A Gait Recognition Method Based on the Combination of Human Body Posture and Human Body Contour

Yue Ma¹, Chenghao Wei²* and Hao Long²

¹ Key Laboratory of Impression Evidence Examination and Identification Technology (Criminal Investigation Police University of China), Ministry of Public Security, China
² Big Data Institute, College of Computer Science and Software Engineering, Shenzhen University, China
Email: chenghao.wei@szu.edu.cn

Abstract. Aiming at the issues that affect the gait recognition, such as outfit changes and carry-on-objects during gait recognition, this paper proposes a gait recognition method based on the improved GaitSet network, which relies on the fusion of human posture and human contour. This method introduces key points of human posture, and improves the precision of human contour extraction by introducing the key points of human posture. At the same time, the human posture map composed of the key points of human posture is used as the synchronous attribute of the contour map to extract gait features. Because key points focus on the inherent walking characteristics of human body, they are not affected by the external information such as clothing and carrying objects. Combined with the rich gait change attributes of human contour, it can effectively improve the accuracy and robustness of the gait recognition model. In scenes of human body wearing coats, our experimental results show that the accuracy of the proposed method in CASIA-B gait data set and self-built database can reach over 73.5% accuracy.

1. Introduction
At present, a variety of biometrics such as the human face, human palm-print, and so on have been widely used in identity authentication. These biometric authentication methods need the active cooperation of the tested object to obtain basic information. Although the existing face recognition technology has been applied in many scenes, its recognizable distance is very short, which can only reach to 7 to 8 meters. It is easy to be blocked and affected by camera focus. Iris recognition requires the camera to focus on the human eye and is easily affected by contact lens products. Fingerprint and palm-print recognition also depends on equipment and distance limitation. Gait recognition is one of the fast rising applications in the field of biometrics in the future. It can collect the target information from a long distance without the cooperation of the subject under any conditions [1].

There are two kinds of gait recognition methods. The first method is to treat gait as an image, that is, to compress all gait contour images into a gait energy map [2], and then to treat gait as an image matching problem. This method ignores the temporal information in gait and cannot describe the spatial information. The second method is to take gait as an image sequence, considering the extraction of human feature directly from the silhouette, and use LSTM [3] or 3D-CNN [4] or other methods to model the spatiotemporal information in gait. Right now, although gait as an image sequence cannot achieve 100% recognition accuracy, its recognition accuracy has reached the application requirements in specific scenarios, e.g., information retrieval of violent and terrorist criminals.
Although the current gait recognition algorithm based on deep learning has achieved some results, gait recognition faces many challenges now. At first, the recognition accuracy is low in multi-view condition [5], because the general positions of cameras are fixed, when pedestrians enter into the camera acquisition area from different directions, it will cause the problem of different target attitude with multi-view difficulty of identification. To solve this problem, Liu et al. [6] proposed to represent samples from different perspectives as a linear combination in the original corresponding view and extract feature representation coefficients for final classification. Kusakunniran [7] proposed a method based on viewpoint invariant gait feature and it realised gait similarity measurement based on shape distance. Makihar et al. [8] proposed a model using frequency-domain characteristics and view conversion. On this basis, Kusakunniran et al. [9] further applied linear discriminant analysis (LDA) to simplify the calculation. Compared with the former two methods, constructing the perspective conversion model can achieve a higher accuracy recognition effect at a lower cost. Secondly, under the scene of human outfit, the recognition rate is low. The human outfit will change the contour of the human body to be a large extent. Because the criminal suspects collected in the crime video often wrap the crime tools in their coats, the extracted gait features belong to the gait features of abnormal walking, which increases difficulty to the gait recognition work. When the clothing is thick or too long, it will block the posture of the human body and affect the recognition effect of gait. Thirdly, the recognition rate of human with carried objects is low. In the process of foreground image separation, the objects carried by human are likely to be extracted as a part of human body, which is actually very common in various types of crime scene. All kinds of criminal suspects often put the crime tools into handbags or backpacks, thus affecting the accuracy of gait features. In addition, the current human body contour extraction scheme in most papers uses the background subtraction method based on the mixed Gaussian model [10]. This method can obtain the human body contour under the face of a fixed single background video condition. In actual scenes, the background image is usually more complex, and the background subtraction method usually requires more post-processing, which affects the application and promotion of gait recognition technology.

