Optimization of GaitSet for Gait Recognition

Jicai Pan, Hao Sun, Yi Wu, Shi Yin, and Shangfei Wang

University of Science and Technology of China
panjicaiz@gmail.com, 1535040450@qq.com, wy221711@mail.ustc.edu.cn,
davidyin@mail.ustc.edu.cn, sfwang@ustc.edu.cn

Abstract. The basic goal of gait recognition is to predict the identity of a subject from his/her walking recording. Since the gait detection process is non-inductive, non-contactive, and does not require the participation of other human beings, gait recognition has great application potential for biometric identification. GaitSet is an efficient, flexible gait recognition algorithm with low computational cost, and is widely used in the field of gait recognition. GaitSet processes the frames of a video as an unordered set and may lose some sequential patterns. To address this, we apply the Micro-motion Capture Module to embed complex temporal patterns for gait recognition. Data augmentation is also used to improve the robustness of the recognizer. Experimental results show the effectiveness of our method.

1 Introduction

Compared with other biological characteristics, such as face, fingerprint and iris, gait information has great advantages to identify a person, since gait is hard to imitate and can be obtained in a non-inductive and non-contactive way. Gait recognition [1, 2] takes either of the two kinds of methods, i.e., model-based methods or model-free methods. Model-based methods typically extract features from the model of body structure. These features include pendulum [3], the inclination of leg [4] and the movement of leg joints [5]. Instead of using features defined in a body model, model-free methods [6, 7, 10] extract features directly from the input frames by machine learning techniques. From these methods, Gaitset [10] achieves state-of-the-art performance since it learns informative representations through a deep neural network model. Therefore, we choose Gaitset as our backbone model. Gaitset is built based on the hypothesis that the frame sequence of a walking video can be recovered from the unordered set of frames. Therefore, GaitSet regards the frames from a video as an unordered set and may lose some sequential patterns. However, such hypothesis may not be satisfied in real-world data with complex spatial and temporal patterns. To promote the capability of temporal modeling, we add Micro-motion Capture Module (MCM)[13] into the original GaitSet model. In addition, we use image augmentation operations, such as flipping, blurring and occlusion, to avoid overfitting and improve generalization performance. Experimental results show that the improved GaitSet method achieves an accuracy of 77.4% on the CASIA-E validation set, a higher performance than 69.41% for the original GaitSet method.

2 Method

2.1 GaitSet

GaitSet[10] is an efficient and flexible gait recognition algorithm with low computational cost. It firstly extracts frame-level features from a sequence of silhouettes by CNNs independently. Then features from each frame are fused to a single feature map, and the Horizontal Pyramid Mapping (HPM) operation is used to capture the information of each body part of the pedestrian from the fused feature map. The whole framework of Gaitset is shown in Fig. 1.

Formally, GaitSet takes a set of m body silhouettes extracted from a person's walking video as the input, denoted as $S = \{s_i\}, i = 1, 2, ..., m$. Gait recognition is dependent on the representation of walking posture, which is extracted by CNNs in Gaitset. Specifically, three convolution modules are adopted, as shown in Fig. 2. The encoded feature map f_i^l is formulated as:

$$f_i^l = \Phi_l(f_i^{l-1}) \tag{1}$$

where $l = 1, 2, 3$, $i = 1, 2, 3, ..., m$, Φ_l denotes the convolution network of module l, and $f_i^0 = s_i$. Tab. 1 shows the network settings. Each frame of silhouette shares the parameters of the convolution modules.

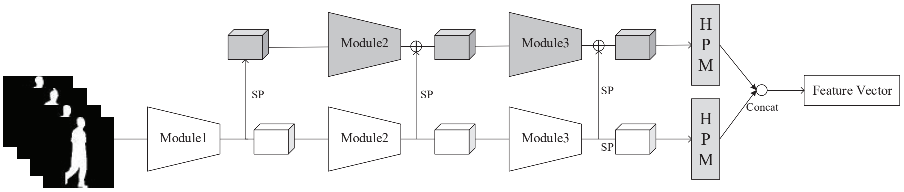

Fig. 1. The framework of GaitSet.

