Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 11, No. 1, 2024

Research on Gait Recognition Method Based on LFE-MGPA Network

Meiqi Zhao, Hua Huo

School of Information Engineering, Henan University of Science and Technology, China

Abstract: As a non-contact long-distance biometric recognition technology, gait recognition is tasked to realize the identification of long-distance pedestrians according to their walking patterns. However, gait recognition is greatly affected by external factors such as what the subject is wearing and carrying. To solve this problem, this paper focuses on gait feature extraction and proposes LFE-MGPA network based on convolutional neural network (CNN). A local feature extractor (LFE) was constructed to extract fine-grained features of gait through SConv, and MGPA module was integrated to extract collection-level information of different layers. In order to improve the recognition accuracy of the network model, a joint loss function is constructed. The experiment was carried out on the open gait dataset CASIA-B, and the recognition accuracy reached 97.53%, 94.5% and 81.0% respectively under the three walking states of normal walking, backpack and overcoat, which is higher than the current mainstream algorithms. In particular, the recognition accuracy is significantly improved in the walking state of backpack and overcoat, which proves that the method has strong robustness.

Keywords: Gait recognition; Convolutional neural network; Feature extraction; Gait contour diagram.

1. Introduction

In recent years, gait recognition research has gradually become a popular research direction in the field of computer vision and biometric recognition. Compared with general biometric recognition technology (such as fingerprints[1], faces[2,3], irises[4], etc), gait recognition has the advantages of no need for subject cooperation, easy acquisition, non-contact, difficult to camouflage, and low image resolution requirements. The above characteristics mean that gait recognition technology has a wide range of application prospects and economic value in social security and medical diagnosis. However, as the basic information of gait recognition, the change of the pedestrian's posture during walking is easily affected by external factors, such as the walking speed of the pedestrian, clothing, carrying items, and the camera's Angle of view and frame rate. These factors make gait recognition very challenging, especially cross-perspective gait recognition, that is, recognizing gait information obtained from different angles.

In the early stage of gait recognition research, static or dynamic features related to gait were manually extracted from contour sequence, and features were dimensionality reduced or matched by machine learning. Alternatively, a template is constructed to maintain the temporal and spatial information in the gait sequence directly through the gait contour diagram, and the discriminative representation is learned by machine learning. Han[5]et al. proposed a spatial-temporal gait representation called gait Energy diagram (GEI), which has low computational complexity and can effectively maintain the spatial information in the gait sequence, but loses the temporal information. Wu[6] et al. proposed a gait recognition method based on deep convolutional neural networks. In this method, three different network structures with GEI as input were designed, and finally these networks were fused to obtain recognition results. Compared with the existing methods, this method has obvious advantages when the Angle of view changes greatly. In order to solve this problem, Wang[7]et al proposed a timing template called Chrono-Gait Image(CGI), which can effectively maintain the timing information in gait.

At present, most contour-based gait recognition methods are based on deep learning[8], among which convolutional neural network has received great attention. Shiraga[9] et al. proposed GEINet, which takes GEI as the input of CNN network, and completes identification through two continuous convolution layers, pooling layers, normalization layers, and finally two fully connected layers. Fan[10] et al. proposed a component-based gait recognition method, GaitPart, which improved the gait recognition rate from the perspective of space and time, extracted fine-grained spatial information from a single image by using focus conver (FConv), and extracted and aggregated short-time gait information from each component by using a micro-motion capture module. Ghorbani[11] et al. proposed a novel deep convolutional Network that uses a Capsule Network (CapsNet) to learn deeper partial and global relationships and assign weights to related features to obtain a gait representation that is more robust to changes in appearance. Chao[12] et al. propose an end-to-end gait recognition method, GaitSet, which takes gait sequence as input, uses convolutional network to directly extract spatial features from the original gait contour, and then adopts Set Pooling (SP) operation to conduct a deep set of spatial features. The Multilayer Global Pipeline with Attention (MGPA) module is used to extract the collection-level information of each convolution layer.

The contour-based gait recognition method is seriously disturbed in the actual complex scene (such as cross-viewing Angle), and can not achieve ideal results. Some researchers[13-15] use matrix transformation to normalize gait images from different perspectives to a specific perspective, and then carry out identity authentication under this perspective. However, information loss in the conversion process will lead to a decline in recognition accuracy, because this method is usually affected by changes in perspective, changes in objects and other factors. As a result, some effective information is lost during the walking process, which will reduce the

recognition rate. To solve the above problems, this paper proposes a convolutional neural network-based LFE-MGPA network model, in which Specific Proportional Segmentation Convolution (SConv) is proposed. The input pedestrian contour map is segmented according to a specific ratio (1:2:1:2), and on this basis, MGPA module in GaitSet is integrated to construct a joint loss function, and experiments are conducted on the open gait dataset CASIA-B[16]. The results show that, The recognition accuracy of the proposed method is superior to the existing mainstream algorithms in the cross-view and multi-walk state.

