16th International Conference on Machine Vision Applications (MVA)
National Olympics Memorial Youth Center, Tokyo, Japan, May 27-31, 2019.

Human identification by gait from event-based camera

Anna Sokolova[1,2], Anton Konushin[1,3]

1. Samsung-MSU Laboratory, Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, Russia
2. National Research University Higher School of Economics, 20 Myasnitskaya str., Moscow Russia
3. Samsung AI Center, 5c Lesnaya str., Moscow, Russia

Abstract

Gait recognition is a computer vision problem complementing other identification problems such as face or iris recognition. Unlike other identifiers, gait is based on the motion of body points and thus it can be captured by Dynamic Vision Sensor (DVS). In this work, we explore the possibility of gait recognition in the event stream by visualizing it and applying the existing method which achieves state-of-the-art results on several benchmarks. During the investigations several related problems such as moving object detection and human pose estimation are considered as auxiliary ones. Several algorithm settings are evaluated and compared. The obtained results show that all the problems can be solved in event-based data with high quality, which is close to the quality achieved on conventional colored videos.

1 Introduction

Gait recognition is a challenging problem comprising human identification in a video by their walking manner. Similar to the face or iris, gait is a biometrical index that is very difficult to fake which makes it an additional characteristic a person can be recognized by. However, unlike many other biometrical identifiers, the gait can be captured from afar and does not require the perfect exactness: the motion of points is of greater importance than particular appearance characteristics. This feature makes gait recognizable not only in conventional videos with colored frames changing 25-30 times per second but in other data formats.

One of such data formats is event stream obtained from Dynamic Vision Sensor (DVS) [11]. Similar to retina, it captures the pixel intensity changes ignoring the points of constant brightness. Instead of the frame sequence, it outputs the stream of events each reflecting the fact that the intensity of the point with spatial coordinates (x, y) has increased or decreased at time t. Once the magnitude of intensity change gets beyond some threshold, a positive or negative event is generated. If the sensor is static the background changes slowly, and the events at corresponding positions are rarely generated preventing unnecessary information transfer. The intensity is measured about 10^3 times per second which leads to asynchronous capture of all the important changes. The reconstruction of an initial scene from such data is a challenging problem, but the flow itself contains much information that is enough for many computer vision applications. Thereby, the event stream is suitable for many motion-based problems such as action, gesture or gait recognition as the static scenes are not considered and all the attention is drawn to points dynamics. Nevertheless, there are still no approaches proposed for gait recognition in such data.

In this work, we investigate the recognizability of the event flow by applying and developing one of state-of-the-art gait recognition methods and compare it with conventional videos. Although the main goal is to identify human, during the investigation we solve several auxiliary problems, such as moving object detection and human pose estimation in DVS-based video sequences, causing interest by themselves. The result obtained by the whole model shows that the intermediate steps work quite successfully, as well.

2 Related work

Since the gait recognition problem has not been considered in terms of dynamic vision sensors, let us discuss the main approaches used for conventional videos. As we will see below the events captured by the sensor can be transformed to images which traditional approaches could be applied to.

All the approaches to gait recognition can be separated into two sets: traditional structural methods where features are created manually based on biometrical and visual considerations, and deep methods where the descriptors are trained automatically by neural networks. Although the neural methods are currently very successful and popular, non-deep methods are developing as well and achieve really high quality, thus we shall discuss both approaches. Most of non-deep methods use the binary silhouettes of moving person to get features and compute various descriptors from them. Gait Energy Image (GEI [8]), the averaged over the gait cycle binary mask, is the most popular gait descriptor which the great number of methods are based on. Binary silhouettes are also used in many neural approaches. Starting from [24] where masks are directly used as network input, the methods have developed greatly during recent years. Siamese networks [19] are proposed to compare pairs of gaits and various fusion and aggregation ways are considered by different researchers [21]. Besides this, similar to other problems of video analysis, gait recognition is proposed to be solved by recurrent networks [7, 12, 20].