Table 1. The convolution network settings

module	layer	kernel size	stride
module1	conv1-1	5×5	1
	conv1-2	3×3	1
	pooling1	2×2	2
module2	conv2-1	3×3	1
	conv2-2	3×3	1
	pooling2	2×2	2
module3	conv3-1	3×3	1
	conv3-2	3×3	1

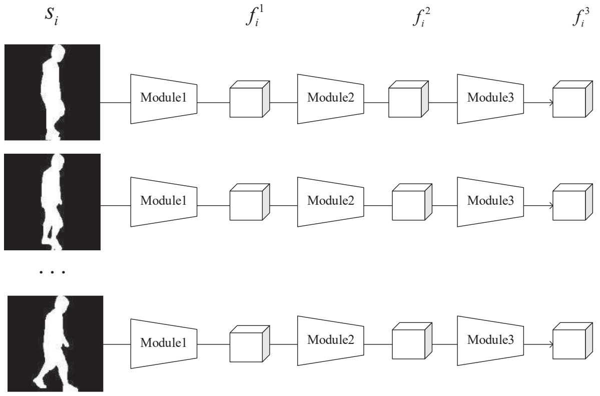

Fig. 2. The convolution network of GaitSet.

The lengths of different silhouette sequences may be variant. In order to integrate information from all frames in the sequence, GaitSet uses a set aggregation operation, i.e., Set Pooling (SP), to fuse features from each frame in the sequence. The SP operation regards the silhouette sequence as a set, and the fused features, as shown in Equation (2), is embedded with set-level patterns by the SP operation.

$$F^l[i,j,k] = SP(\{f_1^l[i,j,k], f_2^l[i,j,k], ..., f_m^l[i,j,k]\}) \tag{2}$$

In Equation (2), i, j, k respectively represent the index of row, column and channel of a feature map, and SP is instantiated as one of these operations, i.e., $\{Max, Median, Mean, JointFunc, etc\}$. $JointFunc$ is formulated as:

$$JointFunc(*) = 1\_1Conv(concat[Max(*), Median(*), Mean(*)]) \tag{3}$$

where $concat$ denotes the channel-wise concatenation operation, 1_1Conv denotes 1×1 convolution layer. The SP operation is performed after each convolution module. Fig. 3 visualizes the fused feature map from the $Mean$ operation. This feature map can be interpreted as a Gait Energy Image (GEI).

In the walking process, different body parts may have different moving patterns. In order to encode fine-grained gait features for these body parts, Gaitset adopts HPM to divide the input feature map into multiple parts evenly in the horizontal direction. If the division scale is n, then the feature map is equally divided into 2^{n-1} parts horizontally. Each divided feature is processed by Global Average Pooling (GAP) operation, Global Max Pooling (GMP) operation and Fully-Connected (FC) network, as shown in Fig. 4.

2.2 Our solution

The SP operation in Gaitset is adopted based on the hypothesis that the frame sequence of a walking video can be recovered from the unordered set of frames.

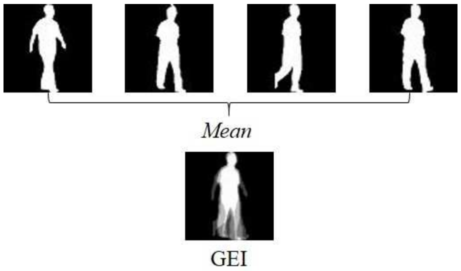

Fig. 3. Fused feature map by $Mean$ operation

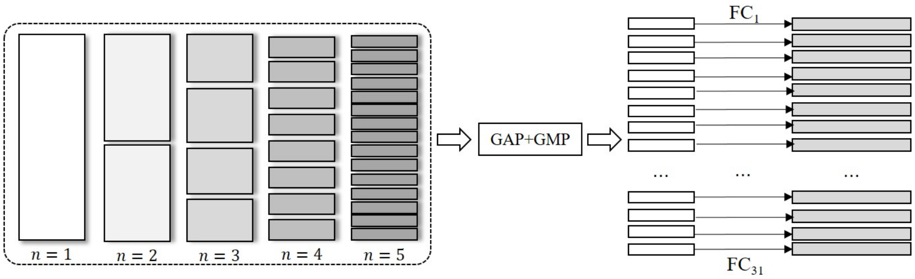

Fig. 4. The Horizontal Pyramid Mapping module.

