SILHOUETTE-BASED VIEW-EMBEDDINGS FOR GAIT RECOGNITION UNDER MULTIPLE VIEWS

Tianrui Chai Xinyu Mei Annan Li∗ Yunhong Wang

State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China.
{trchai,xymei,liannan,yhwang}@buaa.edu.cn

ABSTRACT

Gait recognition under multiple views is an important computer vision and pattern recognition task. In the emerging convolutional neural network based approaches, the information of view angle is ignored to some extent. Instead of direct view estimation and training view-specific recognition models, we propose a compatible framework that can embed view information into existing architectures of gait recognition. The embedding is simply achieved by a selective projection layer. Experimental results on two large public datasets show that the proposed framework is very effective.

Index Terms— Gait recognition, silhouette-based, view-embedding, cross-view, multi-task

1. INTRODUCTION

Gait is a biometric presenting the walking style of people and has an edge over other biometrics such as face, fingerprint because it can be recognized at a distance with much less cooperation. Recently, due to the growing demand of intelligent surveillance, gait recognition attracted more attentions.

Variations like carrying conditions, coat-wearing and viewpoint differences may cause changes in gait appearance and bring significant challenges to gait recognition. Solving these problems is of great significance to improve the performance of gait recognition. Among these problems, viewpoint differences is a very tricky problem, because it may bring greater visual differences than the identity.

Over the years many methods have been proposed to solve the problem of multiple view. Gait recognition methods include model-based [21, 22] and appearance-based methods. Appearance-based methods can be divided into two categories, i.e., regarding gait as a single image [1, 2, 3, 9, 13, 20, 23, 24] and regarding gait as a video or an image sequence [4, 5, 7, 6, 10]. The first category often use gait energy images (GEIs) [8] as the representation of gait, while the second category directly makes use of the gait silhouette images. Between these two categories, silhouettes based methods show better performance and become a main trend.

Although the aforementioned methods work well on representing gait in a multi-view scenario, and the deep neural network can somehow learn view-robust feature from mixed views, view itself, i.e. the explicit view estimation and view-specific modeling, is overlook and underrated. We argue that explicit embedding view information can effectively improve the performance of existing approaches.

In this paper, we proposed a general framework for multi-view gait recognition by explicit view angle embedding, based on which, two state-of-the-art gait recognition backbones, i.e. Gaitset [4] and GaitGL [14] are enhanced. Compared with the original ones, the enhanced ones, improve the performance. The effectiveness is well demonstrated by the experiments on CASIA-B [11] and OUMVLP [12] datasets.

The rest of this paper is organized as follows. In Section 2 we will briefly describe relevant studies. The details of proposed methods are introduced in Section 3, while the experimental validation is described in Section 4. And the conclusion is given in Section 5.

2. RELATED WORK

Gait Recognition. With the emergence of deep learning, convolution neural networks (CNNs) have been adopted to gait recognition and gain a great success. Shiraga et al. [1] and Wu et al. [3] both used CNN to extract feature from GEIs. However, the simple averaging operation of GEI results in a serious detail loss and limits the learning effects of CNN. To this end, Wolf et al. [7] extracted feature directly from original silhouettes by three dimensional convolutions. Chao et al. [4] treated gait silhouette sequence as a set and proposed a novel model named Gaitset. Fan et al. [5] further improved silhouette gait feature by introducing part-based representations. Lin et al. [6, 14] proposed a comprehensive model named GaitGL by integrating both global and local feature. It is recognized as the state-of-the-art gait recognition method.

Multi-task Learning. Although success, all the aforementioned approaches overlook the view information which has
3. PROPOSED METHOD

What we propose is not a specific model, but a general and compatible framework for multi-view gait recognition. As shown in Figure 1, the input is a gait silhouettes sequence $X_{in} \in \mathbb{R}^{T \times H \times W}$ and a backbone model $E$ is used to extract feature map $X_f \in \mathbb{R}^{C_f \times H_f \times W_f}$. The backbone can be any silhouettes-based network, such as Gaitset [4], Gaitpart [5], MT3D [16] and GaitGL [14].

