Multimedia Tools and Applications (2021) 80:6065–6078
https://doi.org/10.1007/s11042-020-09935-x

Non-local gait feature extraction and human identification

Xiuhui Wang[1] · **Wei Qi Yan**[2]

Received: 22 December 2019 / Revised: 15 September 2020 / Accepted: 17 September 2020 /
Published online: 12 October 2020

Abstract
As a new human identification technology, gait recognition is receiving more and more attention in recent years. However, traditional gait recognition techniques are limited by the challenges of feature representation and extraction algorithms. In this paper, by utilizing the self-attention mechanism, we propose a novel gait-based human identification solution. Firstly, we utilize non-local neural networks (NLNN) to extract non-local features from a pair of randomly selected gait energy maps (GEIs). Secondly, based on the relationship between GEIs and various parts of the human body, the output of NLNN is horizontally segmented into three sections, i.e., strong-dynamic region, weak-dynamic region and micro-dynamic region, respectively. Thirdly, the segmented gait features are weighted ensembled by three two-class classifiers. Finally, two experiments are carried out with the OU-ISIR large population dataset and the CASIA dataset B to evaluate the proposed approach.

Keywords Human identification · Non-local features · Gait recognition · Self-attention

1 Introduction

As a prominent human identification technology, gait recognition is receiving more and more attention in recent years [9, 26]. Gait recognition can be used in intelligent video surveillance systems and has also been investigated as new means in human leg rehabilitation medical diagnosis. Compared with other biometric methods, such as face recognition and fingerprint recognition, gait recognition has the advantage of easy remote recognition, and the recognition process usually does not need to be deliberately cooperated by the recognized object [33]. However, due to changes in external factors during gait data collection,

✉ Xiuhui Wang
wangxiuhui@cjlu.edu.cn

Wei Qi Yan
wyan@aut.ac.nz

[1] Key Laboratory of Electromagnetic Wave Information Technology and Metrology of Zhejiang Province, College of Information Engineering, China Jiliang University, No. 258, Xueyuan Street, Hangzhou 310018 China

[2] Auckland University of Technology, No. 2-14, Wakefiled Street, Auckland 1010, New Zealand

such as lighting, road conditions, camera resolution as well as clothing, weight-bearing, and carrying conditions of a pedestrian, gait variance of the same person may be much obvious than the differences from different persons. The traditional method to resolve the problems is to construct well-designed gait features and reduce the influence of interference on gait recognition by setting a group of constraints [17, 22, 23]. Nevertheless, though this kind of methods can well solve the problem in a specific environment, it is difficult to apply them to other applications. Fortunately, deep learning techniques provide better support for gait feature representation using end-to-end technology [5, 19].

This paper proposes a novel gait-based identification method, which combines non-local and regionalized features to better extract fine-grained gait features. The main contributions of this paper include the following two aspects:

(1) A regionalized gait feature representation is proposed. Considering the pre-alignment operation in the generation process of gait energy map (GEI) [7], we segment a human body contour directly into three parts, i.e., strong-dynamic region, weak-dynamic region, and micro-dynamic region. Three two-class Softmax functions are employed as classifiers to be trained separately.
(2) A gait-based human identification method by combining non-local and regionalized features is presented. After extracting non-local features, we input them into two channels of the network separately and obtain the relevant non-local features so as to improve the non-locality of the regionalized features, thus to enhance the inter-class discrimination of gait features.
(3) Comparative experiments are carried out based on two well-known gait databases, which demonstrate the effectiveness of the proposed method. The first database, namely, CASIA Gait Dataset B [40], is used to evaluate the proposed method in a cross-view environment; whilst the second one, i.e. OU-ISIR Large Population Gait Dataset [11], is used to test the proposed method under large-scale datasets.

The rest of this paper is structured as follows. In Section 2, we will review and discuss those existing gait recognition methods. Then, the solution proposed in this paper will be described on details in Section 3. Next, in Section 4, comparative experiments will be carried out based on two well-known gait datasets. Finally, in Section 5, our conclusion and future work of this paper will be addressed.

2 Related work

According to the categories of learning models, the existing gait recognition methods can be roughly categorized into two types: discriminative methods and generative methods [29]. In a typical supervised machine learning task, to predict the label Y from the feature set X, a discriminative method finds $P(Y|X)$, namely, the posterior probability, while a generative method finds $P(X, Y)$, to wit, the joint probability.

