GaitPart: Temporal Part-based Model for Gait Recognition

Chao Fan[1,3], Yunjie Peng[2,3], Chunshui Cao[3], Xu Liu[3], Saihui Hou,[3]
Jiannan Chi[1]*, Yongzhen Huang[3], Qing Li[1], Zhiqiang He[4,2]
[1] University of Science and Technology Beijing,
[2] Beihang University, [3] WATRIX.AI, [4] Lenovo Ltd
s20180566@xs.ustb.edu.cn, YunjiePeng@buaa.edu.cn, {chunshui.cao, xu.liu, saihui.hou}@watrix.ai,
ustbjnc@ustb.edu.cn, hyz@watrix.ai, liqing@ies.ustb.edu.cn, hezq@lenovo.com

Abstract

Gait recognition, applied to identify individual walking patterns in a long-distance, is one of the most promising video-based biometric technologies. At present, most gait recognition methods take the whole human body as a unit to establish the spatio-temporal representations. However, we have observed that different parts of human body possess evidently various visual appearances and movement patterns during walking. In the latest literature, employing partial features for human body description has been verified being beneficial to individual recognition. Taken above insights together, we assume that each part of human body needs its own spatio-temporal expression. Then, we propose a novel part-based model GaitPart and get two aspects effect of boosting the performance: On the one hand, Focal Convolution Layer, a new applying of convolution, is presented to enhance the fine-grained learning of the part-level spatial features. On the other hand, the Micro-motion Capture Module (MCM) is proposed and there are several parallel MCMs in the GaitPart corresponding to the pre-defined parts of the human body, respectively. It is worth mentioning that the MCM is a novel way of temporal modeling for gait task, which focuses on the short-range temporal features rather than the redundant long-range features for cycle gait. Experiments on two of the most popular public datasets, CASIA-B and OU-MVLP, richly exemplified that our method meets a new state-of-the-art on multiple standard benchmarks. The source code will be available on https://github.com/ChaoFan96/GaitPart.

1. Introduction

Gait is a kind of physical and behavioural biometric characteristic that depicts the walking patterns of a person. Compared with other biometric modalities, *e.g.*, face, fingerprint and iris, it can be easily captured at a long-distance and requires no explicit co-operation of interest-subjects. Thus, gait recognition has enormous potential in crime investigation, access control and social security. As an identification task in vision, the essential goal of gait recognition is to learn the unique and invariant representations from the temporal changing characteristics about human body shape. However, in real-world scenarios, variations like bag-carrying, coat-wearing and camera viewpoints cause dramatic changes in gait appearance, which bring significant challenges to gait recognition.

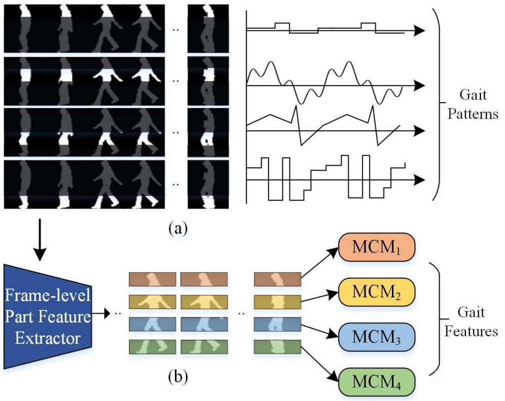

Figure 1. (a): Different parts of human gait possess evidently different shapes and moving patterns during walking. (b): Overview of the GaitPart, consisting of the Frame-level Part Feature Extractor(FPFE) and Micro-motion Capture Module(MCM).

To alleviate these issues, lots of deep-learning based methods have provided promising solutions[30, 25, 5, 26, 18, 29, 21, 14]. Thomas *et al.*[25] applied 3D-CNN to extract the spatio-temporal information, trying to find a general descriptor for human gait. GaitNet[30] proposed an Auto-Encoder framework to extract the gait-related features from raw RGB images and then used LSTMs to model the temporal changes of gait sequence. And GaitSet[5] assumed that the appearance of a silhouette has contained its

[*] Corresponding Author

position information and thus regarded gait as a set to extract temporal information.

These prior methods treat the whole human body shape as a unit to extract the spatio-temporal information for final identification. However, we observe that the different parts of human body possess evidently various shapes and moving patterns during walking, as shown in Fig. 1(a). More evidences from [6, 15, 20, 5, 12, 13, 19, 28] have implied that partial features for human body description can offer fine-grained information, which is conducive to individual identification task. For the temporal changing characteristics, some of these state-of-art methods do not model temporal features explicitly, which leads to the loss of important invariant features in time series[23, 7, 3, 26]. Some other methods model the long-range dependencies to represent the global understanding of gait sequence, via deeply stacking 3D-convolutional or recurrent operations[25, 30]. However, these methods are believed to retain unnecessary long-range sequential constraints for the periodic gait and thus lose the flexibility of gait recognition[5].

Motivated by above findings, we assume that each part of human body needs its own expression, in which the local short-range spatio-temporal features (micro-motion patterns) are the most discriminative characteristics for human gait. Therefore, we propose a novel temporal part-based framework called **GaitPart**. As shown in Fig. 1(b), GaitPart consists of two novel well-designed components, *i.e.*, the Frame-level Part Feature Extractor(**FPFE**) and Micro-motion Capture Module(**MCM**).

