Lagrange Motion Analysis and View Embeddings for Improved Gait Recognition

Tianrui Chai[1], Annan Li[1]*, Shaoxiong Zhang[1], Zilong Li[2], Yunhong Wang[1]
[1] State Key Laboratory of Virtual Reality Technology and Systems,
School of Computer Science and Engineering, Beihang University, China.
[2] School of Engineering Science, University of Chinese Academy of Sciences, China
{trchai,liannan,zhangsx}@buaa.edu.cn, lizilong@imech.ac.cn, yhwang@buaa.edu.cn

Abstract

Gait is considered the walking pattern of human body, which includes both shape and motion cues. However, the main-stream appearance-based methods for gait recognition rely on the shape of silhouette. It is unclear whether motion can be explicitly represented in the gait sequence modeling. In this paper, we analyzed human walking using the Lagrange's equation and come to the conclusion that second-order information in the temporal dimension is necessary for identification. We designed a second-order motion extraction module based on the conclusions drawn. Also, a light weight view-embedding module is designed by analyzing the problem that current methods to cross-view task do not take view itself into consideration explicitly. Experiments on CASIA-B and OU-MVLP datasets show the effectiveness of our method and some visualization for extracted motion are done to show the interpretability of our motion extraction module.

1. Introduction

Gait is a biometric presenting the walking pattern of pedestrian for identity recognition and has an edge over other biometrics such as face, iris or fingerprint since it can be recognized without touch and at a distance. Although it has been studied for years, there are still some challenges in gait recognition. For example, variations like carrying conditions [4, 15, 16, 42, 46], coat-wearing and viewpoint differences [41, 45] may cause changes in gait appearance and make it hard to distinguish pedestrian.

Existing appearance-based approaches for gait recognition rely heavily on the visual appearance of silhouettes. However, when the view angle is close, the appearance difference between two different person can be smaller than that of the same people but viewed from two different angles.

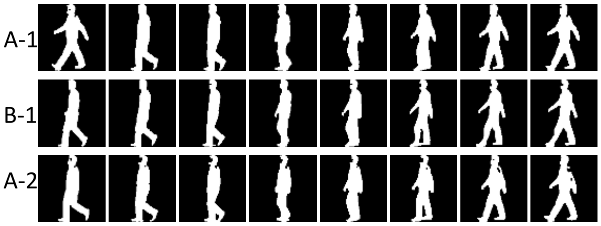

Figure 1. Three samples from CASIA-B dataset, where *A* and *B* denote ID 39 and ID 77 respectively. *A-1* and *A-2* are two samples of *A* selected from different sequences. It can be found that it is difficult to find the difference between *A* and *B* visually. Even on some frames, *A-2* are more similar to *B-1* than *A-1*.

A common approach to address the aforementioned issue is learning viewpoint-invariant or robust features [6, 10, 20, 21, 35, 37, 38, 43]. However, these works focus on how to extract apparent information and the fusion of spatial or temporal features. The detection or estimation of viewpoint is overlook and there is few model explicitly making use of viewpoint. In other words, the viewpoint-robustness of these methods is solely based on the coverage of data, which is a well-known ill-posed problem.

Even when the viewpoint is close, the apparent information is still not very reliable. As shown in Figure 1, it is difficult to distinguish the identity of the three samples only from the body shape.This phenomenon explains why pure appearance-based method such as Gait Energy Images (GEIs) [34] cannot achieve ideal performance. The similar situation will also occur in state-of-the-art Gaitset method [6] which does not use temporal information either.

We argue that the ultimate solution to the problem shown in Figure 1 is the gait motion. Recently, some methods making use of temporal features are proposed [7, 20, 20, 28, 39]. Although these models show a stronger edge in recognition accuracy, They do not discuss the motion information in gait to the extent that some discriminatory biological information may be missing.

In this paper, by mathematical modeling analysis, we argue that it is difficult to distinguish people using only first-

*Corresponding author.

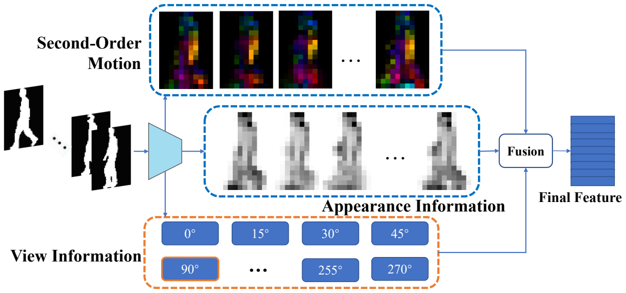

Figure 2. Overview of our multi-branch framework.

order temporal information [1]. To effectively modeling the walking pattern of a pedestrian, second-order motion is necessary. To verify this idea, a novel motion-assisted gait recognition method is proposed. To further reduce the negative impact of viewpoint difference, a view-aware embedding method is also introduced. It results in a multi-branch framework which combined the view, appearance and intrinsic motion of silhouette sequences. Experimental results show that the proposed model can effectively narrow the intra-class distance caused by view variance.

The major contribution of this paper can be summarized as the following four aspects:

- We model human walking by *Lagrange's* Equation and come to the conclusion that we need to use second-order motion features to represent the gait in addition to the first-order motion features.
- Based on the conclusion of *Lagrange* motion analysis, we propose a second-order *Motion Extraction Module* to extract features on the high-level feature maps.
- We proposed a novel and light-weighted view embedding to narrow the difference caused by changes of view.
- We apply our proposed method to the widely used CASIA-B and OU-MVLP datasets and the effectiveness of our method is verified. Some visualizations are conducted to further prove the validity of our idea.

2. Related Works

Gait Recognition Gait recognition methods can be divided into two categories, i.e. model-based methods and appearance-based methods respectively.

Model-based gait recognition methods [2, 17, 19, 24, 33] make use of pose information to model human pose-invariant identity information. This kind of method is naturally robust to interference items such as clothing variation and carrying articles. However, model-based methods are greatly affected by the accuracy of pose estimation. Pedestrian pose estimation itself still remains a challenging problem [14, 18], especially for cross-domain pose estimation [44], which is a closer scenario to gait recognition.

Nowadays, with the development of deep learning, the performance of appearance-based methods have made a greater breakthrough. Wu et al. [37] and Chao et al. [6] proposed networks suitable for gait recognition firstly. Wolf et al. [36], Lin et al. [20] and Huang et al. [12] used three dimensional convolution on gait recognition. Fan et al. [7] and Huang et al. [11] take temporal models into consideration.

View-invariant Modeling Viewpoint change is a challenging problem in biometrics including face recognition and gait recognition. Compared with face, there are fewer methods take view into consideration in gait recognition. He et al. [9] proposed a multi-task GAN and use view labels as the supervision to train the GAN. Chai et al. [5] take different projection matrices as view embedding methods and approach high growth on several backbones. However, these models are complex and have too many parameters.

Optical Flow and Motion Optical flow is one representation of motion and optical flow estimation is a task which predicts the pixel-to-pixel correspondence between two adjacent frames. Recently, many deep learning methods [29, 40] are used for optical flow estimation. Among these methods, RAFT [32] is the one with perfect performance and the fastest speed now. Optical flow has been used in many areas including action recognition [3, 26] and video generation [1].

