Learning Discriminative Representations by Fusing Multimodal Features for Gait Recognition

Binyuan Huang, Yongdong Luo, Jiahui Xie, Zhiwen Li, Chengju Zhou, Jiahui Pan
School of Software, South China Normal University
cjzhou@scnu.edu.cn and panjiahui@m.scnu.edu.cn

1 Introduction

Comparing with other physiological biometrics (e.g., DNA, fingerprints, irises, and faces), gait is a unique one that can be perceived at a long distance without subject cooperation. Therefore, it is a popular behavioral modality for person identity authentication and finds its important role in many applications such as surveillance, forensics, and criminal investigation.

However, gait recognition suffers from various covariates, including view, clothing, and carrying status. To alleviate the effect from covariates, three key components were used to fuse multimodal features to learn more discriminative representations. In the following, we first introduce our overall structure, and then three modules of the network, including Lateral Connection Feature Aggregator (LCFA), Multi-Scale Feature Extractor (MSFE) and Global and Local Feature Modual (GLFM), will be described respectively.

2 Method

2.1 Overview

The overall structure of the network is shown in Figure 1. In order to extract discriminative gait features from the network, three key components are used to realize the comprehensive integration of gait information from various aspects. Particularly, the LCFA module employs a serial multi-scale feature fusion method to combine gait features of different depth layers and different receptive fields. The MSFE module applies a parallel multi-scale feature fusion method to aggregate features of different scales in gait images. The GLFM is used to combine global and local information in gait feature.

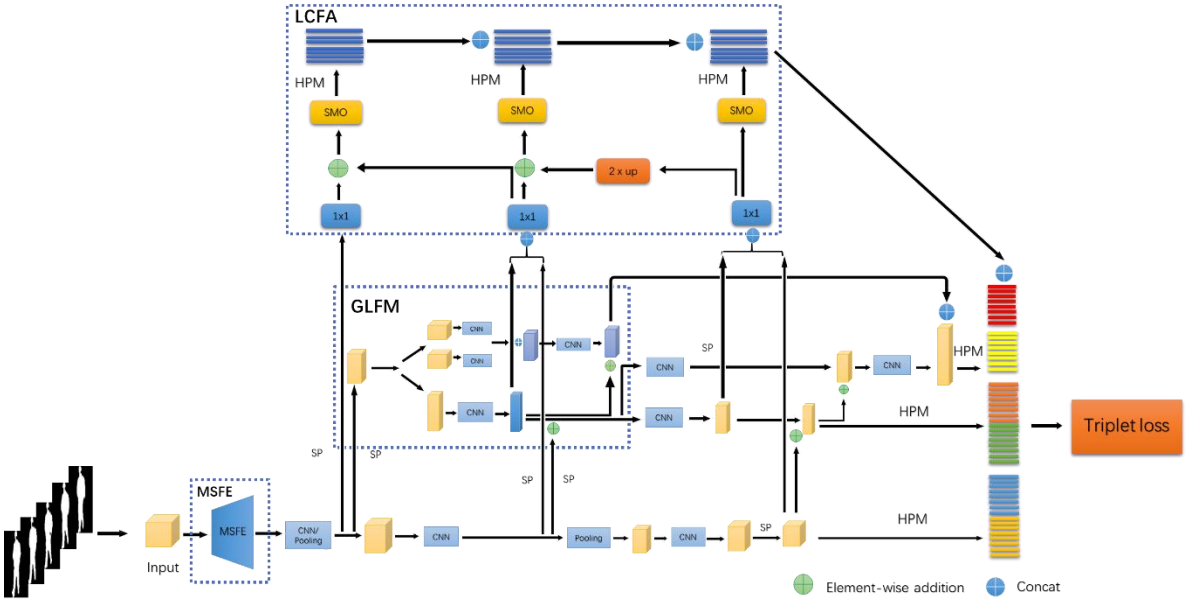

Figure 1. Overview of the proposed gait recognition framework. MSFE means Multi-Scale Feature Extraction, 1x1 means convolution with kernel size 1, SP [1] means Set Pooling, and HPM [1] means Horizontal Pyramid Pooling. 2xup means double the length and width of the feature maps. SMO [3]means Smooth Layer.

