Gait Recognition by Jointing Transformer and CNN

Mingyu Cai, Ming Wang, and Shunli Zhang(✉)

Beijing Jiaotong University, Beijing, China
{23126412,slzhang}@bjtu.edu.cn

Abstract. Gait recognition is a biometric technology based on the human walking state. Unlike other biometric technologies, gait recognition can be used for remote recognition and the human walking pattern cannot be imitated. Gait recognition has wide applications in the field of criminal investigation, security and other fields. Most of the current mainstream algorithms use Convolutional Neural Network (CNN) to extract gait features. However, CNN only captures the local image features in most cases which may not inherently capture global context or long-range dependencies. In order to solve the above problems and to extract more comprehensive and precise feature representations, we propose a novel Gait recognition algorithm jointing Transformer and CNN by introducing the attention mechanism, called GaitTC. The framework consists of three modules, including the Transformer module, CNN module and feature aggregation module. In this paper, we conduct the experiments on CASIA-B dataset. The results of the experiments show that the proposed gait recognition method achieves relatively good performance.

Keywords: Gait recognition · Deep Learning · Transformer · CNN

1 Introduction

At present, with the continuous development of the computer vision, researchers have proposed many gait recognition methods. Most of the existing methods are bulit based on the CNN and can be roughly divided into two categories. One category is the template-based gait recognition framework. It mainly uses some statistical functions, including Max, Mean, etc., to calculate the gait statistics whithin a gait sequence cycle. These methods first extract the temporal features of the gait sequence, and then extract the spatial features through the CNN. The CNN has primarily been designed for local feature extraction, which may not effectively capture global information. As a result, there can be limitations or potential inaccuracies in recognition results when relying solely on CNN-based approaches. The other category mainly extracts the temporal and spatial features of the gait sequence with fixed input length. These methods may greatly limit the length of the input gait sequence and reduce the robustness of the model. Therefore, this paper proposes a novel gait recognition model by jointing Transformer and CNN, GaitTC, which has the following advantages:

W. Jia et al. (Eds.): CCBR 2023, LNCS 14463, pp. 312–321, 2023.
https://doi.org/10.1007/978-981-99-8565-4_30

(1) The Transformer can reduce the number of operations on sequence by using parallel computing, which greatly improves the efficiency.
(2) The Transformer restores positional dependence among image blocks by encoding the positions of segmented image blocks. This allows the model to better capture the spatial features.

In view of the existing problems of the existing gait recognition methods and the advantages of Transformer itself, this paper develops a new Transformer-CNN-based gait recognition framework to better extract the spatio-temporal features of gait sequences and to achieve higher performance. The main work and contributions of this paper are as follows:

(1) This paper proposes a new gait recognition framework based on Transformer and CNN, which can effectively extract the global feature of gait sequence by introducing Transformer. Compared with traditional Recurrent Neural Network(RNN) and CNN based methods, the proposed method with Transformer can improve the efficiency and better extract the global features.
(2) After the Transformer module, CNN is used to further extract the gait features of each frame. Then, the frame-level features are aggregated into sequence-level features to improve the representation ability of the gait features and the accuracy of the recognition.
(3) The proposed method is experimented on CASIA-B dataset, and compared with other gait recognition methods such as ViDP [1], CMCC [2], CNN-LB [3], GaitSet [4] in different wearing conditions and perspectives. The experimental results show that the the proposed method achieves good performance in most conditions.

2 Related Work

In this section, we provide a brief overview of the two important types of gait recognition methods: appearance-based methods and model-based methods.