3. Why second-order motion?

Gait is recognized as the walking pattern that can distinguish pedestrian [23]. In the early years However, appearance-based methods with convolutional neural networks now mainly focus on the two dimensional feature of silhouettes. Even Gaitset [6], which is one of the-state-of-the-arts, does not rely on any temporal feature. It is hard to prove whether current state-of-the-art methods depends on human body shape or traditional "gait". In the early years, some methods [8, 27, 30] have explored the effects of motion as well as acceleration (second-order motion) on gait recognition, but they have not looked deeper into the theory and the physics behind it.

Therefore, in order to explore the essential information, we propose to use the *Lagrange's* Equation [13] to analyze the walking of human. As shown in Figure 3, we assume that the human thighs and legs are rigid and model them mechanically. The length and mass of the two thighs and two legs are denoted as l_1, l_2, m_1, m_2 and l_3, l_4, m_3, m_4 respectively. θ_i represents the angle between them and vertical lines. Also, the human body is assumed to move forward with a small distance x.

[1] In this paper, we call items in the form of $\frac{\mathrm{d}x}{\mathrm{d}t}$ first-order information and items like $\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2}$ second-order information.

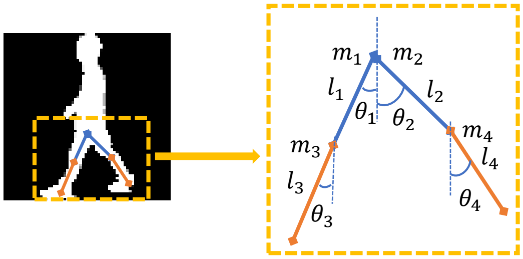

Figure 3. Analysis diagram of pedestrian walks.

Then we can obtain the kinetic energy T as:

$$\begin{aligned} T = \frac{1}{2}(m_1+m_2+m_3+m_4)(\frac{\mathrm{d}x}{\mathrm{d}t})^2 + \frac{1}{6}(m_1 l_1^2(\frac{\mathrm{d}\theta_1}{\mathrm{d}t})^2 \\ + m_2 l_2^2(\frac{\mathrm{d}\theta_2}{\mathrm{d}t})^2 + m_3 l_3^2(\frac{\mathrm{d}\theta_3}{\mathrm{d}t})^2 + m_4 l_4^2(\frac{\mathrm{d}\theta_4}{\mathrm{d}t})^2) \end{aligned}, \quad (1)$$

and the potential energy V as:

$$\begin{aligned} V = -\frac{1}{2}m_1 g l_1 \cos\theta_1 - m_3 g(l_1\cos\theta_1 + \frac{l_3}{2}\cos\theta_3) \\ -\frac{1}{2}m_2 g l_2 \cos\theta_2 - m_4 g(l_2\cos\theta_2 + \frac{l_4}{2}\cos\theta_4) \end{aligned} \quad (2)$$

Let us calculate $L = T - V$. Then with the *Lagrange's Equation*[2], the system can be formulated with $x, \theta_1, \theta_2, \theta_3, \theta_4, t$ as:

$$\begin{aligned} (m_1+m_2+m_3+m_4)\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = Q_0 \\ \frac{1}{3}m_1 l_1^2\frac{\mathrm{d}^2\theta_1}{\mathrm{d}t^2} - \frac{1}{2}(m_1+m_3)g l_1\sin\theta_1\frac{\mathrm{d}\theta_1}{\mathrm{d}t} = Q_1 \\ \frac{1}{3}m_2 l_2^2\frac{\mathrm{d}^2\theta_2}{\mathrm{d}t^2} - \frac{1}{2}(m_2+m_4)g l_2\sin\theta_2\frac{\mathrm{d}\theta_2}{\mathrm{d}t} = Q_2, \\ \frac{1}{3}m_3 l_3^2\frac{\mathrm{d}^2\theta_3}{\mathrm{d}t^2} - \frac{1}{2}m_3 g l_3\sin\theta_3\frac{\mathrm{d}\theta_3}{\mathrm{d}t} = Q_3 \\ \frac{1}{3}m_4 l_4^2\frac{\mathrm{d}^2\theta_4}{\mathrm{d}t^2} - \frac{1}{2}m_4 g l_4\sin\theta_4\frac{\mathrm{d}\theta_4}{\mathrm{d}t} = Q_4 \end{aligned} \quad (3)$$

where Q_0, Q_1, Q_2, Q_3, Q_4 are generalized force, including the force from human muscles and resistance. These force is the essence of pedestrian and they change gradually and continuously in a gait cycle.

It can be observed that in Equation (3), to maintain this dynamical system, we need the second-order derivatives $\frac{\mathrm{d}^2x}{\mathrm{d}t^2}, \frac{\mathrm{d}^2\theta_1}{\mathrm{d}t^2}, \frac{\mathrm{d}^2\theta_2}{\mathrm{d}t^2}, \frac{\mathrm{d}^2\theta_3}{\mathrm{d}t^2}, \frac{\mathrm{d}^2\theta_4}{\mathrm{d}t^2}$, in addition to the first-order derivatives $\frac{\mathrm{d}\theta_1}{\mathrm{d}t}, \frac{\mathrm{d}\theta_2}{\mathrm{d}t}, \frac{\mathrm{d}\theta_3}{\mathrm{d}t}, \frac{\mathrm{d}\theta_4}{\mathrm{d}t}$. If we only have the first-order variables, the system is not unique.

It's not surprising that methods based on three dimensional convolutions [12,20,21,36] show better performance since cascaded 3D convolution layers can extract second-order information at the best situation. We believe that 3D convolution can extract the temporal information, but it is difficult to prove whether the cascaded 3D convolution layers can necessarily extract the second-order motion information. There is no way to know whether the 3D convolution is taking the motion or just summing the feature maps.

With reference to the conclusions drawn from the human motion system, we have designed a module for extracting second-order motion features according to methods used in optical flow estimation [32]. In contrast to 3D convolution, it can explicitly extract the motion between adjacent frames.

4. Methods

In this section, we proposed a novel framework called *LagrangeGait* [3]. As shown in Figure 4, the framework consists of three branches. The upper branch is the motion branch which extracts the second-order motion feature according to the conclusion drawn from Section 3. The middle branch is the main branch to extract the appearance feature and can be any backbone such as Gaitset [6] or GaitGL [21]. The feature maps calculated by shallow layers in main branch are used in motion branch. The bottom branch is the view branch, in which the view of input silhouette sequence is predicted and the learnable view embedding are produced.

