GaitSpike: Event-based Gait Recognition With Spiking Neural Network

Ying Tao[1], Chip-Hong Chang[1,2], Sylvain Saïghi[2,3], Shengyu Gao[1,2]
[1]*School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798*
[2]*CNRS@CREATE, 1 Create Way, 08-01 Create Tower, Singapore 138602*
[3]*Univ. Bordeaux, CNRS, Bordeaux INP, IMS, UMR 5218, F-33400 Talence, France*
Email: TAOY0006@e.ntu.edu.sg, ECHChang@ntu.edu.sg

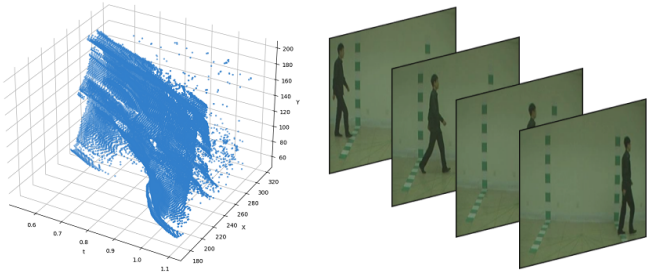
***Abstract*—Existing vision-based gait recognition systems are mostly designed based on video footage acquired with RGB cameras. Appearance-, model- and motion-based techniques commonly used by these systems require silhouette segmentation, skeletal contour detection and optical flow patterns, respectively for features extraction. The extracted features are typically classified by convolutional neural networks to identify the person. These preprocessing steps are computationally intensive due to the high visual data redundancies and their accuracies can be influenced by background variations and non-locomotion related external factors. In this paper, we propose GaitSpike, a new gait recognition system that synergistically combines the advantages of sparsity-driven event-based camera and spiking neural network (SNN) for gait biometric classification. Specifically, a domain-specific locomotion-invariant representation (LIR) is proposed to replace the static Cartesian coordinates of the raw address event representation of the event camera to a floating polar coordinate reference to the motion center. The aim is to extract the relative motion information between the motion center and other human body parts to minimize the intra-class variance to promote the learning of inter-class features by the SNN. Experiments on a real event-based gait dataset DVS128-Gait and a synthetic event-based gait dataset EV-CASIA-B show that GaitSpike achieves comparable accuracy as RGB camera based gait recognition systems with higher computational efficiency, and outperforms the state-of-the-art event camera based gait recognition systems.**

I. Introduction

Gait recognition is a biometric technology to identify people by the way they walk. It is contactless, unobtrusive, non-cooperative and the biometric data can be collected from long distances with low resolution. As these merits are not simultaneously possessed by other human biometrics, gait recognition found versatile applications in visual surveillance, access control, smart home, neurological disorders diagnosis, rehabilitation therapy, criminal investigation, forensic identification, etc.

For gait recognition on frame-based images captured by conventional RGB cameras, appearance-, model- and motion-based methods are most commonly used. Appearance-based methods [1]–[3] obtain human silhouettes via background segmentation. Model-based methods [4]–[6] extract human skeletons using pose estimation models. Motion-based methods [7]–[9] compute optical flow for motion information. These RGB-based gait recognition methods are computationally intensive in gait features extraction. Due to the high spatial and spectral redundancies, backdrop variations can influence the preprocessing and cause a significant degradation on the classifier performance. To minimize the computational inefficiency and sensitivity of preprocessing pipelines, we use event camera for gait recognition.

Event camera [10] uses dynamic vision sensor (DVS), which is an emerging CMOS sensor designed to mimic the sensing and visual-processing characteristics of living organisms. DVS pixels only respond asynchronously to relevant changes in illumination intensity. The outputs are sparse events expressed digitally in "timestamp, address, polarity" format to record the precise timing and locations of the firing pixels. A pixel fires when the light intensity it sensed increases above a positive threshold or decreases below a negative threshold. The event-driven mechanism and succinct output format enable DVS to significantly reduce power consumption, spurious computations, and transmission bandwidth of conventional image sensors. DVS has high temporal resolution (less than 10 μs). It can produce thousands of events per second, which allow high-speed movement or detailed motion phase to be captured without motion-blur or rolling shutter effects. Since only brightness variations (edges) instead of static imagery are captured, DVS imaging also preserves personal privacy. Figs. 1(a) and 1(b) show the event data and four frames from a video footage of a human walking from one end to the other captured by a DVS camera and a RGB camera, respectively. The person and the activity cannot be made out by looking at the 3D plot of event data in Fig. 1(a) but they can be easily identified from Fig. 1(b).



