EV-Gait: Event-based Robust Gait Recognition using Dynamic Vision Sensors

Yanxiang Wang[1]*, Bowen Du[3]*, Yiran Shen[1,2]†, Kai Wu[4],
Guangrong Zhao[1], Jianguo Sun[1], Hongkai Wen[3]†
[1]Harbin Engineering University, [2]Data61 CSIRO, [3]University of Warwick, [4]Fudan University
Email: yiran.shen@csiro.au; hongkai.wen@dcs.warwick.ac.uk

Abstract

*In this paper, we introduce a new type of sensing modality, the Dynamic Vision Sensors (Event Cameras), for the task of gait recognition. Compared with the traditional RGB sensors, the event cameras have many unique advantages such as ultra low resources consumption, high temporal resolution and much larger dynamic range. However, those cameras only produce noisy and asynchronous events of intensity changes rather than frames, where conventional vision-based gait recognition algorithms can't be directly applied. To address this, we propose a new Event-based Gait Recognition (**EV-Gait**) approach, which exploits motion consistency to effectively remove noise, and uses a deep neural network to recognise gait from the event streams. To evaluate the performance of EV-Gait, we collect two event-based gait datasets, one from real-world experiments and the other by converting the publicly available RGB gait recognition benchmark CASIA-B. Extensive experiments show that EV-Gait can get nearly 96% recognition accuracy in the real-world settings, while on the CASIA-B benchmark it achieves comparable performance with state-of-the-art RGB-based gait recognition approaches.*

1. Introduction

Inspired by the principles of biological vision, Dynamic Vision Sensors (DVS) [27, 7, 35] are considered as a new sensing modality for a number of tasks such as visual odometry/SLAM [22, 19, 36], robotic perception [10, 31, 9, 8] and object recognition [39, 24]. Unlike the RGB cameras which produce synchronised frames at fixed rates, the pixels of DVS sensors are able to capture microseconds level intensity change independently, and generate a stream of asynchronous "events". The design of DVS sensors enables many unique benefits over the conventional RGB cameras. Firstly, DVS sensors require much less resource including energy, bandwidth and computation as the events are sparse and only triggered when intensity changes are detected. For example, the DVS128 sensor platform only consumes 150 times less energy than a CMOS camera [27]. Secondly, the temporal resolution of DVS sensors is tens of microseconds which means the DVS sensors are able to capture detailed motion phases or high speed movements without blur or rolling shutter problems. Finally, DVS sensors have significantly larger dynamic range (up to 140dB [27]) than RGB cameras (∼60dB), which allows them to work under more challenging lighting conditions. These characteristics make DVS sensors more appealing over RGB cameras for vision tasks with special requirements on latency, resources consumption and operation environments.

In this paper, we investigate the feasibility of using DVS to tackle the classic gait recognition problem. Specifically, it aims to determine human identities based on their walking patterns captured by the sensors. This is a fundamental building block for many real-world applications such as activity tracking, digital healthcare and security surveillance. In those context, DVS sensors have unique advantages over the standard RGB cameras because i) their low energy and bandwidth footprint makes them ideal for always-on wireless monitoring; and ii) the high dynamic range allows them to work under challenging lighting conditions without dedicated illumination control.

However as shown in Fig. 1 (a), DVS operates in a completely different way than the RGB cameras, which generates asynchronous and noisy events rather than frames when capturing human motion. Therefore, the conventional RGB-based image processing and gait recognition approaches can't be applied directly on the event data. In this paper, we propose a new Event-based Gait Recognition approach, EV-Gait, which is able to work with the noisy event streams and accurately infer the identities based on gait. Concretely, the technical contributions of this paper are as follows:

- To the best of our knowledge, this is the first work investigating event-based gait recognition under practical settings.

*Co-primary authors
†Corresponding authors

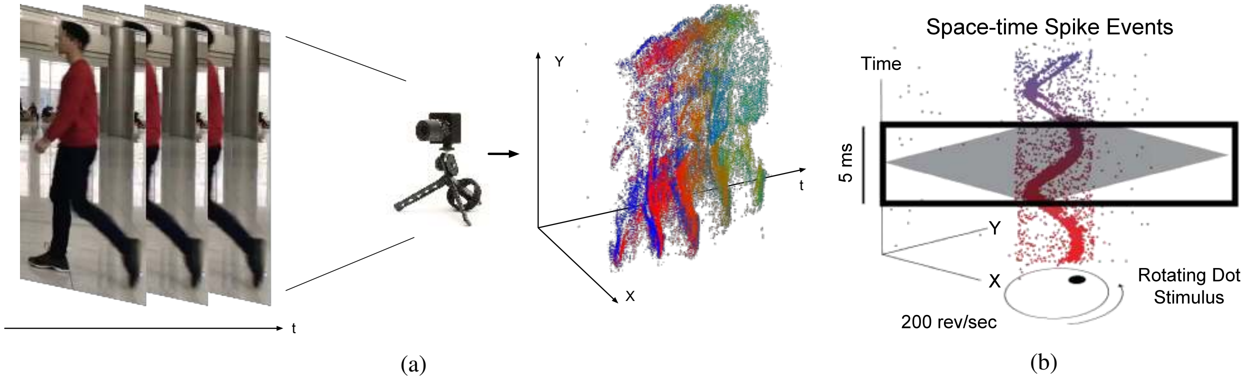


Figure 1. (a) DVS sensor generates asynchronous event stream when a subject is walking in front of it. The positive intensity changes (+1) are denoted in red and negative intensity changes (-1) are in blue. (b) Noisy events stream caused by a rotating dot (adapted from [29]).

- We propose a novel event-based gait recognition approach EV-Gait, which is specifically designed for the dynamic vision sensors. It is able to effectively remove noise in the event streams by enforcing motion consistency, and employs a deep neural network to recognise gait from the asynchronous and sparse event data.