The discriminative gait recognition method is to learn gait models from the historical walking data and judge the probability that the query sample belongs to each known class using the learned model. Typical discriminative methods [6, 8, 15, 27, 34, 38] include support vector machine (SVM), artificial neural networks (ANN), decision tree, conditional random fields (CRF), and linear discriminant analysis, etc. Shiraga et al. [27] presented a gait recognition approach based on convolutional neural network (CNN) and demonstrated it in cross-view gait recognition problems.This approach significantly outperforms many

existing ones on OU-ISIR LP Dataset, but more evaluations on gait datasets with wider view variation have not beed conducted. In [34], a Gabor wavelets-based gait recognition method was proposed, in which an effective dimension reduction algorithm $(2D)^2$PCA was used to preprocess the input GEIs and a multi-class SVM classifier was employed to achieve gait classification.One limitation of the above method is that the process of generating GEI may lose some dynamic gait features, since they are calculated by averaging a series of images. To resolve gait-based human identification problems, Wu et al. [38] proposed three CNN structures which can be trained using a small group of labelled cross-view gait videos. Three network architectures were investigated to compute the differences between two GEIs and achieve gait-based human identification. In [8], gait recognition is achieved using SVM and neural networks by simultaneously extracting model-based and model-free gait features from each frame in a gait cycle. Their experiments demonstrated that the SVM-based method was approximately equal in term of recognition rate with neural network-based approaches when the input was geometry gait features; but the recognition rate of the SVM-based method was less than that of neural network-based ones, when the input is texture gait features. Nomm et al. [15] used decision tree to analyse the gait of patients so as to support modelling and diagnostics of Parkinsons disease. Hagui and Mahjoub [6] utilized a hidden CRF model to combine a spatial classifier and a temporal classifier, the former assigned a label to a local feature and the latter used a motion history image. This method assumed that only one moving object existing in the gait videos which is similar with many other methods. In [37], a gait recognition scheme was proposed, which utilized a new gait feature representation by using consecutive gait silhouette pictures, and constructed a multichannel CNN network to tackle a set of sequential images in parallel.This method only utilized the generated side view image as the matching feature, instead of exploring intermediate layer features neither others effective feature extraction methods. In short, discriminative gait recognition methods are to establish a discriminant function under the condition with a finite number of samples so as to find the optimal classification plane between various categories. The primary advantage of this type of methods is that it can clearly distinguish the differences between multiple classes, or one and other classes. The main disadvantage is that it does not reflect the characteristics of the training data itself.

The generative gait recognition method is to learn a gait model for each known person in a given database, and then use the corresponding model of all classes to determine the matching probability of the query sample. In particular, generative adversarial network (GAN) [4], variational autoencoder (VAE) [13], and their variations, e.g. information maximizing GAN (InfoGAN) [2], GAN using Divided Z-Vector (DzGAN) [31], least squares GAN (LSGAN) [21], Wasserstein GAN (WGAN) [1], Ladder VAE(LVAE) [28], are important approaches in pattern recognition and artificial intelligence, and their outstanding data generation ability has been widely concerned [24]. GAN [4] is a novel generative model, in which are two networks, one is generator, and the other is discriminator. The role of the generator is to create as realistic data as possible to deceive the discriminator. The role of discriminator tries to distinguish fake samples from real ones. InfoGANs [2] apply an information-theoretic extension to the GAN that is able to learn disentangled representations in a completely unsupervised manner.DzGANs [31] implement conditional learning using not images but one-hot vector by dividing the range of z-vector. In the DzGAN, the discriminator is fed by the images with label using one-hot vector and the generator is fed by divided z-vector with corresponding label fed into the discriminator. LSGANs [21] adopt the least squares loss function for the discriminator and minimize the objective function of LSGAN yields minimizing the Pearson chi(2) divergence, which can generate higher quality

images than regular GANs and perform more stable during the learning process. By contrast, WGAN [1] minimizes a Wasserstein-1 distance between the two distributions making progress toward stable training of GANs. On the other hand, a VAE [13, 28] is an autoencoder whose encodings distribution is regularized during the training in order to ensure that its latent space has good properties allowing us to generate some new data. Moreover, the term "variational" comes from the close relation there is between the regularization and the variational inference method in statistics.

Representative gait recognition methods based on generative models [12, 14, 20, 25, 32, 36] include hidden Markov model (HMM), naive Bayesian model, Gaussian mixture model (GMM), latent Dirichlet allocation (LDA), probabilistic latent semantic analysis (PLSA), etc. Wang et al. [36] proposed a method for gait recognition based on a self-adaptive hidden Markov model (SAHMM), which uses a small number of samples with similar acquisition conditions to the target environment. In [32], a cross-view gait recognition method based on ensemble learning was proposed, which combined several basic HMM classifiers to increase the robustness of gait classification in different views. But this method did not utilize the deep learning technologies, which was powerful in feature extraction and representation process. Kozlow et al. [14] proposed a Bayesian network-based approach to achieve gait classification, which represented human gait by using the body joint coordinates from stride length and various joint angles. Manap et al. [20] demonstrated the potential of naive Bayes classifier as an normal gait pattern detection. In [25], a gait-based person identification method based on Gaussian mixture model and universal background model was proposed which used inertial signals from a smartphone. Kanwar and Upadhyay [12] presented an appearance-based gait identification method which is insensitive to view, clothing and lightning conditions. Compared with several other approaches, this method performed better in related experiments, but the results were obtained with small datasets. In a word, generative gait recognition methods represent the distribution of data from a statistical viewpoint. The main advantage of generative gait recognition methods is that their learning methods converge faster; i.e., when the sample number increases, the learned models can converge to the real model more quickly; when there is a hidden variable, the generation method can still be used. The shortcoming of these methods is that the learning and calculation process is more complicated.