The input of GaitPart is a sequence of gait silhouettes. The FPFE, a special but concise stacked CNN, firstly takes each frame as input and then conducts pre-defined horizontal partition on the output feature map. In this way, we can obtain several sequences of part-level spatial feature colored in Fig. 1(b), each of which corresponds to a certain pre-defined part of human body and its micro-motion patterns will be captured by the corresponding MCM. Note that the parameters among these parallel MCMs are independent, which reflects GaitPart is a part-independent approach. And the final gait representations is formed by simply concatenating all the output of these MCMs. More specifically, we make the following three major contributions.

- In FPFE, we propose a simple yet effective applying of convolution, called Focal Convolution (**FConv**), to enhance the fine-grained learning of the part-level spatial features. Its core idea is to enable top convolution kernel focus on more local details inside each certain part of the input frame, intuitively exploiting more fine-grained partial information.
- In MCM, we argue that the local short-range spatio-temporal features(micro-motion patterns) are the most discriminative features for periodic gait while the long-range dependencies are lengthy and inefficient. And more, an attention-based MCM is proposed to model the local micro-motion features and the global understanding of entire gait sequence.
- We propose a novel temporal part-based framework for gait recognition, called GaitPart. Extensive experiments conducted on the widely used gait databases, CASIA-B[17] and OU-MVLP[21], demonstrate that GaitPart outperforms prior state-of-the-art methods by a large margin, showing its superiority. Lots of rigorous ablation experiments conducted on CASIA-B[17] further prove the effectiveness of each component within GaitPart.

2. Related Work

Gait Recognition. Most state-of-art works have taken spatial feature extraction and temporal modeling as the focus[16, 23, 7, 3, 26, 2, 1]. For the first issue, prior CNN-based studies often performed the regular either 2-D[30, 5] or 3-D[25, 27] convolution operation on entire feature map. While this uniform operation in spatial dimension is naturally and widely employed, these methods ignore the significant differences among human body parts in gait task.

To get the spatio-temporal representations, many works tend to explicitly model the temporal changes[30, 25] or directly compress the whole gait sequence into one frame[5, 23, 7, 3, 26, 29]. However, RNN-based methods are believed to retain unnecessary sequential constrains for the periodic gait[5], while another kind of GEI-based methods[23, 7, 3, 26] are sensitive to the variations in real-world scenarios, despite the simplicity advantage of them.

Part-based model. Splitting the feature map into strips and pooling them into column vectors have been commonly used in the very related field. *e.g.*, person Re-ID[6, 20, 15]. With ignoring the spatial alignment, these methods assume that each column vector could represent a certain corresponding part of human body[20].

Different from the field of person Re-ID, we argue that the part-based schemas applied in gait task should be designed in a part-dependent way. Since there are significant differences among human body parts in terms of appearance and moving patterns in gait task, while it is entirely possible that different parts of human body share the common attributes, *e.g.*, color and texture in person Re-ID [1]. Thus, GaitPart have been designed as a part-dependent approach, whose parameters are part-dependent at the stage of generating the spatio-temporal representations.

And more, this paper proposes the Focal Convolution (FConv), which is a novel applying of convolution and has made up the FPFE. More specifically, the FConv first slices the input feature map into several parts and then performs the regular convolution over these parts separately. When

[1] We guess that's part of the reason why prior works[6, 20, 15] tended to use part-shared kernels in the field of person Re-ID.

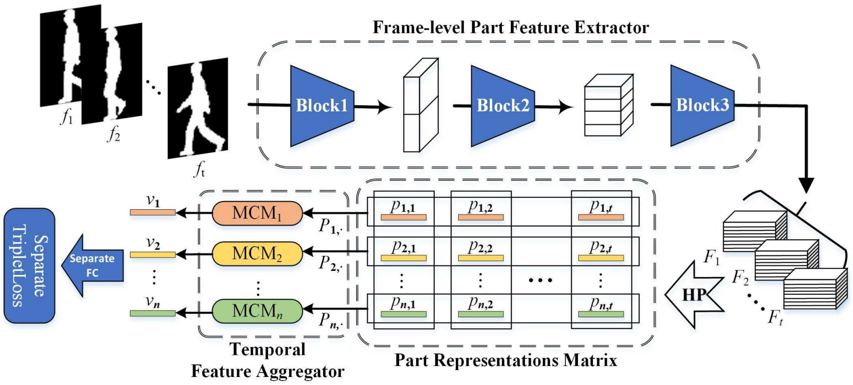

Figure 2. The framework of GaitPart. The Block1, 2 and 3 are composed of FConvs and pooling layers. The HP represents Horizontal Pooling and MCM represents Micro-motion Capture Module. In particular, MCM$_j$ is responsible for aggregating all the vectors at row j in Part Representation Matrix to generate the spatio-temporal feature v_j for the final identification.

deeply stacking the FConv, the receptive field of top-layer neurons will be restricted and is expected to focus on more local details inside the corresponding part of input frame.

Temporal model. The approaches of modeling temporal changes of gait can be generally divided into three categories: 3DCNN-Based [25], LSTM-based[30, 14] and Set-based[5]. Among them, the 3DCNN-based methods[2, 1, 22, 25] directly extract the spatio-temporal features for gait recognition, but these methods are usually difficult to train and can't bring much considerable performance. Zhang *et al*.[30] proposed a novel Auto-Encoder framework to extract the pose features from raw RGB video, and used a three-layer LSTM to aggregate those pose features in time series to generate the final gait feature[30]. However, the LSTM-based methods are believed to retain unnecessary sequential constraints for periodic gait[5]. By assuming the appearance of a silhouette has contained its position information, GaitSet[5] proposed to regard the gait as a set and extracted the spatio-temporal features in a temporal pooling way. This way is concise and effective enough but doesn't model the temporal changes explicitly.