- We collect two event-based gait datasets DVS128-Gait and EV-CASIA-B from both real-world experiments and public gait benchmarks, which will be made available to the community.

- Evaluation on the two datasets shows that the proposed EV-gait can recognise identities up to 96% accuracy in real-world settings, and achieve comparable (even better in some viewing angles) performance with the state-of-the-art RGB-based approaches.

2. Related Work

Gait recognition has been intensively studied for decades in computer vision community [14, 26, 42, 40] and deep learning has been proven to provide state-of-the-art performance on gait recognition without tedious feature engineering [44, 38, 43, 3]. One classic approach for gait recognition proposed in [42] was based on extracted silhouette from background subtraction and modelled the structural and transitional characteristics of gait. Han et al. [15] further improved the silhouette-based approach by extracting scale-invariant features from the gait template. Though template and feature based approaches were widely investigated [40, 30, 41], designing optimal features are still difficult tasks. Deep learning became popular in recent years to solve classification problems in an end-to-end and featureless way. It had been introduced in solving gait recognition problem and produced state-of-the-art performance [44, 38, 43, 3]. Convolutional Neural Networks (CNNs) are known to work well on extracting features from images. Wu et al. [44] proposed different CNN-based architectures for gait recognition and produced state-of-the-art recognition accuracy on CASIA-B dataset. The proposed EV-Gait also uses CNNs, but our network is adapted to process the event data instead of the standard RGB frames.

The excessive noise within the event data has been one of the major challenges for event-based vision. Most of the existing work considered the noise in event data as ad-hoc and sparse. Liu et al [28] searched the eight neighbouring pixels of an incoming event. If there was no other previous events captured within a certain period, it would be marked as noise. Kohoda et al [18] further improved the noise cancellation by recovering events that were mistakenly determined as noise. The work proposed by Padala et al [33] considered a two layers filter. The first layer filtered exploited the fact that two events happened at the same place can't be too close in time domain. The second layer removed the events that lacked spatio-temporal support which was similar to Liu et al [28] approach. However, in this paper, we propose a novel event noise cancellation technique from a new perspective, i.e., the motion consistency in the event stream caused by moving object and show that it outperforms the existing methods by orders of magnitude.

We also review the related work of using DVS sensors for recognition or classification tasks. In [4], the authors applied CNN for identifying gestures, like hand-wave, circling and air-guitar actions. Lagorce at el. [24] proposed a new representation for event data called time-surface then a classification model was built to classify 36 characters(0-9, A-Z). Park et al. [34] employed a shallow neural network to extract the spatial pyramid kernel features for the hand motion recognition using DVS sensor. In addition, Gao at el.[11] used the DVS sensor to track the special markers equipped on the ankle joints of the subjects for gait analysis. However, unlike our approach it did not aim for recognising the identities and required attaching special markers to human bodies which was intrusive.

3. Noise Cancellation for Event Streams

3.1. Dynamic Vision Sensors

Unlike the conventional CMOS/CCD cameras which produce synchronised frames at fixed rate, dynamic vision sensors (DVS) are a class of neuromorphic devices that can capture microsecond level pixel intensity changes as "events", asynchronously at the time they occur. Therefore they are often referred to as the "event cameras", whose output can be described as a stream of quadruplet, (t, x, y, p), where t is the timestamp of an event happens, (x, y) is the location of the event in the 2D pixel space, and p is the polarity. Without loss of generality, we often use $p = +1$ to denote the increase in pixel intensity and -1 as decrease. In practice, the DVS sensors only report such an event when the intensity change at a pixel exceeds certain threshold, i.e.,

$$log\left(I^{x,y}_{now}\right) - log\left(I^{x,y}_{previous}\right) > \theta \tag{1}$$

where $I^{x,y}_{now}$ and $I^{x,y}_{previous}$ are the current and previous intensity at the same pixel (x, y).

Fig. 1 shows an example of how the DVS sensors operate. When an object of interest is moving in the camera field of view, e.g. the rotating dot as in Fig. 1, rather than image frames, the DVS sensor generates an event stream, i.e. the spiral-like shape in the spatial-temporal domain. The asynchronous and differential nature of the DVS sensors brings many unique benefits. For instance, they can have a very high dynamic range (140dB vs. 60dB of standard cameras), which allow them to work under more challenging lighting conditions. The event streams produced by those sensors are at microseconds temporal resolution, which effectively avoids the motion blur and rolling shutter problems. In addition, they are extremely power efficient, consuming approximately 150 times less energy than standard cameras, and have very low bandwidth requirement.

However, one of major challenges of the DVS sensors is that the generated event streams are very noisy. In practice, those sensors are very sensitive to illumination changes or perturbation in the background, and often report large amount of events that are not relevant to the objects of interest. For example, as we can see in Fig. 1, although there is only a rotating dot in the scene, the resulting event stream contains many ad-hoc events that are detached from the desired spiral. This tends to have significant negative impact on the performance of various applications (Sec. 5 will show examples of such impact on gait recognition), which hiders the wide adoption of the dynamic vision sensors. To unlock the full potential of DVS sensors, in the next section we present a novel noise cancellation algorithm, which exploits the spatio-temporal features within the event streams to effectively remove such noise events.

3.2. Noise Cancellation via Motion Consistency

In the context of gait recognition, we are only interested in the people walking (or generally objects moving) within the camera field of view, while the other information captured are considered as noise. As we discussed above, for DVS sensors such noise in the event streams often cause by the subtle illumination changes in the background, or the unstable nature of the electronic circuits. Therefore, the key challenge of noise cancellation is how we can distinguish if an event is triggered by the moving people/objects of interest or not. This is not a straightforward task, since an event stream spans over both spatial and temporal axis and noise can appear arbitrarily. Most of the existing approaches (e.g. [28, 18, 33]) rely on the simple assumption that the noise in the event streams are ad-hoc and sparse, i.e. they should appear in a random fashion and isolated from the the events caused by object motion. However this is not always true, because when the overall lighting condition is not stable, the amount of noise many dominate the stream and bury the events of interest.