In order to obtain the intrinsic features with high discrimination ability, we present a regionalized gait feature representation. Then, we propose a gait-based human identification method by combing non-local and regionalized features. Finally, the proposed methods are evaluated based on CASIA Gait Dataset B and OU-ISIR Large Population Gait Dataset.

3 Our proposed methods

As shown in Fig. 1, a gait-based human identification method mainly consists of three modules:

(1) Gait data preprocessing. This module mainly includes two basic operations, i.e., GEI construction and generation of GEI pairs. In the GEI construction, we calculate GEIs for each person, which reflect the spatiotemporal motion characteristics of the walking-related parts of a human body in a gait cycle. Then, in the second operation, positive and negative GEI pairs are sampled from the obtained GEIs of each person, which will be used as the inputs of next module.

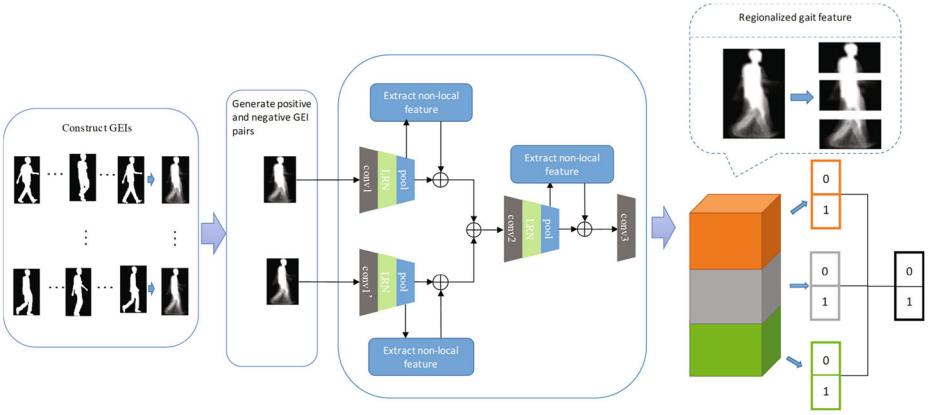

Fig. 1 The flowchart of a gait-based human identification method

(2) Non-local gait feature extraction and feature fusion. In this module, there are three processes for extracting non-local features, two of which are based on the output of two convolutional channels and the other is based on the fusion results. Besides, all four information fusions are occurred in this module using a proportional fusion method; namely, each fusion component is counted as 50%.

(3) Regionalized gait feature extraction and gait classification. Taking consideration of the correspondence between GEI and various parts of a human body and the information capacity in various parts of the human body, we cut the feature map output into three sections by using the third convolutional layer, namely, using static region, micro-dynamic region and strong-dynamic region to represent the motion of different body parts. The theoretical basis of this feature representation is that the convolution and pooling layers of a CNN network do not change the spatial distribution of features. Finally, the regionalized gait feature maps are fed into the three Softmax classifiers to achieve gait classification.

3.1 Gait data preprocessing

Gait energy map (GEI) is the most widely used gait representation, which can reflect the spatiotemporal motion characteristics of the walking-relevant parts of the human body. GEI is calculated by scaling, aligning, averaging, etc. of a sequence of contour images in a gait cycle, as defined by

$$GEI = \frac{1}{T}\sum_{t=1}^{T} I_t\,(x, y) \tag{1}$$

where I_t is the binarized silhouette image at the t-th frame, (x, y) is the coordinates of each pixel in I_t, and T is the number of frames in a gait cycle.

After obtaining the GEIs of each person in a given gait database, we will focus on the second step of the gait data preprocessing, i.e., generation of positive and negative GEI pairs. First, we randomly select a person from the training set, one of the GEIs as the reference g_b. Then we randomly select a GEI from the remaining GEIs of this person as the positive GEI g_p, we can get a positive GEI pair (g_b, g_p), which consists of a reference sample and a positive sample. Similarly, we randomly select another person from the remaining in the

training set and randomly extract one as the negative GEI g_n to construct a negative GEI pair (g_b, g_n). Repeating the above process, we can get a set of positive and negative GEI pairs for each person:

$$G_{ij} = \left\{ \left(g_{bi}, g_{pi}\right)_j, \left(g_{bi}, g_{ni}\right)_j \right\} \tag{2}$$

where $i = 1, 2, \cdots, M$, $j = 1, 2, \cdots, N$, M is the total number of persons in the training set, N is the number of positive and negative GEI pairs. In the subsequent training process, the two GEIs of the positive GEI pair are firstly fed into the two channels of a NLNN, with a training label '0'. Then, the two GEIs of the negative GEI pair are respectively input into the two channels of NLNN, with a training label '1', thereby completing the input of a pair of positive and negative GEIs.