(a) 3d visualization of event data (b) Images from a video

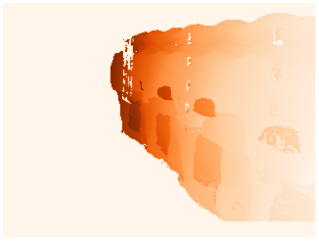
Fig. 1. Comparison of data produced by DVS and RGB cameras

Using event camera for gait recognition has aroused interest recently. Existing methods [11], [12] convert the events into images or graphs and input them to the convolutional neural network (CNN) or graph neural network (GNN) accordingly for classification. These methods achieve a correct recognition rate of 87.3% and 94.9% on DVS128-Gait dataset, and 89.9% and 82.2% on EV-CASIA-B dataset converted from RGB-based CASIA-B [13], respectively. These results are still worse than the state-of-the-art performance of CNN and GNN using RBG-based images of CASIA-B directly for gait detection. To improve the performance, we propose to apply spiking neural network (SNN) to preserve the spatial-temporal relationship and develop a domain-specific representation to extract invariant features of human locomotion for event-based gait recognition.

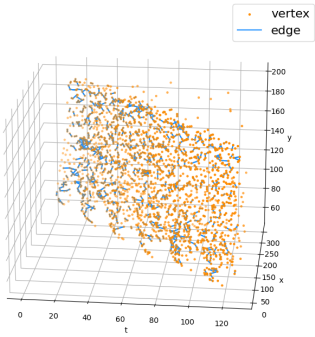
SNNs [14] closely mimic biological neural networks by including time in the computation. They are intrinsically sensitive to the temporal characteristics of information transmission as in the biological neural systems. In SNNs, neurons communicate through binary signals (also called spikes) and the inter-neuron communication is event-driven and asynchronous. Although DVS and SNN fit hand in glove due to their common working principle, some application context adaptive features and sparsity-driven learning paradigm needs to be taken into consideration to make the most of them.

Inspired by the gait feature in [15] which extracts the boundary points vector related to centroid, we design a new representation to express event data captured by DVS camera. The aim is to reduce the intra-class variance of the relative distances from the motion center to the head, shoulders, hip and limbs of a human body that are crucial for gait analysis. We find the motion center every certain time and calculate the distances and angles between the motion center and other event points. We use these distances and angles as floating spatial coordinates to replace the original address of DVS event and we call this representation the locomotion-invariant representation (LIR). Our proposed GaitSpike, is an event-based gait recognition system that uses SNN to process LIR data converted from the raw DVS outputs. To the best of our knowledge, this is the first attempt to use SNN for gait recognition. Extensive experiments are conducted on a real event-based gait dataset DVS128-Gait and a synthetic event-based gait dataset EV-CASIA-B. The results show that our

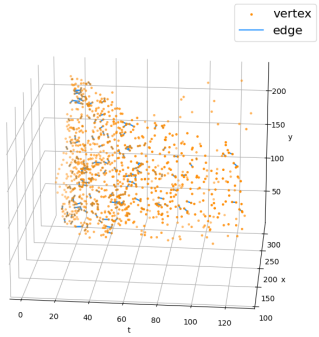
(a) Event image of id 001 at 90°



(b) Event image of id 001 at 162°



(c) Event graph of id 001 at 90°



(d) Event graph of id 001 at 162°

Fig. 2. Visualization of event images and event graphs

method outperforms the state-of-the-art gait recognition methods, particularly for the bag-carrying and cloth-changing conditions.

II. Preliminaries

A. Dynamic Vision Sensor

Each pixel element of a DVS mimics a biological neuron, which generates a spike event upon sensing a temporary log intensity change of magnitude above a specific threshold ϑ [10]. An event is fired when

$$|\log(I_t) - \log(I_{t-1})| > \vartheta, \tag{1}$$

where I_t denotes the instantaneous intensity at time t.