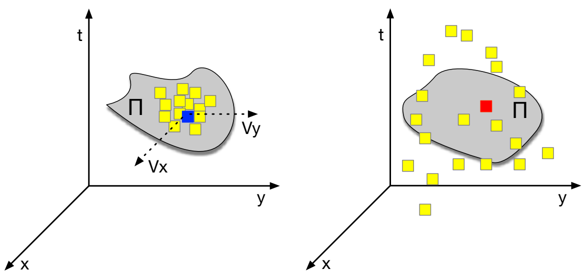
To overcome this problem, we consider a new noise cancellation approach by exploiting the motion consistency within the event streams. The intuition is that if an event is caused by the genuine motion of the objects (human body in our gait recognition case), in the near future there should be another events appear at locations that are consistent with the object motion. In other words, within a local region, the events caused by object motion should be able to form a consistent "moving plane" in the spatial-temporal domain, while the noise event should not. Fig. 2 demonstrates an example of this idea. We see that in Fig. 2(a), for a valid event (the blue dot), there should be a number of previous events that fired in its close vicinity (the yellow dots), since they are triggered by the motion of object across both space and time. Therefore, these events should be able to modelled as a consistent plane Π with velocity (v_x, v_y). On the other hand, as shown in Fig. 2(b), if an event is noise (the red dot), the recently appeared events (the yellow dots) typically have no or little spatial correlation, i.e. they can not be described as a consistent plane. In our approach, we exploit this property by looking at the optical flow within the event streams [6], which can naturally assess motion consistency.

Concretely, to compute the optical flow of an event e_i, we drop its polarity, and express it in the three dimensional space as e_i = (t_i, x_i, y_i). Then the plane where e_i is on can be described as

$$ax_i + by_i + ct_i + d = 0 \tag{2}$$

where a unique $(a, b, c, d) \in \mathbb{R}^4$ defines a unique plane Π.

The for those events that are within close proximity of e_i in both spatial and temporal axis, we fit a plane via least squares:



(a) Valid events (b) Noise events

Figure 2. An example of our noise cancellation approach based on motion consistency. (a) A valid event (blue) and its neighbour events (yellow) should be able to co-locate on the same plane in spatial-temporal domain with reasonable velocity. (b) A noise event (red) can not be fitted on a plane of reasonable velocity with its neighbour events.

$$\hat{\Pi} = \underset{\Pi \in \mathbb{R}^4}{argmin} \sum_{j \in \mathcal{S}_\rangle} \left| \Pi^T \begin{pmatrix} x_j \\ y_j \\ t_j \\ 1 \end{pmatrix} \right|^2 \tag{3}$$

where $\mathcal{S}_i$ is the event set including both e_i and the events appear within the 3×3 neighbourhood of $(x_i,\, y_i)$, and the time window $[t_i - \Delta_t, t_i + \Delta_t]$. In our experiments we set Δ_t to 1ms.

Let us assume that a unique plane $\hat{\Pi}\left(\hat{a}, \hat{b}, \hat{c}, \hat{d}\right)$ is obtained. Then we calculate its velocity at the event e_i as:

$$v = \begin{bmatrix} v_i^x \\ v_i^y \end{bmatrix} = -\hat{c} \begin{bmatrix} \frac{1}{\hat{a}} \\ \frac{1}{\hat{b}} \end{bmatrix} \tag{4}$$

where v_i^x and v_i^y are the velocity of event e_i along the x and y axes respectively. Then we validate the motion consistency by checking the velocity v. If $0 < |v| < V_{max}$, we accept e_i, since a valid event caused by genuine motion should be moving, and the speed should be within certain reasonable range. Otherwise, we declare e_i as noise, and remove it from the event stream. We do this iteratively for each event until all the events in the stream are considered as valid.

4. Event-based Gait Recognition

As shown in Figure 3, Ev-Gait starts from capturing asynchronous raw event stream while the subject is walking through the view. Then the raw event stream is preprocessed through event noise cancellation and represented according to the design of the input layer of the deep neural network for gait recognition. At last, we train our deep network and apply it to recognise the identities of the subjects based on event streams.

4.1. Event Stream Representation

Different from conventional RGB camera, DVS sensors produce asynchronous event streams which can't be directly fitted into state-of-the-art CNN-based structure. In this paper, we adopt the same event stream representation proposed in [46]. Event streams are converted to image-like representation with four channels, termed as event image, for our deep neural networks. The first two channels accommodate the counts of positive or negative events at each pixel respectively. These heatmap-like distributions can effectively describe the spatial characteristics of the event stream. Then the other two channels hold the ratios describing the temporal characteristics. The ratio $r_{i,j}$ at pixel (i, j) is defined as,

$$r_{i,j} = \frac{t_{i,j} - t_{begin}}{t_{end} - t_{begin}} \tag{5}$$

where $t_{i,j}$ is the timestamp of the last event at pixel (i, j), t_{begin} is the timestamp of the first event and t_{end} is the last event of the whole stream. These ratios estimate the lifetime of object of interest at different locations.

After the above processes, the event streams are represented as event images ready for training the deep neural network.

4.2. Deep Recognition Network

Our deep neural network for event-based gait recognition can be vastly divided into two major components: convolutional layers with Residual Block (ResBlock) layers are responsible for feature extraction and fully-connected layers with softmax associate the features to different identities. The convolutional layers have been proved an effective way to extract features and popularly applied in image classification tasks [21, 37, 12]. The ResBlock layers [16] are able to deal with the vanishing features problem when the network goes deeper so that features extracted by convolutional layers can be better integrated. The fully-connected layers decode the features and pass them to the softmax functions to execute classification tasks.

