network: MoveNet

The methodology described in Sec. 3 can be applied to any deep learning architecture as the EROS can be considered analogous to a greyscale image. The MoveNet [1] architecture is a HPE network and was selected, specifically the Lightning version because it is lightweight and can run at high frequencies even without a dedicated GPU.

MoveNet is a single person, 2D, HPE deep learning model published by Google in May 2021 [1]. The model architecture is inspired by the multi-headed CenterNet [10]. It consists of a MobileNetV2 [17] feature extractor backbone with a Feature Pyramid Network (FPN). Designed for sports

applications, the original version is trained on COCO [23] image dataset and a proprietary dataset.

Since the original API does not allow further training of the model, we trained our model from scratch. The original MoveNet also boasts a number of library-based optimisations that have not been incorporated but could be considered in future work.

4. Implementation

4.1. Datasets

In this work we used a variety of datasets to pre-train and train the proposed models:

- Pre-training: MPII [2, 6] converted into EROS-like representation;
- Finetuning: Event-Human 3.6million (eH36m) converted from images to events by us from [20];
- Finetuning: DHP19 [3];

For the comparative analysis, the two event-based datasets, eH36m and DHP19, allow for a wider comparison to other methods.

4.1.1 MPII

MPII [2] is an image based large scale, multi-person HPE dataset with 25k images from a large variety of viewpoints and cluttered, "in-the-wild" scenarios, that promote generalisation [40].

MoveNet is designed for single-person HPE, while MPII contains samples with multiple people. Therefore, the training samples were cropped to segment each person into an individual sample, thereby maximising the number of samples available. Images in which multiple people were still present in the resulting samples were kept to provide further generalisation of the trained models. In a pre-training dataset, the small amount of the increased environment clutter should be constructive to the training paradigm.

The dataset was cropped and converted into the EROS-like representation as described in Section 3.2.1. Samples of MPII after the conversion are shown in Fig. 3.

4.1.2 Event-Human 3.6million

The Human 3.6m dataset [20] is a widely used large scale, multi-view, benchmark video dataset for single person 2D HPE. Each sequence is captured from 4 video cameras (resolution 1000×1000 pixels; frequency 50 Hz), 2 facing the front and 2 behind each subject. It was recorded on 11 subjects and contains 17 scenarios. We converted this dataset to events with the resolution of the target camera: 640×480, employing v2e, a state of the art method [18]. v2e first generates a set of intermediate frames with a slow-motion model, then creates the events based on the changes between these new consecutive frames. It also adds synthetic noise to the output with a noise model derived from the event camera. The dataset provides a pose annotation frequency of 50 Hz.

The following steps are followed to crop and convert each sample to events: The bounding box (BB) defines the minimum region in the image that captures all joint locations over the full sample video. The boundary is then enlarged by 10 pixels in all directions. If the resulting bounding box is smaller than the target camera resolution, it is set to the target camera resolution maintaining the central position. If the bounding box is larger than the target camera resolution the image is resized (maintaining the aspect ratio). The procedure ensures that:

Algorithm 1 Conversion of video sample to events

Require: $vid = RGB\_video$
Require: $gt = MoCap\_annotation$
$bounding\_box \leftarrow find\_min\_max\_x\_y(gt)$
$bounding\_box \leftarrow add\_margin(bounding\_box)$
$bounding\_box \leftarrow smart\_resize\_to\_cam\_res(bounding\_box)$
$vid\_cropped \leftarrow crop(vid, bounding\_box)$
$event\_stream \leftarrow v2e(vid\_cropped)$

- No joints of the actor are cropped out;
- There is a single cropping area for the sample. Different cropping area between consecutive frames can create an discontinuity between frames resulting it artifacts in the resulting events.
- Joints may move to the edges of the view, mitigating data bias of centrally located objects, thus promoting better training for a deep network.

The original dataset is openly available, and the code for conversion is provided. The choice of timestamp resolution is (the highest possible), 1ms. eH36m has a spatial resolution of 640×480, the maximum allowed by the v2e model, and identical to the camera used to test this method online (Prophesee ATIS Gen3).

The front and back view of subjects look very similar in the EROS representation. Therefore, we only consider the front cameras in this study.