Another set of deep methods considers not the silhouettes in each frame but the motion of all the points of the body. This information is contained in the optical flow and can be computed easily. In [5, 17] the optical flow is proposed as the main source of information. This approach is developed in [18] where the consideration of human pose is added. This method is chosen as a basic method since it does not use the silhouettes which are hardly computable in event-based

images and obtains the state-of-the-art results on several benchmarks.

Returning to the subject of DVS, they have become very popular recently. Unlike conventional vision sensors, they do not capture the redundant information from the static areas and hence do not waste memory, time, and computational resources for its processing and storage. Thus a lot of related computer vision tasks are discussed and solved for event-data captured by these sensors. For example, many different approaches are proposed for object detection and tracking problems [13, 23]. [10] addresses the problem of continuous and sparse responses for fast and slow moving points, respectively. Depending on movement speed, the length of a time interval for dynamic information aggregation is selected to obtain features invariant to movement speed. Another problem close to gait recognition is action recognition which is also considered in terms of DVS data. In [2] three two-dimensional projections of the event stream are considered: the natural x-y projection consisting of the events aggregated over the time interval together with x-t and y-t projection where events are accumulated over horizontal and vertical intervals. The features are extracted from these maps using Speeded Up Robust Features (SURF) [3] and than clusterized and classified. Pose estimation which is one of the auxiliary tasks in this work is also considered directly for DVS data in [16]. However, there are no gait recognition approaches for event-based data, thus, in this work we propose and develop our algorithm for this task.

3 Proposed method

The proposed algorithm of human recognition consists of the following consecutive steps:

1. Visualization of event stream;
2. Human figure detection;
3. Estimation of optical flow;
4. Human pose estimation;
5. Gait recognition based on neural features.

The main idea of our algorithm is drawn from the method proposed for gait recognition in RGB videos [18]. Such an approach achieves high recognition quality on several benchmarks and we suggest that it can be applied to event-based data, as well. The details of the original method will be discussed further.

3.1 Data visualization

Since we aim to recognize the person by the walking manner we need to know the mutual arrangement of points and their relative changes rather than discrete events in distinct points. Thus, we visualize the generated event stream to be able to deal with them similarly to conventional video frames. To get an event-based image we consider the temporal window of a certain length (it is usually set to 0.04 seconds to get 25 fps frame rate) and calculate the sum of all the events in each pixel in this time interval. Acting this way, we aggregate all the events occurred in each spatial position inside the temporal window. Thus, the obtained frames can be interpreted as the measure of happenings. Such an aggregation seems the most natural and is often used for DVS signal visualization [2]. The example of visualized frame is shown on the left side of Fig. 2.

Unlike the optical flow which reflects the spatial movements of the points between the frames of the video, the event stream does not show where a point has moved. The events are generated due to intensity change which does not always follow the motion. This change occurs in one pixel and the events are generated independently in each pixel, whether these pixels correspond to the same point in different frames or not. Thus, the event stream visualization does not provide information about motion directly.

Nevertheless, having these event-based images we can apply different image-based methods of detection, pose estimation, and gait recognition to them and check the transferability of these methods to event streams.

3.2 Pose-based gait recognition in RGB videos

Let us shortly remind the main idea of the original method. The authors suggest that the optical flow (OF) can be used for gait recognition in order to consider the motion but not the appearance of the object (since the appearance is much easier to fake). Thus, the maps of OF are calculated for each pair of consecutive frames and used as primary features. Then the deep neural network is trained to recognize the human by such maps. The trained network is used as a feature extractor: the outputs of the last hidden layer are considered as gait descriptors and classified by the Nearest Neighbour method. Besides, the assumption is made, that the motion of some parts of human body influence on the gait more than the others. Hence, several body parts of different sizes are considered and corresponding patches of OF maps are used for identification. The pipeline of the described method is presented in Fig. 1.

