IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 August 2023, accepted 14 September 2023, date of publication 20 September 2023,
date of current version 25 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3317437

RESEARCH ARTICLE

Effective Fusion Method on Silhouette and Pose for Gait Recognition

YANG ZHAO[1], RUJIE LIU[1], WENQIAN XUE[1], MING YANG[1], MASAHIRO SHIRAISHI[2], SHUJI AWAI[2], YU MARUYAMA[2], TAKAHIRO YOSHIOKA[2], AND TAKESHI KONNO[2]

[1]Fujitsu Research and Development Center Company Ltd., Beijing 10020, China
[2]Fujitsu Ltd., Tokyo 105-7123, Japan

Corresponding author: Yang Zhao (zhaoyang.frdc@fujitsu.com)

ABSTRACT Silhouette and pose are two common features to extract the descriptive and unique patterns of a person's gait, and good performance has been already achieved driven by the deep learning techniques. However, some issues still exist, the silhouette is known to be sensitive to the changes of the appearance while pose is not so discriminative as silhouette even though it is considered as being more robust. Therefore, it is advantageous to fuse the two features into one model to achieve both the accuracy as well as the robustnesss. In this paper, we propose a simple yet effective fusion model to combine both the features, where the two features are first scaled by normalisation and then combined by the Compact Bilinear Pooling to model the higher order and fine-grained information. The superiority of the proposed method is verified through experiments on benchmark datasets CASIA-B, OUMVLP, and SOTON-small. In CASIA-B, we achieved SOTA results with an average of 96.9% rank-1 accuracy. In addition, cross data experiments are conducted to demonstrate the robustness of our method.

INDEX TERMS Gait recognition, pattern recognition, multi-modal, deep learning.

I. INTRODUCTION

Gait referring to a person's walking pattern is commonly used for human identification. It has an advantage over other biometric traits, such as face, iris, and fingerprint as it can be collected from a distance with less intrusion. Due to these advantages, gait has wide-ranging applications in society, including forensic identification, social security and health monitor [1], [2], [3]. However, two main challenges for using gait as an identification technique are its sensitivity to extraneous factors and limited discriminative ability. Specifically, factors such as appearance and clothing style may introduce bias into the gait feature, and the coarse grained nature of gait features limit their ability to distinguish between individuals. To tackle these problems, different modalities of gait features like RGB image, mesh, optical flow, depth image silhouette, and skeleton have been developed [4].

Among the modalities, silhouette and skeleton are the most commonly used body representations in recent gait recognition literature. Silhouette is the human body mask by removing the background, which can be obtained by background subtraction [6] or deep learning based segmentation methods [7]. The silhouette-based feature is usually extracted by Convolution Neural Network (CNN) [8], [9], [10]. Silhouette by nature contains more person extrinsic information, such as the contour of the hairstyle and the clothing. These person extrinsic information can be useful for better identity recognition in certain scenarios while at the same time, the variation of it could easily lead to misclassification. On the other hand, skeleton based gait feature is generated by modeling the inner body topology. The skeleton is obtained by pose estimation algorithms, such as HRNet [11], and OpenPose [12]. Some promising results

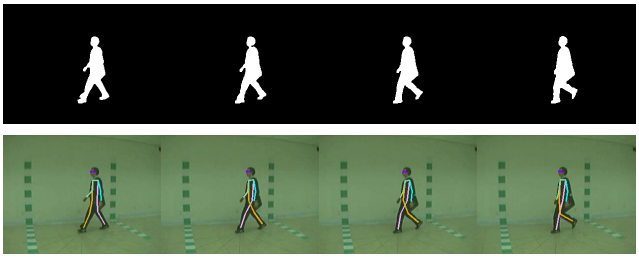


FIGURE 1. The silhouette and pose for gait recognition from CASIA-B [5].