The detailed design of our network is shown in Figure 4. It starts from a special input layer to accommodate the event images presented in Sec. 4.1. The input image is passed through four convolutional layers whose filter size is 3×3 and stride is 2. The number of channels of the four convolutional layers are 64, 128, 256 and 512 respectively. After the convolutional layers, the resultant activations of the ReLu [32] functions are passed through two ResBlock layers to deal with the vanishing gradient problem and keep the features extracted from lower layers when our network goes deeper. The two ResBlock layers share the same parameters: the filter size is 3×3, the stride is 1 and the number of channels are 512. Then, two fully-connected layers with

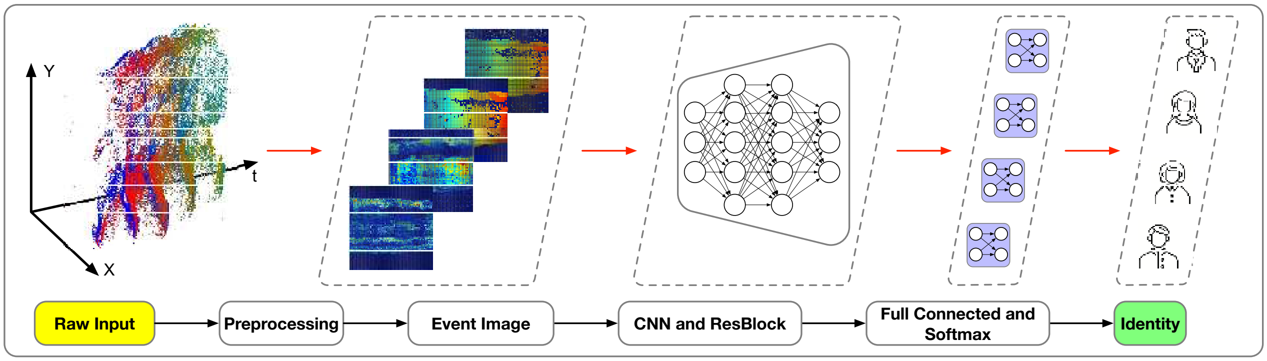


Figure 3. Workflow of the proposed EV-Gait.

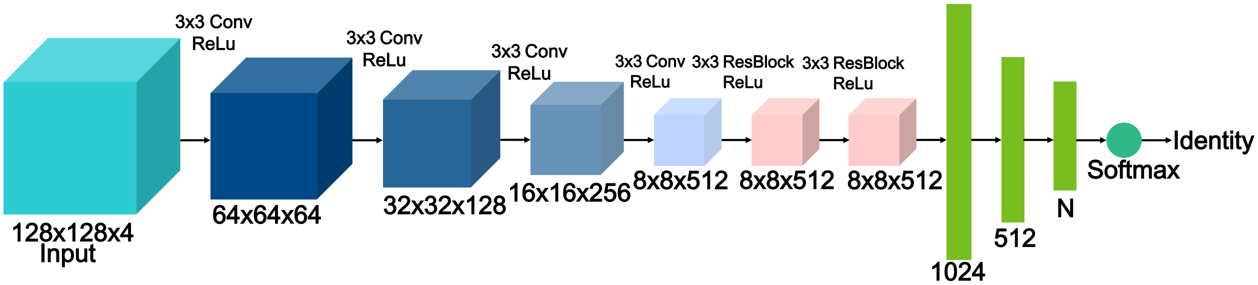


Figure 4. Network architecture of the proposed EV-Gait.

1024 and 512 nodes respectively are connected to the ResBlock layers and softmax functions are stacked to finalise the whole network. At last, the cross entropy loss function and Adam optimizer [20] are adopted to train the network.

5. Evaluation

In this section, we evaluate EV-Gait with both data collected in real-world experiments and converted from publicly available RGB gait databases. In our experiments, we use a DVS128 Dynamic Vision Sensor from iniVation [1] operating at 128×128 pixel resolution. The event data is streamed to and processed on a desktop machine running Ubuntu 16.04, and the deep network (discussed in Sec. 4) is trained on a single NVIDIA 1080Ti GPU. In the following, we first evaluate the performance of event noise cancellation of EV-Gait in Sec. 5.1, and then present the gait recognition performance of our approach in Sec. 5.2.

5.1. Event Noise Cancellation

We compare the proposed noise cancellation technique in EV-Gait against the following three state of the art approaches:

(1) **Liu et al** [28], which discards an event as noise if there is no other event captured at its eight neighbour pixels within a certain time period;

(2) **Khoda et al** [18], which improves Liu's approach by recovering events that are mistakenly classified as noise;

(3) **Padala et al** [33], which filters noise in the event stream by exploiting the fact that two events fire at the same location can't be too close in time domain.

To fully investigate the noise cancellation performance of EV-Gait, we consider two experiment scenarios, where the DVS sensor is configured to capture: i) a static background with nothing moving; and ii) an artificial object moving upon the background.

5.1.1 Noise Cancellation with Static Background

In this experiment setting, we configure the DVS128 camera to face white walls and continuously capture the event streams for fixed time intervals. The environments are controlled and there is no moving object or shadow within the camera field of view, so that the scene captured by the camera is purely static background. We consider two different lighting sources, i) the light-emitting diode (**LED**) and ii) fluorescent tube light (**FTL**), both of which are AC powered. However, the flicker frequency of the fluorescent light is relatively slow (100Hz or 120Hz), and thus can be easily picked up by the DVS sensors, causing more noise in the event streams. On the other hand, the LED lights used in our experiments are more stable, since they use rectifiers to convert the AC to DC and smooth the output with capacitors. Fig. 6(a) and Fig. 5(a) show the the recorded events accumulated within a 20ms window under the two different lighting sources respectively. Clearly in this case, all the events (white dots) should be noise, since the DVS sensor is only capturing the static white wall. We then apply the event noise cancellation technique used in EV-Gait and the competing approaches to the recorded event streams, and Table 1 shows their performance in removing noise.