Followed by the backbone, the feature map will be fed into two branches. The first one performs Horizontal Pyramid Mapping (HPM) $f_{HPM} \in \mathbb{R}^{n \times D}$ will be obtained where $D$ is the dimension of output feature. In the second branch, the feature map is pooled to get the view classification feature $f_v \in \mathbb{R}^{D_v}$, and the projection matrices $\{W_1, W_2, W_3, ..., W_n\}$ ($W_i \in \mathbb{R}^{D \times D}$) are selected according to the predicted view, where $n$ is the number of discrete views in the HPP Module [4]. Then for each feature in HPM, we will multiply the projection matrix of the corresponding view to get the final view-invariant feature.

3.2. HPP feature projection

For the convenience for explanation, the horizontal pyramid mapping (HPM) $f_{HPM} \in \mathbb{R}^{n \times D}$ is expressed as $f_{HPM,i}, i = 1, 2, 3, ..., n$, where $f_{HPM,i} \in \mathbb{R}^{D}$. Suppose that the view $\hat{y}$ of the input gait silhouettes is predicted to be $\theta$ in Equation (3), then the projected features can be expressed as:

$$f_{final,i} = W_i f_{HPM,i},$$

where $i = 1, 2, ..., n$, $W_i \in Z_0$ and the $f_{final}$ is used as the representation for calculating the similarity between two input gait silhouette sequences.


3.3. Joint losses

In the proposed multi-task framework, our loss consists of cross entropy (CE) and triplet loss. Combining the Equation (3), the CE loss can be expressed as:

\[
L_{CE} = - \sum_{j=1}^{N} \sum_{i=1}^{M} y_{ij} \log(p_{ji}) \quad \text{w.r.t.} \quad p_{ji} = \frac{e^{f_{ij}}}{\sum_{i=1}^{M} e^{f_{ij}}},
\]

where \(N\) is the number of all gait silhouette sequences and \(y_{ij}\) is the discrete ground truth of view of the \(j\)-th sequence.

Let a triplet of gait silhouette groups be \((Q, P, N)\), where \(Q\) and \(P\) are from the same subject and \(Q\) and \(N\) are from two different subjects. Denote \(K\) triplets of fixed identity as \(\{T_i | T_i = (f_{final,i}^{QP}, f_{final,i}^{PN}), i = 1, 2, ..., K\}\), then combining the Equation (4), the triplet loss can be expressed as:

\[
L_{trip} = \frac{1}{K} \sum_{i=1}^{K} \max(m - d_{ij}, 0),
\]

where \(d_{ij} = \|f_{final,j}^{QP} - f_{final,j}^{PN}\|_2\). In this paper we used full mining to make triplets. Combine Equation (5) and (6), the joint loss can be defined as:

\[
\mathcal{L} = \lambda_{CE} L_{CE} + \lambda_{trip} L_{trip},
\]

where \(\lambda_{CE}\) and \(\lambda_{trip}\) are hyper-parameters.

### 4. EXPERIMENT

In order to prove the effectiveness of view embedding in gait recognition, two silhouette-based architectures, i.e. Gaitset [4] and GaitGL [14] are used as the backbone.

#### 4.1. Datasets

CASIA-B dataset [11] contains 124 subjects, each contains 11 views and each view contains 10 sequences. The sequences are obtained in three scenarios: normal (NM) (six sequences per subject), walking with bag (BG) (two sequences per subject) and wearing coat or jacket (CL) (two sequences per subject) respectively. We conduct experiments following the settings in [3]. These three settings are small-sample training (ST), medium-sample training (MT) and large-sample training (LT), in which 24, 62 and 74 subjects are used for training and the rest are used for test respectively. The first four sequences of the NM condition (NM\#1-4) are kept in gallery, and the rest sequences are divided into three probe subsets, i.e. NM subsets containing NM \#5-6, BG subsets containing BG \#1-2 and CL subsets containing CL \#1-2.  

OU-MVLP dataset [12] is the largest public gait dataset, which contains 10,307 subjects. 5,153 subjects are used for training and the rest 5,154 subjects are used for test. Each sub-
Fig. 2. Examples of View projection matrices for strip 0 and strip 20. The Diff column shows the absolute difference between the two matrices of different views at the same strip.