Given a silhouette sequence, we denote it as $\boldsymbol{I} = \{\boldsymbol{I}_1, \boldsymbol{I}_2, \boldsymbol{I}_3, ...\boldsymbol{I}_T\}$ and T is the length of sequence. The feature maps extracted from the shallow layers are denoted as $\boldsymbol{X}_{origin} = [\boldsymbol{X}_1, \boldsymbol{X}_2, ...\boldsymbol{X}_t]$, where $\boldsymbol{X}_i \in \mathbb{R}^{C\times H\times W}$, and $\boldsymbol{X}_{origin} \in \mathbb{R}^{t\times C\times H\times W}$ respectively. t represents the length of feature maps in time dimension after pooling, e.g. $t = T$ in Gaitset [6], and $t = \frac{T}{3}$ in GaitGL [21], since GaitGL has a pooling layer with $3\times 1\times 1$ kernel size and the same size of stride. With the obtained $\boldsymbol{X}_{origin}$, we send it into different branches and then the motion feature map $\boldsymbol{X}_{motion}$, the appearance feature map $\boldsymbol{X}_{appearance}$, and the view feature $\boldsymbol{f}_{view}$ can be respectively calculated as

$$\begin{aligned} \boldsymbol{X}_{origin} &= F_{3d}(\boldsymbol{I}), \\ \boldsymbol{X}_{motion} &= F_{motion}(\boldsymbol{X}_{origin}) \\ \boldsymbol{X}_{appearance} &= F_{backbone}(\boldsymbol{X}_{origin}) \\ \boldsymbol{f}_{view} &= F_{view}(\boldsymbol{X}_{appearance}) \end{aligned}, \quad (4)$$

where $\boldsymbol{X}_{appearance}, \boldsymbol{X}_{motion}, \in \mathbb{R}^{C_2\times H\times W}$, $\boldsymbol{f}_{view} \in \mathbb{R}^{C_3}$ and $F_{backbone}, F_{motion}, F_{view}$ are corresponding branches.

Then we first predicted the view of sequence and then fuse it with $\boldsymbol{X}_{appearance}$ and $\boldsymbol{X}_{motion}$:

$$\begin{aligned} \hat{p} &= F_{predict}(\boldsymbol{f}_{view}), \\ \boldsymbol{f}_{motion} &= F_{fusion_1}(\boldsymbol{X}_{motion}, \hat{p}) \\ \boldsymbol{f}_{appearance} &= F_{fusion_2}(\boldsymbol{X}_{appearance}, \hat{p}) \end{aligned}, \quad (5)$$

where $\hat{p}$ is the predicted view and $\hat{p} \in \mathbb{R}^M$, M is the number of discrete views. $\boldsymbol{f}_{motion}$ and $\boldsymbol{f}_{appearance}$ are the final

[2] Details can be found in supplementary materials

[3] The code will be released at https://github.com/ctrasd/LagrangeGait

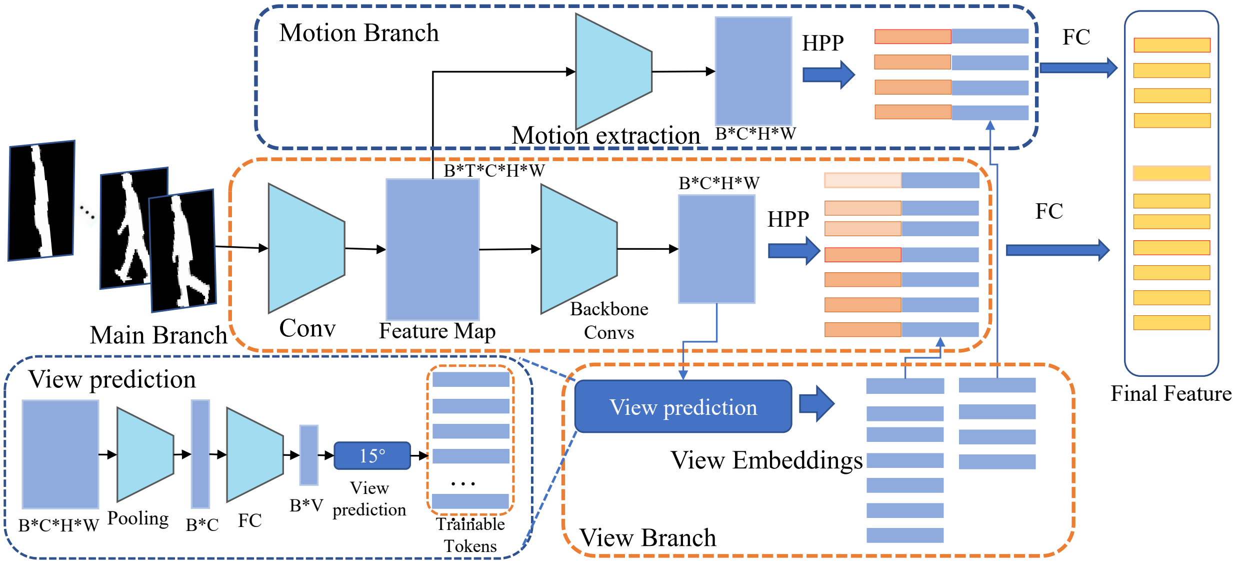

Figure 4. Framework of the proposed *LagrangeGait*.

features of motion and appearance, $\boldsymbol{f}_{motion} \in \mathbb{R}^{n_{motion} \times c_3}$, $\boldsymbol{f}_{appearance} \in \mathbb{R}^{n_{appearance} \times c_3}$. n_{motion}, $n_{appearance}$ denote the number of strips sliced using the HPP module [6] for the motion feature map and appearance feature map. c_3 denote the number of channel of feature maps.

Finally, the feature used for gait recognition can be expressed as

$$\boldsymbol{f}_{final} = [\boldsymbol{f}_{motion}; \boldsymbol{f}_{appearance}] \quad . \qquad (6)$$

4.1. Motion Extraction Module

According to Section 3, we designed a second-order motion extraction module. As shown in Figure 5, 3D convolutions are used as the first-order feature extraction layer. In second-order stage, the structure of RAFT [32] is referenced and adjacent frame response relationships are used.

With $\boldsymbol{X}_{origin}$ obtained from Equation (4), we denote the feature map at time T_i and T_{i+1} as $\boldsymbol{X}_{origin,i}$ and $\boldsymbol{X}_{origin,i+1}$ respectively. Then the correlation of adjacent frames can be calculated as

$$\begin{aligned} \boldsymbol{X_0} &= F_Q(\boldsymbol{X}_{origin,i}) \\ \boldsymbol{X_1} &= F_K(\boldsymbol{X}_{origin,i+1}) \quad , \\ Att(\boldsymbol{X_0}, \boldsymbol{X_1}) &= Softmax(\boldsymbol{X_0}^T \boldsymbol{X_1}) \end{aligned} \qquad (7)$$

where F_Q and F_k are combinations of convolution layers with filter size of 1×1 and dimension merge operation. $Att(\boldsymbol{X_0}, \boldsymbol{X_1}) \in \mathbb{R}^{HW \times HW}$. Then the correlation map should be reshaped to $Cor(\boldsymbol{X_0}, \boldsymbol{X_1}) \in \mathbb{R}^{H \times W \times HW}$. For a pixel in $\boldsymbol{X}_{origin,i}$, its corresponding pixel in the next frame $\boldsymbol{X}_{origin,i+1}$ is assumed not moving too much. So for every pixel $\boldsymbol{x} = (u, v)$ in $\boldsymbol{X}_{origin,i}$, the corresponding point in the feature map $\boldsymbol{X}_{origin,i+1}$ is $\boldsymbol{x}^{'} = (u + f^1(u), v + f^1(v))$. The sampling range is