2. Textual Method

2.1. LFE-MGPA network model architecture

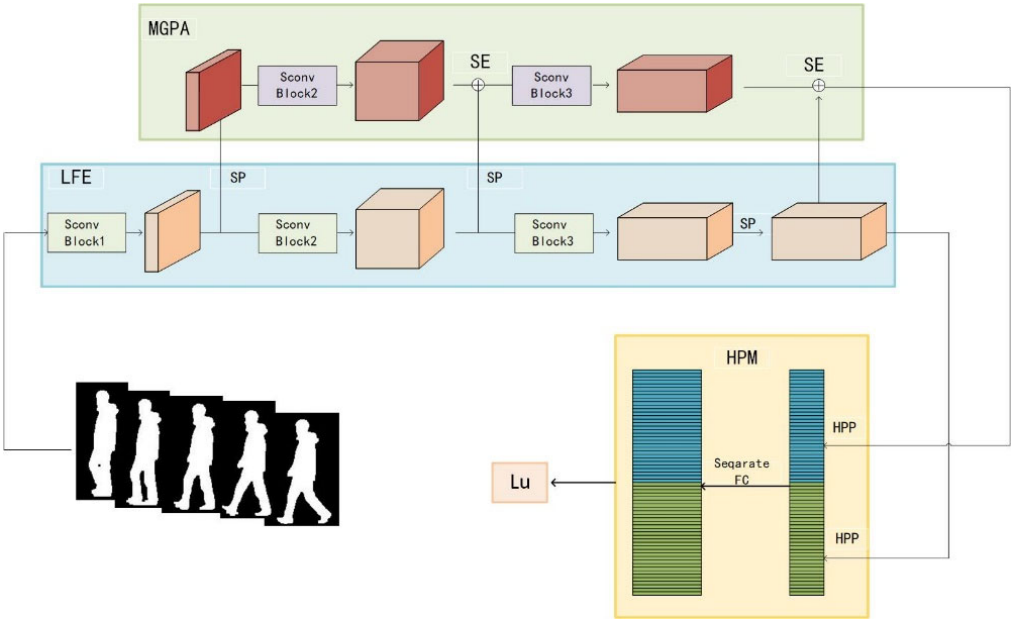


Figure 1. LFE-MGPA network overall frame diagram

Convolutional neural networks can enhance the fitting and generalization ability by learning complex samples, so this paper proposes the LFE-MGPA model based on neural networks, and its overall frame diagram is shown in Figure 1.

The network consists of three parts: the first part is the Local Feature Extraction which contains multiple Specific Proportional Segmentation Convolution (SConv); The second part is MGPA (Multilayer Global Pipeline), which is used to collect the collective-level characteristics of different layers. The third part is the loss function. Firstly, we take the unordered gait sequence as input, and use SP (Set Pooling) to aggregate the frame level features of different LFE layers into the set level features. Secondly, LFE and MGPA are subsampled respectively, and feature fusion is carried out after subsampling. Then, the feature mapping is carried out by Horizontal Pyramid Matching. Finally, we use the joint loss function to train the model. SP and LFE are described in detail in sections 2.2 and 2.3, respectively.

2.2. MGP with Attention（MGPA）

Generally speaking, in convolutional networks, the receptive fields of different convolutional layers are different. As the layer deepens, the receptive field becomes larger. This also means that after the input image passes through the shallow convolution layer, each pixel of the image only extracts the local information of the original image. The extracted feature details are rich, but the context information of the image is little. On the other hand, the input image can be seen in a larger range after the action of the deep convolutional layer, but the details are lost. In view of the above situation, MGPA module is used to extract the collective-level features of different layers. The structure diagram of this module is shown in Figure 2.

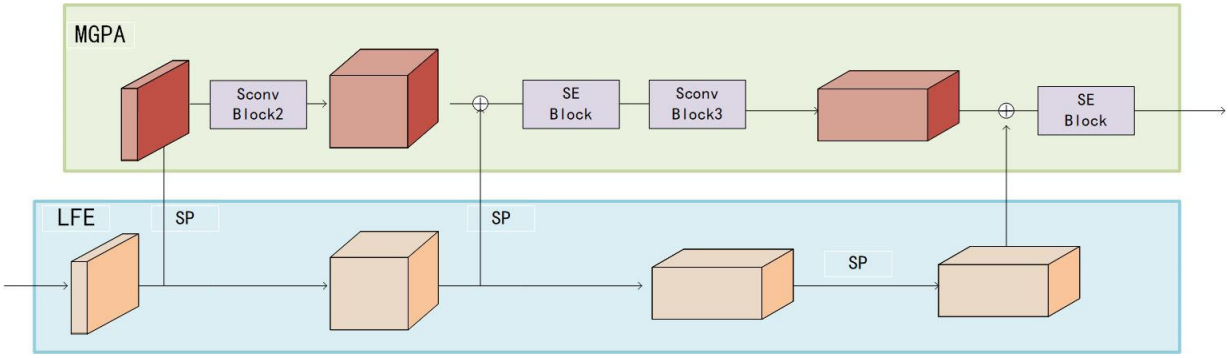


Figure 2. MGPA architecture diagram

The structure of MGPA modules is similar to that of LFE, where SConv Block2 and SConv Block3 have the same structure as SConv Block2 and SConv Block3 in LFE, but the parameters between the two modules are not shared. Firstly, the frame level features extracted by SConv Block1 in LFE are pooled (SP), and the result is input as the first layer. Secondly, MGPA aggregates frame-level features extracted from SConv Block2 in LFE into set-level features through SP operation. The features of the set level are fused with those of the upper layer. Finally, the convolution feature between channels is extracted by SE Block to correct the original output. The third layer is the same as the second layer structure with SP, feature fusion and SE Block respectively.