In this paper, the method of human body segmentation based on the key points of human body posture is used to extract the contour of the human body. This method uses the convolution neural network technology in the excellent performance of object segmentation and combines the positioning information of the key points of human body posture so that the robustness and accuracy of human body segmentation are better. In order to solve the problems of outfit changes and carry objects that affect gait recognition, a gait recognition method based on the fusion of human posture and human contour is proposed. By introducing the key points of human body posture, this strategy not only improves the precision of human body contour extraction, but also extracts the gait features from the human body posture map formed by the key points of human body posture as the synchronous attributes of the contour map. This strategy cannot be affected by external information, such as clothing and carrying objects. Combined with the rich gait change attributes of the human figure contour map, it greatly improves the accuracy and robustness of gait recognition.

2. Human Silhouette and Human Pose

As shown in figure 1, the key points of the human body usually correspond to joints with a certain degree of freedom on the human body, such as neck, shoulder, elbow, wrist, waist, knee, ankle, etc.

Human pose estimation is also called human keypoint detection. Researchers have developed a variety of human pose estimation algorithms. In the early stage, human pose estimation is usually aimed at a single person scene, but in the real situation, the image often contains more than one person. The current multi-person attitude estimation algorithms usually include the top-down method and the low-up method. From the top to the bottom, the strategy is to detect multiple target people at first, and then to estimate each person’s attitude. From the bottom up, the key point is detected first, and then the target person of the key point is determined. Currently, the multi-human body pose estimation algorithm based on deep learning mainly contains bottom-up methods and top-down methods. Methods include
OpenPose [11] and DeepCut [12], while the second one contains RMPE (AlphaPose) [13] and Mask RCNN [14].

Figure 1. Schematic diagram of key points of human posture. Note: 右目: right eye; 左耳: left ear; 右肩: right shoulder;右肘: right elbow;右手首: the head of the right hand; 右腰: right waist; 右膝: right knee; 右足首: head of right foot; 左目: left eye; 左耳: left ear; 左肩: left shoulder;左肘: left elbow;左手首: the head of the left hand;左腰: left waist;左膝: left knee;左足首: head of left foot.

In this paper, the top-down RMPE algorithm is used to extract the key points of human posture. This algorithm is significantly superior to other existing multi-person human posture estimation methods in terms of accuracy and efficiency. The effect of comparing the human figure pose with the original picture is shown in figure 2.

2.1. Human Contour Drawing
In the field of computer vision, human contour extraction can be regarded as an instance segmentation problem. The current mainstream instance segmentation frameworks, such as Deeplabv3+ [15], Mask
R-CNN, etc., are based on strong object detection methods. These methods all adhere to a basic rule, where these methods first generate a large number of candidate regions from the image and then use non-maximum suppression (NMS) algorithm to propose those repeated regions from these thousands of candidate regions. However, when there are two highly overlapped objects in the image, NMS will consider one of the bounding boxes as a duplicate candidate area and delete it, which means that the case segmentation frameworks based on object detection cannot be used. It cannot deal with the situation that the objects in the image are highly overlapped. The detection of human body is a very special category case among all kinds of object detection. Compared with the bounding box, the key points of human pose figure are more suitable to distinguish highly overlapped instances in an image, because it contains accurate positioning information. In the field of computer vision, human skeleton posture has a rich data annotation.

Pose2Seg [15] is a human instance segmentation network based on human bone pose estimation. Its input is a human image and the corresponding key points of human pose. The algorithm first uses a feature extraction network to extract the features of the image, and then uses Affine-Align to separate extract the human instance. The image data of key regions are transformed to a fixed scale, and skeleton features are introduced to improve the segmentation performance. In this paper, the Pose2Seg [15] network model based on bone key points is used to extract the human body contour map. As shown in figure 3, the original gait images and their corresponding human body contour images are provided.