2.2 Lateral Connection Feature Aggregator (LCFA)

Inspired by the successful application of Feature Pyramid Network (FPN) in semantic segmentation, we aggregate the features of different network layers and different receptive fields. In this way, the network can learn a more discriminative gait feature, which is similar to FPN. The feature fusion method using Lateral Connections in GLN [3] attracted our attention, which effectively integrates different semantic information. Similar with GLN [3], we added the LCFA module to the network, using the inherent feature pyramid in deep learning to enhance gait representations. The division of each stage in the LCFA is shown in Figure 2, and the module structure of the LCFA is shown in Figure 3. In stage1 and stage2, LCFA uses Set Pooling to process silhouette-level features, and concatenate its output with the set-level features in the channel dimension. While in stage0, only silhouette-level features can be used, the feature in this stage is also processed in the way of Set Pooling. In each stage of feature extraction, a 1×1 convolution kernel is used to unify features from all the stages into the same dimension. Then, for the feature fusion between stage2 and stage1, we upsample the features generated in stage2 by a factor of 2 and add to the features generated at stage1.While for the feature fusion between stage1 and stage0, we set the size of the feature maps of stage1 and stage0 to 32×32 in the network design to reduce the memory cost, so the features of stage1 and stage0 can be added directly without upsampling.Finally, the merged features of each stage are passed through the smoothing layer in turn to reduce the aliasing effect caused by upsampling and semantic differences between different stages. Every smooth layer is implemented by a 3×3 convolutional layer. Next, we input the features of different stages into the HPM [1] module in turn. In the HPM [1] of LCFA module, we use scales s = {1, 2, 4} to split features horizontally.

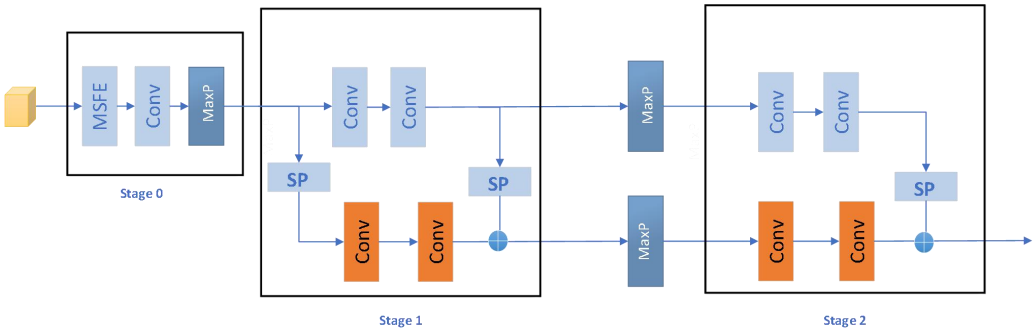

Figure 2. Division of each stage [3]

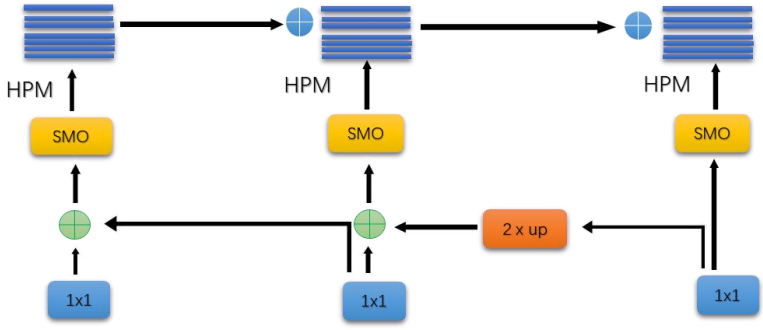

Figure 3. Schematic diagram of LCFA module [3]

2.3 Multi-Scale Feature Extractor（MSFE）

The idea of multi-scale feature extraction has been applied in image classification, target detection, and other fields. Inspired by the idea of multi-scale feature extraction [4], we introduce a parallel multi-scale feature fusion module into the network structure. This module uses 3 convolution kernels of different sizes to extract features so that the network can capture different scales of gait characteristics. Finally, the features on different scales are merged by concatenating the channel dimension, so that the network can learn more discriminative gait features.