In contrast to above, we observe that the frames with a similar visual appearance are likely to appear periodically in the periodic gait, indicating there would be no discriminative information gain after a complete gait cycle. This phenomenon implies that the long-range dependencies(longer than a complete gait cycle, for example) may be redundant and ineffective for gait recognition. Thus, GaitPart turns the attention to local short-range temporal modeling, and proposes the Micro-motion Capture Module. More details will be discussed in Sec.3.3.

3. Proposed Method

In this section, we first present the pipeline of GaitPart, followed by the Frame-level Part Feature Extractor (**FPFE**), and end with the Temporal Feature Aggregator (**TFA**) and implementation details. The framework is shown in Fig.2.

3.1. Pipeline

As shown in Fig.2, a sequence of gait silhouettes containing t frames is fed into GaitPart frame by frame. The Frame-level Part Feature Extractor (**FPFE**), a specially designed convolution network, is used to extract the spatial features F_i for each frame f_i

$$F_i = \mathrm{FPFE}\left(f_i\right) \tag{1}$$

where $i \in 1, 2, ..., t$ denotes the index of frame in gait sequence, and the details of FPFE will be introduced in Sec.3.2. In this way, a sequence of feature maps denoted as $S_F = \{F_i | i = 1, ..., t\}$ can be obtained, where F_i is a three-dimensional tensor, *i.e.*, the channel, height and width dimension.

Then, the Horizontal Pooling(**HP**) module, aiming at extracting the discriminative part-informed features of partial human body, horizontally splits the feature map F_i into n parts. For the j-th part of F_i, $F_{j,i}$, the HP module downsamples it into a column vector $p_{j,i}$ by Global Average Pooling and Global Max Pooling

$$p_{j,i} = \mathrm{Avgpool2d}\left(F_{j,i}\right) + \mathrm{Maxpool2d}\left(F_{j,i}\right) \tag{2}$$

the similar operations are commonly used in [5, 6, 20, 15].

As an intermediate result, each feature map in S_F can be transformed into n part-level feature vectors, from which the Part Representation Matrix (**PR-Matrix**) can be obtained, denoted as $P = (p_{j,i})_{n \times t}$. As shown in Fig.2, obviously, the corresponding row of vectors in PR-Matrix, denoted as $P_{j,\cdot} = \{p_{j,i} | i = 1, 2, ..., t\}$, is expected to represent the gait changes of part j. Thus, it comes naturally that the spatio-temporal feature of part j could be extracted by aggregating $P_{j,\cdot}$ into a feature vector v_j, formulated as

$$v_j = \mathrm{MCM}_j\left(P_{j,\cdot}\right) \tag{3}$$

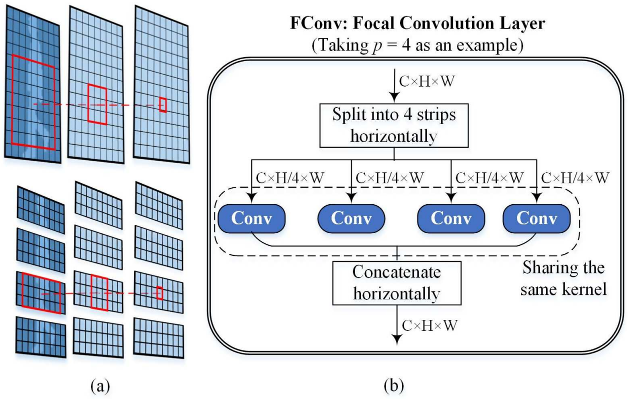

Figure 3. (a): The expansion of top-layer neurons' receptive field in deep network. Top: the regular case. Bottom: using the FConvs. (b): The illustration of FConv, and the feature maps are shown by their dimensions, *e.g.*, C×H×W.

where MCM$_j$ denotes the j-th Micro-motion Capture Module (**MCM**). And there are n parallel MCMs, whose parameters are independent, making up the Temporal Feature Aggregator (**TFA**). In the end, several separate FC layers are employed to map the feature vectors extracted from TFA to metric space for the final individual identification.

3.2. Frame-level Part Feature Extractor

The Frame-level Part Feature Extractor (**FPFE**), aiming at extracting the part-informed spatial features for each frame, is composed of three blocks and each block consists of two Focal Convolution Layers (**FConv**). Next, the FConv will be described in detail first, and followed by the exact structure of FPFE.

Definition. The FConv, a novel applying of convolution, could first split the input feature map into several parts horizontally and then perform a regular convolution over each part, separately. Let p be the number of pre-defined parts, in particular, the FConv is equivalent to the regular convolution layer when $p = 1$.

Motivation. In order to enhance the fine-grained learning of part-informed spatial features, the FConv is developed. As shown in Fig.3(a), with the network going deeper, the receptive field of the top-layer neurons will be restricted to be narrower than the normal case by setting the hyper-parameter p in FConv, which makes it possible for the top-layer neurons to focus on more local details inside the corresponding part of input frame even in a deep network. This constraint on the receptive field is expected to extract more fine-grained and precisely features for each part.

Operation. As shown in Fig.3(b), the input feature map is first split into p pre-defined parts horizontally and then a regular convolution operation will be performed over these part, separately. After that, the output feature maps will be concatenated horizontally and used as the final output of FConv. And more, the exact structure of FPFE are shown in Tab.1 and the ablation study of setting the hyper-parameter p for each FConv will be discussed in Sec.4.3.