Each event is described by a spatio-temporal representation (x, y, t, p), where (x, y) is the event coordinates in the 2D pixel array, t is the event firing time, and p is the event polarity. $p = 1$ if the event is fired by an increase in temporal intensity and -1 if it is due to a decrease in intensity.

B. Event-based Gait Recognition

Event images [11] and event graphs [12] have been proposed for event-based gait recognition. In event image representation, an event stream is converted to a frame-based image-like representation with four channels. The first two channels accumulate the number of events of positive and negative polarities, respectively at each pixel. The other two channels calculate the ratio of the timestamp of the most recent event at each pixel (one channel for positive polarity events and another for negative polarity events) to the total event accumulation time. Each pixel of an event image is defined by a ratio $r_{x,y}$ as follows:

$$r_{x,y} = \frac{t_{x,y} - t_{begin}}{t_{end} - t_{begin}}, \tag{2}$$

where $t_{x,y}$ is the latest event timestamp at pixel (x, y). t_{begin} and t_{end} are the timestamps of the first and last event, respectively of each event stream.

In event image representation, different pixels of an event image may have similar accumulation numbers in the first two channels and $r_{x,y}$ values in the last two channels when they have similar event counts and the same latest timestamp, even if their timestamp distributions are totally different. Event images are susceptible to change in viewing angle. Figs. 2(a) and 2(b) show the event images in the third channel of the same identity 001 in synthetic EV-CASIA-B at 90° and 162° orientations, respectively. As can be seen, most pixels in each event image have similar $r_{x,y}$ values. Also, the event images of the same identity at two different viewing angles look completely different.

In event graph representation, event nodes are used to connect events of the nearest spatio-temporal distance. Two nodes are connected if the distance between the events represented by them falls within a predefined threshold R, i.e.,

$$\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + \alpha(t_i - t_j)^2} < R. \tag{3}$$

where x_i, y_i and t_i refer to the x and y coordinates and the timestamp of the event represented by the i-th node.

As event nodes are connected based solely on their spatio-temporal proximity, different parts of the same target may be connected very differently under different viewing angles, target's motion and velocity. The arms may be connected to the hip when some occlusions happen. As can be seen in Figs. 2(c) and 2(d), where the orange event data points and thin blue lines are nodes and edges of event graphs, the distribution of event data and the graph connections vary for the same person at different angles. The connections provide no valuable clue to gait detection.

The event image and event graph result in significant intra-class variance, which could hinder instead of helping the classifier to learn the distinguishing features. Neither of them provides lucrative invariant biometric features for gait analysis. The strides, and swings of body or arms are not discriminatively reflected in the representations. To better retain the spatial correlation between principal parts of human skeleton so that subtle locomotion differences across subjects will be learnt more efficiently, we propose a locomotion-invariant motion-based representation.

III. GaitSpike: Event-based Gait Recognition by SNN

A. Spiking Neural Network

SNN is a perfect model to process time series data generated by DVS as it computes continuously on spike train that matches well with the asynchronous data output of DVS. Most specifically, the performance of a SNN is contributed by the presence or absence of a signal rather than its numerical value. The signal is transmitted in the form of spikes. The spikes are weighted by the synapses as they propagate to the post-synaptic neurons. The post-synaptic neurons are activated when their membrane potentials are accumulated by the weighted spikes to a certain threshold value. Upon activation, a neuron emits a spike to the subsequent layer and resets its membrane potential to a rest potential. The membrane potential state records and updates the contributions from the spikes as and when they arrive without testing for the activation in every iteration of propagation like a CNN.

The Leaky-Integrate and Fire (LIF) neuron is the most adopted operating principle in SNN computation. A LIF neuron responds to the input spikes as follows:

$$\begin{aligned} V_j(t) =& \beta V_j(t-1) + \sum_i \big(W_{ij} S_i(t-1)\big) \\ & - S_j(t-1) V_\tau, \end{aligned} \tag{4}$$

where $V_j(t)$ is the membrane potential of neuron j at time t, β is the decay rate of the membrane potential, W_{ij} is the synaptic weight between neurons i and j, and $S_j(t) \in \{0, 1\}$ is the spike generated by neuron j when its membrane potential exceeds the threshold V_τ.

$$S_j(t) = \begin{cases} 1, & if\ V_j(t) > V_\tau \\ 0, & otherwises \end{cases} \tag{5}$$

For simplicity, we use the LIF neurons in each layer to compute the convolution of the signal through the synapses.