It can be seen that our Vi-Gaitset is more accurate than the original Gaitset under all of the settings by a large margin. For MT and LT settings, the performance of Vi-GaitGL is close to that of the original GaitGL in NM and BG conditions. And the performance of Vi-GaitGL is much better than that of the original GaitGL in CL conditions by 3.6% and 4.2%. For ST setting, performance of Vi-GaitGL decreased slightly (1.6% and 0.6%) in NM and BG conditions, and increased as high as 4.3% in CL condition. We argue that part of the performance degradation of Vi-GaitGL in ST setting is due to the small size of training data, which is important to multi-task learning. Another possible explanation is the view recognition accuracy. For Vi-GaitGL, it is only 96.2% under ST setting, while it is 97.7% in MT and 97.8% in LT.

Table 2. Rank-1 accuracy (%) on OU-MVLP under 14 probe views excluding identical-view cases.

| Probe angle | GEINet | Gaitset | Vi-Gaitset | GaitPart | GaitGL | Vi-GaitGL |
|-------------|--------|---------|------------|----------|--------|-----------|
| 0°          | 11.4   | 79.5    | 81.8       | 82.6     | 84.3   | 85.6      |
| 15°         | 29.1   | 87.9    | 89.2       | 88.9     | 89.8   | 90.2      |
| 30°         | 41.3   | 89.9    | 90.5       | 90.8     | 90.8   | 91.2      |
| 45°         | 45.5   | 90.2    | 90.5       | 91.0     | 91.0   | 91.5      |
| 60°         | 39.3   | 88.1    | 89.2       | 89.7     | 90.5   | 91.1      |
| 75°         | 41.8   | 88.7    | 89.5       | 89.7     | 90.5   | 90.9      |
| 90°         | 38.9   | 87.8    | 89.0       | 89.9     | 90.3   | 90.4      |
| 180°        | 14.9   | 81.7    | 83.9       | 85.2     | 88.1   | 88.3      |
| 195°        | 33.1   | 86.7    | 88.1       | 88.1     | 87.9   | 88.7      |
| 210°        | 43.2   | 89.0    | 89.7       | 90.0     | 89.6   | 90.6      |
| 225°        | 45.6   | 89.3    | 89.8       | 90.1     | 89.8   | 90.6      |
| 240°        | 39.4   | 87.2    | 88.6       | 89.0     | 88.9   | 90.1      |
| 255°        | 40.5   | 87.8    | 88.5       | 89.1     | 88.9   | 89.9      |
| 270°        | 56.2   | 86.2    | 87.6       | 88.2     | 88.2   | 89.4      |
| mean        | 38.8   | 87.1    | 88.3       | 88.7     | 89.1   | 88.9      |

The results in rank-1 accuracy (%) on OU-MVLP dataset are shown in Table 2. The performance of Vi-Gaitset is better than the original Gaitset [4] under all the probe views. The proposed Vi-GaitGL meets a new state-of-the-art under various cross-view conditions and the mean rank-1 accuracy is 0.8% higher than the original GaitGL [14]. When the data is sufficient, our method is consistently better.

In order to explain the effectiveness of our framework, we compare the projection matrices of different views in Vi-GaitGL (trained on OU-MVLP). As illustrated in Figure 2, their difference has obvious vertical texture, which indicates that the projection matrices of different views have view specificity for feature mapping.

5. CONCLUSION

In this paper, we propose a general view embedding framework for improved multi-view gait recognition, in which the view angle is explicitly estimated and used for model refining. Experimental results on two leading backbone models show that our idea of explicit view embedding is very effective. The proposed framework with GaitGL [14] as the backbone meets the state-of-the-art on two large-scale public gait datasets. It should be pointed out that the proposed framework is not competitive but rather complementary to existing works.
6. REFERENCES

[1] Kohei Shiraga, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi, “Geinet: View-invariant gait recognition using a convolutional neural network,” in *ICB*. IEEE, 2016, pp. 1–8.

[2] Maodi Hu, Yunhong Wang, Zhaoxiang Zhang, James J Little, and Di Huang, “View-invariant discriminative projection for multi-view gait-based human identification,” *IEEE TIFS*, vol. 8, no. 12, pp. 2034–2045, 2013.

[3] Zifeng Wu, Yongzhen Huang, Liang Wang, Xiaogang Wang, and Tieniu Tan, “A comprehensive study on cross-view gait based human identification with deep cnns,” *IEEE TPAMI*, vol. 39, no. 2, pp. 209–226, 2016.