However, this hypothesis may not be satisfied in real-world scenario. For example, when the walking subject moves with high-frequent turning and waving actions, the unordered set may lose some temporal information from the original sequence, and consequently, the accuracy of GaitSet decreases. To address this problem, we introduce a new set pooling operation, i.e., Micro-motion Capture Module (MCM)[13], which models all frames as an ordered sequence instead of an unordered set. The MCM module is shown in Fig. 5. The size of the sliding window is $2r + 1$, where r is a hyperparameter. To extract temporal information, we apply GAP and GMP on the sliding window. These operations can be formulated as:

$$T_i^l = GAP\{f_{i-r}^l, ..., f_i^l, ..., f_{i+r}^l\} + GMP\{f_{i-r}^l, ..., f_i^l, ..., f_{i+r}^l\} \tag{4}$$

Then GMP is used for further fusion of feature maps, as shown in the following equation:

$$F^l = GMP\{T_{r+1}^l, T_{r+2}^l, ..., T_{n-r}\} \tag{5}$$

In order to improve the generalization performance of the model, we adopt a variety of image augmentation strategies, such as horizontal flipping, Gaussian blurring and random occlusions, as shown in Fig. 6, Fig. 7 and Fig. 8.

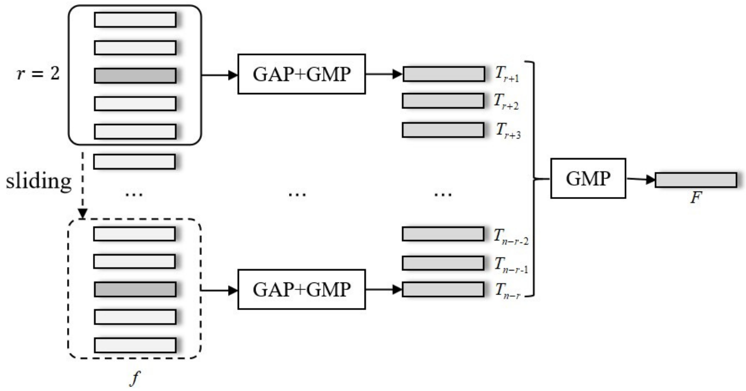

Fig. 5. The Micro-motion Capture Module.

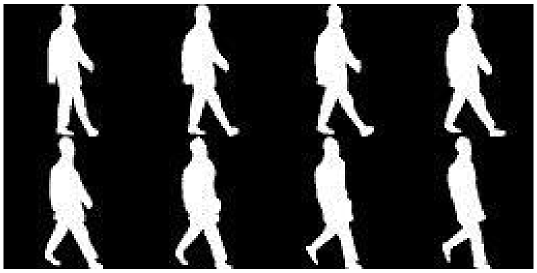

Fig. 6. The flipped frame sequence.

3 Experiments

In this section, firstly, we introduce the training and validating conditions, including the adopted data set and the implementation details. Next, we give the experimental results and make comparisons.

3.1 Data

We adopt the CASIA-E dataset[1] as the experimental data set, which contains 500 videos of walking people. Each video is a frame sequence, where each frame is a silhouette image instead of a RGB image. We randomly split the CASIA-E dataset at a ratio of 9:1 as training and validating data.

[1] https://drive.google.com/drive/folders/1SCcsXfAiWbXGSCb33NQ0ou8Ln1dB44Of?usp=sharinghttps://drive.google.com/drive/folders/1SCcsXfAiWbXGSCb33NQ0ou8Ln1dB44Of?usp=sharing

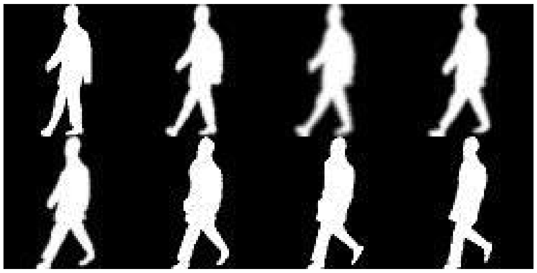

Fig. 7. The frame sequence applied with Gaussian blurring.

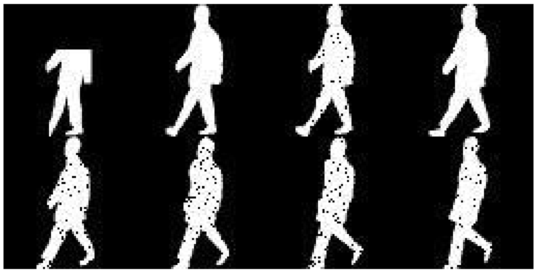

Fig. 8. The frame sequence applied with random occlusions.