B. Locomotion-invariant Representation

To facilitate the learning of robust locomotor movement characteristics of individuals, we propose a locomotion-invariant representation (LIR). The main idea of LIR is to replace the absolute coordinates of address-event data representation by relative coordinates to avoid the problems of inconsistent displacement of human skeleton features with change in direction, viewing angle and occlusion of locomotion.

Given a data sample with N input events $\{(x_i, y_i, t_i, p_i)\}_{i \in [1,N]}$, we split the event data into B time bins. The index of the time bin b of each event i is calculated by:

$$b = (B - 1) \times \frac{t_i - t_1}{t_N - t_1}. \tag{6}$$

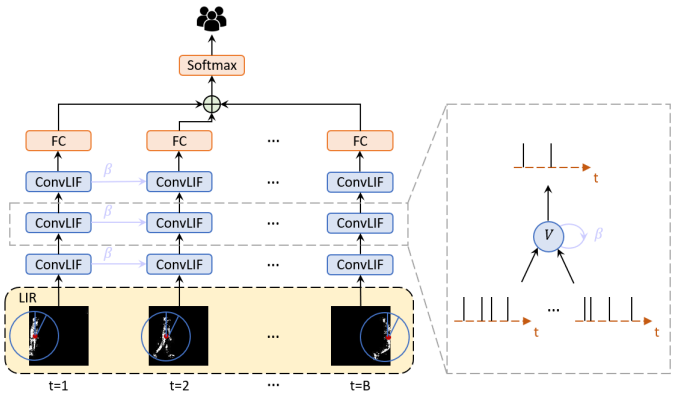


Fig. 3. Architecture of GaitSpike

TABLE I
THE PERFORMANCE COMPARISONS ON DVS128-GAIT

Method	EV-Gait-IMG [11]	EV-Gait-3DGraph [12]	GaitSpike
Accuracy	87.3	94.9	**96.5**

The motion center of each time bin b is first determined by the average coordinates (x_o^b, y_o^b) of all the events that fall within the time bin. Based on the motion center, the polar coordinate of the event i of time bin b is calculated by

$$\rho_i^b = \sqrt{(x_i^b - x_o^b)^2 + (y_i^b - y_o^b)^2}, \tag{7}$$

$$\theta_i^b = arctan\frac{y_i^b - y_o^b}{x_i^b - x_o^b}. \tag{8}$$

Then, a sparse event-based binary image for each time bin is constructed based on the presence ('1') or absence ('0') of an event in the polar coordinate defined in (7) and (8). The image size is $M_\rho \times N_\theta$, which is predefined according to the spatial resolution of the DVS. The row and column numbers of the binary image are the quantized radial distance ρ and quantized polar angle θ, respectively. For simplicity, we round the real value of ρ_i^b to the nearest integer $\lfloor \rho_i^b + 0.5 \rfloor$ and ignore $\lfloor \rho_i^b + 0.5 \rfloor \geq M_\rho$. θ_i^b is uniformly quantized to $\frac{\theta_i^b}{2\pi} N_\theta$. The polarity of the event i has little influence and is not considered in this LIR representation.

LIR focuses on events that happen within a bounding circle centered at the motion center of each bin. The motion center approximates the center of mass of a human body silhouette. Events occur far from the motion center are insignificant for human locomotor action analysis. The relationship between the skeletal movements and the motion center is encoded in the form of distance and angle, which are important features for distinguishing individual gaits in different movements, occlusions and viewing angles.