[4] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng, “Gaitset: Regarding gait as a set for cross-view gait recognition,” in *AAAI*, 2019, vol. 33, pp. 8126–8133.

[5] Beibei Lin, Shunli Zhang, and Feng Bao, “Gait recognition with multiple-temporal-scale 3d convolutional neural network,” in *ACM MM*, 2020, pp. 3054–3062.

[6] Thomas Wolf, Mohammadreza Babaei, and Gerhard Rigoll, “Multi-view gait recognition using 3d convolutional neural networks,” in *ICIP*. IEEE, 2016.

[7] Jinguang Han and Bir Bhanu, “Individual recognition using gait energy image,” *TPAMI*, vol. 28, no. 2, pp. 316–322, 2005.

[8] Yiwei He, Junping Zhang, Hongming Shan, and Liang Wang, “Multi-task gans for view-specific feature learning in gait recognition,” *IEEE TIFS*, vol. 14, no. 1, pp. 102–113, 2018.

[9] Yuqi Zhang, Yongzhen Huang, Shiqi Yu, and Liang Wang, “Cross-view gait recognition by discriminative feature learning,” *IEEE TIP*, vol. 29, pp. 1001–1015, 2019.

[10] Shiqi Yu, Daoliang Tan, and Tieniu Tan, “A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition,” in *ICPR*. IEEE, 2006, vol. 4, pp. 441–444.

[11] Noriko Takemura, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi, “Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition,” *IPSJ TCVA*, vol. 10, no. 1, pp. 4, 2018.

[12] Shuangqun Li, Wu Liu, Huadong Ma, and Shaopeng Zhu, “Beyond view transformation: Cycle-consistent global and partial perception gan for view-invariant gait recognition,” in *ICME*, 2018.

[13] Beibei Lin, Shunli Zhang, Xin Yu, Zedong Chu, and Haikun Zhang, “Learning effective representations from global and local features for cross-view gait recognition,” *arXiv preprint arXiv:2011.01461*, 2020.

[14] Manuel J Marín-Jiménez, Francisco M Castro, Nicolás Gui, F de La Torre, and R Medina-Carnicer, “Deep multi-task learning for gait-based biometrics,” in *ICIP*. IEEE, 2017, pp. 106–110.

[15] Ming Liang, Bin Yang, Yun Chen, Rui Hu, and Raquel Urtasun, “Multi-task multi-sensor fusion for 3d object detection,” in *CVPR*, 2019, pp. 7345–7353.

[16] Xi Yin and Xiaoming Liu, “Multi-view convolutional neural network for pose-invariant face recognition,” *IEEE TIP*, vol. 27, no. 2, pp. 964–975, 2017.

[17] Jifeng Dai, Kaiming He, and Jian Sun, “Instance-aware semantic segmentation via multi-task network cascades,” in *CVPR*, 2016, pp. 3150–3158.

[18] Shaoxiong Zhang, Yunhong Wang, and Annan Li, “Gait-based age estimation with deep convolutional neural network,” in *ICB*. IEEE, 2019, pp. 1–8.

[19] Worapan Kusakunniran, Qiang Wu, Jian Zhang, Hongdong Li, and Liang Wang, “Recognizing gaits across views through correlated motion co-clustering,” *IEEE TIP*, vol. 23, no. 2, pp. 696–709, 2013.

[20] Rijun Liao, Edel B Garcia, Shiqi Yu, and Yongzhen Huang, “Pose-based temporal-spatial network (ptsn) for gait recognition with carrying and clothing variations,” in *CCBR*. Springer, 2017, pp. 474–483.

[21] Rijun Liao, Shiqi Yu, Weizhi An, and Yongzhen Huang, “A model-based gait recognition method with body pose and human prior knowledge,” *Pattern Recognition*, vol. 98, pp. 107069, 2020.

[22] Xianye Ben, Chen Gong, Peng Zhang, Xitong Jia, Qiang Wu, and Weixiao Meng, “Coupled patch alignment for matching cross-view gaits,” *IEEE TIP*, vol. 28, no. 6, pp. 3142–3157, 2019.

[23] Xianye Ben, Chen Gong, Peng Zhang, Rui Yan, Qiang Wu, and Weixiao Meng, “Coupled bilinear discriminant projection for cross-view gait recognition,” *IEEE TCSVT*, vol. 30, no. 3, pp. 734–747, 2019.