3.3 Human figure detection

In all our experiments we consider the videos with a static background and use the fact that the points of such background do not generate the events. Such an assumption may seem strong, but actually, it appears from natural reasons. In practice, the most prevalent application of dynamic vision sensors is in video surveillance, for example, as a part of a smart home or for high-speed motion capture. The video surveillance sensors are fixed which leads to static background on the captured data.

Due to this motionless background, the human figure is the only object leading to event generation, so, human figure detection can be reduced to the separation of a non-uniform human from an almost monophonic background. Thereby, we constructed a basic blob model considering the average relative deviation of pixel brightness in each row and column of visualized stream from the median value and detecting when

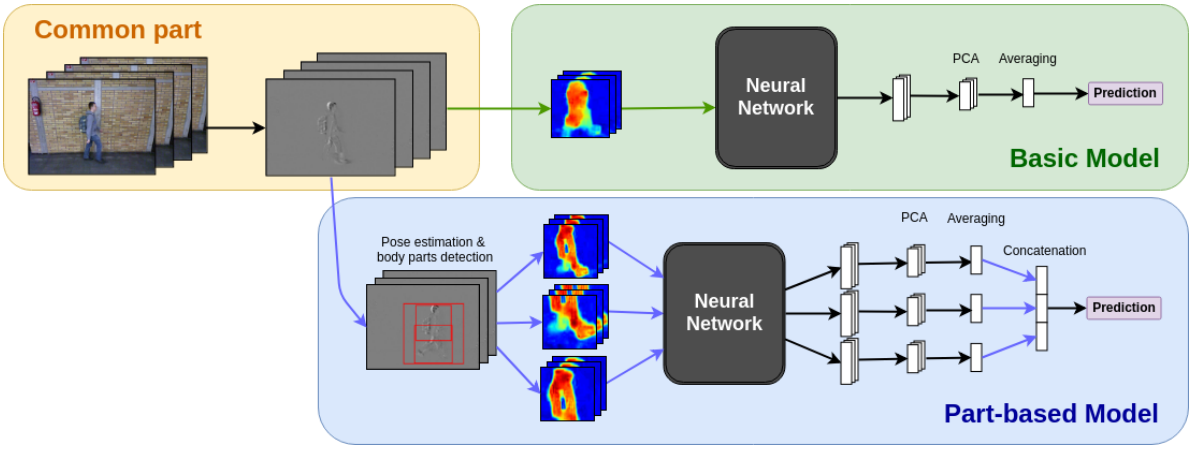

Figure 1. The scheme of basic and part-based models. Green and blue arrows correspond to the steps that differ in two models while black ones define common transformations.

these deviations exceed some threshold. The example of detection is shown on the right side of Fig. 2. One can see that the legs are not easily separable from the background, but nonetheless, such approximate bounding boxes are enough for further human pose estimation.

Figure 2. The examples of event-based image (left) and blob detection result (right).

3.4 Pose Estimation

The experiments conducted in [18] have shown that more precise consideration of body parts improves recognition quality compared to the full-body classification. Thus, we assumed that such an approach can increase the identification accuracy for event images, as well. Although the human figure is noticeable on event-based images, the state-of-the-art methods for pose estimation in RGB images do not cope with event-based ones. Hence, we have fine-tuned a pre-trained model on the event-based data. We have chosen the Stacked Hourglass Network [15] since it is well-trainable and has an adequate number of parameters which makes it easily and quickly applicable.