![Figure 3. Human body contour images.](image)

2.2. Two-Channel Gait Feature Map

The above-mentioned key points of human pose images and human body contour images are all binary images, which are the same size. In this paper, the binary key points of human pose image and the human body contour image are combined to form a two-channel gait feature model as shown in figure 4. Two images are merged into a dual-channel gait feature model. Assuming that the original image size is 240 * 240 three-channel image, we can initially generate a 240 * 240 two-channel target matrix with 0 values. Then, we copy the corresponding values of the binary human body contour figure to the first channel matrix. After that, we will assign the corresponding values of the human body posture figure to the matrix of the second channel. By using such methods, the two-channel gait feature model is obtained.
3. Gait Recognition Based on Combination Feature Model

3.1. Gaitset Network Model

In 2018 CVPR, research team of Fudan University proposed Gaitset network model [16]. Based on the human visual perception of gait, the relationships of gait sequence can be recognised easily. Instead of modelling the relationship of the gait contour sequence, this research regards the gait contour sequence as an image set without relationship and lets the depth neural network optimize itself to extract and utilize this relationship.

For a $N$ pedestrian dataset $y_i, i \in 1, 2, ..., N$, we assume that a person’s gait profile sequence belongs to the distribution $P_i$. Therefore, all contours in the sequences of a pedestrian can be regarded as $n$ collections of silhouettes $X_i$, which can be defined as:

$$X_i = \{x^j_i \mid j = 1, 2, ..., n\} \quad x^j_i \in P_i \quad (1)$$

Gaitset model uses three steps to solve the gait recognition task, which can be defined as:

$$f_i = H\left(G(F(X_i))\right) \quad (2)$$

Here $F$ is a convolutional neural network used to extract frame-level features. $G$ is a permutation invariant function, which is used to map a set of frame-level features to set-level features and latter can be implemented by set pooling. $H$ is used to extract discriminative discriminants from set-level features and is implemented with a horizontal pyramid map (HPM) structure. Based on the original Gaitset network model, the network structure of $F$ module is modified, so that it can directly process the sequence of two channel fusion gait. The improved GaitSet network structure is shown in figure 5.

![Figure 5. Improved GaitSet network structure.](image-url)
3.2. Network Model Training and Testing
In the experiment, each frame of image needs to be transformed into a 64*64 size dual-channel image. The initial size of the network is set to 30. By using the Adam optimiser, the number of horizontal pyramid map (HPM) size is set to 5, which means s is equal to 5, the output of the improved GaitSet network is \(2 \times \sum_{i=1}^{5} 2^i-1\), which is a 62 dimension vector, the batch of all triplet loss is set to 0.2. A p*k sampling strategy is used, here p is the number of characters, we set to 8, k is the number of samples for each person in a batch that is set to 16. Each sample here contains multiple contour maps collected from the sequence. In the frame-level feature extraction network of figure 5, the number of channels in the C1 and C2 layers is set to 32, the number of channels in the C3 and C4 layers is set to 64, and the number of channels in the C5 and C6 layers is set to 128. The learning rate is set to 1e-4, The number of iterations is set to 80,000 times.

For the testing of our model, we divide all images set into training set and testing set firstly. We use the training set to generate the training model. Then, we divide the testing set into two parts, the first part is used as reference while the second one is used as testing. All feature vectors were generated by using the training model with those images in testing set. We will calculate the Euclidean distance between the feature vector of the test sample image and all the feature vectors of the test images. Then, we use Rank-1 rule as the final recognition rate.