Firstly, we find that the amount of noise caused by fluorescent tube light (FTL) is much more than that of the LED light (1,082,840 vs. 19,009 noise events), which confirms that DVS sensors are very sensitive to different lighting conditions. On the other hand, we see that our technique can effectively remove most of the noise in the event streams, up to 97.79% and 99.73% under LED and FTL. This significantly outperforms all the completing approaches (see Fig. 5 and Fig. 6 for visualisation of the remaining noise events), where the best one (Khoda [18]) keeps almost 78 times (21.06% v.s. 0.27%) more noise events than ours under the unstable FTL lighting. This is expected as the competing approaches only use spatial and temporal inconsistencies to filter out noise events, while the proposed EV-Gait exploits moving surfaces based on optical flow, which is inherently more robust.

	# of Noise	EV-Gait	Liu [28]	Khoda [18]	Padala [33]
LED	19,009	**2.21%**	29.3%	5.13%	15.56%
FTL	1,082,840	**0.27%**	48.25%	21.06%	47.37%

Table 1. Noise cancellation performance of the proposed and competing approaches under LED (1st row) and FTL lights (2nd row). First column shows the total numbers of noise events under the two lighting conditions, while the rest show the percentage of noise events left after applying individual approaches.

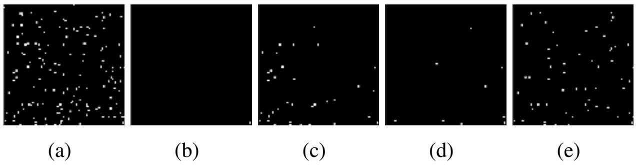


Figure 5. Visualisation of events (400ms) captured for a static background under FTL lighting by (a) no processing; (b) EV-Gait; (c) Liu [28]; (d) Khoda [18] and (e) Padala [33].

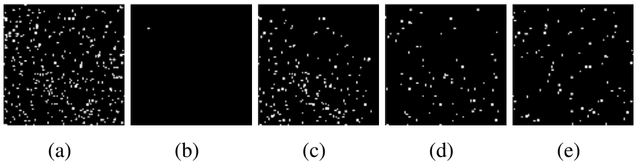


Figure 6. Visualisation of events (400ms) captured for a static background under FTL lighting by (a) no processing; (b) EV-Gait; (c) Liu [28]; (d) Khoda [18] and (e) Padala [33].

5.1.2 Noise Cancellation with Moving Objects

The second set of experiments investigate the performance of different noise cancellation approaches in the presence of moving objects. We again configure the DVS sensor to face the white walls in both LED and FTL lighting conditions, but rather than capturing the background in this case we use a red laser pointer to generate a moving dot on the wall. This moving dot can be captured by the DVS sensor as series of events, as well as the noise. Intuitively, an ideal noise cancellation approach should only extract the events corresponding to that moving dot and discard all the others, forming the complete and clean trajectories. Fig. 7(a) and Fig. 8(a) show the visualisation of events captured by the DVS sensor under LED and FTL lighting. We can see that although there are trajectories visible, the noise events still occupy most of the scene, especially in the FTL case where the lighting source is not very stable (flickering). Fig. 7(b)-(e) and Fig. 8(b)-(e) show the visualisation of events produced by EV-Gait and the competing approaches under LED and FTL lighting respectively. We see that clearly the proposed EV-Gait performs the best, in the sense that it can reject most of the noise events spread across the scene while retaining the positive events corresponding to the moving dot, i.e. preserving the complete and clean trajectories. On the other hand, the competing approaches performs significantly inferior: only Liu [28] and Kohoda [18] could achieve acceptable results under the stable LED lighting (see Fig. 7(c)-(d)), but they immediately fail under the unstable FTL condition (see Fig. 8(c)-(d)).

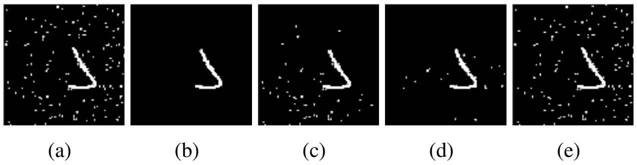


Figure 7. Visualisation of events (400ms) captured for a moving object under LED lighting by (a) no processing; (b) EV-Gait; (c) Liu [28]; (d) Khoda [18] and (e) Padala [33].

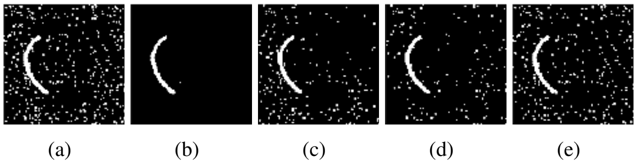


Figure 8. Visualisation of events (400ms) captured for a moving object under FTL lighting by (a) no processing; (b) EV-Gait; (c) Liu [28]; (d) Khoda [18] and (e) Padala [33].

5.2. Gait Recognition

Now we are in a position to present the gait recognition performance of the proposed EV-Gait approach. We evaluate our approach on two event-based gait datasets: i) the **DVS128-Gait** dataset, which is collected in real-world settings with a cohort of 21 volunteers over three weeks; and ii) the **EV-CASIA-B** dataset, which is converted from the

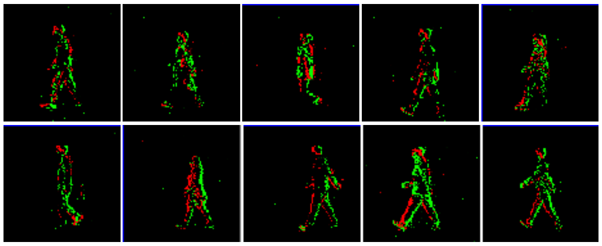


Figure 9. Visualisation of the event streams (accumulated over 20ms) of 10 different identities in the DVS128-Gait dataset.