Table 1. The exact structure of Frame-level Part Feature Extractor. In_C, Out_C, Kernel and Pad represent the input channels, output channels, kernel size and padding of the FConv, respectively. In particular, p indicates the number of pre-defined parts in FConv.

Frame-level Part Feature Extractor						
Block	Layer	In_C	Out_C	Kernel	Pad	p
Block1	FConv1	1	32	5	2	1
	FConv2	32	32	3	1	1
	MaxPool, kernel size =2, stride=2					
Block2	FConv3	32	64	3	1	4
	FConv4	64	64	3	1	4
	MaxPool, kernel size =2, stride=2					
Block3	FConv5	64	128	3	1	8
	FConv6	128	128	3	1	8

3.3. Temporal Feature Aggregator

As mentioned in Sec.3.1, the Temporal Feature Aggregator (**TFA**) is composed of n parallel Micro-motion Capture Modules (**MCMs**) and each MCM is responsible for modeling the short-range spatio-temporal representations of the corresponding part. Next, we take the details of MCM as focus and thus ignore the index of pre-defined parts.

The MCM contains two parts: the Micro-motion Template Builder (**MTB**) and Temporal Pooling (**TP**). Let $S_p = \{p_i | i = 1, 2, ..., t\}$ be a row of the PR-Matrix, which is a two-dimensional tensor with the sequence and channel dimension. The MTB is designed to map the sequence of part-level feature vectors S_p into the sequence of micro-motion feature vectors S_m, formulated as $S_m = \mathrm{MTB}(S_p)$. After that, by aggregating the sequence S_m, the TP module will extract the most discriminative motion feature vector v, formulated as $v = \mathrm{TP}(S_m)$, for the final identification. Next, the MTB module will be described in detail first, and followed by the TP module.

Micro-motion Template Builder

Description. Map the frame-level part-informed feature vectors into the micro-motion feature vectors.

Motivation. Assume the short-range spatio-temporal representations (micro-motion features) are the most discriminative features for the cycle gait, and think the micro-motion patterns at any certain moment should be totally determined by itself and its neighbor frames.

Operation. Let $R\left(p_i, r\right) = \{p_k | k = i - r, ..., i, ..., i + r\}$ represent the sub-sequence composed of p_i and its r-neighbor frames, and then the micro-motion feature at moment i can be defined as

$$m_i = \mathrm{TempFunc}\left(R\left(p_i, r\right)\right) \tag{4}$$

where the TempFunc denotes micro-motion template function, and aim at compressing the sub-sequence $R\left(p_i, r\right)$.

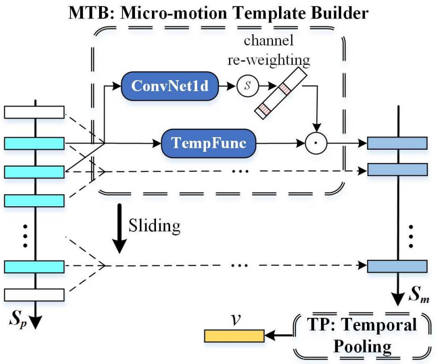

Figure 4. The detailed structure of Micro-motion Capture Module (MCM, including the MTB and TP module). The MTB module slides on sequence dimension, and aggregates each adjacent $2r+1$ column vectors into a micro-motion feature vector. And then, the TP module applies a simple max function to gather the most discriminative micro-motion features among frames and channels dimension for the final recognition.

Refer to the GEI's practice of taking the average of all frames in the sequence as the spatio-temporal representation for gait[7], we join two statistical functions as the instantiation of template function. As shown in Fig.4, applying template function to every moment of S_p, we are actually performing 1-D Global Average Pooling and 1-D Global Max Pooling with the kernel size of $2r+1$. In this way, the sequence of micro-motion feature vectors, S_m, will be obtained, and can be formulated as

$$S_m = \text{Avgpool1d}\left(S_p\right) + \text{Maxpool1d}\left(S_p\right) \tag{5}$$

Further, in order to obtain the more discriminative representations for micro-motion, the channel-wise attention mechanism is introduced to re-weight the feature vector at each moment[9, 24, 12, 4]. In practice, 1-D convolutional kernel is employed and the re-weighted micro-motion sequence S_m^{re} can be formulate as

$$\begin{aligned} S_{logits} &= \text{Conv1dNet}\left(S_p\right) \\ S_m^{re} &= S_m \cdot \text{Sigmoid}\left(S_{logits}\right) \end{aligned} \tag{6}$$

where the Conv1dNet denotes a small network composed of two 1-D convolution layers.

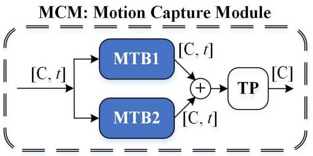

Figure 5. The abstract structure of Micro-motion Capture Module in practice, containing the TP and two parallel MTBs module with different window size (3 and 5).

Table 2. The exact structure of MTB1 and MTB2. In_C, Out_C, Kernel and Pad represent the input channels, output channels, kernel size and padding of the 1-D convolution layer, respectively. In particular, C and s represent the channels of input feature map and the squeeze ratio, respectively. Note that the values around '|' represent the setting of MTB1 and MTB2, respectively.