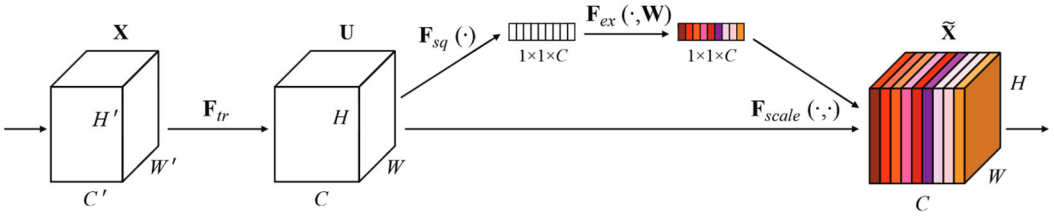


Figure 3. SE Block Overall structure

The set level features aggregated by SP operation are fused with the features of the previous layer, and the fused features are represented by X as input. Suppose $V=[v_1,v_2,...,v_C]$ represents the learned set of filters, where v_C represents the parameter of the c filter. F_{tr} is a function transformation, X through F_{tr} output $U=[u_1,u_2,...,u_C]$. The dimension of X is $H'\times W'\times C'$, and the dimension of U is $H\times W\times C$. The formula for F_{tr} is as follows: (1). Where * represents convolution, $v_c=[v_c^1,v_c^2,...,v_c^{C'}]$, $X=[x^1,x^2,...,x^{C'}]$, and the offset term is omitted here for convenience. v_c^s is a 2D space kernel, and v_c represents some channel acting on the corresponding channel in X.

$$\mathrm{u}_c=v_c*X=\sum_{s=1}^{C'}v_c^s*x^s \tag{1}$$

SE Block proposes a branching structure with two parts: Squeeze and Excitation. The Squeeze operation corresponds to the function in the figure. In order to make full use of the relationship between channels, each channel is compressed, and its formula is shown as equation (2). Where is the c channel in feature U, representing the c vector after compression. It is not difficult to see from the formula that each feature mapping channel with size H×W is compressed into a pixel point. Thus, a feature map of size H×W×C, after passing through, outputs a vector of 1×1×C.

$$z_c=F_{sq}(u_c)=\frac{1}{H\times\mathrm{W}}\sum_{\mathrm{i}=1}^{H}\sum_{j=1}^{W}u_c(i,j) \tag{2}$$

The Excitation operation corresponds to the F_{ex} function in the figure. After the channel features are compressed by Squeeze operation, the Excitation operation is needed to capture the dependency between channels. In view of the above, the function needs to satisfy two conditions: First, the function must be able to learn nonlinear relationships between channels; Second, the function must be able to learn mutually exclusive relationships between channels. The function formula is shown in equation (3). Where δ represents the ReLu function, W_1 and W_2 are the weight matrix, $W_1\in R^{(C/r)\times C}$, $W_2\in R^{C\times(C/r)}$.In simple terms, Excitation is the two fully connected layers. The first layer is the dimensionality reduction layer containing parameters, transforming a 1×1×C vector into a 1×1×C/r vector. The excitation layer is then connected to the ReLu. The second layer is an ascending layer containing parameters, which restores the 1×1×C/r vector to 1×1×C vector, followed by the activation layer of Sigmoid.

$$s=F_{ex}(z,W)=\sigma(g(z,W))=\sigma(W_2\delta(W_1z)) \tag{3}$$

The results obtained by F_{sq} and F_{ex} functions need to be merged with the original feature U, which corresponds to the F_{scale} function in Figure 3, whose formula is shown in equation (4). Let $\tilde{X}=[\tilde{x}_1,\tilde{x}_2,...,\tilde{x}_C]$ be the final output of the SE Block and s be the weight vector of the channel. $\tilde{x}_c$ represents the output of channel c. F_{scale} represents the channel-level product of feature u_c and scalar s_c. In popular terms, it is to multiply each element in s with the channel corresponding to the feature U by the number of channels.

$$\tilde{x}_c=F_{scale}(u_c,s_c)=s_c\cdot u_c \tag{4}$$

In this study, in order to better control the size of feature maps, Adaptive pooling [] is adopted to achieve Squeeze operation. The core idea of this strategy is that, for inputs of different sizes, Adaptive pooling can automatically average

each position, so as to achieve accurate processing of feature maps. Compared with the traditional average pooling method, this technique does not need to specify the size of the pooling kernel, but adaptively adjusts according to the size of the required output, and the output size can be manually specified according to the demand. When Adaptive pooling is adopted, the specific application scheme of SE Block in this paper is shown in Figure 4.