4.1.3 DHP19

The DHP19 dataset [3] is the first and, at the time of writing, the only large scale HPE dataset acquired from event cameras. It is acquired in an indoor, un-cluttered environment, using 4 synchronised cameras (346×280 pixels) placed at

-90, -45, 45 and 90 degrees from the front face of the subject. It provides 3D annotation, and the camera parameters required to calculate the 2D projections. 17 subjects were recorded, performing 33 movements each.

The 13 joint annotations refer to the position of the location of the markers, however the standard practice instead provides the center of the joints using a biomechanical model. The difference in annotation results in DHP19 annotations being different from eH3.6m and MPII. Correcting joint locations to the standard positions should be solved at dataset creation, and is non-trivial in post processing. For example the head (the marker is on top of the head, instead of the centre of the head) could possibly be shifted by a fixed offset, but the same is not true for other joints, in which the joint orientation is required to be known to shift the joint correctly.

Therefore, a model that has been trained on DHP19, would give low accuracy values on another dataset in this study. In contrast, the annotations of MPII and event-human 3.6m are consistent with each other, and with most benchmark datasets and state of the art models like OpenPose and MoveNet.

4.2. Metrics

The primary metric presented in this work is the PCK (percentage of correct keypoints), it is a percentage accuracy measure and is well suited for a single person HPE scenario. PCK is defined as:

$$\frac{100}{N}\sum_{i=1}^{N}\delta(T-d_i) \quad where \quad \delta(x)=\begin{cases}1 & \text{if } x>0\\ 0 & \text{if } x<0\end{cases}$$

where d_i is the distance between the predicted and Ground Truth positions of the i^{th} joint, $N=13$ is the number of keypoints, and T is the percentage of the Threshold, defined by a limb of the subject in the scene. Therefore, the PCK metric is independent of the resolution, the height of the subject, the orientation or viewpoint of the camera, or the distance of the subject from the camera. Instead MPJPE, used in most HPE benchmarks, is biased by the camera resolution, but we still report values for the sake of comparison with other manuscripts. In this work, the PCK threshold is the diagonal torso length of the subject. We use all 13 joints in the calculation of the PCK.

4.3. Technical specifications

All the results presented in this section were performed using a Dell Alienware m15 R3, Intel Core i9-10980HK @ 2.40GHz x 16, NVIDIA GeForce RTX 2070, 32 GB DDR4 @ 2667MHz. The event camera used for online experiments is Prophesee's evaluation kit EVK1 with array size 640x480, pixel pitch 15 μm, optical format 3/4", typical latency 200 μs and events temporal resolution 1 μs.

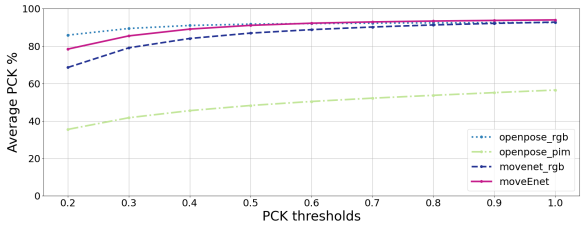


Figure 5. PCK results on Event-Human 3.6m dataset as compared to other methods.

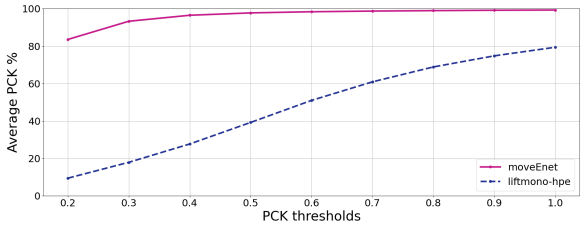


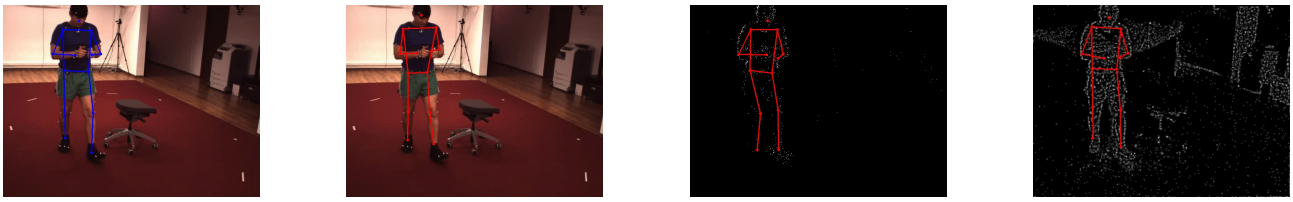
Figure 6. PCK results on DHP19 dataset as compared to other methods.