4. Experimental Results and Analysis

4.1. Experimental Platform and Gait Data Set
Table 1 shows the hardware and software of our experiment.

| CPU CORE | RAM  | GPU number and model | Cuda version | OS version | Deep learning framework |
|----------|------|----------------------|--------------|------------|------------------------|
| 48       | 256G | 10 RTX2080Ti         | 10.1         | Ubuntu 16.04 | pytorch                |

The CASIA-B [17] gait database from the Chinese Academy of Sciences is used. The database has 124 pedestrians, and each pedestrian includes 11 viewing angles from 0 degrees to 180 degrees at intervals of 18 degrees. Three walking scenes includes wearing coat (cl#01- #02), backpack (bg#01- #02) and normal walking (nm #01- #06). There are 11 angle changes in each walking scene, therefore, there are 124 × 10 × 11 × 80 = 1091200 pictures. This experiment also uses a self-built gait database, which includes 100 pedestrians, each pedestrian includes 9 scenes, from 0 degrees to 160 degrees at 20 degree intervals, their outfits are more complicate.

For evaluating the effectiveness of the proposed algorithm, the above two image datasets are used. At first, the training image set (train set), reference image set (gallery set), and test image set (probe) are all obtained from CASIA-B data set. More specifically, all the gait data of pedestrians (0-73) are used for the training set. The gait data of pedestrians (74-123) are divided into 2 groups, nm#02, bg #02 and cl#02 walking scene data are used as the network’s reference image set, mm#01, bg#01 and cl#01 walking scene data are used as the network’s test image set. Secondly, all the gait data of CASIA-B are used for training the model. For testing the model, we use 5 perspective data of 0-80 degrees of all pedestrians from the self-built gait library as the reference image set, and the total images of 100-160 degrees of all pedestrians from the self-built library are used as the test image set. The description of data is shown in table 2.

4.2. Experiments and Results
In order to verify the effectiveness of the method, the experiment uses the average rank-1 accuracy rate to evaluate the results. The average rank-1 accuracy rate is obtained by averaging the results of the test image set against the reference image set with 10 runs.
Table 2. Experimental data based on CASIA-B library and self-built gait library.

| Parameter                  | Training set | Reference image set (gallery set) | Test set                        |
|----------------------------|--------------|-----------------------------------|---------------------------------|
| Same data environment      | CASIA-B, 0-73, 74 | CASIA-B, 74-123, 50 PEOPLE nm#02, bg#02, cl#02 | CASIA-B, 74-123, 50 PEOPLE nm#01, bg#01, cl#01 |
| All gait data              | Self-built library 0-99,100 people 0-80 degrees, 5 viewing angles | CASIA-B, 74-123, 50 PEOPLE nm#01, bg#01, cl#01 |
| Cross-data environment     | CASIA-B, 0-123,124 | Self-built library 0-99,100 people 100-160 degrees, 4 viewing angles |

Table 3. Average rank-1 accuracy.

| Parameter                  | Reference image set | Test image set | Average rank-1 accuracy % |
|----------------------------|---------------------|----------------|---------------------------|
| The first test             | CASIA-B, 74-123, 50 PEOPLE nm#02(10) | CASIA-B, 74-123, 50 PEOPLE bg#02(10) | 95.581                      |
|                            | bg#02(10)           | cl#02(10)      | 89.026                     |
|                            | Self-built library 0-99,100 people 0-80 degrees, 5 viewing angles | Self-built library 0-99,100 people 100-160 degrees, 4 viewing angles | 73.198                      |
| The second test            |                      |                |                            |

For the first test, compared with the performance of the original Gaitset model [16], the test results in the normal walking scene have a certain improvement in the scene of wearing coats and backpacks, which usually affect the gait contour effect. This is mainly due to the introduction of the human posture figure as an additional attribute. The human posture figure focuses on the inherent walking characteristics, excluding the clothing and carrying objects to block the contour of pedestrians. Combined with the rich gait change attributes of the human contour figure, it effectively improves the accuracy and robustness of gait recognition.

For the second test, it still can get 73.5% rank-1 accuracy. In the self-built database, the clothing changes of the tested people are very complicate. The actual scene can be equivalent to CASIA-B’s coat-wearing scene, and the test results are very close to the coat-wearing scene in the first test. Thus, experiments prove the effectiveness of our method.