C. Architecture of GaitSpike

The whole architecture of GaitSpike is shown in Fig. 3. A four-layer SNN is used to process the succinct LIR images derived from the DVS event stream. The SNN consists of three convolutional layers of LIF neurons and a fully connected layer. The output layer has the same number of neurons as the number of subjects to be identified. Each subject is labeled by a unique identity number in the training dataset. Cross entropy spike rate [16] is used as the loss function for backpropagation during training. We accumulate the spikes throughout the timesteps in the output neurons to obtain the spike count vector $\vec{c}$, where the i^{th} element of $\vec{c}$ is denoted as c_i, $i = 1, 2, \cdots, N_C$, and N_C is the number of output neurons. The spike count rate can then be obtained by feeding these N_C spike counts into a softmax function as follows:

$$p_i = \frac{e^{c_i}}{\sum_{i=1}^{N_C} e^{c_i}}. \tag{9}$$

The cross entropy between p_i and the target label $y_i \in \{0, 1\}^{N_C}$ is obtained by:

$$L_{CE} = \sum_{i=1}^{N_C} y_i \log(p_i). \tag{10}$$

The problem with backpropagation through time (BPTT) is the intractable gradient of SNN. The emitted spike is a heaviside step function. When the neuron is activated, the derivative of the spike with respect to the membrane potential is infinity, which leads to the 'exploding gradients' problem. When the neuron is inactive, the derivative of the spike with respect to the membrane potential is zero, which leads to the 'dead neuron' problem. A recommended solution is to use the step function in the forward pass but a surrogate gradient function [17] for backpropagation. In our experiment, sigmoid function is used as the surrogate function.

IV. EXPERIMENTS

A. Datasets and Training Details

Event-based DVS128-Gait [12] and synthetic EV-CASIA-B are used for the evaluation of GaitSpike.

DVS128-Gait. DVS128-Gait [12] consists of the footages of 20 subjects (14 males and 6 females) walking normally in front of a DVS camera. The viewing angle is 90° with respect to the walking direction. Each identity has 200 event streams. Each event stream lasts for 3 seconds. There are 4000 samples in total. We use 2000 samples for training and 2000 samples for testing.

EV-CASIA-B. CASIA-B [13] consists of the footages of 128 subjects walking under 11 different camera viewpoints uniformly distributed in [0°,180°]. Each subject in each viewing angle has 10 samples of 3 seconds each under 3 different conditions, including 6 normal walking (NM), 2 walking with bags (BG), and 2 walking with a coat worn (CL). CASIA-B is a video-based dataset. We use a DVS simulation tool [20] to convert CASIA-B to a synthetic event database EV-CASIA-B. We use four NM samples for training and the other six samples for testing each subject in each viewing angle.

Implementation Details: We implement our network in snnTorch [21]. The input size $M_\rho \times N_\theta$ is set to half the resolution of the event camera. So, for DVS128-Gait, the input size is 64×64. For EV-CASIA-B, the input size is 160×120. Following the setting of [12], the time bin number is set as 128. During training, we set the batch size as 64 and the learning rate as 1e-4. Adam is chosen as the optimizer. Both the training and testing are performed on NVIDIA GeForce RTX 3090.

B. Comparisons with Related Methods

Evaluation on DVS128-Gait. We compare our GaitSpike with two event-based gait recognition methods, EV-Gait-IMG [11] and EV-Gait-3DGraph [12]. The subject identification accuracies of the three methods are shown in Table I, which show that GaitSpike outperforms EV-Gait-IMG and EV-Gait-3DGraph.

Evaluation on EV-CASIA-B. We also compare GaitSpike with four latest RGB-based gait recognition methods, GaitSet [2], GaitPart [3], LagrangeGait [18] and DANet [19]. The results in Table II show that GaitSpike outperforms these state-of-the-art methods in all walking conditions. Although the models are all trained using only samples in NM condition, the accuracy of our model deteriorates less than other methods in BG and CL conditions. An interesting phenomenon is observed in Table II. While all the other methods have the worst performance at a viewing angle of 90° and best performance at 36° or 134°, our model achieves the best accuracy of around 90° and superior accuracy at 0° or 180°. The swinging of arms is occluded by the body at the 90° view which causes the poor performance in RGB-based methods. However, DVS captures the dynamic motion of arms even at 90°, which preserves the recognition accuracy. A possible reason for the minimum accuracy of our model at around 36° or 134° view is the larger error in the calculated motion center at these angles.