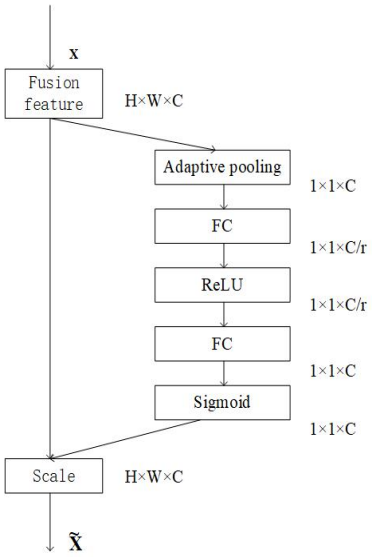


Figure 4. SE usage scheme

LFE extracts the gait features of independent frames, and SP aggregates these feature information into the collection-level information, the formula is z=S (V), where z represents the collection-level feature, V represents the frame-level feature, and the formula is $V=\{v_j \mid j=1,2,...,n\}$, n represents the number of gait frames contained in the current set, and v_j represents the feature map of frame j. It should be noted that the length of a video is not fixed, so the input set contains no fixed number of frames, which means that the function S should be able to take any set of frames. Under this premise, two statistical functions, max and median, are considered, and a joint function composed of max and median is constructed as an instance of function S, as shown in equation (5). In this article, we will compare these functions in Section 3 to get more useful examples.

$$S_{union}=1\_1C(cat(\max, median)) \tag{5}$$

Where cat represents the connection of the channel dimension, max statistical function and median statistical function are applied to the set dimension, and the 1×1 convolution layer represented by 1_1C can learn appropriate weights to fuse the information extracted by the two statistical functions.

2.3. Local Feature Extractor (LFE)

The frame diagram of the local feature extractor is shown in Figure 5. The input is a set of disordered gait profiles. LFE extracts the gait features of individual frames from each profile, and then aggregates the frame level features into set level features through SP operation.

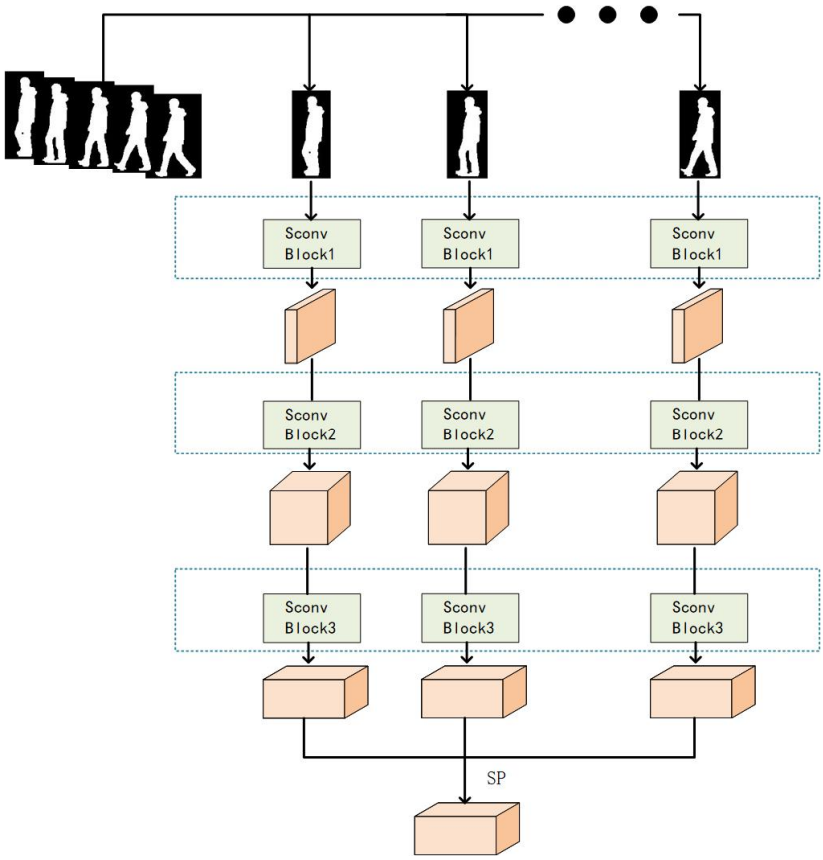
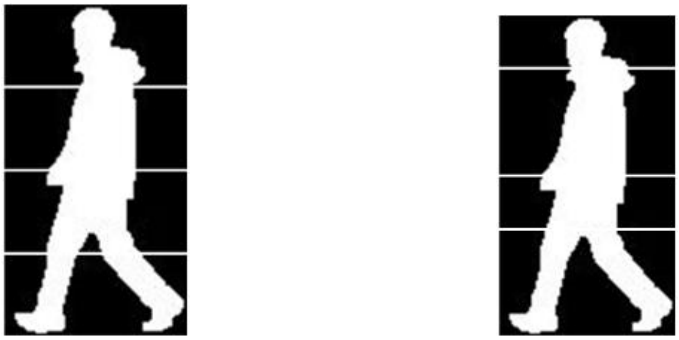


Figure 5. LFE structure diagram

The Focal Convolution Layer (FConv) in the GaitPart model divides the feature map horizontally, as shown in Figure 6(a). Inspired by his ideas, this paper proposes SConv, which splits the input gait feature map horizontally in a specific proportion, as shown in Figure 6 (b).