The associate editor coordinating the review of this manuscript and approving it for publication was Yongjie Li.

have been obtained by Graph Convolution Network (GCN) [13], [14]. Skeleton based methods are generally more robust to the appearance variations than those based on silhouettes since the model only consider the inner body structure. The accuracy of the pose based models are usually lower than the silhouette based ones. One reason might be that silhouette contains more human related information which CNN models can utilise for identification.

From the above analysis, it can be seen that the two representations are complementary. The silhouette focuses on the body shape and neglects the inner body structure; on the contrary, the skeleton preserves the inner body structure while ignores the body shape. Therefore, their combination is expected to improve the representation ability of gait. In this paper, we propose a simple yet effective fusion network to exploit both the advantage of skeletons and silhouettes. Unlike other fusion methods, which combine the features through simple concatenation, we focus on constructing reasonable fusion method to exploit the strength of both features through Bilinear Pooling (BP). BP can be considered as a linear projection on the tensor product space of two vectors, which covers the interactions of each dimension of the vector. Thus, it can represent rich and fine-grained multimodal information [15], and has been shown to exceed the performance of attention based models for multi-modal fusion tasks [15], [16]. In addition, to avoid the computational complexity due to the curse of dimension, we apply a Compact Bilinear Pooling (CBP) method [17] which retains full representation of the two features as much as possible at a low dimension. In addition, our method has another effective design which is scale normalization. Since the extracted pose and silhouette features are obtained through different models, the distances between each feature may be incomparable due to scale differences. To ensure the effectiveness of the fusion mechanism, it is critical that the features are normalized to the same scale.

The effectiveness of the proposed method is verified by experiments, e.g. it outperforms the SOTA on CASIA-B, and gets very competitive results on OUMVLP. Furthermore, cross data experiments are also conducted to further confirm the robustness of our method. The contribution of this paper is summarized as follows,

- A simple and effective fusion method is proposed where CBP and scale normalization are adopted for better accuracy.
- Our method achieves competitively high performance on open benchmark data CASIA-B, OUMVLP and SOTON-small. Meanwhile promising results are also achieved on cross data experiments, which shows the robustness of our method.

II. RELATED WORK

According to the survey [4], the development on gait recognition goes through three stages. The earliest phase started in the early 1990s [18], mainly exploring the feasibility of human recognition at distances. The methods achieve reasonable performance on small-scale datasets [19], [20]. The second stage research contains more details into consideration. It can be classified according to the inputs into two categories: appearance-based methods which exploit the surface profile of the subject and model-based methods which depend on the modeling of the underlying human body structure. The third stage research leverages the deep learning method which extracts the temporal spatial information of the gait through large amount of data [4].

For **appearance-based methods**, before the wide application of the deep learning, the temporal dimension in a video sequence is usually depressed by averaging the silhouettes over a walking cycle into a gait energy image (GEI) [21], [22], [23], [24]. Due to the computational efficiency and the effectiveness, the methods are widely applied. However, the compression of the temporal information hinders the further improvement of the accuracy, and thus deep learning methods are utilised to model the spatial and temporal information in fine level. The first deep model with significant improvement on the standard benchmark dataset CASIA-B is GaitSet [9], which explores the spatial information by 2DCNN and models the temporal information as unordered set. GaitPart [10] enhanced GaitSet by extracting the local part information of the silhouette and the temporal information accordingly instead of taking the temporal information as an unordered set. Another set-based gait sequence silhouette approach utilizes a Set Residual Network (SRN) [25], which can effectively integrate silhouette and set information to extract more discriminate features. GaitGL [8] integrates the global and local information at the same time to further enrich the spatial representation. In addition CSTL [26] models the temporal information by introducing multi-scale temporal relations and obtains the spatial representation by selecting the most discriminate parts among the whole sequence. The above mentioned models CSTL and GaitGL employ 3DCNN to model spatial-temporal information, and Local3D model also shares the similar merit. Other models, such as LSTM [27] is also employed in combination with 2DCNN to extract the temporal and spatial information respectively. To enhance the temporal representation ability, the temporal representation, a multi-scale model [28] is proposed where the small-temporal-scale branch is used to model slow changes, while the larger-temporal-scale one is designed to grasp the rapid gait changes. Recent development seen a lot new ideas emerging, for instance, instead of extracting the spatial-temporal information, Gaithop [29] focuses on the channel information by switching channel of different frames. In addition, an end-to-end model is proposed by GaitEdge [30] which combines the silhouette extraction and the gait recognition model into one pipeline and achieves significant improvement over the separate pipeline. To address the view change problem, [31] incorporates view information explicitly through 'View projection matrix selection', allowing the model to include view information in the final expression.