2.1 Model-Based Methods

The model-based gait recognition methods mainly build the human model for gait recognition. These methods usually require a structure model to capture the static characteristics of human, and a motion model to capture the dynamic characteristics [5]. The structure model describes the structure of a person's body, including the stride, height, trunk and other body parts. The motion model is used to simulate the motion trend and trajectory of different body parts of a person during walking. The existing model-based gait recognition methods can be divided into two categories. One is to capture the evolution of these parameters over time by fitting a model. In these body parameter estimation methods, the angle of the body skeleton joints during walking is mainly obtained, such as the angular movement of knees and legs at different stages; the other is to estimate the parameters of the body (length, width, step frequency, etc.) directly from the original video. Gait recognition based on three-dimensional human body modeling belongs to this type. By analyzing video, image and other data, the motion parameters of the human

body are obtained, and a complete gait sequence is constructed. Then the sequence is converted into the corresponding coordinate information to realize the extraction of human motion features, so as to reconstruct the 3D model of the human body. Zhao et al. [6] constructed a human skeleton model with 10 joints and 24 degrees of freedom by using multiple cameras to capture the movement process of the human body. In order to obtain better performance, several features extracted from different directions are combined into a complete set of features for recognition, which can improve the stability of this method. At the same time, a gait recognition method based on geometric description [7] has also been proposed. This method mainly learns the deep features of the gait sequence by locating the skeletal joint coordinates.

Although the above methods can provide more complex gait feature information and can effectively perform gait recognition in complex environments, in actual scenes such as shopping malls and banks, due to the inability to deploy a large number of cameras, it is not possible to shoot gait sequences from multiple angles of the camera at the same time; at the same time, realizing the 3D human body model requires a lot of computing resources and a lot of computer computing power, which is not conducive to the training and development of the model. How to meet the low-cost sequence extraction without consuming a lot of computer resources is one of the main problems.

2.2 Appearance-Based Methods

With the development and maturity of deep learning algorithms, many gait recognition methods based on deep learning have emerged. At present, most of the networks used in gait recognition are CNN and RNN.

Since CNNs have excellent image classification capabilities, gait recognition based on CNNs has also occured. Shiraga et al. [8] proposed the GEINet network structure. This network consists of three modules. The first two modules include a convolutional layer, a pooling layer, and a normalization layer, respectively. The last module is composed of two fully connected layers. At the same time, the input of the network is a gait energy map. The gait feature reflects the accumulation of gait energy during a person's walking process. Compared with other methods, GEINet focuses on subtle inter-subject differences in the same action sequence. Liao et al. [9] proposed a posture-based spatio-temporal network through the GEI, which has better effect on gait recognition in complex states. In addition, Huang et al. [10] proposed to extract the local features of human gait sequence according to the parts of the body. The human body composition is defined as six local paths, i.e. the head, left arm, right arm, trunk, left leg and right leg, and features are extracted from each path. At the same time, a 3D local CNN network is introduced into the backbone. The backbone contains three network blocks, and each block is composed of two CNN layers. Finally, the ReLU function is used as the activation function to output the obtained features. Wolf et al. [11] proposed a gait recognition method based on 3D CNN. This method captures the spatiotemporal information of the gait in multiple sequence frames. This method can well summarize the gait characteristics in a variety of perspective changes.

3 Method

In this section, we will introduce the implementation of the proposed method in detail. Firstly, we overview the proposed Transformer and CNN based gait recognition framework. Secondly, we explain the Transformer model in detail. Finally, we describe the CNN Module and Feature Aggregation Module.

3.1 GaitTC

We propose a Transformer-CNN-based gait recognition method built upon the traditional Transformer model. The overview structure of the proposed method is shown in Fig. 1 which is designed to generate more robust gait feature representations. To address the issue of limited effectiveness of the traditional Transformer model on small-scale datasets, we incorporate the CNN module after the Transformer model. Firstly, the Transformer module is used to extract the global features, and the corresponding attention weights of each image block are obtained. Then, CNN is used for local feature extraction. At the same time, CNN also makes up for the defect that the Transformer model has poor effect on the feature extraction in small-scale datasets.