(a) Average level segmentation (b) Specific proportional segmentation

Figure 6. Gait silhouette segmentation method

When people walk, the movements of different parts of the body show obvious differences. For example, the movement of the head and the center of gravity of the human body is relatively small, while the arms and legs have a large swing amplitude, showing a periodic pendulum movement. This difference is of great significance in gait recognition. Through detailed analysis and modeling of the movement characteristics of these different parts, individuals' gait patterns can be identified more accurately. Among them, the movement pattern of the head and the center of gravity of the human body may be more suitable for recognizing specific identity information or emotional states, while the swing of the arms and legs is more conducive to judging the stability and speed of gait parameters. Therefore, the in-depth study of the motion characteristics of various parts of the human body and the correlation between these characteristics and gait recognition will help to improve the accuracy and applicability of gait recognition technology.

In view of the above, this paper divides the gait feature map horizontally into four parts: head, trunk, center of gravity and lower limb, and sets the ratio of these four parts as 1:2:1:2.

The LFE contains three SConvs, each with the same structure, as shown in Figure 7. The input gait contour is divided into four parts in the height dimension, and the four parts are convolution operation respectively, and then the height dimension is spliced. Suppose $F_{SC}(input) = \{F_{SC}^{i}(input) \mid i = 1,2,3,4\}$, SConv can be expressed as formula (3-6).

$$F_{SC}(output) = cat\begin{pmatrix} 3\_3C(F_{SC}^{1}(input)) \\ 3\_3C(F_{SC}^{2}(input)) \\ 3\_3C(F_{SC}^{3}(input)) \\ 3\_3C(F_{SC}^{4}(input)) \end{pmatrix} \tag{6}$$

Where 3_3C represents a 3×3 convolution and cat represents a concatenation of height dimensions.

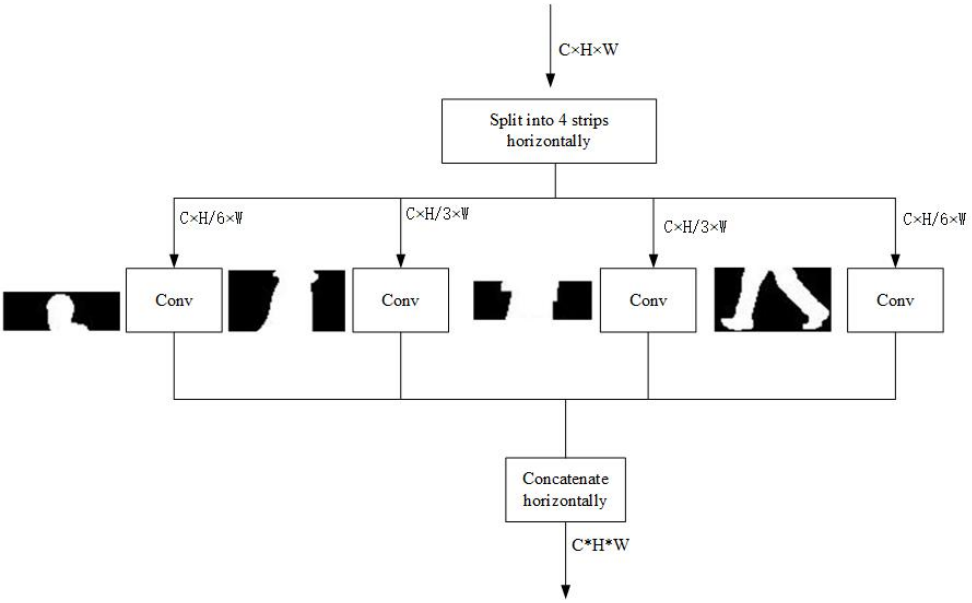


Figure 7. SConv structure diagram

Given a dataset containing N individuals' gait profiles, all profiles in one or more sequences for each individual can be viewed as a set of n profiles, expressed as $X_i = \{x_i^j \mid j = 1, 2, \ldots, n\}$, where $i \in 1, 2, \ldots, N$. With X as the input, the characteristics of the LFE output are expressed by $F_L(out)$, which can be expressed as formula (3-7).

$$F_L(out) = S(F_S(x_i^1), F_S(x_i^2), \cdots, F_S(x_i^n)) \tag{7}$$

Where F_S is a convolutional network consisting of three SConvs and two pooling operations designed to extract frame-level features from each gait profile. S is used to map a set of frame-level features to a set level feature based on Set Pooling (SP).

2.4. Joint loss function (L_U)

The joint Loss function used in this paper is derived from the combination of Cross Entropy Loss and Center Loss.