For **model-based methods**, the features are usually extracted by leveraging the prior knowledge of human body. For example, the length of the stride and cadence are used as learning features [32], [33]. There are also works modeling gait motion as Fourier series and transforming the pattern into spectral space for recognition [34]. Other handcrafted features [35], [36], [37] are developed to represent the structural and dynamic characteristics of pedestrians. However, the handcrafted features can only provide limited representations, and thus, the deep learnings are also actively studied in the model-based area. On one hand, accurate pose estimation algorithms, such as HR-Net [11], and OpenPose [12], serve as the foundation for pose estimation; on the other hand, the graph convolution network (GCN) [14] provides a basic tool to describe the pose structure. ST-GCN [14] combines GCN and CNN to model the spatial and temporal information respectively. Classic two stream framework is adopted in [38] to extract the pose pattern, where the second order information is modeled by adaptive GCN. GaitGraph [13] enhances the structure of the model by adding residual connections and bottleneck blocks to achieve better performance. In addition, Multi-scale Gait Graph (MSGG) network is proposed by [39], leveraging on the inherent hierarchical structure of the body joints. A gait recognition system, HEATGait [40], is proposed to improve the performance of existing multi-scale graph convolutions by using an efficient hop-extraction technique to alleviate weight bias issues. Utilizing spatial-temporal graph convolution layers (ST-GCN), Gait-D [41] is designed to improve the performance of gait recognition by eliminating the redundant information using canonical polyadic decomposition (CPD). Besides, attention scheme is introduced recently. In GaitMixer [42], self-attention mixer is used to learn spatial information and large-kernel convolution mixer to extract temporal information. While in [43], a multi-stream strategy is proposed to simultaneously model joint, bone, and motion dynamics using GCN with channel-wise attention. The proposed strategy merges part-level information by capturing features from the skeleton graph and its subgraphs concurrently.

In summary, for both types of methods, the up-to-date mainstream are deep learning due to the superior performance. Regarding to deep learning, both methods extract temporal and spatial information, and in terms of spatial modeling, they employ both local and global information. For comparison, model-based methods directly use the human skeleton to describe the gait, without any other information. Therefore, these methods tend to be robust and computational efficient. Appearance-based models extract features from the contour of the subject, which can capture more direct human related information, such as the hairstyle and the shape of the clothes. Meanwhile, the input is easy to be affected by the variance of the appearance of the subject, and thus lead to wrong classification. These characters make the two features complementary to each other.