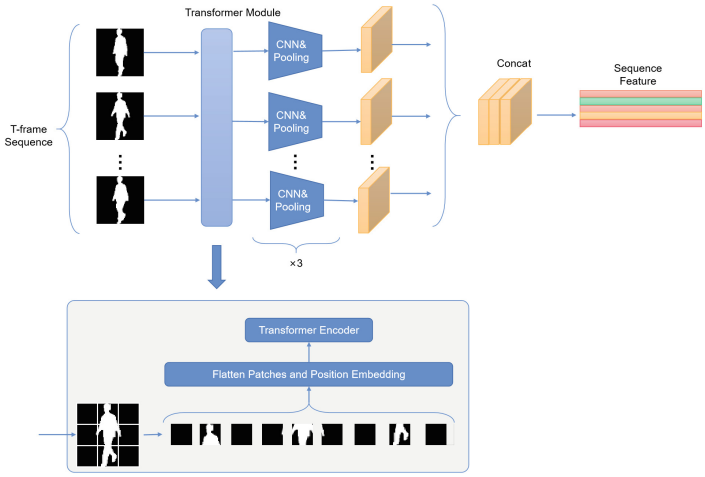

Fig. 1. The proposed gait recognition framework with Transformer and CNN

3.2 Network Structure

This section will mainly introduce the model of GaitTC. The model is mainly divided into three modules, namely Transformer module, CNN module and feature aggregation module. The Transformer module is mainly used to extract the most useful global information from the input gait sequence. Then, the output of the Transformer module is put into the CNN module to extract more comprehensive gait features. Finally, we use the feature aggregation module to fuse the features.

3.2.1 Transformer Module

This module uses the Transformer to operate the input gait silhouette. Firstly, the gait silhouette map is divided into image blocks and linearly projected, and then input into the Transformer module for processing. In the linear projection process, each input image block is mapped to a d-dimensional vector, and each vector needs to be multiplied by a linear matrix E. At this time, the image block is a vector with a dimension of d after smoothing. Next, in order to enable the model to encode the position of the image block vector, before entering the model, we embed the position information of each image block into the vector of the corresponding image block, and then embedded vector is connected with a learnable class marker. The internal value of the vector can be learned and adjusted during the model training process to obtain a feature representation with attention weight.

In the Transformer, the encoder module has two important sub-modules, which are Multi-Head Self-attention (MSA) and Multi-Layer Perceptron (MLP) modules. The encoder will receive the image block of the gait sequence as input, and the input image block will first pass through the normalization layer. In the normalization layer, the input values of all neurons are normalized in the feature dimension, which greatly reduces the training time and improves the training performance. Subsequently, the output of the normalization layer is input to the multi-head attention module, and then the output corresponding to the multi-head attention module is connected with the original input through the residual network. The output after the normalization layer will be sent to the multi-layer perceptron layer to simulate more complex nonlinear function relationships. The residual network will be used for two modules in the Transformer encoder to retain the gradient information of the module during the training process, avoiding the problem that the gradient disappears during the training process.

In the multi-head self-attention module, the multiple self-attention operations will be performed according to the number of heads in the attention module. In each attention head, the d-dimensional flattened image block vector p will be multiplied by the multiple attention weight matrices W_q, W_k, and W_v to obtain Query, Key, Value, as shown in Eq. (1):

$$[q, k, v] = [p \cdot W_q, p \cdot W_k, p \cdot W_v], (W_q, W_k, W_v \in \mathrm{R}^{d \times d_H}) \tag{1}$$

MSA captures the information from different aspects at different positions of each head, which also allows the model to encode more complex features in gait sequences in parallel. At the same time, due to the use of parallel computing mechanism, the time cost of multi-head attention calculation is similar to that of single-head attention mechanism, which improves the performance of the model to a certain extent and reduces the consumption of computing resources.

The multi-layer perceptron module contains two fully connected layers and a GeLU function. Finally, the Transformer module utilizes a residual structure to connect the output of the multi-layer perceptron with the original vector output through the multi-head attention mechanism, output the attention value between each image block and other image blocks, and then pass it to the next module for further feature extraction.

3.2.2 CNN Module and Feature Aggregation Module

The CNN module extracts features by extracting feature blocks with attention weights output of the Transformer. The module mainly includes three convolution pooling layers, and the kernel size, and step size of the convolution kernels in each layer are equal. In order to extract more detailed information, the convolution with the kennel size of 3 * 3 * 3 is used to extract the features of each frame. The feature contains the spatial information of each frame and the time information of the gait sequence, so that the feature representation is more complete. The higher-level features extracted by the convolution operation will be put into the feature aggregation module.