Cross entropy is an effective measure used to assess the degree of difference between the true probability distribution and the predicted probability distribution. It quantifies the similarity or difference between two distributions by calculating the amount of cross-information between them, thereby helping to understand the predictive effect of the model. In machine learning and deep learning, cross entropy is often used as a loss function to optimize the training process of the model so that it better fits the real data distribution. The formula is as follows:

$$L_{ce} = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_{ij}) \log(q(x_{ij})) \tag{8}$$

Where m is the number of samples, n is the number of classes, and $p(x_{ij})$ is used to indicate whether sample i belongs to class j in the real distribution. If sample i belongs to class j in the true distribution, then $p(x_{ij}) = 1$; If not, then $p(x_{ij}) = 0$. $q(x_{ij})$ is used to represent the probability that sample i belongs to class j in the prediction distribution.

The cross-entropy loss function adopts the inter-class competition mechanism and is good at learning inter-class information, while the central loss function is dedicated to constraining intra-class compactness. Its formula is as follows:

$$L_C = \frac{1}{2}\sum_{i=0}^{n} \| x_i - c_{y_i} \|_2^2 \tag{9}$$

Where, n represents the number of training samples, x_i represents the output value of the y_i pedestrian passing through the network, and c_{y_i} represents the feature center of the y_i category. The Center Loss function has no obvious improvement on the model with a small number of classifications, and it is more suitable for the model with a large number of classifications.

After using this loss function, the samples of each class become more aggregated, so the distance between classes increases. With the increasing loss weight value, the class spacing becomes larger and larger. It achieves the effect of increasing the class distance and narrowing the distance within the class.

The specific formula of the joint loss function L_U, which consists of the cross entropy loss function and the central loss function, is shown in equation (10), where α is the weight parameter.

$$L_U = \alpha L_{ce} + L_C \tag{10}$$

3. Experimental Results and Analysis

3.1. CASIA-B data set

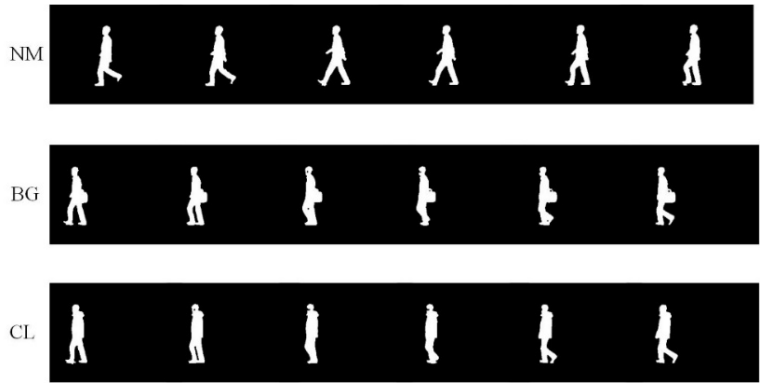


Figure 8. Silhouette at 90° of partial gait

The CASIA-B dataset is a large-scale, multi-view gait library provided by the Institute of Automation, Chinese Academy of Sciences. It contains raw video data and contours of 124 subjects. Each subject had 11 viewing angles and three walking conditions, namely normal walking (NM), backpack walking (BG), and coat walking (CL). Each subject had 6 video sequences in NM state, 2 video sequences in BG state, and 2 video sequences in CL state. Therefore, each subject has 11×(6+2+2) = 110 gait sequences. The partial contours of the three walking states of the CASIA-B dataset are shown in Figure 8.

3.2. Experimental setup

In this paper, the samples in CASIA-B dataset were preprocessed, and the gait profile with size 128×88 was obtained after normalization. During training, 30 consecutive gait contour diagrams were randomly selected from the sequence as input, the learning rate was set to 1e-4, and the batch size was set as: p×k, so that p=8, k=16, where p represented the number of classes, and k represented the number of training samples for each class in the batch.

In the experiment, the large sample (LT) training method was used in this paper. The data of the first 74 subjects were used for training, and the data of the remaining 50 subjects were used for testing. In the test set, the first 4 gait sequences in the gait data of 50 subjects in the NM walking state were taken as the registration set, expressed as: NM#01-NM#04. The remaining sequences constitute the verification set, that is, the verification set includes the remaining 2 gait sequences of the subject in the NM walking state, which are expressed as: NM#05, NM#06; The two gait sequences in the BG walking state are represented as BG#01, BG#02; The two gait sequences in CL walking state are represented as: CL#01, CL#02. Each gait sequence mentioned above contains gait data from 11 angles, and the specific Settings of the registration set are shown in Figure 9.

Figure 9. Gallery Settings

3.3. Contrast test

Table 3-1, 3-2, and 3-3 show the recognition results between the proposed gait model and four existing gait models (SRN[17], GaitSet, CapsNet, and GaitPart) in NM, BG, and CL walking states on the CASIA-B data set under the large-sample (LT) training mode. It should be noted that in the table, 0°, 18°... The data in the 162° and 180° columns are the average recognition rate excluding the same viewing Angle. For example, the recognition rate of the viewing Angle 108° is the average recognition rate of the other 10 viewing angles except 108°. The last column shows the average recognition rate for 11 different viewing angles.