Training The MoveNet architecture is first pre-trained from scratch on EROS-like representation of the MPII dataset until convergence. The resulting model is trained on a subset of the target dataset, *i.e.* either eH36m or DHP19. The architecture is designed to take inputs of spatial resolution 192×192. Thus, the inputs are resized to fit this dimension, and the resulting predictions are re-scaled to the original size. The method is tested with a 13-joint skeleton. A random hyper-parameter search was executed with eH36m. The optimal parameters found are used for both datasets.

5. Experiments

5.1. Comparison to state-of-the-art

Few methods can be compared directly to our work without placing too many assumptions. We compare to OpenPose-RGB and MoveNet-RGB as we can use the original RGB images of Human 3.6m. As a baseline comparison for an event-based method, we ran OpenPose on polarity-integrated images (PIM, see Sec. 2), which operated reasonably for the conditions present in H3.6m. To facilitate a 2D comparison, LiftMono-HPE [32] was projected onto a single image plane. Considering it produces a 3D pose from a monocular input, the projected joints would be less precise but the comparison is valid for larger PCK thresholds. The model presented with DHP19 [3] is also compared, with additional use of MPJPE_{2D} calculated only on



(a) Input RGB frame and GT (b) Movenet_rgb (c) Movenet_fixedcount (d) MoveEnet

Figure 7. Qualitative results on a sample from eH36m. (a) Frame from original Human 3.6m with superimposed pose annotation; (b) inference of original MoveNet on an RGB frame; (c) fixed count representation and pose estimated by movenet_fixedCount and (d) Result from MoveEnet. The sample has movement in the entire nody simultaneously. All methods are effective in this scenario.

Metric	dhp19 [3]	LiftMono [32]	MoveEnet
PCK@0.4	–	0.28	0.97
PCK@0.6	–	0.51	0.98
$MPJPE_{2D}$	7.03	26.79	6.28

Table 1. Results for DHP19 dataset on available models with multiple metrics. PCK is an accuracy out of 1. MPJPE is an average error in pixels.

Method	Representation	PCK@0.4	PCK@0.8
openpose_rgb	RGB	0.91	0.92
movenet_rgb		0.84	0.91
movenet_fixedCount	fixed Count	0.87	0.93
openpose_pim	PIM	0.46	0.54
movenet_wo_finetune	EROS	0.3	0.52
movenet_wo_pretrain		0.59	0.81
MoveEnet (ours)	EROS	0.89	0.93

Table 2. Comparison of accuracy with method for events-Human 3.6m dataset. 0.4 and 0.8 are the Threshold values.

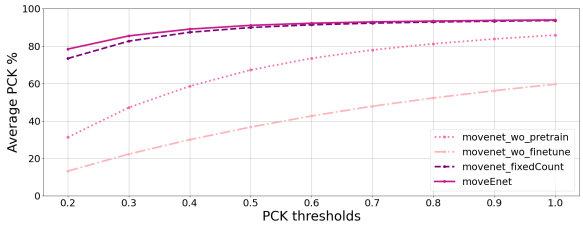


Figure 8. Ablation Study on Event-Human 3.6m dataset.

the front cameras of DHP19 dataset.

Results are shown in Figs. 5 and 6 for eH36m and DHP19 datasets respectively. Select results are also reported in Tabs. 1 and 2.

5.2. Ablation Studies

The presented model, MoveEnet, has 3 different components that contribute to its accuracy and robustness: the EROS representation, the pre-training, and the targeted fine-tuning. The contribution of each of these components can be understood from Fig. 8. Since the pre-training was made possible only due to the relationship between EROS and EROS-like representations, the $movenet\_fixedCount$ which is trained on a fixed-count representation, accumulation of n events, also did not have fine tuning. The associated ground truth is defined as the pose at the average timestamp within the integration time window of the $n = 7500$ used for each input. Results are shown in Fig. 8.