$$N(x)_r = \{\boldsymbol{x} + \boldsymbol{dx} | \boldsymbol{dx} \in \mathbb{Z}^2, \|\boldsymbol{dx}\|_1 \leq r\}, \qquad (8)$$

where $\boldsymbol{dx}$ is the sampling offset and r is the sampling radius. For each pixel x on $Cor(\boldsymbol{X_0}, \boldsymbol{X_1})$ we sample it according to $N(x)_r$ and $X^{'}_{corr,i} \in \mathbb{R}^{H \times W \times (2r+1)^2}$ can be obtained. Then we exchange the channel and form it into $\boldsymbol{X}_{corr,i} \in \mathbb{R}^{(2r+1)^2 \times H \times W}$.

Finally, the second-order feature maps are integrated in time dimension to obtain the feature map of sequence:

$$X_{corr} = [X_{corr,1}; X_{corr,2}; ...; X_{corr,t-1}]. \qquad (9)$$

Here $X_{corr} \in \mathbb{R}^{(2r+1)^2 \times t-1 \times H \times W}$. and we use 3D convolution to extract the final feature:

$$\boldsymbol{X}_{motion} = F_{3dconv}(X_{corr}), \qquad (10)$$

where F_{3dconv} are convolution layers with kernel size $3 \times 3 \times 3$. $\boldsymbol{X}_{motion} \in \mathbb{R}^{C_2 \times T \times H \times W}$.

4.2. View Embedding

For gait recognition, there are few methods taking view itself into consideration. In this paper we propose a more light-weighted view embedding method.

First we calculate the view feature of input sequence with feature map $\boldsymbol{X}_{origin}$ obtained in Equation (4) as

$$\begin{aligned} \boldsymbol{X}_{appearance} &= P_{Max}(\boldsymbol{X}_{origin}) \\ \boldsymbol{f}_{view} &= P_{Global\_Avg}(\boldsymbol{X}_{appearance}) \end{aligned} , \qquad (11)$$

where P_{Max} is the max pooling at time dimension and P_{Global_Avg} is the global average pooling.

Then with $\boldsymbol{f}_{view}$, the prediction of view can be formulated as

$$\begin{aligned} \hat{p} &= W_{view} \boldsymbol{f}_{view} + B_{view} \\ \hat{y} &= \arg\max_i \hat{p}_i \end{aligned} \quad . \qquad (12)$$

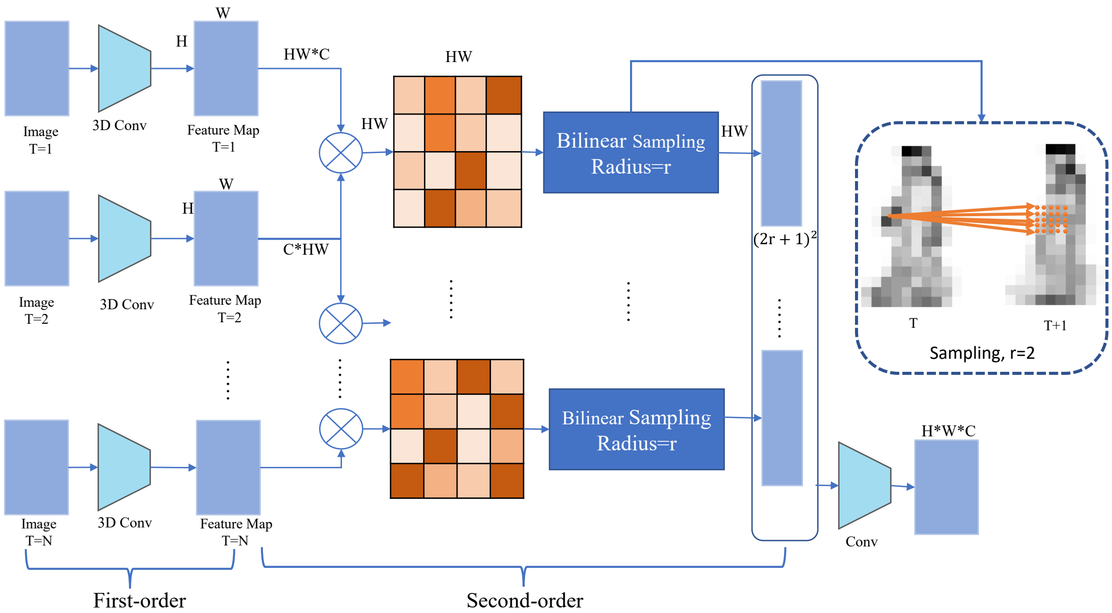

Figure 5. Structure of the second-order motion extraction module. We use 3D convolution as the first-order motion extraction module, and calculate the pixel-to-pixel corresponding matrix as the second-order motion feature. Since a pixel cannot move too far away from the origin location, we do the bi-linear sampling to reduce the computation cost.

Here M is the number of views, $M = 11$ for CASIA-B [41] and for OUMVLP [31] $M = 14$. $W_{view} \in \mathbb{R}^{M \times C_2}$ is the weight of FC layer and B_{view} is the bias of FC layer. $\hat{y} \in \{0, 1, 2, ..., M-1\}$ is the result of view prediction.

For every discrete view $\hat{y}$, we will train two embedding $E_{m,\hat{y}} \in \mathbb{C}_0, E_{a,\hat{y}} \in \mathbb{C}_0$ for motion and appearance feature and they will be used in the horizontal pyramid pooling module [6]. C_0 is the dimension of feature map obtained from the first convolution layer in Figure 4.

4.3. HPP with View Embedding

In gait recognition, horizontal pyramid Pooling (HPP) [6] is a widely used module. In this paper, in addition to using HPP on appearance feature maps, we also do the same operation on the motion feature maps. After pooling, the features are connected with the proposed view embedding to make the final feature projection.

For appearance feature map obtained after horizontal pyramid pooling, we denote them as:

$$\boldsymbol{f}_{app,1}, \boldsymbol{f}_{app,2}, ...\boldsymbol{f}_{app,n}, \tag{13}$$

where n is the number of strips to split, $\boldsymbol{f}_{app,i} \in \mathbb{R}^{C_2}$. For appearance branch and motion branch, the number of strips are $n_{appearance}$ and n_{motion}

Assuming the prdicted view of $\boldsymbol{X}_{appearance}$ is z. Then the procedure of $\boldsymbol{F}_{fusion1}$ can be formulated as:

$$\begin{aligned} \boldsymbol{f}_{av,i} &= [\boldsymbol{f}_{app,i}; E_{a,z}] \\ \boldsymbol{f}_{finala,i} &= W_{p,i}\boldsymbol{f}_{av,i}, i = 1, 2, ...n_{appearance} \\ \boldsymbol{f}_{app} &= [\boldsymbol{f}_{finala,1}, \boldsymbol{f}_{finala,2}, ..., \boldsymbol{f}_{finala,n_{app}}] \end{aligned} . \tag{14}$$

Here $\boldsymbol{f}_{av,i} \in \mathbb{R}^{C_2+C_0}$, $\boldsymbol{f}_{finala,i} \in \mathbb{R}^{C_2}$, $\boldsymbol{f}_{app} \in \mathbb{R}^{n_{app} \times C_2}$.