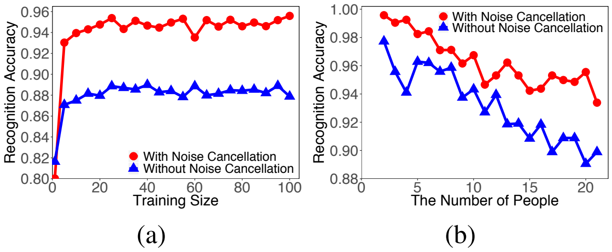


Figure 10. (a) recognition accuracy of EV-Gait (with and without noise cancellation) vs. different training samples per identity. (b) recognition accuracy of EV-Gait (with and without noise cancellation) vs. different number of identities considered.

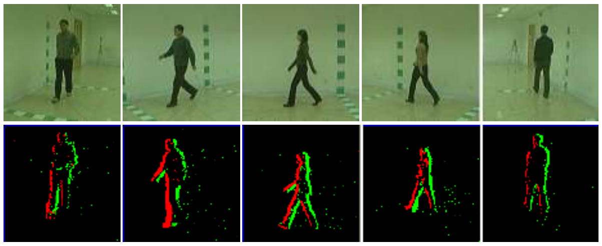


Figure 11. Examples from the original CASIA-B dataset (top row) and visualisation of the corresponding event streams (accumulated over 20ms) in our converted EV-CASIA-B dataset (bottom row).

state-of-the-art RGB camera-based gait recognition benchmark CASIA-B [45].

5.2.1 Performance on DVS128-Gait Dataset

Data Collection: We recruited a total number of 21 volunteers (15 males and 6 females) to contribute their data [1] in two experiment sessions spanning over three weeks time. In each session, the participants were asked to walk normally in front of a DVS128 sensor mounted on a tripod, and repeat walking for 100 times. The sensor viewing angle is set to approximately 90 degrees with respect to the walking directions. The second experiment session was conducted after a week since the end of first session to include potential variances in the participants gait. Therefore, in total we collected 4,200 samples of event streams capturing gait of 21 different identities. Fig. 9 shows visualisation of the data from 4 different identities (events accumulated within 20ms), where the colour of pixels indicate polarity (red for +1, green for -1).

Implementation Details: We implement the proposed deep network in EV-Gait (discussed in Sec. 4) with TensorFlow [2]. The data collected in the first session is used for training, while for testing we use data from the second session. During training we set the batch size as 64 and learning rate as 3e-6. Both training and testing were performed on a 12GB NVIDIA 1080Ti GPU.

Results: The first set of experiments investigate the recognition accuracy of EV-Gait with respect to the amount of training samples per identity. In particular, we use data from all 21 participants, but randomly select different numbers of training samples for each of them, varying from 1 to 100. For each case, we retrain EV-Gait for 30 times and report the averaged recognition accuracy. Fig. 5.2.1 (a) shows the results, and we see that as more samples are used in training, the recognition accuracy of EV-Gait increase immediately, while after 25 samples per identity the accuracy tends to be stable (approximately >94%). This indicates that EV-Gait doesn't require massive training data to converge, and the recognition accuracy is reasonably good even with data collected from practical settings. On the other hand, we also observe that there is a significant performance gap between using vs. not using the noise cancellation technique, e.g. removing the noise in the event stream using our approach can improve recognition accuracy up to 8%. This confirms that the proposed noise cancellation approach in EV-Gait is crucial, and have very positive knock-on effect on the overall gait recognition performance.

We then study the impact on recognition accuracy when the number of identities considered vary. We randomly select a subset of identities (i.e. participants) in the dataset, from 1 to 21 respectively, and use all samples of the selected identities in the training set (data from the first session) to train EV-Gait. We again retrain the model and report the averaged recognition accuracy over 30 inference on the test set, and Fig. 5.2.1 (b) shows the results. We see that as the number of identities increases, the recognition accuracy drops accordingly. This is expected because although we have extra data for training, it is more challenging to distinguish more identities. However, we see that even with 20 identities, EV-Gait can still achieve almost 96% recognition accuracy. In addition, similar with the previous case we observe that the noise cancellation technique in EV-Gait helps a lot, e.g. increasing the accuracy up to 8%.

5.2.2 Performance on EV-CASIA-B Dataset

We have showed that EV-Gait performs well in data collected from real-world settings, and now we show that it

[1] IRB approval for the experiments has been granted

Methods	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	mean
EV-Gait	77.3%	89.3%	94.0%	91.8%	**92.3%**	**96.2%**	**91.8%**	91.8%	91.4%	87.8%	**85.7%**	89.9%
3D-CNN	87.1%	93.2%	97.0%	94.6%	90.2%	88.3%	91.1%	93.8%	96.5%	96%	85.7%	92.1%
Ensemble-CNN	88.7%	95.1%	98.2%	96.4%	94.1%	91.5%	93.9%	97.5%	98.4%	95.8%	85.6%	94.1%

Table 2. Gait recognition accuracy of EV-Gait (evaluated on EV-CASIA-B dataset) and two competing RGB based approaches (evaluated on CASIA-B dataset). Note that for viewing angles 72°, 90° and 108°, EV-Gait even performs better than the RGB based approaches.

could also achieve comparable performance with the state-of-the-art gait recognition approaches that are designed for RGB images. Since those approaches do not work on event streams, for fair comparison, we convert the widely used CASIA-B [45] benchmark into its event version EV-CASIA-B. Then we run EV-Gait on the converted EV-CASIA-B dataset, and compare the resulting recognition accuracy with that of the state-of-the-art approaches on the original CASIA-B dataset.