Through observation, it is found that the proposed method has no obvious advantage compared with other methods when the pedestrian is in NM state. The reason may be that the information contained in the input gait contour diagram of the proposed method is limited, or the proposed method does not extract enough useful information. As shown in Table 3-2, when pedestrians are in the BG state, the proposed method has obvious advantages compared with GaitSet, CapsNet and GaitPart, while the performance is not significantly improved compared with SRN. According to Table 3-3, when pedestrians are in CL state, the recognition rate of the proposed method reaches 81.0%. Compared with SRN, GaitSet, CapsNet and GaitPart, the average Rank-1 recognition rate has been significantly improved by 3.3, 10.7, 8.6 and 2.3 percentage points respectively. Except that the average recognition rate at 108° and 126° is slightly lower than SRN, the average recognition rate at the other nine angles is higher than the other four methods.

Table 1. The average Rank-1 accuracy of different algorithms on the CASIA-B data set at NM

Method	Recognition rate under different perspectives /（%）											
	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Average
SRN	94.4	99.3	99.4	98.7	96.8	96.8	97.5	98.5	99.5	98.8	92.3	97.45
GaitSet	91.1	99.0	99.9	97.8	95.1	94.5	96.1	98.3	99.2	98.1	88.0	96.10
CapsNet	91.8	98.3	99.0	98.0	94.1	92.8	96.3	98.1	98.4	96.2	89.2	95.65
GaitPart	94.1	98.6	99.3	98.5	94.0	92.3	95.9	98.4	99.2	97.8	90.4	96.22
Ours	94.8	99.2	99.8	98.8	96.8	96.9	97.1	98.6	99.4	98.7	92.7	97.53

Table 2. The average Rank-1 accuracy of different algorithms on the CASIA-B data set at BG

Method	Recognition rate under different perspectives /（%）											
	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Average
SRN	91.5	97.4	98.4	97.1	92.2	89.7	93.1	96.2	97.5	96.5	88.0	94.32
GaitSet	86.7	94.2	95.7	93.4	88.9	85.5	89.0	91.7	94.5	95.9	83.3	90.80
CapsNet	87.3	93.7	94.8	93.1	88.1	84.5	88.8	93.5	96.3	93.3	83.9	90.66
GaitPart	89.1	94.8	96.7	95.1	88.3	94.9	89.0	93.5	96.1	93.8	85.8	92.46
Ours	92.9	98.5	99.7	97.0	91.9	86.1	90.0	96.6	99.7	98.7	88.8	94.54

Table 3. The average Rank-1 accuracy of different algorithms on the CASIA-B data set at CL

Method	Recognition rate under different perspectives /（%）											
	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Average
SRN	69.2	82.5	84.0	81.0	78.6	76.3	78.6	82.8	80.5	76.8	64.7	77.72
GaitSet	59.5	75.0	78.3	74.6	71.4	71.3	70.8	74.1	74.6	69.4	54.1	70.28
CapsNet	63.4	77.3	80.1	79.4	72.4	69.8	71.2	73.8	75.5	71.7	62.0	72.41
GaitPart	70.7	85.5	86.9	83.3	77.1	72.5	76.9	82.2	83.8	80.2	66.5	78.69
Ours	74.8	88.5	90.8	84.9	79.1	77.6	78.5	82.4	84.9	80.9	69.3	81.06

By comprehensively observing these three tables, it is not difficult to find that the proposed method has better robustness compared with SRN, GaitSet, CapsNet and GaitPart when the walking state of pedestrians changes. Especially in the CL state, the method presented in this paper shows greater advantages.

3.4. Ablation experiment

3.4.1. Local feature extractor (LFE) ablation experiment

In order to verify the advantages of SConv over FConv and find out the function most suitable for SP, ablation experiments were conducted on the CASIA-B dataset, and the experimental results were shown in Table 4. √ means use the current convolution or function. Conv1 in Conv represents ordinary convolution; FConv stands for focal convolution in GaitPart; SConv represents a specific proportional partition convolution proposed in this paper. Max, Median, and Union in Set Poling represent the Max function, Median function, and union function respectively. By comparing the results of the first three groups of experiments, it is found that the SConv proposed in this paper can extract more useful gait features and obtain better experimental results, indicating the effectiveness of SConv. After comparing the results of the first Set of experiments with the latter two sets, it is found that neither the Median function nor the joint function composed of 1×1 convolution has a better effect than the Max function. Therefore, Max function is finally selected as the example of Set Pooling.