3.2 Non-local gait feature extraction

In the process of gait feature extraction, a non-local operation is defined as:

$$y_i = \frac{1}{C} \sum_j f(x_i, x_j) g(x_j) \tag{3}$$

where x is the input gait image and y is the output image of the same size as x, i is the index of an output position and j is the index that enumerates all possible positions, function $f(.)$ computes a scalar between i and all j, function $g(.)$ computes a representation of the input image at the position j, and the response is normalized by a factor C.

As GEIs have already contained the temporal and spatial gait characteristics in a gait cycle, in this paper we directly extract non-local information from each GEI. More specifically, as shown in Fig. 1, after the feature extraction process from a set of GEI pairs through the two channels of a NLNN, we extract two types of non-local information, *viz*, the non-local information of a single GEI and that between two GEIs. The NLNN includes a convolutional layer, a pooling layer, and a normalized layer. We use LBNet [38] as the basic CNN unit, and apply the local response normalization (LRN) [16] to construct the normalized layer. The specific settings of each CNN module in Fig. 1 are shown in Table 1.

To simplify the training process of NLNN [35], in the CNN3 module, we remove the normalization layer and pooling layer, and add dropout to prevent over-fitting. The dot-product similarity $f(.)$ is defined as:

$$f(x_i, x_j) = g(x_i)^T g(x_j)) \tag{4}$$

Table 1 setting of the CNN modules in Fig. 1

Module	Layer	Kernels	Size	Step	Activation function
CNN1	conv1	16	$7 \times 7 \times 1$	1	ReLU
	pool		2×2	2	
CNN1′	conv1	16	$7 \times 7 \times 1$	1	ReLU
	pool		2×2	2	
CNN1	conv1	16	$7 \times 7 \times 1$	1	ReLU
	pool		2×2	2	
CNN1	conv1	16	$7 \times 7 \times 1$	1	ReLU
	pool		2×2	2	

where $g(x_i) = W_g x_i$,and W_g is the weight matrix to be trained.

Finally, the 256 feature maps generated by using the CNN3 module are directly grouped into blocks and connected to three binary classifiers. Unlike the non-local neural network, in NLNN, non-local operations are simplified as:

$$z_i = G_i + W_z y_i \tag{5}$$

where z_i is a combination of two features, G_i represents the output of GEI after the convolution module, y_i represents the output after the non-local module, W_z is used to enlarge y_i to the same dimension as G_i.

To implement the first layer in a non-local network, as shown in Fig. 2, we obtain 16 feature maps with size 61×41, and use three convolution kernels with the size 1×1, namely, θ, Φ and g, to reduce the number of channels and the calculation amount.

Then, the transposition of the output matrix of the θ channel is multiplied with the output matrix of the Φ channel to calculate the similarity of corresponding inputs. On this basis, the Softmax function is used to obtain the combining attention of these channels, the normalized correlation between each pixel in the current feature map and all other position pixels is calculated. At the same time, the input data of the g channel is similarly operated, and the result is multiplied with the Softmax result of the first two channels.

Finally, after dealt with by an output channel with a 1×1 convolutional kernel, the input and output scales are adjusted to the same; the output is combined with the gait features of the first CNN module. This step obtains global information through the operations of non-local feature extraction and uses residuals [10] to integrate the local and non-local output.

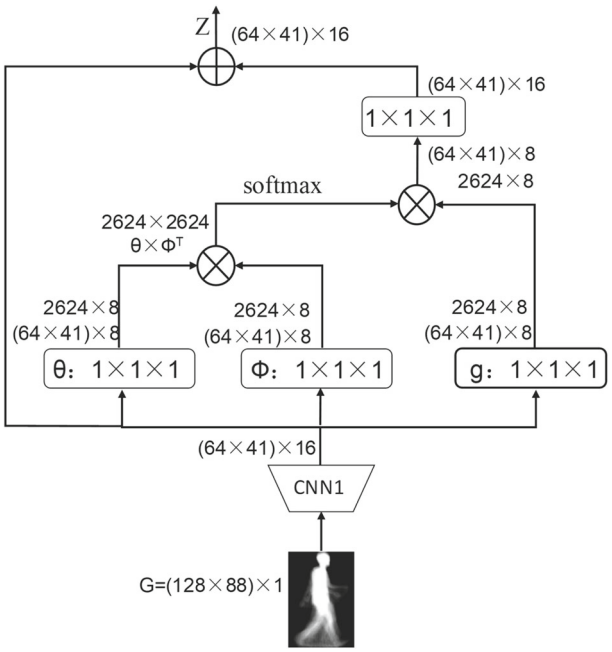

Fig. 2 The flowchart of extracting non-local gait features

3.3 Gait-based human identification algorithm through self-attention

Fine-grained classifications are to accurately classify subcategories in a category [16]. Taking the picture-based bird classification as an example, it is necessary to detect the presence of birds in a picture, and to detect which kind of bird it is. This can be summarized as a classification task that utilizes both global features and local features. Thus, in a fine-grained classification of birds, the global feature is the whole picture, while the local features are the local characteristic or important parts.