Data Collection: CASIA-B is one of the most popular benchmark for RGB camera-based gait recognition methods [25, 13, 5, 23]. It contains data from 124 subjects, each of which has 66 video clips recorded by RGB camera from 11 different view angles (0° to 180°), i.e., 6 clips for each angle. The view angle is the relative angle between the view of the camera and walking direction of the subjects. To convert the CASIA-B dataset to event format, we use a similar approach as in [17] and use a DVS128 sensor to record the playbacks of the video clips on screen. In particular, we use a Dell 23 inch monitor with resolution 1920×1080 at 60Hz. Fig. 11 shows some examples from the original CASIA-B dataset (top row) and the visualisation of the corresponding event streams in our converted EV-CASIA-B dataset.

Implementation Details: We consider the same deep network structure as in the previous experiments on the DVS128-Gait dataset. For training, we use the data of the first 74 subjects to pre-train the network. Then for the other 50 subjects, for each viewing angle we use the first 4 out of 6 clips to fine-tune the network, and the rest 2 clips are used for testing. We implement two competing approaches that work on RGB images: i) **3D-CNN** [44] and ii) **Ensemble-CNN** [44], which can achieve state-of-the-art gait recognition performance on the original CASIA-B benchmark.

Results: Table 2 shows the gait recognition accuracy of the proposed EV-Gait with the competing approaches 3D-CNN and Ensemble-CNN. It is worth pointing out that the frame rate of the video clips in CASIA-B dataset is only 25 FPS, with a low resolution at 320×240. As a result when converting such data into event format via playback on the screen, the DVS sensor will inevitably pick up lots of noise. In addition, unlike the original RGB data, the event streams inherently contain much less information (see Fig. 11). However, as we can see from Table 2, the proposed EV-Gait can still achieve comparable gait recognition accuracy (89.9%) with the competing RGB camera based approaches overall (94.1%). For some viewing angles, especially when the walking directions of the subjects are perpendicular with the camera optical axis (e.g. around 90°), the proposed EV-Gait even outperforms the state-of-the-art 3D-CNN and Ensemble-CNN (96.2% vs. 88.3% and 91.5%). This is because in such settings the event streams captured by the DVS sensor can preserve most of the motion features, while removing the gait irrelevant information in RGB images such as cloth texture. On the other hand, for the viewing angles that the subjects walk towards/away from the camera (e.g. 0° or 162°), the accuracy of EV-Gait is slightly inferior to the RGB-based approaches. This is expected, since in those cases compared to RGB images, the event streams contain fewer informative features on the subjects' motion patterns, and thus struggle to extract their identities.

6. Conclusion

In this paper, we propose EV-Gait, a new approach for gait recognition using DVS sensors. EV-Gait features a new event noise cancellation technique exploiting motion consistency of the moving objects to clean up event streams and can be generally applied on a wide range of applications on tracking, localisation, activities recognition using DVS sensors. Then a deep neural network in EV-Gait is designed for recognising gait from event streams. We collect two event-based gait datasets from both real-world experiments and RGB-based benchmark and will make them available to the community. According to the evaluations on the datasets, EV-Gait achieves up to 96% accuracy in real-world settings and comparable performance with state-of-the-art RGB-based approaches on the benchmark.

Acknowledgement This work is partially supported by National Natural Science Foundation of China under Grant 61702133 and International Exchange program of Harbin Engineering University for Innovation-oriented Talents Cultivation. The authors would like to thank NVIDIA for GPU donations.

References

[1] https://inivation.com/support/hardware/dvs128/. *DVS128, Inivation*. 5

[2] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016. 7

[3] Munif Alotaibi and Ausif Mahmood. Improved gait recognition based on specialized deep convolutional neural network. *Computer Vision and Image Understanding*, 164:103–110, 2017. 2

[4] Arnon Amir, Brian Taba, David J Berg, Timothy Melano, Jeffrey L McKinstry, Carmelo Di Nolfo, Tapan K Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, et al. A low power, fully event-based gesture recognition system. In *CVPR*, pages 7388–7397, 2017. 2

[5] Khalid Bashir, Tao Xiang, and Shaogang Gong. Gait recognition using gait entropy image. 2009. 8

[6] Ryad Benosman, Charles Clercq, Xavier Lagorce, Sio-Hoi Ieng, and Chiara Bartolozzi. Event-based visual flow. *IEEE Trans. Neural Netw. Learning Syst.*, 25(2):407–417, 2014. 3

[7] Raphael Berner, Christian Brandli, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240× 180 10mw 12us latency sparse-output vision sensor for mobile applications. In *VLSI Circuits (VLSIC), 2013 Symposium on*, pages C186–C187. IEEE, 2013. 1

[8] Jörg Conradt, Matthew Cook, Raphael Berner, Patrick Lichtsteiner, Rodney J Douglas, and T Delbruck. A pencil balancing robot using a pair of aer dynamic vision sensors. In *Circuits and Systems, 2009. ISCAS 2009. IEEE International Symposium on*, pages 781–784. IEEE, 2009. 1

[9] Tobi Delbruck and Manuel Lang. Robotic goalie with 3 ms reaction time at 4% cpu load using event-based dynamic vision sensor. *Frontiers in neuroscience*, 7:223, 2013. 1

[10] T Delbruck, Michael Pfeiffer, Raphaël Juston, Garrick Orchard, Elias Müggler, Alejandro Linares-Barranco, and MW Tilden. Human vs. computer slot car racing using an event and frame-based davis vision sensor. In *Circuits and Systems (ISCAS), 2015 IEEE International Symposium on*, pages 2409–2412. IEEE, 2015. 1

[11] Ge Gao, Maria Kyrarini, Mohammad Razavi, Xingchen Wang, and Axel Gräser. Comparison of dynamic vision sensor-based and imu-based systems for ankle joint angle gait analysis. In *Frontiers of Signal Processing (ICFSP), International Conference on*, pages 93–98. IEEE, 2016. 2