The procedure of $F_{fusion2}$ is similar with $\boldsymbol{F}_{fusion1}$:

$$\begin{aligned} \boldsymbol{f}_{mv,i} &= [\boldsymbol{f}_{motion,i}; E_{m,z}] \\ \boldsymbol{f}_{finalm,i} &= W_p\boldsymbol{f}_{mv,i}, i = 1, 2, ...n_{motion}; \\ \boldsymbol{f}_{motion} &= [\boldsymbol{f}_{finalm,1}, \boldsymbol{f}_{finalm,2}, ..., \boldsymbol{f}_{finalm,n_{motion}}] \end{aligned} , \tag{15}$$

where $\boldsymbol{f}_{mv,i} \in \mathbb{R}^{C_2+C_0}$, $\boldsymbol{f}_{finalm,i} \in \mathbb{R}^{C_2}$, $\boldsymbol{f}_{motion} \in \mathbb{R}^{n_{motion} \times C_2}$.

Finally, the final feature can be approached by bring Equation (14) and (15) into (6). where $\boldsymbol{f}_{final} \in \mathbb{R}^{(n_{motion}+n_{appearance}) \times C_2}$.

4.4. Joint Losses

In the proposed framework, our losses include cross entropy (CE) and triplet loss. Combining the Equation (12), the CE loss can be expressed as

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} log(p_{ij}) \;\; w.r.t. \;\; p_{ij} = \frac{e^{\hat{p}_{ij}}}{\sum_{j=1}^{M} e^{\hat{p}_{ij}}}, \tag{16}$$

where N is the number of samples, M is the number of views and y_{ij} is whether the view of ith sample is j.

Let a triplet of gait silhouette sequences group 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 $\{T_i | T_i = (f_{final}^{Q_i}, f_{final}^{P_i}, f_{final}^{N_i}), i = 1, 2, ..., K\}$. Then the triplet loss can be expressed as

$$\mathcal{L}_{trip} = \frac{1}{K}\sum_{i=1}^{K}\sum_{j=1}^{n} max(m - d_{ij}^{-} + d_{ij}^{+}, 0), \tag{17}$$

Table 1. Rank-1 accuracy (%) on CASIA-B [41] under 11 probe views excluding identical-view cases. * denotes the methods are trained with additional cross-entropy loss for identity classification. Ours* are trained without view embedding.

Gallery NM#1-4		0°-180°											
Probe		0°	18°	36°	54°	72°	90°	108°	128°	144°	162°	180°	mean
NM#5-6	CNN-Ensemble [37]	88.7	95.1	98.2	96.4	94.1	91.5	93.9	97.5	98.4	95.8	85.6	94.1
	GaitSet [6]	90.8	97.9	**99.4**	96.9	93.6	91.7	95.0	97.8	98.9	96.8	85.8	95.0
	GaitPart [7]	94.1	**98.6**	99.3	**98.5**	94.0	92.3	95.9	98.4	**99.2**	97.8	90.4	96.2
	GaitGL [22]	94.6	97.3	98.8	97.1	95.8	94.3	96.4	98.5	98.6	**98.2**	90.8	96.4
	Ours	**95.2**	97.8	99.0	98.0	**96.9**	**94.6**	**96.9**	**98.8**	98.9	98.0	**91.5**	**96.9**
	GaitGL* [21]	96.0	98.3	99.0	97.9	96.9	95.4	97.0	98.9	**99.3**	98.8	94.0	97.4
	3DLocal* [12]	96.0	99.0	**99.5**	**98.9**	**97.1**	94.2	96.3	**99.0**	98.8	98.5	95.2	97.5
	CSTL* [11]	**97.2**	**99.0**	99.2	98.1	96.2	**95.5**	**97.7**	98.7	99.2	**98.9**	**96.5**	**97.8**
	Ours*	95.7	98.1	99.1	98.3	96.4	95.2	97.5	**99.0**	99.3	**98.9**	94.9	97.5
BG#1-2	CNN-LB [37]	64.2	80.6	82.7	76.9	64.8	63.1	68.0	76.9	82.2	75.4	61.3	72.4
	GaitSet [6]	83.8	91.2	91.8	88.8	83.3	81.0	84.1	90.0	92.2	94.4	79.0	87.2
	GaitPart [7]	89.1	94.8	96.7	95.1	88.3	84.9	89.0	93.5	96.1	93.8	85.8	91.5
	GaitGL [22]	**90.3**	**94.7**	**95.9**	94.0	91.9	86.5	90.5	95.5	97.2	96.3	87.1	92.7
	Ours	89.9	94.5	**95.9**	**94.6**	**93.9**	**88.0**	**91.1**	**96.3**	**98.1**	**97.3**	**88.9**	**93.5**
	GaitGL* [21]	92.6	**96.6**	96.8	95.5	93.5	89.3	92.2	96.5	**98.2**	96.9	**91.5**	94.5
	3DLocal* [12]	92.8	95.9	**97.8**	**96.2**	93.0	87.8	**92.7**	96.3	97.9	**98.0**	88.5	94.3
	CSTL* [11]	91.7	96.5	97.0	95.4	90.9	88.0	91.5	95.8	97.0	95.5	90.3	93.6
	Ours*	**94.2**	96.2	96.8	95.8	**94.3**	**89.5**	91.7	**96.8**	98.0	97.0	90.9	**94.6**
CL#1-2	CNN-LB [37]	37.7	57.2	66.6	61.1	55.2	54.6	55.2	59.1	58.9	48.8	39.4	54.0
	GaitSet [6]	61.4	75.4	80.7	77.3	72.1	70.1	71.5	73.5	73.5	68.4	50.0	70.4
	GaitPart [7]	70.7	85.5	86.9	83.3	77.1	72.5	76.9	82.2	83.8	80.2	66.5	78.7
	GaitGL [22]	76.7	88.3	90.7	86.6	82.7	77.6	83.5	86.5	88.1	83.2	68.7	83.0
	Ours	**81.6**	**91.0**	**94.8**	**92.2**	**85.5**	**82.1**	**86.0**	**89.8**	**90.6**	**86.0**	**73.5**	**86.6**
	GaitGL* [21]	76.6	90.0	90.3	87.1	84.5	79.0	84.1	87.0	87.3	84.4	69.5	83.6
	3DLocal* [12]	**78.5**	88.9	91.0	89.2	83.7	**80.5**	83.2	84.3	87.9	**87.1**	74.7	84.5
	CSTL* [11]	78.1	89.4	91.6	86.6	82.1	79.9	81.8	86.3	88.7	86.6	**75.3**	84.2
	Ours*	77.4	**90.6**	**93.2**	**90.2**	**84.7**	80.3	**85.2**	**87.7**	**89.3**	86.6	71.0	**85.1**

where $d_{ij}^- = \|f_{final,j}^{Q_i} - f_{fina,j}^{N_i}\|_2^2$ and $d_{ij}^+ = \|f_{final,j}^{Q_i} - f_{final,j}^{P_i}\|_2^2$.