5. Conclusion
In order to solve the problems of outfit changes and carry on objects that affect human gait recognition, this paper proposes an improved Gaitset network model, which introduces a gait recognition strategy based on the combination of human posture and human contour. Such strategy improves the precision of human contour extraction by introducing the key points of human posture. At the same time, the gait feature is extracted from the human pose figure, which is composed of the key points of the human pose as the synchronous attribute of the contour figure. Because the human pose figure focuses on the inherent walking feature of the human body, it is not affected by the information such as clothing and carrying objects. Combined with the rich gait change attribute of the human profile figure, the accuracy and robustness of gait recognition are greatly improved.
Acknowledgment
Supported by Key Laboratory of Impression Evidence Examination and Identification Technology (National Police University of China), Liaoning Collaboration Innovation Center for CSLE.

References
[1] Wang K J, Ding X N and Xing X L 2019 A survey of multi-view gait recognition Acta Automatica Sinica 45 (5) 841-852.
[2] Babae M, Lin W L and Gerhard R 2019 Person identification from partial gait cycle using fully convolutional neural networks Neurocomputing 338 116-125.
[3] Liao R J, Cao C S, Garcia E B, Yu S Q and Huang Y Z 2017 Pose-based temporal-spatial network (PTSN) for gait recognition with carrying and clothing variations Chinese Conference on Biometric Recognition pp 474-483.
[4] Wolf T, Babae M and Rigoll G 2016 Multi-view gait recognition using 3D convolutional neural networks IEEE International Conference on Image Processing pp 4165-4169.
[5] Wan C S, Li W and Vir P 2018 A survey on gait recognition ACM Computing Surveys 51 (5) 11-35.
[6] Liu N N, Ji W L and Yap P T 2011 Joint subspace learning for view-invariant gait recognition IEEE Sig. Pro. Let. 18 (7) 431-434.
[7] Kusakunniran W, Wu Q, Zhang J, Ma Y and Li H 2013 A new view-invariant feature for cross view gait recognition IEEE Trans. Infor. For. and Sec. 8 (10) 1642-1653.
[8] Makihara Y, Sagawa R, Mukaiwawa Y, Echigo T and Yagi Y 2006 Gait recognition using a view transformation model in the frequency domain European Conference on Computer Vision pp 205-207.
[9] Kusakunniran W, Wu Q, Li H and Zhang J 2009 Multiple views gait recognition using view transformation model based on optimized gait energy image IEEE 12th International Conference on Computer Vision Workshops pp 1058-1064.
[10] Godbehere A B, Matsukawa A and Goldberg K 2012 Visual tracking of human visitors under variable lighting conditions for a responsive audio art installation American Control Conference pp 4305-4312.
[11] Cao Z, Tomas S, Shi E W and Yaser S 2017 Realtime multi-person 2D pose estimation using part affinity fields IEEE Conference on Computer Vision and Pattern Recognition pp 1302-1310.
[12] Leonid P, Eldar I, Si Y T, Bjoern A, Mykhaylo A, Peter G and Bernt S 2016 Deepcut: Joint subset partition and labeling for multi person pose estimation Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 4929-4937.
[13] Hao S F, Shu Q X, Yu W T and Ce W L 2017 RMPE: Regional multi-person pose estimation IEEE International Conference on Computer Vision pp 2353-2362
[14] Jiang A, Yan N, Wang F, Huang H, Zhu H and Wei B 2019 Visible image recognition of power transformer equipment based on mask R-CNN IEEE Sustainable Power and Energy Conference pp 657-661.
[15] Chollet F 2017 Xception: Deep learning with depthwise separable convolutions IEEE Conference on Computer Vision and Pattern Recognition pp 1800-1807.
[16] Chao H Q, He Y W, Zhang J P and Feng J F 2018 GaitSet: regarding gait as a set for cross-view gait Proceedings of Computer Vision and Pattern Recognition.
[17] Yu S Q, Tan D L and Tan T N 2006 A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition 18th International Conference on Pattern Recognition pp 20-24.