Module	MTB1 | MTB2			
Layer	Conv1d_1	Conv1d_2	Avgpool1d	Maxpool1d
In_C	C | C	C/s | C/s	×	×
Out_C	C/s | C/s	C | C	×	×
Kernel	3 | 3	1 | 3	3 | 5	3 | 5
Pad	1 | 1	0 | 1	1 | 2	1 | 2

As shown in Fig.4, the MTB is just like a sliding-window detector. On the one hand, all the frame-level feature vectors inside window will be compressed into a micro-motion vector by TempFunc. On the other hand, the channel-wise attention mechanism is introduced to make the model enables to re-weight micro-motion vector according to the features inside window, so that more discriminative motion expressions could be highlighted for the final identification.

In practice, there are two MTBs, using different window size (3 and 5) in MCM, as shown in Fig 5. And the exact structures of Conv1dNet in each MTB are shown in Tab. 2. The purpose of this design is to fuse the multi-scale information in sequence dimension, so as to gather more abundant characteristics of micro-motion. The ablation study will be shown in Sec.4.3.

Temporal Pooling

Description. Aggregate the sequence of micro-motion feature vectors, $S_m^{re}(t) = \{m_i^{re} | i = 1, ..., t\}$, to represent the gait motion patterns, formulated as

$$v = \text{TP}\left(S_m^{re}(t)\right) \tag{7}$$

where v denotes the output of TP module (a column vector).

Principle. As a periodic motion, a complete cycle should be able to thoroughly represent the entire gait sequence under the ideal condition [2]. So, let $S_m^{re}(T) = \{m_i^{re} | i = 1, ..., T\}$ represent the sequence of micro-motion features inside a complete gait cycle (T denotes the period), and the TP module should satisfy the following formulation:

$$\text{For } \forall t > T, \ \exists \ \text{TP}\left(S_m^{re}(t)\right) = \text{TP}\left(S_m^{re}(T)\right) \tag{8}$$

This is the Ground Principle of Gait Temporal Aggregation, whose core idea is that there could be no discriminative information gain after a complete cycle for periodic gait.

Operation. Two natural and simple statistical functions applied on sequence dimension have been taken as the instantiation of TP module, namely the mean$(\cdot)$ and max$(\cdot)$.

When $\text{TP}(\cdot) = \text{mean}(\cdot)$

$$\text{TP}\left(S_m^{re}(t)\right) = \frac{\sum_1^t m_i^{re}}{t} \tag{9}$$

[2] The individual gait is a pure periodic process without any interferences, *e.g.*, view change, gait-unrelated motion and so on.

$$\text{TP}\left(S_m^{re}(T)\right) = \frac{\sum_1^T m_i^{re}}{T} \tag{10}$$

It is obvious that if and only if t is an integral multiple of T, Eq.9=Eq.10 (here m_i^{re} would't be a constant). However, the length of real-world gait video is uncertain, the statistical function of mean$(\cdot)$ seem like a bad choice.

When TP $(\cdot) = \max(\cdot)$

$$\text{TP}\left(S_m^{re}(t)\right) = \max\left(m_1^{re}, ..., m_T^{re}, ..., m_t^{re}\right) \tag{11}$$

$$\text{TP}\left(S_m^{re}(T)\right) = \max\left(m_1^{re}, ..., m_T^{re}\right) \tag{12}$$

Since gait is a periodic action, it is obvious that Eq.11=Eq.12. Thus, the function max $(\cdot)$ is employed in final descision. The ablation study will be discussed in Sec.4.3.

3.4. Implementation Details

Network hyper-parameters. As shown in Tab. 1, the FPFE is made up by FConv Layers, Max Pooling Layers[11] and Leaky ReLU activations. What needs to be pointed out is the setting of pre-defined parts number p in the FConv. When p is larger, the constraint applied on the receptive field is stronger. And when $p = 1$, the FConv is equivalent to regular convolution layer and the constraint would be removed. Therefore, the experience of setting the value of p is increasing as the network going deeper.

Loss and Sampler. As mentioned in Sec3.1, the outputs of GaitPart are n column feature vectors. In this work, the separate Batch All (BA_+) triplet loss[8] is employed to train the network, and the corresponding column feature vectors among different samples will be used to compute the loss. The training batch size is (pr,k), where pr denotes the number of persons and k denotes the number of samples for each person in a training batch. In addition, at the test phase, the raw gait video will be directly fed into the model; at the train phase, for the length of gait video is uncertain, the sampler should collect a fixed-length segment as input: intercept a 30-40 frame-length segment first, and then randomly extract 30 sorted frames for training. Specially, if the length of raw video is less than 15 frames, it will be discarded. And while the length is more than 15 frames but less than 30 frames, it will be repeatedly sampled.

Testing. At the test phase, the distance between gallery and probe is defined as the average of Euclidean distance of the corresponding feature vectors.

4. Experiments

Two open databases have been applied to evaluate the GaitPart, namely CASIA-B[17] and OU-MVLP[21]. In this section, these databases will be described firstly. And then, the performances of GaitPart will be compared with that of other state-of-the-art methods. Finally, the detailed ablation studies will be conducted strictly on CASIA-B[17] to verify the effectiveness of each component in GaitPart.

4.1. Datasets and Training Details

CASIA-B. Composed of 124 subjects, the CASIA-B[17] is a widely applied gait dataset, in which each subject contains 11 views and each view contains 10 sequences. And this 10 sequences are obtained under 3 walking conditions, the first 6 sequences are obtained under normal case (NM), the other 2 sequences are obtained with subjects carrying bags (BG), and the last 2 are obtained with subjects wearing coats or jackets (CL). In other word, each subject contains 11×(6+2+2)=110 sequences. Based on CASIA-B, there are various experimental protocols[30], and for the fairness, this paper strictly follows the popular protocol carried out by [26]. In addition, the first 74 subjects are grouped into train set, and the remaining 50 subjects are reserved for testing. During the test, the first 4 sequences of NM condition(NM#1-4) are regarded as gallery, and the remaining six sequences are divided into three subsets according to the walking conditions, *i.e.*, the NM subset contains NM#5-6, the another BG subset contains BG#1-2, and the last CL subset contains CL#1-2.