Stacked Hourglass Network is a neural model consisting of several "hourglass" residual modules allowing to capture information across different scales. The chosen architecture containing two stacked hourglasses was trained on the MPII Human Pose [1] dataset and achieves $86,95\%$ of correct key points (PCKh). Nevertheless, this network does not find the correct locations of the joint on event-based images, and thus could not be used directly for body parts detection. To fine-tune the model, we have simulated event-based images for CASIA [22] dataset and annotated it automatically using the OpenPose library [4]. The details of the simulation algorithm are described in Section 4. CASIA dataset is collected exactly for gait recognition and consists of full height walks of 124 subjects captured under different viewing angles and carrying and clothing conditions. Having quite a small number of subjects this database contains large number of conditions for each person and thus thousands of frames where human full height figure is depicted. Annotating the CASIA database allowed us to get a large DVS dataset labeled with pose key points that can be used for network training. Since Stacked Hourglass Network requires the approximate location of human as input, we used the results of human figure detection described above for preliminary analysis and fed the obtained bounding boxes to pose estimation model along with the images themselves. The main requirement to the boxes is that they should contain the target human figure inside and not contain any other figures. Since these conditions are usually satisfied, the detector allows estimating pose accurately enough for further body parts finding. We did not evaluate the exact quality of the estimator as it is just an auxiliary problem, but visually the results are close to the ground truth. The examples of several skeletons evaluated for event-based images are shown in Fig. 3. The first two images correspond to almost perfect evaluation, all the key points are located correctly. Two other examples reflect the most common mistakes that occur while pose estimation: leg invisibility leading to "one-legged" skeleton and "noisy thorax" when specifically this key point turns out to be misdetected.

Stacked Hourglass Network estimates the positions of main pose key points and having them obtained we compute the "areas of interest" where the optical flow will be considered, thus we can throw away the thorax key point which is sometimes noisy and should not affect the boundaries of considered areas. In the original method, five OF patches were cropped from each frame: two legs, upper and lower body, and the

whole figure. Additionally to this set of body parts, we considered a reduced set consisting of three patches corresponding to the full body (which contains all the joints), upper (containing the joints from head to hips), and lower body parts (containing the key points from hips to feet). The reason of such choice is that being on the ground the supporting leg does not move and, thus, no events are generated for a while, which makes the leg invisible as shown on the third image of Fig. 3.

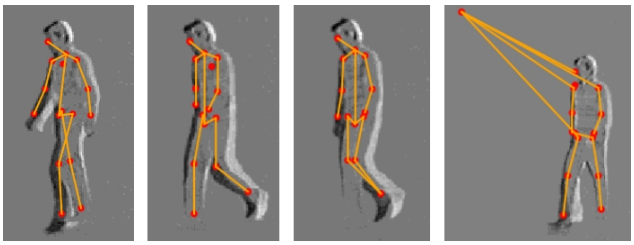

Figure 3. Examples of the skeletons estimated by Stacked Hourglass model.

The experiments show that considering some body parts more precisely the algorithm takes more details into account increasing the recognition quality.

3.5 Optical Flow Estimation

Since we apply method [18] to event-based sequences, the main source of information is optical flow in the surrounding of the human figure.

As well as in [18] we compute the optical flow by Farnebäck algorithm [6] but unlike [18] we do not add the third channel to horizontal and vertical components of optical flow and store the maps in binary format to prevent compression and discretization. Fig. 4 shows horizontal and vertical components of such maps after normalization.

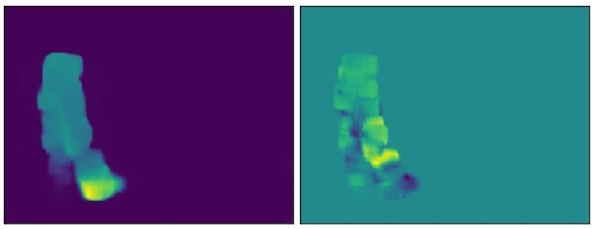

Figure 4. Horizontal (left) and vertical (right) components of optical flow maps computed from event stream.

4 Datasets

Since there are no publicly available gait recognition datasets captured by DVS, we simulated the event stream from the existing databases used for gait recognition. The data obtained from DVS has two main features: each event represents the fact of brightness increase or decrease, but not the value of change, and events are generated asynchronously with a very high frequency (several thousand times per second).

In conventional video sequences where the frame rate is about 25 fps, each pixel shifts a lot between frames so, the events cannot be correctly generated directly. To get real events we approximate the intermediate frames linearly following the approach proposed in [14].