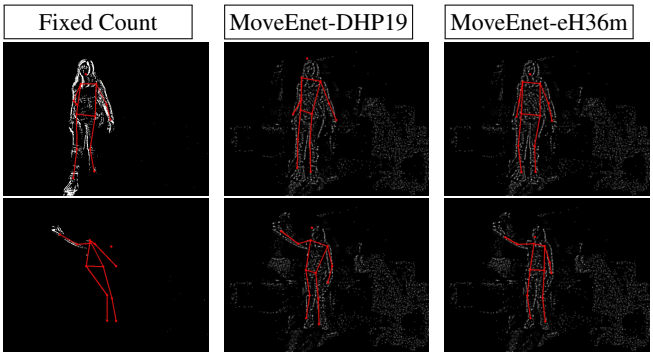


Figure 9. Samples from event stream captured on the event camera, evaluated on (left to right) movenet_fixedCount, movenet_dhp19 (training on lower resolution), and moveEnet. Top: Person is moving their full body: all 3 identify most joints. Bottom: Only the hand moves: The first system does not find the rest of the body.

5.3. On-line Experiments

An event stream was captured from the camera with blind spots, and a single person in the view and played back at regular speed to be tested by MoveEnet, $movenet\_fixedCount$ and $movenet\_dhp19$ (trained on lower resolution). The system was tested to maximum frequency without glitches or system crash. The results are included in a single video in the supplementary material and a few extracted images are shown in Fig. 9.

6. Results and Discussion

Global analysis MoveEnet performs well on eH3.6m dataset and is extremely accurate for DHP19. PCK across the threshold values provides information about the precision of a system, accuracy at lower thresholds requires higher precision. MoveEnet obtains high accuracy at even small threshold values, showing high precision.

With respect to RGB methods processed on RGB values, MoveEnet is less precise than OpenPose, but more precise than MoveNet for low thresholds. At high thresholds, all 3 methods converge. The MoveNet architecture has multiple models. MoveNet has been extensively trained on a large scale proprietary dataset. Thus, like OpenPose, MoveNet is challenging to fine tune, if at all. OpenPose is more accurate than MoveEnet but runs at about 10 Hz as opposed to MoveEnet's 100Hz. In conclusion, MoveEnet provides competitive results to both RGB counterparts, especially for online applications of HPE.

On the DHP19 analysis, MoveEnet performs better than LiftMono-HPE, as expected, especially because of the error added to LiftMono-HPE due to the projection from 2D to 3D and then back. However, this influence is minimal at high threshold values, showing MoveEnet has superior performance. In fact, MoveEnet's performance with DHP19 is extremely accurate at 97% PCK@0.4. Additionally, the DHP19 baseline model shows higher average error for 2D than MoveEnet.

Ablation Without any targeted fine tuning ($movenet\_wo\_finetune$) with EROS, MoveEnet is relatively inaccurate, as expected. Without any pre-training, its precision is lower. $movenet\_fixedCount$, instead, performs well in the structured experiment. This happens for a number of reasons. Firstly, as mentioned in Sec. 5.2 since the fixed-Count representation lacks temporal information, the associated labels are the pose at the average timestamp of the sample, increasing accuracy, but adding latency to the system. Additionally, the dataset has stationary camera. In practice, if the camera moves, the number of events is very high, and an approach based on fixed-Count frames would fail, but the EROS representation is robust to camera movement and therefore is expected to maintain its performance. The same issue can emerge if there is large movement in the scene, with the same results. Moreover, fixed-Count fails to handle blind spots and generalise for camera input. These is evident qualitatively from Fig. 7 where $movenet\_fixedCount$ fails to find the legs of the human agent since the movement is primarily in the torso of the person. Video attached in the supplementary material would clarify this further.

MoveEnet is robust to changes in resolution, as can be seen from Fig. 7: The MoveEnet trained on DHP19 (with resolution of 346×280) performs remarkably on the new data that is acquired with a higher resolution camera (640×480).

With the success of this 2D single pose estimation method, MoveEnet can be run on a Region of Interest, in a larger, multi-person HPE system, by combining it with a person detector. Moreover, with this precision, a stereo system can also be created to get accurate 3D Human Pose.