In the feature aggregation module, the model aggregates the features extracted from each subject under a fixed number of frames, that is, the features of each frame in the gait sequence are aggregated into a sequence set. The module will first calculate the maximum value, average value and median value of each element of the feature of each frame, respectively, and splice the obtained feature. In order to better represent the set-level features of each sequence, the spliced feature will be finally performed. Global average pooling and global maximum pooling are used to aggregate frame-level features, and the sum of the two is used as the feature representation of the final gait sequence.

4 Experimental Results

4.1 CASIA-B Dataset

The experiments in this paper were conducted on the current popular gait dataset CASIA-B, which contains 124 subjects. In order to make the experimental results more rigorous and reliable, we conduct the experiment in different sample scale conditions. According to the different proportions of sample division between the training set and the test set, the experiment is divided into three parts, which include small sample training (ST), medium sample training (MT) and large sample training (LT). The training set of small-scale samples contains 24 subjects, the medium-scale sample training set contains 62 subjects and the large-scale sample training set contains 74 subjects. The rest subjects will taken as the test set. Through different division of LT, MT and ST, the performance of the model under different conditions can be tested, which can better reflect the robustness of the model.

4.2 Results and Analysis

In this section, the experimental results of this model are compared with some excellent gait recognition algorithms, including CNN-LB, GaitSet, MGAN [12], AE [13], ViDP, CMCC and so on.

Small-scale sample data is closer to practical applications for gait recognition tasks, because for recognition tasks, the number of samples to be identified in practical applications(i.e., test data) is much larger than the number of samples during training (i.e., training data), so the accuracy of small-scale samples can better reflect the performance of the proposed method. According to the experimental results conducted in small-scale

sample data (as shown in Table 1), it can be seen that the accuracy varies in different cross-views. Under normal conditions, the experients maintain better accuracy under cross-views such as 36°, 126°, and 144°, which is 13% higher than the 0° under the same condition. In the complex state, it is 10% higher than the 0°. Besides, according to the results, it can be observed that the accuracy of the proposed method is higher than the existing excellent methods in some cases. The results show that the proposed method achieves appealing performance at difficult angles such as 0°, 90° and 180°. However, the accuracy is slightly lower than GaitSet under the 36° view angle.

Table 1. The accuracy of the proposed GaitTC on the CASIA-B under the ST condition.

Gallery NM#1–4			0°–180°											
Probe			0	18	36	54	72	90	108	126	144	162	180	Mean
ST	NM	ViDP	–	–	–	59.1	–	50.2	–	57.5	–	–	–	–
		CMCC	46.3	–	–	52.4	–	48.3	–	56.9	–	–	–	–
		CNN-LB	54.8	–	–	77.8	–	64.9	–	76.1	–	–	–	–
		GaitSet	64.6	83.3	90.4	86.5	80.2	75.5	80.3	86.0	87.1	81.4	59.6	79.5
		GaitTC	**75.8**	**85.9**	**88.9**	**87.4**	**81.9**	**77.9**	**83.4**	**88.5**	**88.4**	**83.6**	**68.9**	**82.8**
	BG	GaitSet	55.8	70.5	**76.9**	**75.5**	**69.7**	63.4	68.0	75.8	**76.2**	70.7	52.5	68.6
		GaitTC	**64.6**	**73.8**	76.7	74.3	67.4	**64.1**	**69.0**	**77.0**	75.4	**70.8**	**58.3**	**70.1**
	CL	GaitSet	29.4	43.1	49.5	48.7	42.3	40.3	44.9	47.4	43.0	35.7	25.6	40.9
		GaitTC	**40.8**	**47.9**	**50.0**	**45.8**	**46.9**	**44.5**	**47.8**	**49.7**	**44.8**	**37.1**	**29.9**	**44.1**

Under the medium sample condition, the experimental results are shown in Table 2. The accuracy is improved compared with the small samples in some cases, but the accuracy maintains small margin compared to the GaitSet under normal conditon(NM) and walking with bag condition(BG). In the case of wearing coat or jacket (CL), the accuracy improvement is significant. Compared with the GaitSet, the accuracy of the proposed is higher by 5% to 10% in the case of wearing a jacket.