We have transformed two gait datasets into event streams for training and evaluation of our model. The main dataset for the experiments is TUM Gait from Audio, Image and Depth (GAID) database [9]. This database contains side-view video sequences captured for 305 subjects. There are 10 RGB videos for each person with different carrying and clothing conditions: 6 normal walks, 2 walks with the backpack and 2 walks in coating shoes. The frame rate of the videos is 30 fps, thus, we have to apply the frame approximation described above to generate the intermediate events.

The second database we used is CASIA Gait Dataset B, which was used for pose estimator training. We have transformed it to event-based form and got about 1.7M labeled frames applicable for pose estimation. To accelerate the training process and make the data less complex, we have used only videos captured under the viewing angle from 54 to 144 degrees. Hence we have got about 770 000 of annotated frames which is quite enough for pre-trained model fine-tuning.

5 Experiments and Results

We have conducted a set of experiments aiming to check if a person can be recognized by the gait in the event stream and how accurate such recognition is. In the experiments, we compared detection methods and the influence of different body parts on the recognition.

Similar to [18] we follow the common evaluation protocol for TUM database: the split for training and testing parts is provided by its authors, thus, we trained the classification network on the data for 150 training subjects and the rest of the data was used for fitting and evaluating the final classifier based on neural features. 10 videos for each subject are separated into two parts: 4 normal walks for Nearest Neighbour classifier fitting and other 6 walks for its evaluation. The metrics of model quality is Rank-k (we calculate Rank-1 and Rank-5) representing the fraction of samples such that their true label is among k most probable classes.

The first basic model we examined makes a prediction based on full body patch of optical flow without pose estimation. The pipeline of his model is presented at the top of Fig. 1 (yellow and green blocks). The accuracy of this model is shown in the first row of Table 1. Since we are the first to recognize the gait from event-based data there are no state-of-the-art methods we can compare our results with. However, regardless of any other approaches we can investigate the recognizability of the data itself and compare it to conventional video. For this reason, we present the quality of the original method in the last row of the table. The accuracy of our model is about 2% less than [18] which demonstrates that tracking the events of brightness change instead of the total scene observation almost does not lose the information about the motion and,

thus, does not deteriorate the recognition.

After constructing the basic model we modified it by adding the pose estimation prior to OF patch cropping. This advanced model is shown in yellow and blue blocks of Fig. 1. As the usefulness of the optical flow around the legs is not obvious we compared two sets of body parts: all five parts and three "big" parts of the body (upper, lower and full body). The second and the third rows of Table 1 represent the comparison of these models with the corresponding approach based on RGB videos. The model based on three body parts achieves high accuracy, while the quality of five-parts-based model turns out to be a bit lower. We suppose that it happened because of the redundant noise in the surroundings of the legs and probable misdetections due to legs' invisibility in many frames.

Table 1. Results of the end-to-end model on simulated TUM-GAID dataset.

Body part set	Rank-1 [%]	Rank-5 [%]
Basic full body	98,0	99,7
Pose-based, three parts	99,0	100,0
Pose-based, five parts	98,8	99,8
Original RGB model [18]	99,8	100

In addition, we investigate the influence of detection quality on the whole model. The mistakes of our simple blob detector can worsen the final result, thus we have conducted the same experiments, but using more accurate bounding boxes, computed from initial RGB videos by background subtraction procedure. All the other steps of the algorithm remain the same. Due to the static background the result of the subtraction is the silhouette mask of moving person and its bounding box is very close to ground truth. Using these boxes we can evaluate the effectiveness of optical flow itself. The results of such "pure" evaluation of the basic model and the best pose-based one are presented in Table 2.

Table 2. Results of the model with perfect detections on simulated TUM-GAID dataset.

Body part set	Rank-1 [%]	Rank-5 [%]
Basic full body	98,3	99,6
Pose-based, three parts	99,1	99,8
Original RGB model [18]	99,8	100

Although the accuracy of both models increased a bit, the results remain quite close to the end-to-end model. It shows that probable mistakes of detector do not decrease the quality greatly.