Fusion methods The multi-modal methods can utilise the properties of different inputs and hence achieve better results than single modal ones. The fusion procedure can be characterised into sensor level, feature level, opinion level and decision level according to the fusion stage, and it is believed that the earlier fusion will yield better results [44]. However, the research on multi-modal fusion method for gait recognition is limited as far as we know. BiFusion model [39] and two-branch model [45] are proposed to combine silhouette and pose features, where two stream architecture is applied in both methods. These works focus on designing complicated pose models by introducing multi-scale as well as attention techniques, while only adopting simple concatenation as the fusion method. Besides silhouette and pose, a pioneer work [46] explores the usage of 3D feature, 3D meshes, and combines the 3D meshes with the silhouette by matrix multiplication for gait recognition. In addition, TransGait [47] utilizes transformer structure to receive silhouette and heatmap as feature inputs and combine the different features at an early stage to capture more complex interaction of the two features. However, this model requires frame level alignment for the two inputs which could be hard since different features are collected with different sensors and thus might not be easy to make exact alignment. The mmGaitSet [48] combines silhouette with pose heatmap through both intra-modal and the inter-modal fusion. The intra-modal fusion method integrates low-level structural features with high-level semantic features, which results in increased discrimination of single modality features. Meanwhile, the inter-modal fusion method aggregates complementary information from different modalities, thereby enhancing the overall pedestrian gait representation. Another fusion model [49] is designed to combine silhouette from CNN and human model information, s.t. joints, limbs, and static joint distances from a fully connected deep-layer structure.

The research suggested that most of the fusion models focus on developing more representative single modal models with little focus on reasonable fusion methods [39], [45]. Differently, our method focuses on constructing reasonable fusion method which aligns the different embeddings to make the final fusion to yield better accuracy. Besides, the architecture of our method is flexible in that our fusion method can be applied to any single modal feature extractor.

III. OUR METHOD

The overall structure of the fusion model is explained in Figure 2. First, the silhouette and pose are fed into the corresponding feature extractors. Then the distribution of the features are scaled to the same magnitude by applying our normalisation technique. Then the normalised features go through the CBP layer to obtain the higher order and fine-grained representation of the gait.

GaitGL and GaitGraph are adopted in our case to extract silhouette and pose features, since they are representative and achieve high accuracy over several bench mark datasets. It is worth noticing that, the feature extractors can be any other models, such as ST-Graph for pose features and Gaitset, Gaitpart for silhouette features.

In the subsections, the feature extractors, GaitGraph and GaitGL are first explained briefly, and then the main part of fusion method, CBP, along with some other application details are presented.

A. FEATURE EXTRACTOR

GaitGraph, which employs GCN, outperforms other GCN-based models, while GaitGL, outperforms most other gait recognition models. Besides delivering strong performance, their architectures are classic. GaitGraph employs GCN which is a popular method for modeling pose, and the mainstream approach for silhouette-based models involves using CNN with temporal-spatial modeling and local and global spatial feature mining. These architectures are classical and representative, and they have served as inspiration for other recent models, such as CSTL. In the following, we explain the two feature extractors briefly.

GaitGL is a silhouette based gait recognition model which achieves high performance on CASIA-B and OUMVLP datasets [8] due to two main advantages. First, GaitGL preserves the temporal information by the Local Temporal Aggregation operation, and in addition, it considers not only the global but also the local features. The local feature is constructed by dividing the silhouette into horizontal stripes to model different body parts, leading to a final $[m, d]$ dimensional feature with m the number of the horizontal stripes and d the dimension of features.

GaitGraph utilises GCN [14] to extract the spatial information from the pose/skeleton of the gait sequences. The topology of the skeleton is naturally modeled as a graph, defined as $G = (V, E)$, where V is the set of vertices and E is the set of edges. Two vertices are adjacent if they are connected in the skeleton, and such relation of the V is formulated by the adjacency matrix $\boldsymbol{A}$, where $\boldsymbol{A}_{i,j} = 1$ if vertex v_i and v_j are connected, and $\boldsymbol{A}_{i,j} = 0$ other wise. As an analogy to CNN, the convolution on graph operates on the adjacency vertices, and the k^{th} order GCN is defined as follows,

$$f_{out} = \sigma(\sum_{j=1}^{k} \boldsymbol{\Lambda}_j^{-\frac{1}{2}} \boldsymbol{A}_j \boldsymbol{\Lambda}_j^{-\frac{1}{2}} f_{in} W_j),$$

where f_{in} is the input feature, f_{out} is the output feature, $\boldsymbol{A}_j := \prod_{i=0}^{j} \boldsymbol{A}$ is the j^{th} order adjacent matrix, $\boldsymbol{\Lambda}_j$ is the diagonal degree matrix for $\boldsymbol{A}_j$ and W is learnable weight parameters. In addition, bottleneck structure is applied in GaitGraph to down size the features and the residual learning is applied for stable training.