OU-MVLP. The OU-MVLP gait database[21] is so far the world's largest public gait dataset. It is composed of 10307 subjects (5153 subjects for training and the rest 5154 subjects for test). In addition, each subject contains 14 views (0,15, ...,90; 180, 195, ..., 270) and each view embodies 2 sequences. At the test phase, the sequences with index #01 are grouped into the galleries while the rest sequences with index #02 are grouped into the probes.

Training details. 1) Common configuration: the input silhouettes are aligned by the approach mentioned in [21] and resized to the size of 64×44. Adam optimizer is used with the learning rate of 1e-4, and the momentum of 0.9[10]. The margin in separate triplet loss is set to 0.2[8]. **2)** In CASIA-B[17], the batch size is set to (8, 16) following the manner introduced in Sec3.4. Moreover, we train the model for 120K iterations. **3)** In OU-MVLP, due to it contains almost 20 times more sequences than CASIA-B, an additional block composed of two FConv Layers is stacked into the FPFE (the output channel is set to 256) and the value of p of each block is set to 1, 1, 3, 3 respectively. The batch size is set to (32, 16), the iterations is set to 250K, and the learning rate would be reduced to 1e-5 at 150k iterations.

4.2. Comparison with State-of-art Methods

CASIA-B. As shown in Tab.3, to ensure the GaitPart can be compared systematically and comprehensively with other state-of-the-art methods, all the cross-view and cross-walking-condition cases are included in the comparison scope. **1)** Expect CNN-LB[26] is GEI-based, other methods shown in Tab.3 are video-based and all of which outperform CNN-LB significantly. This indicates that video-based methods have great potential in extracting more fine-grained and discriminative information from the raw gait se-

Table 3. Averaged rank-1 accuracies on **CASIA-B**, excluding identical-view cases. CNN-LB:[26], GaitSet[5], GaitNet[30].

Gallery NM#1-4		0° – 180°											mean
Probe		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
NM #5-6	CNN-LB[26]	82.6	90.3	96.1	94.3	90.1	87.4	89.9	94.0	94.7	91.3	78.5	89.9
	GaitSet[5]	90.8	97.9	**99.4**	96.9	93.6	91.7	95.0	97.8	98.9	96.8	85.8	95.0
	GaitNet[30]	91.2	92.0	90.5	95.6	86.9	**92.6**	93.5	96.0	90.9	88.8	89.0	91.6
	GaitPart(ours)	**94.1**	**98.6**	99.3	**98.5**	**94.0**	92.3	**95.9**	**98.4**	**99.2**	**97.8**	**90.4**	**96.2**
BG #1-2	CNN-LB[26]	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[5]	83.8	91.2	91.8	88.8	83.3	81.0	84.1	90.0	92.2	**94.4**	79.0	87.2
	GaitNet[30]	83.0	87.8	88.3	93.3	82.6	74.8	**89.5**	91.0	86.1	81.2	85.6	85.7
	GaitPart(ours)	**89.1**	**94.8**	**96.7**	**95.1**	**88.3**	**94.9**	89.0	**93.5**	**96.1**	93.8	**85.8**	**91.5**
CL #1-2	CNN-LB[26]	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[5]	61.4	75.4	80.7	77.3	72.1	70.1	71.5	73.5	73.5	68.4	50.0	70.4
	GaitNet[30]	42.1	58.2	65.1	70.7	68.0	70.6	65.3	69.4	51.5	50.1	36.6	58.9
	GaitPart(ours)	**70.7**	**85.5**	**86.9**	**83.3**	**77.1**	**72.5**	**76.9**	**82.2**	**83.8**	**80.2**	**66.5**	**78.7**

Table 4. Averaged rank-1 accuracies on **OU-MVLP**, excluding identical-view cases. GEINet:[18], GaitSet:[5].

Probe	Gallery All 14 views		
	GEINet[18]	GaitSet[5]	GaitPart(ours)
0°	11.4	79.5	**82.6**
15°	29.1	87.9	**88.9**
30°	41.5	89.9	**90.8**
45°	45.5	90.2	**91.0**
60°	39.5	88.1	**89.7**
75°	41.8	88.7	**89.9**
90°	38.9	87.8	**89.5**
180°	14.9	81.7	**85.2**
195°	33.1	86.7	**88.1**
210°	43.2	89.0	**90.0**
225°	45.6	89.3	**90.1**
240°	39.4	87.2	**89.0**
255°	40.5	87.8	**89.1**
270°	36.3	86.2	**88.2**
mean	35.8	87.1	**88.7**

Table 5. Ablation Study, Group A. Control Condition: the value of p in each block. Results are rank-1 accuracies averaged on 11 views, excluding identical-view cases.