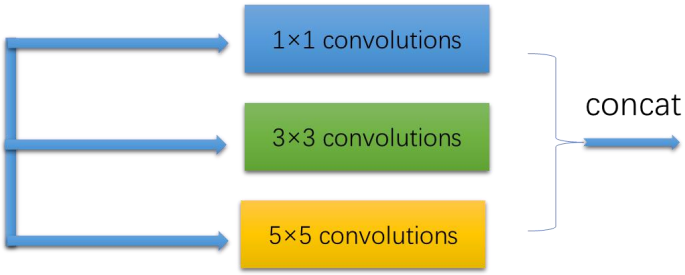

Figure 4. Schematic diagram of MSFE module

2.4 Global and Local Feature Module (GLFM)

Inspired by the idea of global and local feature extraction in gait-based age estimation [2], we use the GLFM to extract and fuse global and local gait features. The GLFM is divided into global and local feature extractors. The global feature extractor uses convolution kernels to extract global information, while the local feature extractor divide the gait feature into multiple parts and execute the different parts separately. Each convolution network in local feature extractor is used to learn the feature of specific parts, and then integrate them by concatenating horizontally. Finally, we fuse the global and local features through element-wise addition.

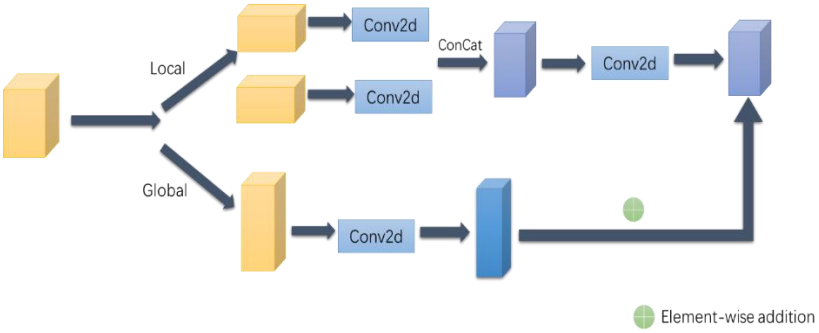

Figure 5. Schematic diagram of GLFM module

3 Experiments

We refer to GaitSet Fine-tune [4] to realize part of the implementation. In detail, in order to reduce the memory consumption, we cut the input silhouette to 64x64. In the training phase, we feed the first 30 silhouettes of gait sequence into the network. While in the test phase, we input the whole gait sequence into the network. The batch size is set as 64 in the CASIA-E dataset. Adam optimizer is adopted throughout our experiments, and the learning rate is fixed at 1e-4. For the CASIA-E data set, the proposed model reaches convergence after 11k iterations in the training. During the experiment, we also conducted some ablation study to verify the LCFA effectiveness. The experimental results are shown in Table 1.

Table 1 Ablation study

Components	Accuracy
without LCFA	66.755
with LCFA	64.468

4 Results

The accuracy of the CASIA-E dataset is 66.8%.

5 Conclusions

In this experiment, we used the LCFA module to fuse the multimodal features between different layers of networks in a serial multi-scale feature fusion way, while the MSFA module is employed to fuse features of different scales in a parallel multi-scale feature fusion way. Through the GLFM, both

the global and local feature are fused. By adopting the idea of fusion with multimodal features, our accuracy rate has been effectively improved.

6 References

[1]. H., Chao, Y.,He, J.,Zhang, J., Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. Proceedings of the AAAI conference on artificial intelligence. 2019, 33(01), pp. 8126-8133.

[2]. Zhu, H., Zhang, Y., Li, G. et al. Ordinal distribution regression for gait-based age estimation. Sci. China Inf. Sci. 63, 120102 (2020). https://doi.org/10.1007/s11432-019-2733-4

[3]. S., Hou, C., Cao, X., Liu, Y.，Huang. Gait Lateral Network: Learning Discriminative and Compact Representations for Gait Recognition. In *European Conference on Computer Vision*. 2020, August. pp. 382-398.

[4]. GaitSet_Fine-tune: https://github.com/dwzhu97/GaitSet_Fine-tune.