Gait classification is typical fine-grained classification problem, which pays more attention to local gait information and focuses on finding the distinguishing regional regions between gaits from different persons. Considering there are aligning operations in the GEI generation process, we split the 256 features map obtained after the CNN module and the non-local module into three parts, as shown in Fig. 1, to represent the strong-dynamic region, the weak-dynamic region and the micro-dynamic region, respectively. Furthermore, three binary Softmax classifiers are used for gait classification, and the three-part classification results are combined to generate the final verification results. In addition, Wu et al. [38] have demonstrated that in CNN-based gait recognition tasks, network performance with three convolutional layers is optimal amongst having two, three or five convolutional layers.

According to the above conclusions, this paper designs an NLNN network with three convolutional layers and proposes a gait-based human identification algorithm through NLNN, as shown in Algorithm 1.

Algorithm 1 Gait-based human identification algorithm through NLNN.

Input: A training dataset $\{H_{pq}|p = 1, 2, \cdots, P;\ q = 1, 2, \cdots, Q\}$ which contains silhouette images of P persons, each person has Q silhouette images from different views.
Output: Identification results.
Step 1: Construct GEIs for each person. Based on the gait silhouette image in training dataset, we use the method shown in (1) to compute GEIs for each person.
Step 2: Sample positive and negative GEI pairs. As described in Section 3.1, a set of positive and negative GEI pairs is constructed by randomly selecting samples from the GEIs of each person.
Step 3: Extract non-local gait feature. By taking the obtained positive and negative GEI pairs as the input, we extract non-local features from both each GEI and the fusion results of two GEIs. Figure 2 depicts an example of extracting non-local gait features.
Step 4: Extract regionalized features. In this step, we seprate the feature map output by using the third convolutional layer of a NLNN into three sections, namely static region, micro-dynamic region and strong-dynamic region, to represent the motion properties of different body parts separately.
Step 5: Perform gait classification. With the obtained regionalized gait features, we use three Softmax classifiers to achieve gait classification.
Step 6: Achieve gait-based human identification, which is derived from the gait classification results of step 5.

4 Experimental results

In this section, we designed two experiments to evaluate our method according to the evaluation criteria provided by the baseline algorithm [11]. The recognition performance was

evaluated using the rank-1 identification rate as an evaluation in a one-to-N matching applications, the rank-1 metric denotes the percentages of correct objects out of all the objects appearing within the first rank.

The first experiment is based on the OU-ISIR LP Dataset that is provided by the Institute of Scientific and Industrial Research (ISIR), Osaka University (OU). The data have been collected since March 2009 through outreach activity events in Japan. As one of the largest gait datasets at present, there are over 4,000 subjects with a large age span and a balanced gender ratio, as shown in Fig. 3. Each individual in OU-ISIR LP dataset is sampled with two video gait sequences, namely the gallery sequence and the probe sequence in the normalized silhouette images with the size of 128 × 88 pixels. Besides, according to the view angle of cameras, each sequence is grouped into four types, i.e., 55°, 65°, 75°, 85°.

The second experiment is based on the CASIA dataset B that is a large multiview gait database. CASIA dataset B was provided by the Institute of Automation, Chinese Academy of Sciences (CASIA) in 2005. There are 124 people, each of them is sampled from 11 view angles, i.e., 0°, 18°, 36°, $\cdots$, 180° as shown in Fig. 4. Three changing conditions i.e. view angle, clothing with or without a bag, are separately considered.

4.1 Experimental configuration

According to the test protocol of gait recognition and evaluation criteria [23], we grouped the 1912 subsets with full view angles in the OU-ISIR LP dataset into two sections, i.e., 956 objects for training and 956 objects for test. Then FAR, FRR, EER and rank1 cross-view recognition rates are collected and compared with the benchmark methods and deep learning-based methods. In the experiment based on CASIA dataset B, we use 7:3 ratio to segment the dataset, i.e., 100 objects are used for training, and the remaining 24 objects are used to test and calculate the average recognition rate of rank1 under various conditions.

In addition, since the proposed NLNN network calculates the similarity of GEI pairs from two channels, it is necessary to ensure the equalization of positive and negative samples. In experiments based on the OU-ISIR LP dataset, we first randomly selected one object from the list, and randomly extracted one gait sample from its GEI set. Then, when constructing a positive GEI pair, we selected the same object as the reference sample, and then randomly extraced a sample with random view angles in the corresponding GEI set to obtain a positive GEI pair by combining it together with the reference sample. As the construction of negative sample pairs, we used the same operation to select a negative sample but chose an object that is different with the reference sample. In addition, for the experiments conducted on the CASIA dataset B dataset, a similar approach was used to construct positive and negative GEI pairs.