Group A	Block1	Block2	Block3	NM	BG	CL
a	1	1	1	95.6	88.4	76.1
b	1	1	8	**96.6**	90.4	77.1
c	1	4	8	96.2	**91.5**	**78.7**
d	2	4	8	95.8	90.7	78.4

quence. **2)** Compared with GaitSet[5], the GaitPart clearly presents better performance with possessing a similar backbone (In fact, the parameters of GaitPart are only about half of GaitSet's [3]). This result experimentally reveals the superiority of the FConv and MCM. **3)** Compared with GaitNet[30], these two methods bear the same purpose but different means. In GaitNet, an Auto-Encoder framework was introduced to obtain more discriminative features, and the multi-layers LSTM was applied for spatio-temporal modeling[30]. And in our model, FConv and M-CM have been proposed, respectively. From the view of experiments, the GaitPart achieves better performance under various walking conditions on CASIA-B.

OU-MVLP. In order to verify its generalization, the evaluation of GaitPart is completed on the worldwide largest public gait dataset[21]. As shown in Tab.4, GaitPart meets a new state-of-the-art under various cross-view conditions. It should be pointed out that the maximum value of rank-1 accuracy cannot reach 100% due to the missing of the sequences in some subjects and this situation is neglected at the test phase. If the subjects in probe without corresponding samples are discarded, the average rank-1 accuracy of all probe views should be 95.1%, instead of 88.7%.

4.3. Ablation Study

To verify the effectiveness of each component in GaitPart, several ablation studies with various settings will be conducted on CASIA-B, including setting the different p values in FConv, setting only one or two MTBs in MCM module, w/ and w/o applying attention mechanism in M-CM module and using the different instances for TP module. The experiments result and analysis are as follows.

Effectiveness of FConv. Following the manners of setting the hyperparameter p in FConv mentioned in Sec.3.4, four controlled experiments (numbered as A-a, b, c and d, respectively) are conducted in experiment Group A, and all the results are shown in Tab.5. It is worth noting that in the backbone of experiment A-a, the p value of all the F-Convs is set to 1, that is, the backbone is totally composed of regular layers. **1)** It is clearly discovered that all the experiments with using FConvs (including A-b, c and d) gain better performance than the experiment A-a. On the one hand, this verifies the effectiveness of FConv. On the other hand, the robust of varying the value p in FConv is also declared in GaitPart. **2)** The comparison between A-d with A-c shows that the performance is negatively affected by using the FConv at Block1 (bottom layers). The possible reason is that on the bottom layers, the edge and contour information between adjacent parts would be damaged by the FConvs. **3)** By comparing the differences among exper-

[3] About 2.56×10^6 for GaitSet[5] while 1.47×10^6 for GaitPart.

Table 6. Ablation Study, Group B. Control Condition: w/ and w/o applying MTB1 or MTB2, w/ and w/o attention mechanism in MTB and different instantiations of TP. Results are rank-1 accuracies averaged on 11 views, excluding identical-view cases.

Group B	MTB			TP		NM	BG	CL
	MTB1	MTB2	Attention	Max	Mean			
a	✓	✓	✓	✓		**96.2**	**91.5**	**78.7**
b	✓	×	✓	✓		95.8	90.8	76.2
c	×	✓	✓	✓		96.1	90.6	77.3
d	✓	✓	×	✓		95.4	89.3	73.1
e	✓	✓	✓		✓	93.6	86.5	70.1

iment A-a, b and c, it can be found that the average rank-1 accuracies first go up and then decrease on the NM subset, but keep in increasing under the BG and CL subset. The cause of this phenomenon is believed that the different receptive field of top-layer neurons could be fit to different walking conditions.

In addition, there is another thing worth mentioning that, the experiment A-a without employing the FConvs at all, achieves the worse performance in Group A but the best performance among other benchmarks mentioned in Tab.3. Because the backbone applied in A-a is lighter and more concise than that of other benchmarks, so it can loosely and partially verify the effectiveness of the MCM module. Final, the experiment A-c with impressive comprehensive performance is selected as the baseline of GaitPart.

Effectiveness of MCM. As shown in Tab.6, there are 5 controlled experiments (numbered as B-a[4], b, c, d and e, respectively) in Group B, among which B-a, b, c and d focus on the design of MTB module while the rest B-e only takes the instantiation of TP module into consideration. **1)** By comparing the differences between experiment B-a, b and c, we find the best performance is achieved by using MTB1 and MTB2 together. This shows the multi-scale design in MCM (mentioned in Sec.3.3) is helpful to capture the discriminative micro-motion features. **2)** By comparing the experiment B-a with B-d, we find the introduction of attention mechanism is necessary. And it indeed enable the model to highlight the most representative micro-motion features. **3)** The comparison between experiment B-a with B-e declares that the instantiation of TP module is of vital importance for GaitPart. When it doesn't satisfy the 'The Ground Principle of Gait Temporal Aggregation', taking instantiated as the function mean$(\cdot)$ as example, the worst performance among all the experiments in Group B and A is obtained.

4.4. Spatio-temporal Study

We generally think that both static appearance features and dynamic temporal information are representative characteristics for individual gait. But many prior approaches have achieved good performances without modeling the temporal features explicitly, in the other word, the order of input frames don't matter in these state-of-the-art methods[26, 5]. So in this section, we aim at openly exploring what roles do the temporal information and appearance features play in GaitPart, respectively.

Table 7. Spatio-temporal Study, Group C. Control Condition: sort/shuffle the input sequence at train/test phase. Results are rank-1 accuracies averaged on 11 views, excluding identical-view cases.