Combining Equation (16) and (17), the final loss can be expressed as:

$$\mathcal{L} = \mathcal{L}_{trip} + \lambda_{CE}\mathcal{L}_{CE}, \tag{18}$$

where λ_{CE} is a hyper-parameter.

5. Experiments and Analysis

In order to prove the effectiveness of our framework, we did experiments on CASIA-B dataset [41] and OU-MVLP dataset [31]. Also we did ablation study to verify the effectiveness of each module.

5.1. Datasets and Evaluation Protocols

CASIA-B [41] is a widely used gait dataset. It contains 124 subjects, each contains 11 views and there are ten sequences for every view. The sequences are obtained in three scenarios: walking under normal situation (NM), walking with bag (BG) and walking wearing coat or jacket (CL) respectively. Experiments in this paper are conducted following the LT setting in [37], in which 74 subjects are used for training and the rest are used for testing. The first four sequences of the NM condition (NM#1-4) are kept in gallery and the rest of 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 [31] 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. There are 14 views for every subject and two sequences for every view. During test, sequences with index#01 are kept in gallery and the rest of which contained in index#00 are used for probe.

5.2. Implementation Details

All models proposed are implemented in Pytorch [25] with four Nvidia 1080Ti/2080Ti GPUs. We pre-process images in CASIA-B [41] and OU-MVLP [31] datasets with the same method in [6]. The image size of each silhouette is 64×44. Adam optimizer is used for training and the margin in separate triplet loss is set to 0.2. The λ_{CE} is set to 0.03 on CASIA-B dataset and 0.3 on OU-MVLP dataset.

The GaitGL [22] is used as our backbone and we add our second-order motion extraction module after the first basic 3D convolution and temporal pooling. The learning

Table 2. Rank-1 accuracy on OU-MVLP dataset [31] (%), * means the methods are trained with additional cross-entropy loss for identity classification.

Method	Probe View														mean
	0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	
GEINet	11.4	29.1	41.5	45.5	39.5	41.8	38.9	14.9	33.1	43.2	45.6	39.4	40.5	36.3	35.8
Gaitset	79.5	87.9	89.9	90.2	88.1	88.7	87.8	81.7	86.7	89.0	89.3	87.2	87.8	86.2	87.1
GaitPart	82.6	88.9	90.8	91.0	89.7	89.7	89.9	85.2	**88.1**	90.0	90.1	89.0	89.1	88.2	88.7
GaitGL	84.3	89.8	90.8	91.0	90.5	**90.5**	**90.3**	88.1	87.9	89.6	89.8	88.9	88.9	88.2	89.1
Ours	**84.5**	**89.8**	**91.0**	**91.2**	**90.7**	**90.5**	90.2	**88.5**	87.9	**89.9**	**90.0**	**89.2**	**89.2**	**88.7**	**89.4**
3DLocal*	86.1	**91.2**	**92.6**	**92.9**	**92.2**	**91.3**	**91.1**	86.9	**90.8**	**92.2**	**92.3**	**91.3**	**91.1**	**90.2**	**90.9**
CSTF*	**87.1**	91.0	91.5	91.8	90.6	90.8	90.6	**89.4**	90.2	90.5	90.7	89.8	90.0	89.4	90.2
GaitGL*	84.9	90.2	91.1	91.5	91.1	90.8	90.3	88.5	88.6	90.3	90.4	89.6	89.5	88.8	89.7
Ours*	85.9	90.6	91.3	91.5	91.2	91.0	90.6	88.9	89.2	90.5	90.6	89.9	89.8	89.2	90.0

rate is set to 3e-4 for CASIA-B dataset and 1e-4 for OU-MVLP dataset at first. The batch size is set to (32,8) for CASIA-B and (20,8) for OU-MVLP dataset. For CASIA-B dataset, we trained our model for 70K iterations. And for OU-MVLP dataset, the number of training iteration is 200K. The learning rate will be reset to 1e-5 at the 160K iteration. 30 frames are sampled from one sequence during training period and all frames are used during test period.

5.3. Comparison with State-of-the-Arts

Evaluation on CASIA-B [41]. The rank-1 accuracy(%) on CASIA-B [41] are shown in Table 1. Methods with * are trained with additional cross-entropy loss for identity classification. We find that joint loss with $\mathcal{L}_{CE}$ and cross-entropy loss for identity classification can lead to severe overfitting according to our experiments and the performance is poor. So the results of "Ours*" are reported without $\mathcal{L}_{CE}$ in Equation (16) and view embedding.

We compared our method with the latest methods, including CNN-Ensemble [37], GaitSet [6], GaitPart [7], GaitGL [21, 22], 3D-Local [12] and CSTL [11]. It can be seen that our method has a large improvement under all conditions compared to all other methods when the cross-entropy loss of identification is not used. The recognition accuracy of our method in conditions NM, BG and CL is 96.9%, 93.5%, 86.6%, which outperforms the original GaitGL [22] by 0.5%, 0.8% and 3.6% respectively. They are big step ups from the already high performance.

Methods using the cross-entropy loss for identity classification have relatively high performances. Our approach is almost equal to the previous SOTA performance at NM and BG settings and the rank-1 accuracy is much higher than that of other methods. The average rank-1 accuracy of our method is 0.56%, 0.3%, 0.53% higher than GaitGL [21], 3DLocal [12] and CSTL [11].

Evaluation on OU-MVLP [31]. The rank-1 accuracy(%) on OU-MVLP dataset are shown in Table 2. As can be seen, our method consistently outperforms SOTA except for the 90° and 195° compared with methods which do not use cross-entropy loss of identification. When compared with methods in which additional cross-entropy loss for identity classification is used, our method make some progress based on the baseline GaitGL [21] which proves the validity of motion extraction module. It can be seen that ours is 0.3% higher than GaitGL [22]. Although, we can not outperform 3DLocal [12] and CSTF [11], we argue that our method does not conflict with the contribution points of them. The code of them are not open source now and their model are complex enough, It's not easy to do experiments with our modules on them. We believe that if we integrate our ideas with theirs, we can get better performance.

Table 3. Ablation study on motion and view(Rank-1,%).

	NM	BG	CL	Mean
Baseline(GaitGL) [22]	96.4	92.7	83.0	90.7
+motion	96.8	93.1	84.7	91.5
+motion+view(Supervision)	96.7	93.5	85.3	91.8
+motion+view embedding	**96.9**	**93.5**	**86.5**	**92.3**

5.4. Ablation Study

In this paper, we propose the second-order motion extraction module and view-embedding. To verify the validity of each module, we performed ablation study on them. In addition, we explored the effect of sampling radius r in Equation (8) and the length C_0 of the view-embedding in Equation (14) and Equation (15). Due to the richness of CASIA-B data type, we performed ablation study on it.