Table 4. Ablation experiments of LFE on CASIA-B dataset

Conv			Set Pooling			NM	BG	CL
Conv1	FConv	SConv	Max	Median	Union			
		√	√			97.5	94.5	81.0
	√		√			96.1	90.9	77.3
√			√			96.0	89.8	70.8
		√			√	96.9	93.7	80.3
		√		√		96.8	91.9	78.7

3.4.2. Global model ablation experiment

In order to verify the effectiveness of each component in this method, ablation experiments were conducted on these components on the CASIA-B dataset, and the experimental results are shown in Table 5. √ indicates the use of this component. In the table, Conv stands for common convolution, SConv stands for the specific proportional segmentation convolution proposed in this paper, MGP stands for Multilayer Global Pipeline, SE stands for SE Block, MGP+SE stands for MGPA module. By dividing the six groups of experimental results into the first three groups and the last three groups for analysis, it can be seen that whether using Conv or SConv, the recognition accuracy of our model is slightly improved by using MGP alone. However, the performance of MGP module is not good in CL walking state. With the addition of SE to MGP, the performance of our models has been significantly improved. Comparing the first and fourth sets of data, the second and fifth sets of data, and the third and sixth sets of data, it is obvious that SConv is more effective than Conv in this model.

Table 5. Ablation experiments of LFE-MGPA on CASIA-B dataset

Conv	SConv	MGP	SE	NM	BG	CL
√				95.3	88.4	73.8
√		√		95.3	88.5	73.9
√		√	√	96.2	91.2	75.3
	√			96.0	90.3	79.6
	√	√		96.2	91.0	79.5
	√	√	√	97.5	94.5	81.1

4. Summary

To solve the problem that gait recognition is greatly affected by external factors such as wearing and carrying, the LFE-MGPA network model is proposed in this paper. Firstly, an LFE containing multiple SConvs is constructed in the model. The SConv splits the input gait contour to extract fine-grained gait features. Secondly, the MGP module is integrated, and the channel attention mechanism is added to the module. Finally, a joint loss function is constructed to improve the recognition rate. Experimental results on the CASIA-B dataset show that compared with the current mainstream

algorithms, the proposed method achieves a higher recognition rate, especially in the disturbed state, the recognition rate is significantly improved, indicating that the method has good robustness in complex scenarios. In spite of this, the influence of the shape change on the gait profile can not be ignored. In the next work, the fusion with model-based methods such as human posture will be considered to obtain a more accurate representation of gait.

References

[1] Zeng F, Hu S, Xiao, K. Research on partial fingerprint recognition algorithm based on deep learning[J]. Neural computing & applications,2019,31(9):4789-4798.

[2] Sepasmoghaddam A , Pereira F M , Correia P L .Face Recognition: A Novel Multi-Level Taxonomy based Survey[J].IET Biometrics, 2019, 9(2).

[3] Sun Y, Wang X, Tang X. Deep Learning Face Representation from Predicting 10,000 Classes[J]. 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014: 1891–1898.

[4] Wang K , Kumar A .Periocular-Assisted Multi-Feature Collaboration for Dynamic Iris Recognition[J].IEEE Transactions on Information Forensics and Security, 2021, 16:866-879.

[5] Han J, Bhanu B. Individual recognition using gait energy image[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(2): 316-322.

[6] Wu Z, Huang Y, Wang L, et al. A Comprehensive Study on Cross-View Gait Based Human Identification with Deep CNNs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(2): 209-226.

[7] Wang C, Zhang J, Wang L, et al. Human identification using temporal information preserving gait template[J]. IEEE transactions on pattern analysis and machine intelligence, 2012, 34(11): 2164-2176.

[8] Alireza S, Ali E. Deep Gait Recognition: A Survey[J]. IEEE transactions on pattern analysis and machine intelligence, 2022, 45(1): 264-284.

[9] Shiraga K, Makihara Y, Muramatsu D, et al. GEINet: View-invariant gait recognition using a convolutional neural network[C]. Proceedings of the 2016 International Conference on Biometrics (ICB), 2016: 1-8.

[10] Fan C, Peng Y, Cao C, et al. GaitPart: Temporal Part-Based Model for Gait Recognition[J]. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2020: 14213-14221.

[11] Sepas-Moghaddam A, Ghorbani S, Troje NF, et al. Gait recognition using multi-scale partial representation transformation with capsules[C]. Proceedings of the 2020 25th international conference on pattern recognition (ICPR), 2021, 8045-8052.

[12] Chao H, Wang K, He Y, et al. GaitSet: Cross-view Gait Recognition through Utilizing Gait as a Deep Set[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021: 1-12.

[13] Liu N, Lu J, Tan Y. Joint subspace learning for view-invariant gait recognition[J]. IEEE Signal Process. Lett, 2011, 18(7): 431-434.

[14] Jean F, Albu AB, Bergevin R. Towards view-invariant gait modeling: Computing view-normalized body part trajectories[J]. Pattern Recognition, 2009, 42(11): 2936-2949.

[15] Goffredo M, Bouchrika I, Carter JN, et al. Self-calibrating view-invariant gait biometrics[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2010, 40 (4): 997-1008.

[16] Yu S, Tan D, Tan T. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition[C]. Proceedings of the 18th international conference on pattern recognition (ICPR'06), 2006: 441-444.

[17] Hou S , Liu X , Cao C ,et al.Set Residual Network for Silhouette-Based Gait Recognition[J].IEEE Transactions on Biometrics, Behavior, and Identity Science, 2021(3-3).DOI:10.1109/TBIOM.2021.3074963.