C. Computation Overheads

We calculate the computation overheads in terms of the number of parameters (#params) and number of floating-point operations (#FLOPs) of GaitSpike and some RGB-based methods, including model-based GaitGraph [6], appearance-based GaitSet [2] and motion-based ResNet with optical flow (OFResNet-34) [9], whose performances are close to the state-of-the art. The results are shown in Table III. 'Feature extraction' refers to LIR for

TABLE II
THE PERFORMANCE COMPARISONS OF FOUR RGB-BASED METHODS ON CASIA-B AND GAITSPIKE ON EV-CASIA-B

	Method	Probe View											Mean
		0°	18°	36°	54°	72°	90°	108°	126°	134°	162°	180°	
NM#5-6	GaitSet [2]	90.8	97.9	**99.4**	96.9	93.6	91.7	95.0	97.8	98.9	96.8	85.8	95.0
	GaitPart [3]	94.1	98.6	99.3	98.5	94.0	92.3	95.9	98.4	99.2	97.8	90.4	96.2
	LagrangeGait [18]	95.7	98.1	99.1	98.3	96.4	95.2	97.5	99.0	99.3	98.9	94.9	97.5
	DANet [19]	96.4	**99.1**	99.2	98.2	96.6	95.5	97.6	99.4	**99.5**	**99.3**	96.9	98.0
	GaitSpike(ours)	**99.3**	**99.1**	99.3	**99.5**	**99.1**	**99.7**	**99.7**	**99.9**	97.8	96.2	**97.7**	**98.9**
BG#1-2	GaitSet [2]	83.8	91.2	91.8	88.8	83.3	81.0	84.1	90.0	92.2	94.4	79.0	87.2
	GaitPart [3]	89.1	94.8	96.7	95.1	88.3	94.9	89.0	93.5	96.1	93.8	85.8	91.5
	LagrangeGait [18]	94.2	96.2	96.8	95.8	94.3	89.5	91.7	96.8	98.0	97.0	90.9	94.6
	DANet [19]	95.0	**97.3**	98.3	97.4	94.7	91.0	93.9	97.4	**98.2**	**97.6**	94.2	95.9
	GaitSpike(ours)	**98.8**	95.2	**98.4**	**97.6**	**98.8**	**99.6**	**99.6**	**99.6**	97.6	92.7	**95.2**	**97.5**
CL#1-2	GaitSet [2]	61.4	75.4	80.7	77.3	72.1	70.1	71.5	73.5	73.5	68.4	50.0	70.4
	GaitPart [3]	70.7	85.5	86.9	83.3	77.1	72.5	76.9	82.2	83.8	80.2	66.5	78.7
	LagrangeGait [18]	77.4	90.6	93.2	90.2	84.7	80.3	85.2	87.7	89.3	86.6	71.0	85.1
	DANet [19]	82.8	**94.8**	96.9	94.3	89.0	83.9	87.9	92.3	95.1	**92.0**	80.3	89.9
	GaitSpike(ours)	**96.4**	93.5	**97.6**	**96.8**	**98.0**	**99.2**	**98.4**	**98.0**	**97.6**	89.1	**91.5**	**96.0**

TABLE III
COMPARISON OF COMPUTATION OVERHEADS

Method	Feature Extraction		Model	
GaitSpike	#Params	#FLOPs	#Params	#FLOPs
	None	**4.3M**	**0.8M**	2.5G
GaitGraph [6]	#Params	#FLOPs	#Params	#FLOPs
	28.5M	11.1G	2M	**0.7G**
GaitSet [2]	#Params	#FLOPs	#Params	#FLOPs
	None	19.0M	2.6M	3.3G
OFResNet [9]	#Params	#FLOPs	#Params	#FLOPs
	None	35.4M	21.5M	3.6G

GaitSpike, pose estimation model HRNet [22] for GaitGraph, silhouette extraction for GaitSet, and optical flow calculation for OFResNet-34. 'Model' refers to SNN for GaitSpike, GNN for GaitGraph, and CNN for both GaitSet and OFResNet-34. It is clear that RGB-based methods incur considerable computational overheads in feature extraction while the event-based feature extraction of GaitSpike is less computationally intensive. Moreover, in terms of model complexity, GaitSpike has the least number of parameters and the second lower number of FLOPs. GaitGraph has the least number of FLOPs but incurs a huge computational cost in feature extraction. Overall, GaitSpike is the most computationally efficient method.