We can observe that as the number of training samples increases, the accuracy improves. However, in the practical application the number of training samples is often less than the number of the test, which requires the model to maintain good recognition ability in small-scale training. By comparing the results with other methods, the average accuracy reached 82.8% under NM conditions, 70.1% under BG conditions, and 44.1% under CL conditions, which are better than the GaitSet. Therefore, the proposed method has better performance and stronger robustness.

Secondly, in the same sample division, the model can work well under the NM condition. In the three divisions, the average accuracy of the NM is higher than BG and CL condition. At the same time, it can be observed from the results that the people in BG condition is easier to be identified than the CL condition. The accuracy in NM condition is 10% to 20% higher than that in complex condition (BG and CL). Furthermore, the gait recognition framework proposed in this paper achieves better gait recognition

Table 2. The accuracy of the proposed method GaitTC on the CASIA-B under the MT condition

Gallery NM#1–4			0°–180°											
Probe			0	18	36	54	72	90	108	126	144	162	180	Mean
MT	NM	AE	49.3	61.5	64.4	63.6	63.7	58.1	59.9	66.5	64.8	56.9	44.0	59.3
		MGAN	54.9	65.9	72.1	74.8	71.1	65.7	70.0	75.6	76.2	68.6	53.8	68.1
		GaitSet	86.8	95.2	98.0	94.5	91.5	89.1	91.1	**95.0**	**97.4**	93.7	80.2	92.0
		GaitTC	**87.2**	**95.4**	**97.5**	**94.7**	**91.1**	**88.4**	**91.9**	94.9	96.5	**93.9**	**83.9**	**92.3**
	BG	AE	29.8	37.7	39.2	40.5	43.8	37.5	43.0	42.7	36.3	30.6	28.5	37.2
		MGAN	48.5	58.5	59.7	58.0	53.7	49.8	54.0	51.3	59.5	55.9	43.1	54.7
		GaitSet	79.9	**89.8**	**91.2**	**86.7**	**81.6**	76.7	81.0	**88.2**	90.3	88.5	73.0	84.3
		GaitTC	**80.3**	**88.3**	**90.8**	**86.3**	**81.3**	**77.3**	**82.0**	87.4	**91.6**	**89.3**	**76.0**	**84.6**
	CL	AE	18.7	21.0	25.0	25.1	25.0	26.3	28.7	30.0	23.6	23.4	19.0	24.2
		MGAN	23.1	34.5	36.3	33.3	32.9	32.7	34.2	37.6	33.7	26.7	21.0	31.5
		GaitSet	52.0	66.0	72.8	69.3	63.1	61.2	63.5	66.5	67.5	60.0	45.9	62.5
		GaitTC	**64.9**	**77.1**	**76.1**	**74.4**	**71.4**	**68.0**	**69.9**	**73.1**	**71.4**	**68.4**	**55.1**	**70.0**

performance than other methods, and the accuracy in different partition is higher than other models. There are two main reasons: First, the Transformer module preferentially extracts the attention value of each image block, and retain the gradient of the original data through the residual network, which will be more convenient for the subsequent CNN pooling module to extract gait features. On the other hand, the Transformer module with global receptive field not only extracts the global feature representations in advance, but also further mines the local feature representations after introducing the CNN pooling module, thus improving the performance of the model (Table 3).

Table 3. The accuracy of the proposed method GaitTC on the CASIA-B under LT condition.