Fig. 3 Examples from OU-ISIR LP Dataset

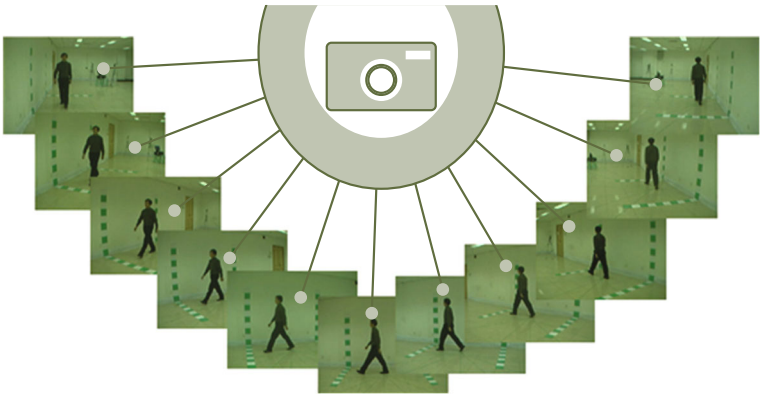

Fig. 4 Examples from CASIA dataset B

4.2 Comparative experiments on the OU-ISIR LP dataset

The methods involved in this experiment include view transformation models [22, 23], GEINet [27], FBW-CNN method [39], FMP method proposed [3] and the proposed method in this paper. The corresponding rank-1 recognition rates of six methods are shown in Table 2. The results reveal that the proposed method performs excellent compared to others.

According to Table 2, we see that the proposed method has the advantage of high recognition rate for various cross-view combinations on the OU-ISIR LP dataset and its correct recognition rate is significantly higher than other methods as the view angles increase. The reason is that the method proposed in this paper not only obtains the fine-grained information of human gait, but also improves the non-locality of segmentation feature by using non-local operations, thus obtaining more distinguishing gait features.

4.3 Comparative experiments on the CASIA dataset B

To further evaluate the methods presented in this paper, we performed a comparative experiment using CASIA Dataset B. This experiment adopts the same experimental configuration as STDNN [30]. The input picture size is 126×126, the rate of recognition of rank-1 is calculated under the viewangle of 36° and 54° respectively. In this experiment, we compared six methods, namely $(2D)^2PCA$ [34], SST-MSCT [18], CNN-CGI [38], STDNN [30], CVGR-EL [32] and our method. The experimental results are shown in Tables 3 and 4, in which experimental data of SST-MSCT [18], CNN-CGI [38] and STDNN [30] are obtained from the published work [30].

Tables 3 and 4 show that out method performs the best compared to others in terms of correct recognition rate with 36° and 54° on CASIA dataset B. On the one hand, compared with shallow learning methods, our method has a very significant improvement. For example, when the gallery data is 36° and the probe data is 72°, the correct recognition rates of $(2D)^2PCA$ [34] and SST-MSCT [18] are 57.1% and 59.7%, while the correct recognition rate of our method is 97.9%. On the other hand, our method performs better than other methods based on deep learning networks.

There are two reasons why our method achieves this performance:

Table 2 Rank1 recognition rate of different methods on OU-ISIR LP Dataset(%), GA- = gallery, PR- = probe

View Angles	Methods	GA-55°	GA-65°	GA-75°	GA-85°
PR-55	WQVTM [23]	/	81.5	70.2	51.1
	TCM [22]	/	81.7	71.9	53.7
	FBW-CNN [39]	86.1	79.8	65.3	51.9
	FMP [3]	95.2	93.6	81.2	62.2
	GEINet [27]	94.7	93.7	90.1	81.4
	Our method	99.1	99.2	98.5	96.9
PR-65	WQVTM [23]	78.3	/	80.0	68.5
	TCM [22]	79.9	/	80.0	73.0
	FBW-CNN [39]	80.1	88.9	84.4	73.7
	FMP [3]	90.9	95.3	95.5	90.2
	GEINet [27]	93.2	95.1	94.1	91.2
	Our method	99.3	99.3	99.1	98.4
PR-75	WQVTM [23]	64.0	79.2	/	79.0
	TCM [22]	70.8	79.5	/	79.4
	FBW-CNN [39]	70.1	86.1	87.8	84.1
	FMP [3]	77.5	94.4	96.0	94.2
	GEINet [27]	89.1	93.8	95.2	94.6
	Our method	98.4	99.1	99.2	98.7
PR-85	WQVTM [23]	48.6	67.5	78.2	/
	TCM [22]	54.5	70.2	79.0	/
	FBW-CNN [39]	55.7	78.1	84.6	86.4
	FMP [3]	55.4	87.1	94.8	94.7
	GEINet [27]	79.9	90.6	93.8	94.7
	Our method	96.5	99.1	99.5	99.5

(1) In contrast to the progressive behavior of recurrent and convolutional modules, non-local modules capture long-range dependencies directly by computing interactions between any two positions, regardless of their positional distance.