D. Ablation Study

TABLE IV
COMPARISONS BETWEEN GAITSPIKE WITH AND WITHOUT LIR EVALUATED WITH DVS128-GAIT AND EV-CASIA-B

Dataset\Method		GaitSpike	GaitSpike w/o LIR
DVS128-Gait [12]		**96.5**	86.6
EV-CASIA-B	NM	**98.9**	97.6
	BG	**97.5**	96.0
	CL	**96.0**	95.5

To verify the effectiveness of our locomotion-invariant representation (LIR), we remove LIR from GaitSpike by inputting the event-based binary frame to the SNN directly for gait recognition. As shown in Table IV, when tested with the real DVS dataset DVS128-Gait, the correct identification rates of GaitSpike with and without LIR are 96.5% and 86.6%, respectively. There is an accuracy improvement of 9.9% with LIR. When tested with the synthetic dataset EV-CASIS-B, the identification accuracies of GaitSpike without LIR are 97.6%, 96.0% and 95.5% in NM, BG and CL conditions, respectively. The accuracies under these three conditions increase by 1.3%, 1.5% and 0.5% with LIR to 98.9%, 97.5% and 96.0%, respectively. The improvements are not as significant in the synthetic DVS dataset. This is because the event streams of EV-CASIS-B dataset are converted from the video frames of RGB sensors, which are less noisy than real outputs of DVS sensors. The event samples of real DVS contain a significant amount of background activity noises alongside the signal, which makes gait recognition more challenging for DVS128-Gait dataset. This ablation study indicates the robustness of LIR in the presence of more complex noises in raw event outputs of DVS cameras.

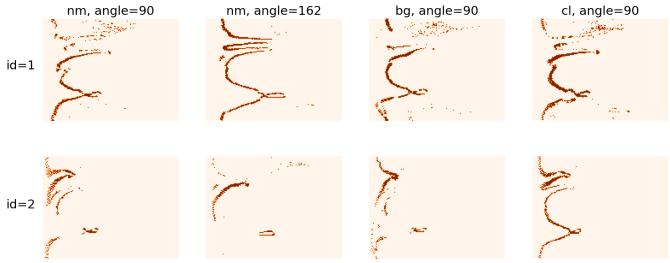


Fig. 4. Visualization of proposed LIR representation of two subjects in normal walking (nm, at 90° and 162°), carrying bag (bg) and clothing change (cl) conditions

We also visualize the LIR of two identities, 001 and 002, from EV-CASIA-B of the same time bin at various conditions like different viewing angles, carrying bag and clothing change. As seen in Fig. 4, the representations of the same identity are invariant at different conditions. The representation at 164° is a stretch of the representation at 90°. The representation of bag carrying or clothing change shows a small perturbation in certain parts over the representation of the normal walking condition. The variations of LIR are smaller and far more consistent than the event images and event graphs shown in Fig. 2. Also, the representations of the two identities have different contours, making it easier for SNN to learn their differences.

V. CONCLUSION

In this paper, we propose GaitSpike, a four-layer SNN for event-based gait recognition featuring a new event representation LIR. It changes the raw context-agnostic event data into biometric-aware features that can be better learnt and differentiated by the SNN. Our experiment results show that GaitSpike outperforms the state-of-the-art gait recognition methods with only a slight accuracy deterioration in complex walking conditions like occlusion and angle changes and is overall more computationally efficient. Our future work will explore new SNN design and learning algorithm to abstract more distinctive temporal features such as the speed variations of different parts of the body from the spatio-temporal information of address-event stream to overcome more complex influences like footwear, fatigue, and injury.

ACKNOWLEDGMENT

This research is supported in part by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) DesCartes programme, and in part by the Ministry of Education, Singapore, under its AcRF Tier 2 Award MOE-T2EP50220-0003.

REFERENCES

[1] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 2, pp. 316–322, 2005.