Analysis of Motion Extraction and View embedding. The effects of second-order motion extraction module and view embedding are shown in Table 3. It can be seen that both motion and view embedding make sense. The motion extraction module can boost from 90.7% to 91.5% on average. The only use of view as the training supervision boost from 91.5% to 91.8% and the whole use of view embedding can boost another 0.5% on average. Among them, the addition of motion improves the rank-1 accuracy under all three conditions. The addition of view can mainly improve the accuracy under BG and CL condition.

Analysis of sample radius and length of embedding.

The result are shown in Table 4. It can be seen that the the influence of length of embedding and the radius are negligible and do not cause serious impact. However, it can be seen that embedding length of 32 is preferable while the radius of 3 is also much better.

Table 4. Ablation study on sample radius and length of embedding (Rank-1, %)

Em_length	Radius	NM	BG	CL	Mean
32	2	96.8	93.3	85.6	91.9
64	2	96.6	93.5	84.7	91.6
32	3	**96.9**	**93.5**	**86.6**	**92.3**
64	3	96.4	93.3	86.5	92.0
32	4	96.6	93.1	85.4	91.7
64	4	96.4	93.3	86.2	92.0

Analysis of model generalization. In order to verify the generalization of our proposed second-order motion extraction module and view embedding, we replaced the backbone in our framework with another widely used open source network GaitSet [6]. The result are shown in Table 5. It can be seen that our method make a huge improvement in the mean rank-1 accuracy. The proposed modules mainly make increase on BG and CL and the motion extraction module make a relatively greater contribution. It is consistent with common sense that motion extraction module improves BG and CL more since sequences under BG and CL condition will have more difference in appearance from probe sequence under NM condition.

Table 5. Rank-1 accuracy(%) on CASIA-B [41] with Gaitset [6] as backbone.

	NM	BG	CL	Mean
Baseline(Gaitset) [6]	95.0	87.2	70.4	84.2
+motion	**95.0**	87.3	73.3	85.2
+motion+view embedding	94.9	**87.8**	**73.3**	**85.4**

5.5. Visualization

In order to intuitively show the functions of our proposed module, we visualize the extracted second-order motion feature map. As shown in Figure 6, the proposed second-order motion extraction module can surely capture the motion information, including the motion direction and the motion distance. Also, we visualized the extracted motions from different pedestrian using the standard optical flow visualization method. Results are shown in Figure 7, it can be seen that the visualization for different pedestrian are discriminable.

6. Limitations and Negative impact

In this paper, though the *Lagrange*'s equation is used to analyze the pedestrian motion and come to the conclusion about second-order motion, we believe there will be a better network matching the Equation 3. Also, due to the lack of open source codes of some SOTA methods at this time, we cannot further demonstrate the superiority of proposed theory on their basis.

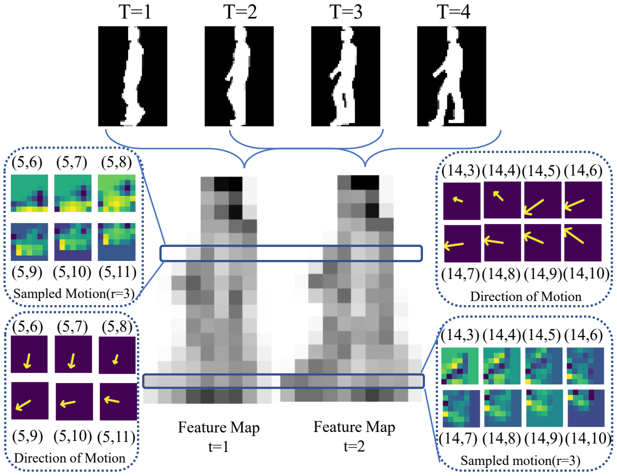

Figure 6. Visualization of extracted motion. (x,y) denotes the location of pixel in feature map t=1. The feature map is visualized with mean operation at the channel dimension. In the sampled corresponding map, the brighter the pixel, the more similar the pixel in the t=2 frame to the frame at the (x, y) position of the t=1 frame. It can be found that the extracted features can well represent the motion direction and distance of each pixel.

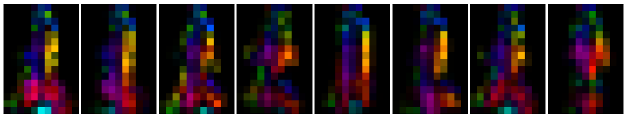

Figure 7. Visualization of extracted motion feature map with optical flow format.

Since this study is an application of biometrics technology to intelligent video surveillance, possible privacy issue may arise if the deployment is not under authorization. Strict monitoring management and personal data protection are the keys to avoiding the negative impact of the proposed method.

7. Conclusion

In this paper, we first model the walking of human with *Lagrange*'s Equation and illustrate that second-order information is necessary to distinguish between two pedestrian with similar appearance. According to this conclusion, we explain why the current methods with 3D convolution can approach better performance and proposed a novel second-order motion extraction module. Besides, we put forward a lightweight view-embedding to reduce the intra-class distance caused by the change of view. Our experiment on two widely-used dataset proved our idea. We hope that proposed modules or just some ideas can bring convenience to future work. Our motion analysis framework may also be used in other fields, such as video-based person re-identification.

8. Acknowledgements

This work was supported by the Key Program of National Natural Science Foundation of China (Grant No. U20B2069).

References

[1] Irene Amerini, Leonardo Galteri, Roberto Caldelli, and Alberto Del Bimbo. Deepfake video detection through optical flow based cnn. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 0–0, 2019. 2

[2] Weizhi An, Shiqi Yu, Yasushi Makihara, Xinhui Wu, Chi Xu, Yang Yu, Rijun Liao, and Yasushi Yagi. Performance evaluation of model-based gait on multi-view very large population database with pose sequences. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(4):421–430, 2020. 2

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6299–6308, 2017. 2

[4] Tianrui Chai, Xinyu Mei, Annan Li, and Yunhong Wang. Semantically-guided disentangled representation for robust gait recognition. In *Int. Conf. Multimedia and Expo*, pages 1–6. IEEE, 2021. 1

[5] Tianrui Chai, Xinyu Mei, Annan Li, and Yunhong Wang. Silhouette-based view-embeddings for gait recognition under multiple views. In *IEEE Int. Conf. Image Process.*, pages 2319–2323, 2021. 2

[6] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *AAAI*, volume 33, pages 8126–8133, 2019. 1, 2, 3, 4, 5, 6, 7, 8

[7] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. Gaitpart: Temporal part-based model for gait recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 14225–14233, 2020. 1, 2, 6, 7

[8] Davrondzhon Gafurov, Kirsi Helkala, and Torkjel Søndrol. Biometric gait authentication using accelerometer sensor. *J. comput.*, 1(7):51–59, 2006. 2