Additionally to traditional classification problem, to make the investigation more thorough we consider the verification task. For a pair of event streams we need to decide whether there are different people whose motion have led to events generation in two sequences or the same one. The evaluation protocol for such verification problem is as follows.

For each pair of probe and gallery videos the Euclidean distance between neural descriptors is calculated and depending on their closeness the decision is made whether these descriptors belong to the same person. The classical metrics used for verification evaluation are ROC AUC and EER (equal error rate). Since we compare three modes of the model (the full body baseline and two pose-based models) and the original RGB-based approach, we have plotted the ROC curves for each model and calculated both metrics (Fig. 5 and Table 3). The curves difference is hardly noticeable, thus, we have made an auxiliary figure in a large scale to show that the curve corresponding to the three-part-based DVS model is closer to the upper-left corner of the plot than all the others.

The numerical results provided in Table 3 also show that event-based model surpasses the original RGB-based one in verification task. Although the difference is quite small both AUC and EER turn out to be higher in our model than in [18].

Table 3. Results of verification on TUM-GAID dataset.

Body part set	ROC AUC	EER [%]
Basic full body	0,9985	1,89
Pose-based, three parts	0,9994	1,10
Pose-based, five parts	0,9992	1,34
Original RGB model [18]	0,9975	1,26

6 Implementation Details

Several public libraries were used to implement the steps of the proposed recognition algorithm. The "perfect detection" and optical flow estimation were computed in the OpenCV library. The automatic joints labeling for pose model fine-tuning was made by OpenPose library [4]. The pose estimator was implemented in PyTorch library, while for the final classification model Lasagne framework with Theano backend was applied to match the original model precisely. It took about 4 hours to fine-tune the pose model and 10 hours to train classification network on NVIDIA GTX 1070 GPU.

7 Conclusion

In this work, we investigated the applicability of existing methods of video analysis to event-based data. While solving the gait recognition problem, we considered object detection and pose estimation problems. The visualization method we use makes the event flow similar to grayscale images both visually and semantically which makes the application of conventional methods natural. The results of conducted experiments show that being applied to the videos containing single moving person modern computer vision methods achieve very high quality close to colored-video-based models and even superior to them in verification task. Hence, the event-based data is not only presumably similar to the conventional one, but this closeness is confirmed experimentally.

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, pages 3686–3693, 2014.

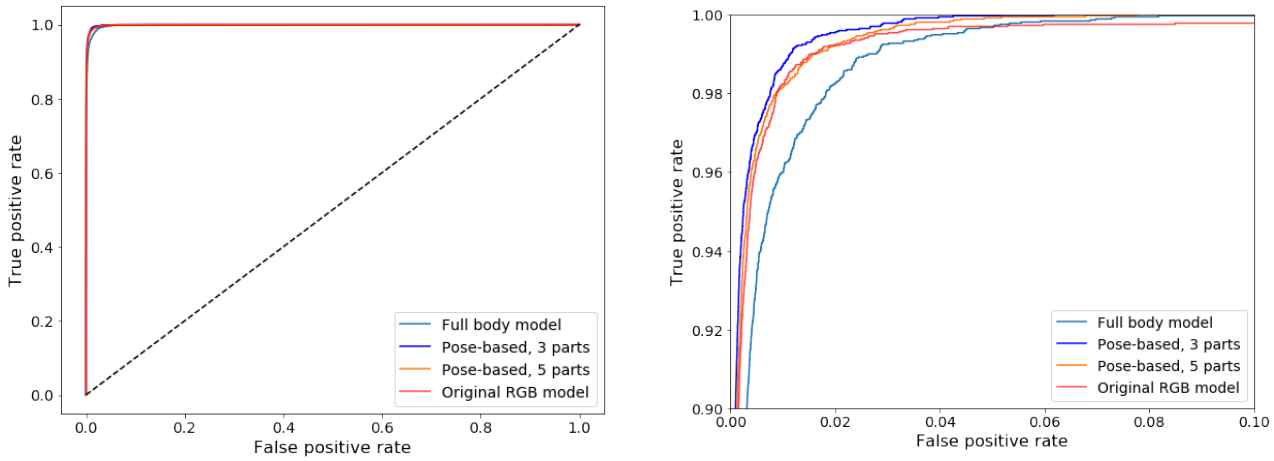

Figure 5. The ROC curves corresponding to three considered models: fully depicted (left) and large-scale (right) to show the difference.