[2] H. Chao, Y. He, J. Zhang, and J. Feng, "Gaitset: Regarding gait as a set for cross-view gait recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 8126–8133.

[3] C. Fan, Y. Peng, C. Cao, X. Liu, S. Hou, J. Chi, Y. Huang, Q. Li, and Z. He, "Gaitpart: Temporal part-based model for gait recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2020, pp. 14 225–14 233.

[4] R. Liao, C. Cao, E. B. Garcia, S. Yu, and Y. Huang, "Pose-based temporal-spatial network (ptsn) for gait recognition with carrying and clothing variations," in *Chinese Conference on biometric recognition (CCBR)*. Springer, 2017, pp. 474–483.

[5] R. Liao, S. Yu, W. An, and Y. Huang, "A model-based gait recognition method with body pose and human prior knowledge," *Pattern Recognition*, vol. 98, p. 107069, 2020.

[6] T. Teepe, A. Khan, J. Gilg, F. Herzog, S. Hörmann, and G. Rigoll, "Gaitgraph: Graph convolutional network for skeleton-based gait recognition," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 2314–2318.

[7] Z. Mahfouf, I. Bouchrika, H. F. Merouani, and N. Harrati, "Gait biometrics via optical flow motion features for people identification," in *International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA)*. IEEE, 2016, pp. 312–321.

[8] T. Sattrupai and W. Kusakunniran, "Deep trajectory based gait recognition for human re-identification," in *TENCON 2018-2018 IEEE region 10 conference*. IEEE, 2018, pp. 1723–1726.

[9] F. M. Castro, M. J. Marín-Jiménez, N. Guil, S. López-Tapia, and N. P. de la Blanca, "Evaluation of cnn architectures for gait recognition based on optical flow maps," in *International conference of the biometrics special interest group (BIOSIG)*. IEEE, 2017, pp. 1–5.

[10] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128× 128 120 db 15 μs latency asynchronous temporal contrast vision sensor," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, 2008.

[11] Y. Wang, B. Du, Y. Shen, K. Wu, G. Zhao, J. Sun, and H. Wen, "Ev-gait: Event-based robust gait recognition using dynamic vision sensors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6358–6367.

[12] Y. Wang, X. Zhang, Y. Shen, B. Du, G. Zhao, L. Cui, and H. Wen, "Event-stream representation for human gaits identification using deep neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3436–3449, 2021.

[13] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *International conference on pattern recognition (ICPR)*, vol. 4. IEEE, 2006, pp. 441–444.

[14] F. Ponulak and A. Kasinski, "Introduction to spiking neural networks: Information processing, learning and applications." *Acta neurobiologiae experimentalis*, vol. 71, no. 4, pp. 409–433, 2011.

[15] L. Wang, T. Tan, H. Ning, and W. Hu, "Silhouette analysis-based gait recognition for human identification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 25, no. 12, pp. 1505–1518, 2003.

[16] J. Wu, Y. Chua, M. Zhang, Q. Yang, G. Li, and H. Li, "Deep spiking neural network with spike count based learning rule," in *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–6.

[17] E. O. Neftci, H. Mostafa, and F. Zenke, "Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks," *IEEE Signal Processing Magazine*, vol. 36, no. 6, pp. 51–63, 2019.

[18] T. Chai, A. Li, S. Zhang, Z. Li, and Y. Wang, "Lagrange motion analysis and view embeddings for improved gait recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 20 249–20 258.

[19] K. Ma, Y. Fu, D. Zheng, C. Cao, X. Hu, and Y. Huang, "Dynamic aggregated network for gait recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 22 076–22 085.

[20] D. Gehrig, M. Gehrig, J. Hidalgo-Carrió, and D. Scaramuzza, "Video to events: Recycling video datasets for event cameras," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3586–3595.

[21] J. K. Eshraghian, M. Ward, E. O. Neftci, X. Wang, G. Lenz, G. Dwivedi, M. Bennamoun, D. S. Jeong, and W. D. Lu, "Training spiking neural networks using lessons from deep learning," *Proceedings of the IEEE*, 2023.

[22] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2019, pp. 5693–5703.