3.2 Implementation details

On each frame, we resize the silhouette to 64 × 64, then feed it into the network. The sequence of silhouettes is shown in Fig. 9. Our method is optimized by Adam[16] with an initial learning rate of 1e-4, r in Equation (4) is set as 2, the batch size is set as 64, the total number of iteration is set as 200000.

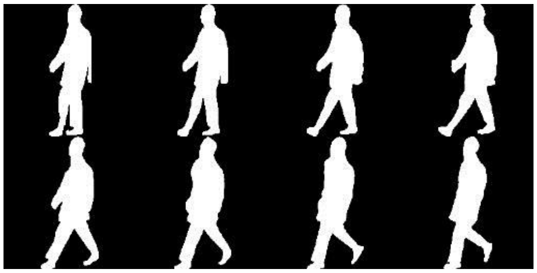

Fig. 9. The sequence of silhouettes.

3.3 Experimental results

The recognition accuracy of different experimental settings is shown in Tab. 2. From this table, we find that the MCM and the data augmentation strategies we adopt improve the performance of Gaitset significantly.

Table 2. Recognition accuracy on the validation set of CASIA-E.

Method	Validation accuracy
original gaitset	0.6941
gaitset + imgaug	0.7673
gaitset + MCM	0.7562
gaitset + MCM + imgaug	**0.7696**

4 Conclusion

We add the Micro-motion Capture Module and some data augmentation strategies, such as horizontal flipping, Gaussian blurring and random occlusions, into the Gaitset method, and significantly improve the performance of Gaitset on the CASIA-E data set in gait recognition accuracy .

References

1. Wan, Changsheng, Li Wang, and Vir V. Phoha, eds. "A survey on gait recognition." ACM Computing Surveys (CSUR) 51.5 (2018): 1-35.
2. Wang J, She M, Nahavandi S, et al. A review of vision-based gait recognition methods for human identification[C]//2010 international conference on digital image computing: techniques and applications. IEEE, 2010: 320-327.
3. David Cunado, Mark S. Nixon, and John N. Carter. 1997. Using gait as a biometric, via phase-weighted magnitude spectra. In International Conference on Audio-and Video-Based Biometric Person Authentication. Springer, 93–102.
4. Chiraz BenAbdelkader, Ross Cutler, and Larry Davis. 2002. Stride and cadence as a biometric in automatic person identification and verification. In Proceedings of 5th IEEE International Conference on Automatic Face and Gesture Recognition, 2002. IEEE, 372–377.
5. Liang Wang, Huazhong Ning, Tieniu Tan, and Weiming Hu. 2004. Fusion of static and dynamic body biometrics for gait recognition. IEEE Transactions on Circuits Systems for Video Technology 14, 2, 149–158.
6. Ju Man and Bir Bhanu. 2006. Individual recognition using gait energy image. IEEE Transactions on Pattern Analysis and Machine Intelligence 28, 2, 316–322.
7. Jianyi Liu and Nanning Zheng. 2007. Gait history image: A novel temporal template for gait recognition. In 2007 IEEE International Conference on Multimedia and Expo. IEEE, 663–666.

8. Zheng S , Zhang J, Huang K , et al. Robust view transformation model for gait recognition[C]// IEEE International Conference on Image Processing. IEEE, 2011.
9. Yiwei H , Junping Z , Hongming S , et al. Multi-Task GANs for View-Specific Feature Learning in Gait Recognition[J]. IEEE Transactions on Information Forensics Security, 2018, PP:1-1.
10. Chao H, He Y, Zhang J, et al. Gaitset: Regarding gait as a set for cross-view gait recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33: 8126-8133.
11. Yu S, Tan D, Tan T. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition[C]//18th International Conference on Pattern Recognition (ICPR'06). IEEE, 2006, 4: 441-444.
12. Takemura N, Makihara Y, Muramatsu D, et al. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition[J]. IPSJ Transactions on Computer Vision and Applications, 2018, 10(1): 4.
13. Fan C, Peng Y, Cao C, et al. GaitPart: Temporal Part-Based Model for Gait Recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 14225-14233.
14. Makihara Y, Suzuki A, Muramatsu D, et al. Joint intensity and spatial metric learning for robust gait recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 5705-5715.
15. Zhang Z, Tran L, Yin X, et al. Gait recognition via disentangled representation learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 4710-4719.
16. Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.