Table 3 The cross-view correct recognition rate achieved by different methods with 36° on CASIA dataset B

Gallery	Probe	$(2D)^2$PCA [34]	SST-MSCT [18]	CNN-CGI [38]	STDNN [30]	CVGR-EL [32]	Our methods
36°	0°	56.3	63.4	71.5	87.3	89.8	97.9
36°	72°	57.1	59.7	74.6	90.5	91.4	97.9
90°	54°	59.0	61.2	76.6	91.2	90.7	97.9
90°	126°	64.6	66.3	74.3	92.1	91.1	91.7
144°	108°	61.9	57.8	72.9	90.1	89.9	95.8
144°	180°	63.9	64.7	70.2	89.7	89.7	91.7

Table 4 The cross-view correct recognition rate achieved by different methods with 54° on CASIA dataset B

Gallery	Probe	$(2D)^2$PCA [34]	SST-MSCT [18]	CNN-CGI [38]	STDNN [30]	CVGR-EL [32]	Our methods
54°	0°	48.5	47.8	54.7	83.7	82.2	83.3
54°	108°	51.7	49.2	58.5	86.3	86.9	97.9
90°	36°	59.0	62.5	62.9	86.4	88.4	91.7
90°	144°	59.1	57.8	61.1	87.5	85.1	85.4
126°	72°	63.7	61.5	57.4	84.3	89.3	97.9
126°	180°	57.5	51.8	54.9	82.4	85.0	87.5

(2) Because different parts of human body in the vertical direction bear different degrees of gait characteristics in the process of walking, we divide the human gait information into different blocks to improve the recognition rate.

5 Conclusion and future work

Considering that the existing deep learning-based gait recognition approaches mainly extract global features by stacking multiple convolutional layers and neglect most fine-grained gait features, this paper proposed a novel gait-based human identification method that integrates non-local and regionalized gait features. The proposed method effectively improves the global nature of the obtained gait features by extracting the non-local feature of each GEI and the relative non-local features of the merged gait data in subsequent processing. Then, using the geometric characteristics of GEIs, the output feature map is horizontally divided into three sections for training three binary classifiers and presenting fine-grained information in the gait data, which can enhance the distinguishing ability of obtained features.

In addition, since the proposed gait recognition method takes GEIs as input, one possible limitation of this method is that the construction process of GEI may face tremendous challenges in practical applications. This is mainly due to the fact that the construction of GEI highly depends on the extraction accuracy of gait silhouette images and the fact that GEIs retain only small part of the temporal-spatial characteristics of gait data. Therefore, our future work will focus on optimizing the robustness and generalization capabilities of the proposed algorithm to improve its performance in complex scenarios.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grant No.61602431 and Zhejiang Provincial Natural Science Foundation of China under Grant No.Y20F020113, as well as a scholarship from the China Scholarship Council.

Compliance with Ethical Standards

Conflict of interest We declare that we have not financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled “Non-local Gait Feature Extraction and Human Identification”.