[9] Yiwei He, Junping Zhang, Hongming Shan, and Liang Wang. Multi-task gans for view-specific feature learning in gait recognition. *IEEE Transactions on Information Forensics and Security*, 14(1):102–113, 2018. 2

[10] Maodi Hu, Yunhong Wang, Zhaoxiang Zhang, James J Little, and Di Huang. View-invariant discriminative projection for multi-view gait-based human identification. *IEEE Transactions on Information Forensics and Security*, 8(12):2034–2045, 2013. 1

[11] Xiaohu Huang, Duowang Zhu, Hao Wang, Xinggang Wang, Bo Yang, Botao He, Wenyu Liu, and Bin Feng. Context-sensitive temporal feature learning for gait recognition. In *Int. Conf. Comput. Vis.*, pages 12909–12918, October 2021. 2, 6, 7

[12] Zhen Huang, Dixiu Xue, Xu Shen, Xinmei Tian, Houqiang Li, Jianqiang Huang, and Xian-Sheng Hua. 3d local convolutional neural networks for gait recognition. In *Int. Conf. Comput. Vis.*, pages 14920–14929, October 2021. 2, 3, 6, 7

[13] J.L. Lagrange. *Mécanique analytique*. Number 1 in Mécanique analytique. Ve Courcier, 1811. 2

[14] Jiefeng Li, Siyuan Bian, Ailing Zeng, Can Wang, Bo Pang, Wentao Liu, and Cewu Lu. Human pose regression with residual log-likelihood estimation. In *Int. Conf. Comput. Vis.*, pages 11025–11034, October 2021. 2

[15] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, and Mingwu Ren. Joint intensity transformer network for gait recognition robust against clothing and carrying status. *IEEE Transactions on Information Forensics and Security*, 14(12):3102–3115, 2019. 1

[16] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, and Mingwu Ren. Gait recognition via semi-supervised disentangled representation learning to identity and covariate features. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13309–13319, 2020. 1

[17] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, Shiqi Yu, and Mingwu Ren. End-to-end model-based gait recognition. In *Proceedings of the Asian conference on computer vision*, 2020. 2

[18] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *Int. Conf. Comput. Vis.*, pages 11313–11322, October 2021. 2

[19] Rijun Liao, Shiqi Yu, Weizhi An, and Yongzhen Huang. A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognition*, 98:107069, 2020. 2

[20] Beibei Lin, Shunli Zhang, and Feng Bao. Gait recognition with multiple-temporal-scale 3d convolutional neural network. In *ACM Multimedia*, pages 3054–3062, 2020. 1, 2, 3

[21] Beibei Lin, Shunli Zhang, and Xin Yu. Gait recognition via effective global-local feature representation and local temporal aggregation. In *Int. Conf. Comput. Vis.*, pages 14648–14656, October 2021. 1, 3, 6, 7

[22] 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. 6, 7

[23] M Pat Murray, A Bernard Drought, and Ross C Kory. Walking patterns of normal men. *JBJS*, 46(2):335–360, 1964. 2

[24] Thanh Trung Ngo, Yasushi Makihara, Hajime Nagahara, Yasuhiro Mukaigawa, and Yasushi Yagi. The largest inertial sensor-based gait database and performance evaluation of gait-based personal authentication. *Pattern Recognition*, 47(1):228–237, 2014. 2

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inform. Process. Syst.*, 32:8026–8037, 2019. 6

[26] AJ Piergiovanni and Michael S Ryoo. Representation flow for action recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9945–9953, 2019. 2

[27] Liu Rong, Zhou Jianzhong, Liu Ming, and Hou Xiangfeng. A wearable acceleration sensor system for gait recognition. In *IEEE Conference on Industrial Electronics and Applications*, pages 2654–2659, 2007. 2

[28] Chunfeng Song, Yongzhen Huang, Yan Huang, Ning Jia, and Liang Wang. Gaitnet: An end-to-end network for gait based human identification. *Pattern Recognition*, 96, 2019. 1

[29] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8934–8943, 2018. 2

[30] Yan Sun, Jonathon S Hare, and Mark S Nixon. Detecting acceleration for gait and crime scene analysis. In *7th International Conference on Imaging for Crime Detection and Prevention*, pages 1–6. IET, 2016. 2

[31] 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*, 10(1):4, 2018. 5, 6, 7

[32] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 2, 3, 4

[33] Torben Teepe, Ali Khan, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Gaitgraph: Graph convolutional network for skeleton-based gait recognition. *arXiv preprint arXiv:2101.11228*, 2021. 2

[34] Chen Wang, Junping Zhang, Liang Wang, Jian Pu, and Xiaoru Yuan. Human identification using temporal information preserving gait template. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(11):2164–2176, 2011. 1

[35] Yanyun Wang, Chunfeng Song, Yan Huang, Zhenyu Wang, and Liang Wang. Learning view invariant gait features with two-stream gan. *Neurocomputing*, 339:245–254, 2019. 1

[36] Thomas Wolf, Mohammadreza Babaee, and Gerhard Rigoll. Multi-view gait recognition using 3d convolutional neural networks. In *IEEE Int. Conf. Image Process.*, 2016. 2, 3

[37] 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 Trans. Pattern Anal. Mach. Intell.*, 39(2):209–226, 2016. 1, 2, 6, 7

[38] Chi Xu, Yasushi Makihara, Xiang Li, Yasushi Yagi, and Jianfeng Lu. Cross-view gait recognition using pairwise spatial transformer networks. *IEEE Trans. Circuit Syst. Video Technol.*, 31(1):260–274, 2020. 1

[39] Wei Xue, Hong Ai, Tianyu Sun, Chunfeng Song, Yan Huang, and Liang Wang. Frame-gan: Increasing the frame rate of gait videos with generative adversarial networks. *Neurocomputing*, 380:95–104, 2020. 1

[40] Zhichao Yin, Trevor Darrell, and Fisher Yu. Hierarchical discrete distribution decomposition for match density estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6044–6053, 2019. 2

[41] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *Int. Conf. Pattern Recog.*, volume 4, pages 441–444. IEEE, 2006. 1, 5, 6, 7, 8

[42] Kaihao Zhang, Wenhan Luo, Lin Ma, Wei Liu, and Hongdong Li. Learning joint gait representation via quintuplet loss minimization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4700–4709, 2019. 1

[43] Shaoxiong Zhang, Yunhong Wang, and Annan Li. Cross-view gait recognition with deep universal linear embeddings. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9095–9104, June 2021. 1

[44] Xiheng Zhang, Yongkang Wong, Xiaofei Wu, Juwei Lu, Mohan Kankanhalli, Xiangdong Li, and Weidong Geng. Learning causal representation for training cross-domain pose estimator via generative interventions. In *Int. Conf. Comput. Vis.*, pages 11270–11280, October 2021. 2

[45] Yuqi Zhang, Yongzhen Huang, Liang Wang, and Shiqi Yu. A comprehensive study on gait biometrics using a joint cnn-based method. *Pattern Recognition*, 2019. 1

[46] Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Xiaoming Liu, Jian Wan, and Nanxin Wang. Gait recognition via disentangled representation learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4710–4719, 2019. 1