Gallery NM#1–4			0°–180°											
Probe			0	18	36	54	72	90	108	126	144	162	180	Mean
LT	NM	CNN-3D	87.1	93.2	97.0	94.6	90.2	88.3	91.1	93.8	96.5	96.0	85.7	92.1
		GaitSet	90.8	97.9	**99.4**	96.9	93.6	91.7	95.0	**97.8**	**98.9**	96.8	85.8	95.0
		GaitTC	**93.9**	**98.0**	98.9	**97.9**	**95.0**	**93.4**	**95.1**	97.2	97.6	**98.0**	**92.0**	**96.1**
	BG	CNN-LB	64.2	80.6	82.7	76.9	64.8	63.1	68.0	76.9	82.2	75.4	61.3	72.4
		GaitSet	83.8	91.2	91.8	88.8	83.3	81.0	84.1	90.0	92.2	94.4	79.0	87.2
		GaitTC	**89.4**	**94.0**	**97.9**	**95.7**	**94.7**	**91.2**	**92.3**	**95.5**	**95.4**	**93.7**	**88.5**	**93.5**
	CL	CNN-LB	37.7	57.2	66.6	61.1	55.2	54.6	55.2	59.1	58.9	48.8	39.4	54.0
		GaitSet	61.4	75.4	80.7	77.3	72.1	70.1	71.5	73.5	73.5	68.4	50.0	70.4
		GaitTC	**79.8**	**90.4**	**88.8**	**85.8**	**85.2**	**81.6**	**81.9**	**87.7**	**87.4**	**85.2**	**70.2**	**84.0**

5 Conclusion

In this work, we propose a novel Transformer-based gait recognition framework, GaitTC. It includes the Transformer model, CNN pooling module and feature aggregation module. The proposed model not only capture global context to extract the global feature representations, but also can obtain the local feature representation using the CNN pooling module. The multi-head self-attention mechanism in this model has good robustness to image noise and incompleteness. At the same time, the residual structure and the layer normalization structure further improve the performance of the algorithm. In the comparative experiments with other models, the accuracy of the model in this paper is higher than other models in most perspectives. The experimental results show that the model also shows excellent performance under three different sample scales.

References

1. Hu, M., Wang, Y., Zhang, Z., Little, J., et al.: View-invariant discriminative projection for multi-view gait-based human identification, pp. 2034–2045 (2013)
2. Kusakunniran, W., Wu, Q., Zhang, J., et al.: Recognizing gaits across views through correlated motion co-clustering. IEEE Trans. Image Process. **23**(2), 696–709 (2014)
3. Wu, Z., Huang, Y., Wang, L., et al.: A comprehensive study on cross-view gait based human identification with deep CNNs. IEEE Trans. Pattern Anal. Mach. Intell. **39**, 209–226 (2016)
4. Chao, H., Wang, K., He, Y., et al.: GaitSet: cross-view gait recognition through utilizing gait as a deep set. Cornell University – arXiv:2102.03247v1 (2021)
5. Rida, I., Almaadeed, N., Almaadeed, S.: Robust gait recognition: a comprehensive survey. IET Biomet. **8**, 14–28 (2018)
6. Zhao, G., Liu, G., Li, H., et al.: 3D gait recognition using multiple cameras, pp. 529–534 (2006)
7. Zheng, X., Li, X., Xu, K., et al.: Gait identification under surveillance environment based on human skeleton. arXiv preprint arXiv:2111.11720 (2021)

8. Shiraga, K., Makihara, Y., Muramatsu, D., et al.: GEINet: view-invariant gait recognition using a convolutional neural network. In: 2016 International Conference on Biometrics (ICB), Halmstad, Sweden, pp. 1–8 (2016)
9. Liao, R., Cao, C., Garcia, E.B., Yu, S., Huang, Y.: Pose-based temporal-spatial network (PTSN) for gait recognition with carrying and clothing variations. In: Zhou, J., et al. (eds.) CCBR 2017. LNCS, vol. 10568, pp. 474–483. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-69923-3_51
10. Huang, Z., Xue, D., Shen, X., et al.: 3D local convolutional neural networks for gait recognition. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada (2022)
11. Wolf, T., Babaee, M., Rigoll, G.: Multi-view gait recognition using 3D convolutional neural networks. In: 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, pp. 14920–14929 (2016)
12. He, Y., Zhang, J., Shan, H., et al.: Multi-task GANs for view-specific feature learning in gait recognition. IEEE Trans. Inf. Forensics Secur. 102–113 (2018)
13. Yu S, Wang, Q., Shen, L., et al.: View invariant gait recognition using only one uniform model. In: 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, vol. 239, pp. 81–93 (2017)