Group C	Train		Test		NM	BG	CL
	Shuffle	Sorted	Shuffle	Sorted			
a	✓			✓	95.6	89.9	71.5
b		✓		✓	**96.2**	**91.5**	**78.7**
c		✓	✓		92.5	85.8	65.1

To this end, experiment Group C is conducted and all the results are shown in Tab.7. As you can see, the worse performances are achieved by shuffling the input frames at both train (C-a) and test phase (C-c), but the accuracy degradation is not so serious. It reveals that even under the scrambled temporal information of input sequence, the model can still achieve not bad performance. This phenomenon indicates the static appearance features indeed play a vital role in gait recognition. But we don't think the temporal information is trivial or inessential, because the model gained a considerable accuracy boost under cross-wearing condition, where the gait appearance changes a lot in the real-world scenarios. Tab.7 shows that temporal information is also very important robust features in GaitPart.

5. Conclusion

In this paper, we present a novel insight that each part of human body needs its own spatio-temporal modeling, owing to the different visual appearance and moving patterns among human body during walking. Thus, GaitPart is proposed, which includes the Frame-level Part Feature Extractor composed of FConv and the Temporal Feature Aggregator consisting of several parallel and dependent Micro-motion Capture Modules. The core goal of these two parts is to enhance the fine-grained learning of part-level features and extract the local short-range spatio-temporal expressions, respectively. In final, experiments conducted on the well-known public databases, CASIA-B[17] and OU-MVLP[21], experimentally demonstrate the superiority of GaitPart as well as all its components.

Acknowledgement

This work was supported in part by National Key R&D Program of China (Grant No.2017YFB1002804) and in part by Scientific and Technological Innovation Foundation of Shunde Graduate School, USTB.

[4] This experiment is identical to the Group A-c.

References

[1] G. Ariyanto and M. S. Nixon. Model-based 3d gait biometrics. In *2011 International Joint Conference on Biometrics (IJCB)*, pages 1–7, Oct 2011. 2, 3

[2] G. Ariyanto and M. S. Nixon. Marionette mass-spring model for 3d gait biometrics. In *2012 5th IAPR International Conference on Biometrics (ICB)*, pages 354–359, 2012. 2, 3

[3] K. Bashir, T. Xiang, and S. Gong. Gait recognition using gait entropy image. In *3rd International Conference on Imaging for Crime Detection and Prevention (ICDP 2009)*, pages 1–6, 2009. 2

[4] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. *arXiv:1904.11492*, 2019. 5

[5] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. *AAAI*, 2019. 1, 2, 3, 7, 8

[6] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal pyramid matching for person re-identification. *AAAI*, 33, 04 2018. 2, 3

[7] Ju Han and Bir Bhanu. Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 28(2):316–322, 2006. 2, 5

[8] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *ArXiv*, abs/1703.07737, 2017. 6

[9] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019. 5

[10] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014. 6

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *NerulPS*, pages 1097–1105. Curran Associates, Inc., 2012. 6

[12] S. Li, S. Bak, P. Carr, and X. Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *CVPR*, pages 369–378, 2018. 2, 5

[13] W. Li, X. Zhu, and S. Gong. Harmonious attention network for person re-identification. In *CVPR*, pages 2285–2294, 2018. 2

[14] Rijun Liao, Chunshui Cao, Edel B. García Reyes, Shiqi Yu, and Yongzhen Huang. Pose-based temporal-spatial network (PTSN) for gait recognition with carrying and clothing variations. In *Biometric Recognition - 12th Chinese Conference, CCBR 2017, Shenzhen, China, October 28-29, 2017, Proceedings*, pages 474–483, 2017. 1, 3

[15] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPR Workshops*, June 2019. 2, 3

[16] Sudeep Sarkar, P. Jonathon Phillips, Zongyi Liu, Isidro Robledo Vega, Patrick Grother, and Kevin Bowyer. The humanid gait challenge problem: Data sets, performance, and analysis. *IEEE transactions on pattern analysis and machine intelligence*, 27:162–77, 03 2005. 2

[17] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 4, pages 441–444, 2006. 2, 6, 8

[18] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi. Geinet: View-invariant gait recognition using a convolutional neural network. In *2016 International Conference on Biometrics (ICB)*, pages 1–8, June 2016. 1, 7

[19] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian. Pose-driven deep convolutional model for person re-identification. In *ICCV*, pages 3980–3989, 2017. 2

[20] Yifan Sun, Liang Zheng, Yi Yang, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling. *ECCV*, 11 2017. 2, 3

[21] 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 Transactions on Computer Vision and Applications*, 10, 12 2018. 1, 2, 6, 7, 8

[22] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, Dec 2015. 3

[23] Chen Wang, Junping Zhang, Liang Wang, Jian Pu, and Xiaoru Yuan. Human identification using temporal information preserving gait template. *IEEE transactions on pattern analysis and machine intelligence*, 34, 12 2011. 2

[24] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *CVPR*, 11 2017. 5

[25] T. Wolf, M. Babaee, and G. Rigoll. Multi-view gait recognition using 3d convolutional neural networks. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 4165–4169, Sep. 2016. 1, 2, 3

[26] 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 Transactions on Pattern Analysis & Machine Intelligence*, 39(2):209–226, 2016. 1, 2, 6, 7, 8

[27] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, 2017. 2

[28] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian. Deep representation learning with part loss for person re-identification. *IEEE Transactions on Image Processing*, 28(6):2860–2871, June 2019. 2

[29] Kaihao Zhang, Wenhan Luo, Lin Ma, Wei Liu, and Hongdong Li. Learning joint gait representation via quintuplet loss minimization. In *CVPR*, pages 4700–4709, 2019. 1, 2

[30] Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Jian Wan, Nanxin Wang, and Xiaoming Liu. Gait recognition via disentangled representation learning. In *CVPR*, 2019. 1, 2, 3, 6, 7