On-line The system can operate at up to 100Hz without glitch on the GPU enabled system employed. On the CPU-only this value was 20Hz. Fig. 9 shows the results of 3 models on event streams acquired by an event camera. The full video with results from the 3 models is in the supplementary results. MoveEnet generalises to this real world input. The model trained on DHP19 with the lower resolution still performs reasonably well on the higher resolution input. When some parts of the body are in blind spot, MoveEnet with EROS input continues to work accurately, while the fixed Count based model fails.

7. Conclusions

In this work, we presented a system for online, high frequency, 2D, single person Human Pose Estimation. To this end, we created a training paradigm that leveraged the visual similarity of the EROS event representation, with edge detection on images to widen access to annotated training data, thereby obtaining more data for pre-training. This pre-training is followed by a training on the event-based conversion of the widely used baseline Human 3.6m dataset. The proposed system is robust to different resolution in event cameras. Additionally, we demonstrate additional quantitative results on DHP19 and qualitative results on a data sample with a different setting, acquired in house. Our analysis shows that MoveEnet performs competitively to state-of-the art methods, while achieving low-latency online inference, that is crucial in natural interaction of machines with humans in application that requires fast decision making and actuation, for example in ensuring safety in industrial physical human-robot collaboration. MoveEnet can be used in a variety of scenarios, especially robot-machine interaction and collaboration, safety or for sports applications. Expanding MoveEnet to multi-person HPE will entail the development of a low-latency person detector, that could be based on a parallel frame-based pipeline, or rely solely on events.

Acknowledgements

This work was funded by the VOJEXT project (952197) of European Union's Horizon 2020 research and innovation programme.

References

[1] Movenet: Ultra fast and accurate pose detection model. https://www.tensorflow.org/hub/tutorials/movenet. Accessed: 2022-05-20.

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[3] Enrico Calabrese, Gemma Taverni, Christopher Awai Easthope, Sophie Skriabine, Federico Corradi, Luca Longinotti, Kynan Eng, and Tobi Delbruck. DHP19: Dynamic vision sensor 3D human pose dataset. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2019-June:1695–1704, 2019.

[4] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[6] Nicolò Carissimi, Gaurvi Goyal, Franco Di Pietro, Chiara Bartolozzi, and Arren Glover. [wip] unlocking static images for training event-driven neural networks. In *2022 8th International Conference on Event-Based Control, Communication, and Signal Processing (EBCCSP)*, pages 1–4. IEEE, 2022.

[7] Jiaan Chen, Hao Shi, Yaozu Ye, Kailun Yang, Lei Sun, and Kaiwei Wang. Efficient human pose estimation via 3d event point cloud. *arXiv preprint arXiv:2206.04511*, 2022.

[8] Yucheng Chen, Yingli Tian, and Mingyi He. Monocular human pose estimation: A survey of deep learning-based methods. *Computer Vision and Image Understanding*, 192:102897, 2020.

[9] Qi Dang, Jianqin Yin, Bin Wang, and Wenqing Zheng. Deep learning based 2d human pose estimation: A survey. *Tsinghua Science and Technology*, 24(6):663–676, 2019.

[10] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6569–6578, 2019.

[11] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[12] Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jorg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-Based Vision: A Survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2022.

[13] Luna Gava, Marco Monforte, Chiara Bartolozzi, and Arren Glover. How late is too late? a preliminary event-based latency evaluation. In *2022 8th International Conference on Event-Based Control, Communication, and Signal Processing (EBCCSP)*, pages 1–4, 2022.

[14] Daniel Gehrig, Henri Rebecq, Guillermo Gallego, and Davide Scaramuzza. Asynchronous, photometric feature tracking using events and frames. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 750–765, 2018.

[15] Arren Glover, Aiko Dinale, Leandro De Souza Rosa, Simeon Bamford, and Chiara Bartolozzi. luvHarris: A Practical Corner Detector for Event-cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8828(c), 2021.

[16] Gaurvi Goyal, Nicoletta Noceti, and Francesca Odone. Cross-view action recognition with small-scale datasets. *Image and Vision Computing*, page 104403, 2022.

[17] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[18] Y Hu, S C Liu, and T Delbruck. v2e: From video frames to realistic DVS events. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2021.

[19] Massimiliano Iacono, Stephan Weber, Arren Glover, and Chiara Bartolozzi. Towards Event-driven Object Detection with Off-the-shelf Deep Learning. In *IEEE International Conference on Intelligent Robots and Systems*, Madrid, Spain, 2018.

