Multi-grid Spatial and Temporal Feature Fusion for Human Identification at a Distance

Panfeng Zhang[1]#, Zhiqiang Song[1]#, Xianglei Xing[1] *

[1] Harbin Engineering University, School of Intelligent Systems Science and Engineering

839323876@qq.com

Song34782@163.com

xingxl@hrbeu.edu.cn

Abstract. Gait, as a unique biological feature that can identify human information at a long distance, is widely used in crime prevention, forensic identification and social security because of its unique security, non-contact and difficult to camouflage. In order to describe the gait, the existing gait recognition methods are mainly divided into two types, one is to use the gait template, in which the time information is difficult to be saved, the other is to use the gait sequence, which must maintain unnecessary sequence order constraints, thus losing the flexibility of gait recognition. In this competition, we propose a multi-layer network inspired by the GaitSet[1] method, which can ignore the order of gait sequence. The spatial and temporal gait features are fused from multiple scales to obtain the deep fusion of the identity-oriented features. We have equipped the prototype network with the ResNet backbone and high-resolution data flow, to achieve good results in the CASIA-E dataset for the competition.

1 Introduction

Unlike other biometric, such as face, fingerprint, and iris, gait is a unique biometric that can be recognized at a distance without subject cooperation and intrusion. Therefore, it is widely used in crime prevention, forensic identification and social security. However, gait recognition is affected by many uncertain factors, especially at a distance, such as walking speed, dressing and carrying conditions, camera viewpoint and frame rate. At present, the normal gait recognition methods are divided into two ways, one is to regard gait as image, the other is to regard gait as video sequence. In the first method, as input, all gait silhouettes are compressed into an image or a gait template for gait recognition. This method is simple and easy to implement, but it is easy to lose temporal and fine-grained spatial information. The second method directly extracts features from the original gait silhouettes sequence, but this method also has the disadvantage that it is easily affected by external factors.

2 The Proposed Method

2.1 The Original GaitSet method

After analyzing the data set provided by the competition, we improved the GaitSet method proposed by Chao et al.[1], so that the improved version can better adapt to the competition to achieve higher recognition accuracy. The authors of the GaitSet method judged that the order information of gait sequence is not necessary in the gait recognition task, so the input of GaitSet network is regarded as

silhouette sequence composed of independent frames, which can be disordered. First, each frame image is input into a CNN to extract frame-level features. Second, an operation called set pooling is utilized to aggregate frame-level features into a single set-level feature. And then send frame-level and set-level features to HPM (horizontal pyramid mapping) together, to obtain more discriminative features. Since this operation is applied to high-level feature maps instead of the original silhouettes, it can preserve spatial and temporal information better than gait template.

2.2 Our improved method

We found that the CNN part of the original network has three layers, each layer has two convolution modules, the number of channels is 32, 64, 128, respectively. From this observation, we modified the network structure. In the CNN part, we add one layer, also including two convolution modules. After the modification, the depth of the network increased, and the number of channels in each layer became 32, 64, 128, 256, so that we can get more detailed information. Generally, the deeper the network structure is, the more information we can obtain and extract. To train the deeper network easier and more stable, and to work on the smooth solution space, we employ the ResNet[2] as our backbone. After the network is deepened, we do cross layer connection processing for convolution. We connect input and output across the two convolution blocks of each layer. The input of each layer is not processed, but directly superimposed on the output port. The stacked result is taken as the input of the next layer. However, the cross-layer stacking may lead to the problem of channel number mismatch. We use a 1 * 1 convolution operation to match the channel number. The overall network structure is shown in Figure 1 and the ResBlock is shown in Figure 2.