References

1. Arjovsky M, Chintala S, Bottou L (2017) Wasserstein generative adversarial networks. In: Proceedings of the 34th international conference on machine learning, vol 70
2. Chen X, Duan Y, Houthooft R, Schulman J, Sutskever I, Abbeel P (2016) InfoGAN: interpretable representation learning by information maximizing generative adversarial nets. In: Advances in neural information processing systems, vol 29
3. Chen Q, Wang Y, Liu Z, Liu Q, Huang D (2017) Feature map pooling for cross-view gait recognition based on silhouette sequence images. In: IEEE international joint conference on biometrics (IJCB), pp 54–61
4. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ (eds) Advances in neural information processing systems 27. Curran Associates, Inc, pp 2672–2680
5. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Cambridge
6. Hagui M, Mahjoub MA (2016) Hidden conditional random fields for gait recognition. In: International image processing, applications and systems, pp 1–6
7. Han J, Bhanu B (2006) Individual recognition using gait energy image. IEEE Trans Pattern Anal Mach Intell 28(02):316–323
8. Hanon AlAsadi A (2014) Gait recognition using support vector machine and neural network. J Basrah Res 40:68–78
9. He Y, Zhang J (2018) Deep learning for gait recognition: a survey. Pattern Recognit Artif Intell 31(05):442–451
10. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 770–778
11. Iwama H, Okumura M, Makihara Y, Yagi Y (2012) The OU-ISIR gait database comprising the large population dataset and performance evaluation of gait recognition. IEEE Trans Inf Forensics Secur 7(5):1511–1521
12. Kanwar A, Upadhyay P (2014) An appearance based approach for gait identification using infrared imaging. In: International conference on issues and challenges in intelligent computing techniques (ICICT), pp 719–724
13. Kingma DP, Welling M (2014) Auto-encoding variational bayes. In: 2nd International conference on learning representations, vol 1
14. Kozlow P, Abid N, Yanushkevich SN (2018) Gait type analysis using dynamic bayesian networks. Sensors 18(10):3329–3338
15. Krajushkina A, Nõmm S, Toomela A, Medijainen K, Tamm E, Vaske M, Uvarov D, Kahar H, Nugis M, Taba P (2018) Gait analysis based approach for parkinson's disease modeling with decision tree classifiers. In: IEEE International conference on systems, man, and cybernetics, vol 10, pp 3720–3725
16. Krizhevsky A, Sutskever I, Hinton GE (2017) ImageNet classification with deep convolutional neural networks. CACM 60(6):84–90
17. Kusakunniran W, Wu Q, Li H, Zhang J (2010) Multiple views gait recognition using view transformation model based on optimized gait energy image. In: IEEE International conference on information and automation, pp 1058–1064
18. Lam T, Cheung KH, Liu J (2011) Gait flow image: a silhouette-based gait representation for human identification. Pattern Recognit 44:973–987
19. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7):436–445
20. Manap HH, Tahir NM, Abdullah R (2012) Anomalous gait detection using naive bayes classifier. In: IEEE symposium on industrial electronics and applications, pp 378–381
21. Mao X, Li Q, Xie H, Lau RYK, Wang Z, Smolley SP (2017) Least squares generative adversarial networks. In: 2017 IEEE international conference on computer vision (ICCV), pp 2813–2821
22. Muramatsu D, Makihara Y, Yagi Y (2015) Cross-view gait recognition by fusion of multiple transformation consistency measures. IET Biom 4(2):62–73
23. Muramatsu D, Makihara Y, Yagi Y (2016) View transformation model incorporating quality measures for cross-view gait recognition. IEEE Trans Cybern 46(7):1602–1615
24. Pan Z, Yu W, Yi X, Khan A, Yuan F, Zheng Y (2019) Recent progress on generative adversarial networks (GANs): a survey. IEEE Access 7:36322–36333
25. San-Segundo R, Cordoba R, Ferreiros J, D'Haro-Enríquez LF (2016) Frequency features and GMM-UBM approach for gait-based person identification using smartphone inertial signals. Pattern Recogn Lett 73(C):60–67

26. Sarkar S, Phillips P, Liu Z (2005) The humanid gait challenge problem: data sets, performance, and analysis. IEEE Trans Pattern Anal Mach Intell 27(02):162–177
27. Shiraga K, Makihara Y, Muramatsu D, Echigo T, Yagi Y (2016) Geinet: view-invariant gait recognition using a convolutional neural network. In: International conference on biometrics (ICB), vol 1, pp 1–8
28. Sonderby CK, Raiko T, Maaloe L, Sonderby SK, Winther O (2016) Ladder variational autoencoders. In: Advances in neural information processing systems, vol 29
29. Takemura N, Makihara Y, Muramatsu D, Echigo T, Yagi Y (2018) On input/output architectures for convolutional neural network-based cross-view gait recognition. IEEE Trans Circ Syst Video Technol 1(1):1–1
30. Tong S, Fu Y, Yue X, Ling H (2018) Multi-view gait recognition based on a spatial-temporal deep neural network. IEEE Access 6:57583–57596
31. Tsunashima H, Hoshi T, Chen Q (2018) DzGAN: improved conditional generative adversarial nets using divided Z-vector. In: 2018 International conference on computing and big data. International conference on computing and big data, Coll Charleston, Charleston, SC, SEP 08-10, 2018, pp 52–55
32. Wang X, Yan WQ (2019) Cross-view gait recognition through ensemble learning. In: Neural computing and applications
33. Wang X, Yan WQ (2020) Human gait recognition based on frame-by-frame gait energy images and convolutional long short term memory. Int J Neural Syst 30(1):1–12
34. Wang X, Wang J, Yan K (2018) Gait recognition based on Gabor wavelets and $(2D)^2$PCA. Multimed Tools Appl 77(10):12545–12561
35. Wang X, Girshick R, Gupta A, He K (2018) Non-local neural networks. In: IEEE/CVF conference on computer vision and pattern recognition, pp 7794–7803
36. Wang X, Feng S, Yan WQ (2019) Human gait recognition based on self-adaptive hidden Markov model. In: IEEE transactions on computational biology and bioinformatics, pp 1–10
37. Wang X, Zhang J, Yan WQ (2019) Gait recognition using multichannel convolution neural networks. In: Neural computing and applications, pp 532–539
38. Wu Z, Huang Y, Wang L, Wang X, Tan T (2017) A comprehensive study on cross-view gait based human identification with deep cnns. IEEE Trans Pattern Anal Mach Intell 39(02):209–226
39. Wu H, Weng J, Chen X, Lu W (2018) Feedback weight convolutional neural network for gait recognition. J Vis Commun Image Represent 55:424–432
40. Yu S, Tan D, Tan T (2006) A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In: International conference on pattern recognition, pp 441–444