[20] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014.

[21] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Multiposenet: Fast multi-person pose estimation using pose residual network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 417–433, 2018.

[22] Xavier Lagorce, Garrick Orchard, Francesco Galluppi, Bertram E. Shi, and Ryad 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.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[24] Wu Liu and Tao Mei. Recent advances of monocular 2d and 3d human pose estimation: A deep learning perspective. *ACM Computing Surveys (CSUR)*, 2022.

[25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.

[26] Jacques Manderscheid, Amos Sironi, Nicolas Bourdis, Davide Migliore, and Vincent Lepetit. Speed invariant time surface for learning to detect corner points with event-based

cameras. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:10237–10246, 2019.

[27] Elias Mueggler, Christian Forster, Nathan Baumli, Guillermo Gallego, and Davide Scaramuzza. Lifetime Estimation of Events from Dynamic Vision Sensors. In *The IEEE International Conference on Robotics and Automation*, 2015.

[28] David Pascual-Hernández, Nuria Oyaga de Frutos, Inmaculada Mora-Jiménez, and José María Cañas-Plaza. Efficient 3d human pose estimation from rgbd sensors. *Displays*, 74:102225, 2022.

[29] Lingteng Qiu, Xuanye Zhang, Yanran Li, Guanbin Li, Xiaojun Wu, Zixiang Xiong, Xiaoguang Han, and Shuguang Cui. Peeking into occluded joints: A novel framework for crowd pose estimation. In *European Conference on Computer Vision*, pages 488–504. Springer, 2020.

[30] Umer Rafi, Andreas Doering, Bastian Leibe, and Juergen Gall. Self-supervised keypoint correspondences for multi-person pose estimation and tracking in videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 36–52. Springer, 2020.

[31] Beanbonyka Rim, Nak-Jun Sung, Jun Ma, Yoo-Joo Choi, and Min Hong. Real-time human pose estimation using rgb-d images and deep learning. *Journal of Internet Computing and Services*, 21(3):113–121, 2020.

[32] Gianluca Scarpellini, Pietro Morerio, and Alessio Del Bue. Lifting monocular events to 3d human poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1358–1368, 2021.

[33] Simon Schaefer, Daniel Gehrig, and Davide Scaramuzza. Aegnn: Asynchronous event-based graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12371–12381, 2022.

[34] Cedric Scheerlinck, Nick Barnes, and Robert Mahony. Asynchronous spatial image convolutions for event cameras. *IEEE Robot. Autom. Lett.*, 4(2):816–822, April 2019.

[35] Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. HATS: Histograms of Averaged Time Surfaces for Robust Event-Based Object Classification. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (Figure 1):1731–1740, 2018.

[36] Kai Su, Dongdong Yu, Zhenqi Xu, Xin Geng, and Changhu Wang. Multi-person pose estimation with enhanced channel-wise and spatial information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5674–5682, 2019.

[37] Hui Tang, Qing Wang, and Hong Chen. Research on 3d human pose estimation using rgbd camera. In *2019 IEEE 9th international conference on electronics information and emergency communication (ICEIEC)*, pages 538–541. IEEE, 2019.

[38] Jinbao Wang, Shujie Tan, Xiantong Zhen, Shuo Xu, Feng Zheng, Zhenyu He, and Ling Shao. Deep 3d human pose estimation: A review. *Computer Vision and Image Understanding*, 210:103225, 2021.

[39] Lan Xu, Weipeng Xu, Vladislav Golyanik, Marc Habermann, Lu Fang, and Christian Theobalt. Eventcap: Monocular 3d capture of high-speed human motions using an event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4968–4978, 2020.

[40] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

[41] Hong Zhang, Xuzhong Yan, and Heng Li. Ergonomic posture recognition using 3d view-invariant features from single ordinary camera. *Automation in Construction*, 94:1–10, 2018.

[42] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:989–997, 2019.

[43] Shihao Zou, Chuan Guo, Xinxin Zuo, Sen Wang, Pengyu Wang, Xiaoqin Hu, Shoushun Chen, Minglun Gong, and Li Cheng. Eventhpe: Event-based 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10996–11005, 2021.