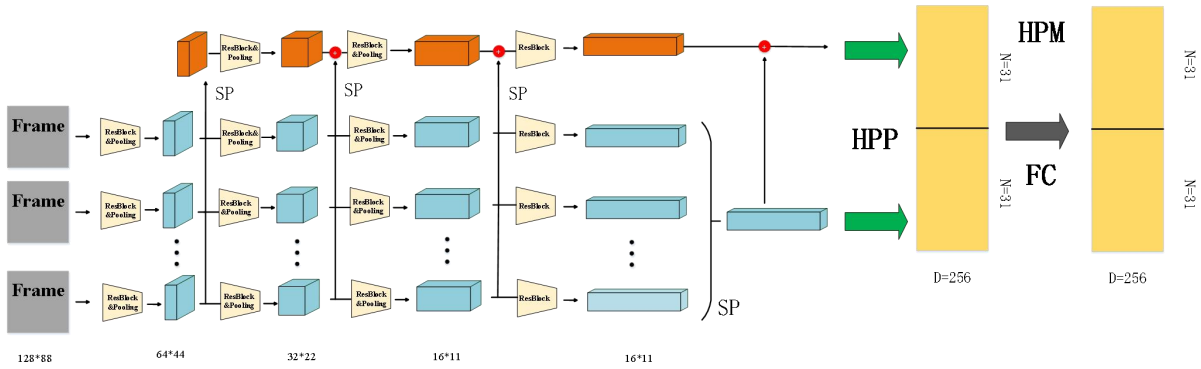


Figure 1: The structure of the network. SP means set pooling.

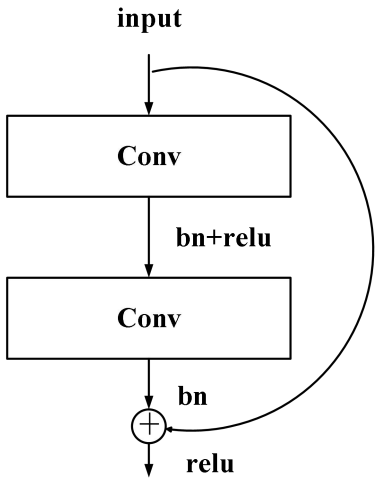
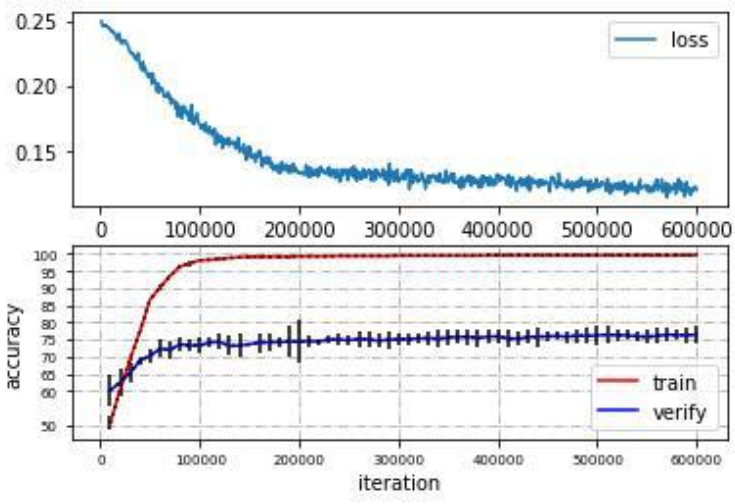


Figure 2: The structure of the ResBlock. Note that, in the first layer the kernel size of the convolution is 5*5 and 3*3 respectively, and the kernel size in other layers is both 3*3.

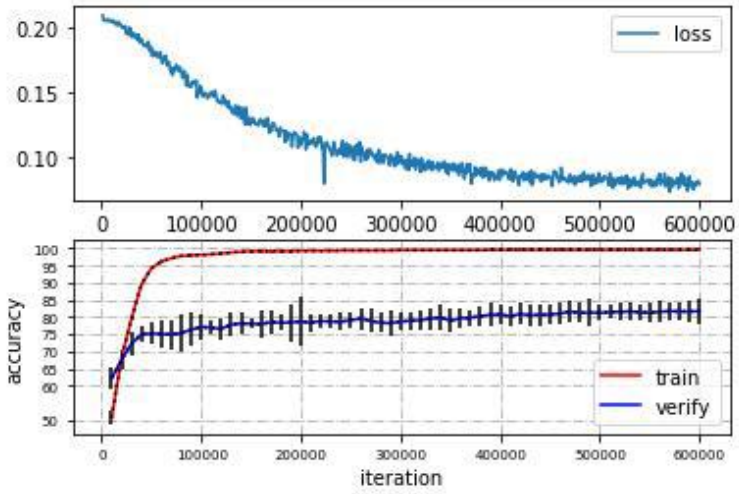
Furthermore, considering that the input resolution of the original network is 64 * 64, while the resolution of the data set given by the competition is 128 * 128, we adjust the input resolution of the network to 128 * 128 to keep more detailed information from the original data. It has been verified in the next section that high-resolution data flow can improve the performance of the network.