[2] S. A. Baby, B. Vinod, C. Chinni, and K. Mitra. Dynamic vision sensors for human activity recognition. *CoRR*, abs/1803.04667, 2018.

[3] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *Computer Vision – ECCV 2006*, pages 404–417, 2006.

[4] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[5] F. M. Castro, M. J. Marín-Jiménez, N. Guil, and N. Pérez de la Blanca. Automatic learning of gait signatures for people identification. In *Advances in Computational Intelligence*, pages 257–270, 2017.

[6] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Image Analysis*, pages 363–370, 2003.

[7] Y. Feng, Y. Li, and J. Luo. Learning effective gait features using LSTM. In *International Conference on Pattern Recognition*, pages 325–330, 2016.

[8] J. Han and B. Bhanu. Individual recognition using gait energy image. *IEEE TPAMI*, 28(2):316–322, 2006.

[9] M. Hofmann, J. Geiger, S. Bachmann, B. Schuller, and G. Rigoll. The TUM Gait from Audio, Image and Depth (GAID) database: Multimodal recognition of subjects and traits. *J. of Visual Com. and Image Repres.*, 25(1):195 – 206, 2014.

[10] J. Li, F. Shi, W. Liu, D. Zou, Q. Wang, H. Lee, P.-K. Park, and H. E. Ryu. Adaptive temporal pooling for object detection using dynamic vision sensor. In *British Machine Vision Conference (BMVC)*, 2017.

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

[12] D. Liu, M. Ye, X. Li, F. Zhang, and L. Lin. Memory-based gait recognition. In *BMVC*, pages 1–12, 2016.

[13] A. Mitrokhin, C. Fermüller, C. Parameshwara, and Y. Aloimonos. Event-based moving object detection and tracking. *CoRR*, abs/1803.04523, 2018.

[14] E. Mueggler, H. Rebecq, G. Gallego, T. Delbruck, and D. Scaramuzza. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam. *The International Journal of Robotics Research*, 36(2):142–149, 2017.

[15] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016.

[16] H. Rehbinder and B. K. Ghosh. Pose estimation using line-based dynamic vision and inertial sensors. *IEEE Transactions on Automatic Control*, 48:186–199, 2003.

[17] A. Sokolova and A. Konushin. Gait recognition based on convolutional neural networks. In *ISPRS Archives*, volume XLII-2/W4, pages 207–212, 2017.

[18] A. Sokolova and A. Konushin. Pose-based deep gait recognition. *IET Biometrics*, 8(2):134 – 143, 2018.

[19] N. Takemura, Y. Makihara, and D. Muramatsu. On input/output architectures for convolutional neural network-based cross-view gait recognition. In *IEEE Trans Circuits Syst Video Technol 11*, 2017.

[20] S. Tong, Y. Fu, H. Ling, and E. Zhang. Gait identification by joint spatial-temporal feature. In *Biometric Recognition*, pages 457–465, 2017.

[21] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan. A comprehensive study on cross-view gait based human identification with deep cnns. *IEEE TPAMI*, 39, 2016.

[22] S. Yu, D. Tan, and T. Tan. A Framework for Evaluating the Effect of View Angle, Clothing and Carrying Condition on Gait Recognition. In *Proc. of the 18th ICPR*, volume 4, pages 441–444, 2006.

[23] W. Yuan and S. Ramalingam. Fast localization and tracking using event sensors. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4564–4571, 2016.

[24] X. Zhang, S. Sun, C. Li, X. Zhao, and Y. Hu. Deepgait: A learning deep convolutional representation for gait recognition. In *Biometric Recognition*, pages 447–456, 2017.