[12] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 4

[13] Michela Goffredo, Imed Bouchrika, John N Carter, and Mark S Nixon. Self-calibrating view-invariant gait biometrics. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 40(4):997–1008, 2010. 8

[14] Jinguang Han and Bir Bhanu. Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (2):316–322, 2006. 2

[15] Jinguang Han and Bir Bhanu. Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (2):316–322, 2006. 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[17] Yuhuang Hu, Hongjie Liu, Michael Pfeiffer, and Tobi Delbruck. Dvs benchmark datasets for object tracking, action recognition, and object recognition. *Frontiers in neuroscience*, 10:405, 2016. 8

[18] Alireza Khodamoradi and Ryan Kastner. O (n)-space spatiotemporal filter for reducing noise in neuromorphic vision sensors. *IEEE Transactions on Emerging Topics in Computing*, 2018. 2, 3, 5, 6

[19] Hanme Kim, Stefan Leutenegger, and Andrew J Davison. Real-time 3d reconstruction and 6-dof tracking with an event camera. In *European Conference on Computer Vision*, pages 349–364. Springer, 2016. 1

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 4

[22] Beat Kueng, Elias Mueggler, Guillermo Gallego, and Davide Scaramuzza. Low-latency visual odometry using event-based feature tracks. In *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, pages 16–23. IEEE, 2016. 1

[23] Worapan Kusakunniran, Qiang Wu, Jian Zhang, and Hongdong Li. Gait recognition under various viewing angles based on correlated motion regression. *IEEE transactions on circuits and systems for video technology*, 22(6):966–980, 2012. 8

[24] 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. 1, 2

[25] Toby HW Lam, King Hong Cheung, and James NK Liu. Gait flow image: A silhouette-based gait representation for human identification. *Pattern recognition*, 44(4):973–987, 2011. 8

[26] Lily Lee and W Eric L Grimson. Gait analysis for recognition and classification. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 155–162. IEEE, 2002. 2

[27] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120 db 15μ s latency asynchronous temporal contrast vision sensor. *IEEE journal of solid-state circuits*, 43(2):566–576, 2008. 1

[28] Hongjie Liu, Christian Brandli, Chenghan Li, Shih-Chii Liu, and Tobi Delbruck. Design of a spatiotemporal correlation filter for event-based sensors. In *Circuits and Systems (ISCAS), 2015 IEEE International Symposium on*, pages 722–725. IEEE, 2015. 2, 3, 5, 6

[29] Shih-Chii Liu and Tobi Delbruck. Neuromorphic sensory systems. *Current opinion in neurobiology*, 20(3):288–295, 2010. 2

[30] Zongyi Liu and Sudeep Sarkar. Simplest representation yet for gait recognition: Averaged silhouette. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 4, pages 211–214. IEEE, 2004. 2

[31] Elias Mueggler, Nathan Baumli, Flavio Fontana, and Davide Scaramuzza. Towards evasive maneuvers with quadrotors using dynamic vision sensors. In *ECMR*, pages 1–8, 2015. 1

[32] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010. 4

[33] Vandana Padala, Arindam Basu, and Garrick Orchard. A noise filtering algorithm for event-based asynchronous change detection image sensors on truenorth and its implementation on truenorth. *Frontiers in neuroscience*, 12:118, 2018. 2, 3, 5, 6

[34] Paul KJ Park, Kyoobin Lee, Jun Haeng Lee, Byungkon Kang, Chang-Woo Shin, Jooyeon Woo, Jun-Seok Kim, Yunjae Suh, Sungho Kim, Saber Moradi, et al. Computationally efficient, real-time motion recognition based on bio-inspired visual and cognitive processing. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 932–935. IEEE, 2015. 2

[35] Christoph Posch, Daniel Matolin, and Rainer Wohlgenannt. A qvga 143 db dynamic range frame-free pwm image sensor with lossless pixel-level video compression and time-domain cds. *IEEE Journal of Solid-State Circuits*, 46(1):259–275, 2011. 1

[36] Henri Rebecq, Timo Horstschaefer, and Davide Scaramuzza. Real-time visualinertial odometry for event cameras using keyframe-based nonlinear optimization. In *British Machine Vis. Conf.(BMVC)*, volume 3, 2017. 1

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 4

[38] Kohei Shiraga, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. Geinet: View-invariant gait recognition using a convolutional neural network. In *Biometrics (ICB), 2016 International Conference on*, pages 1–8. IEEE, 2016. 2

[39] Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. Hats: Histograms of averaged time surfaces for robust event-based object classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1731–1740, 2018. 1

[40] Dacheng Tao, Xuelong Li, Xindong Wu, and Stephen J Maybank. General tensor discriminant analysis and gabor features for gait recognition. *IEEE transactions on pattern analysis and machine intelligence*, 29(10), 2007. 2

[41] Liang Wang, Tieniu Tan, Weiming Hu, Huazhong Ning, et al. Automatic gait recognition based on statistical shape analysis. *IEEE transactions on image processing*, 12(9):1120–1131, 2003. 2

[42] Liang Wang, Tieniu Tan, Huazhong Ning, and Weiming Hu. Silhouette analysis-based gait recognition for human identification. *IEEE transactions on pattern analysis and machine intelligence*, 25(12):1505–1518, 2003. 2

[43] Thomas Wolf, Mohammadreza Babaee, and Gerhard Rigoll. Multi-view gait recognition using 3d convolutional neural networks. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 4165–4169. IEEE, 2016. 2

[44] Zifeng Wu, Yongzhen Huang, Liang Wang, Xiaogang Wang, and Tieniu Tan. A comprehensive study on cross-view gait based human identification with deep cnns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1. 2, 8

[45] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 4, pages 441–444. IEEE, 2006. 7, 8

[46] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Ev-flownet: Self-supervised optical flow estimation for event-based cameras. *arXiv preprint arXiv:1802.06898*, 2018. 4