3 Experiments

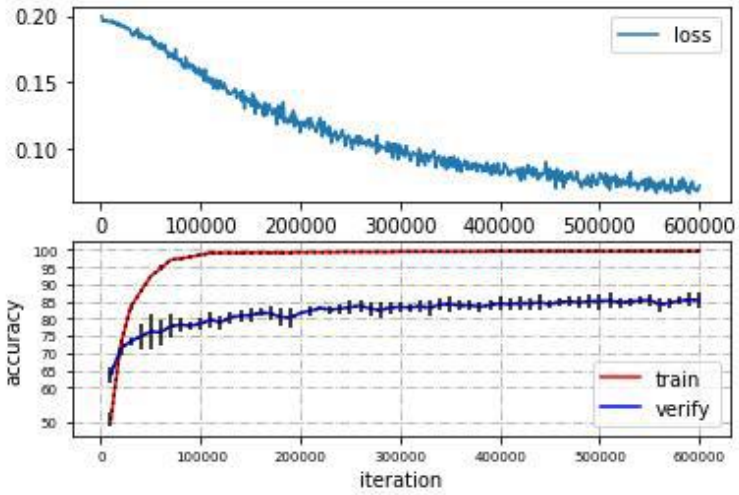
We mainly carry experiments on three models, the first is the original GaitSet network without any modification, the second is the improved method with the input image resolution is 128*128, the third is our proposed model. The validation results are shown in Figure 3.



(a) GaitSet



(b) High-resolution data flow



(c) Our method

Figure 3: The accuracy of different models. (a): GaitSet. The result of the GaitSet model without modification. (b): High-resolution data flow. Improving the GaitSet method with high-resolution data 128*128. (c): Our method: ResNet backbone+ High-resolution data flow.

We divided the training set provided by the organizer into two parts, 80% as the training set and 20% as the validation set. We use 5-fold cross validation to make sure the reliability of experimental results. It can be seen from the Figure 3 that the original GaitSet converges the fastest and the variance is small, but the recognition accuracy is relatively low. Increasing the network depth can improve the recognition accuracy, but it has large variance. We think that increasing the network depth will increase the recognition accuracy. However, due to the simplicity of input, increasing depth may miss useful information. So we used ResBlock to prevent this problem and obtain the most effective features. The accuracy has been further improved, as shown in (c) in Figure 3. We tested and verified the above three models, experimental results are shown in Table 1.

By the split ratio mentioned in the last paragraph, we first trained and verified the original GaitSet network. After 450000 iterations, we found that the experimental results were not ideal. The accuracy on validation set was only 75.3%, and on test set was only 49.2%. We increased the iterations to

600000, that improved the accuracy by about 1.1% on the validation set, and 0.7% on the test set only. This proves that the original GaitSet network cannot generalize to the data set provided by this competition, and its generalization ability was limited.

Table 1: The accuracy of different model and different iterations. V and T means validation set and test set respectively.

Method	Iterations	Accuracy	
GaitSet	450000 times	V:75.3%	T:49.2%
GaitSet_high-resolution	450000 times	V:80.6%	T:51.2%
Ours	450000 times	V:85.1%	T:53.2%
GaitSet	600000 times	V:76.4%	T:49.9%
GaitSet_high-resolution	600000 times	V:81.8%	T:51.6%
Ours	600000 times	V:85.4%	T:53.4%

According to the same data set split ratio, we tested GaitSet_high-resolution network. We found that the performance of the deeper network has been greatly improved. After the same 450000 iterations, the accuracy of the validation set achieved 80.6%, and the accuracy of the test set was 51.2%. Then we observed that the performance of the network reached the peak when it has been trained for 600000 times. The accuracy of validation set and test set was 81.8% and 51.6%, respectively. Finally, we tested our final network. We found that the use of ResBlock not only increased the accuracy, but also increased the stability. After the same 450000 iterations, our method attains an 85.1% accuracy on the validation, and accuracy on the test set was 53.2%. After 600000 iterations, the accuracy on validation set was 85.4% and the accuracy on test set has been increased to 53.4%.

4 Conclusion and the research in the future

In this paper, we presented a novel network that deepened the original GaitSet network, and integrated the idea of ResNet into it, our proposed network achieves a higher the recognition accuracy of the data set provided by the competition. But there is still much room for improvement in this accuracy rate. If we further deepen the network, it is reasonable to believe that its performance will be better. However, it will also cause the increase of the number of parameters, then more training resources will be occupied and generalization ability may be weakened. Furthermore, the proposed method can also be combined with a variety of classifiers, such as SVM, AdaBoost and so on, that is the key to our research in the future.

5 References

[1] Chao H, He Y, Zhang J, et al. GaitSet: Regarding gait as a set for cross-view gait recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33: 8126-8133.
[2] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.