B. THE FUSION METHOD

Instead of direct concatenation or summation, we apply BP to learn the fused representation [50]. Let X_s and X_p be the silhouette and pose gait feature obtained from GaitGL and GaitGraph respectively, where $X_s \in \mathcal{R}^{m \times d_1}$, and $X_p \in \mathcal{R}^{d_2}$ with m being the number of the parts of the silhouette features, and d_1, d_2 the feature dimension of the silhouette and pose. The fused feature through bilinear pooling becomes

$$\begin{aligned} X^{fuse} &= W(X_s \otimes X_p) \\ &= W[X_s(1)X_p(1), X_s(1)X_p(2), \ldots, X_s(d_1)X_p(d_2)]^T, \end{aligned}$$

where $W \in \mathcal{R}^{d_1 \times d_2, d_3}$ with d_3 the dimension of the fused feature. As we can see, BP can be considered as a linear transform on $X_s \otimes X_p \in \mathcal{R}^{d_1 \times d_2}$ with $\otimes$ being the tensor product. Thus, with BP, the second order interaction of every dimension of the feature can be modeled, and the resulting representation contains rich and fine-grained information [15].

However, the dimension of the W grows quadratically as the dimension of the underlying feature grows. Therefore, the method is not practical unless a more compact expression is constructed, and the problem is addressed by CBP [17]. For CBP, the count sketch projection $\phi(.)$ is applied to project $X_s \otimes X_p$ to a lower dimensional space. In addition, $\phi(.)$ is featured by the property of converting outer product into convolution operation, that is, the count sketch of the outer product of two features can be expressed as convolution of both count sketches projections: $\phi(X_s \otimes X_p, \boldsymbol{h}, \boldsymbol{v}) = \phi(X_s, \boldsymbol{h}) * \phi(X_p, \boldsymbol{v})$, where $*$ refers to the convolution, and $\boldsymbol{h}, \boldsymbol{v}$ are two parameter vectors [51]. To further reduce the computational complexity, the fast Fourier transformation (FFT) is adopted to realize heavy convolution by light point-wise product [51], as follows:

$$\begin{aligned} &\phi(X_s, \boldsymbol{h}) * \phi(X_p, \boldsymbol{v}) \\ &\quad = FFT^{-1}(FFT(\phi(X_s, \boldsymbol{h})) \odot FFT(\phi(X_p, \boldsymbol{v}))), \end{aligned}$$

where $\odot$ denotes the pointwise product. With the combination of count sketch projection and FFT, the fused feature through CBP can be represented as:

$$\begin{aligned} X^{fuse} &= W(X_s \otimes X_p) \sim W\left(\phi(X_s \otimes X_p, \boldsymbol{h}, \boldsymbol{v})\right) \\ &= W\left(\phi(X_s, \boldsymbol{h}) * \phi(X_p, \boldsymbol{v})\right) \\ &= W\left(FFT^{-1}(FFT(\phi(X_s, \boldsymbol{h})) \odot FFT(\phi(X_p, \boldsymbol{v})))\right), \end{aligned}$$

and $X^{fuse} \in \mathcal{R}^{d_3}$. While applying CBP, it is a common practice to add one more dimension to the feature to make CBP represent not only the second order interactions but also the first order ones. Specifically, the input features are augmented from X_p and X_s to $[1, X_p]$ and $[1, X_s]$ respectively.

Given the GaitGraph and GaitGL as the feature extractors, the resulting feature dimensions do not match. The dimension of silhouette feature is $[m, d_1]$ due to the partition of a silhouette into m horizontal strides, but pose feature is a

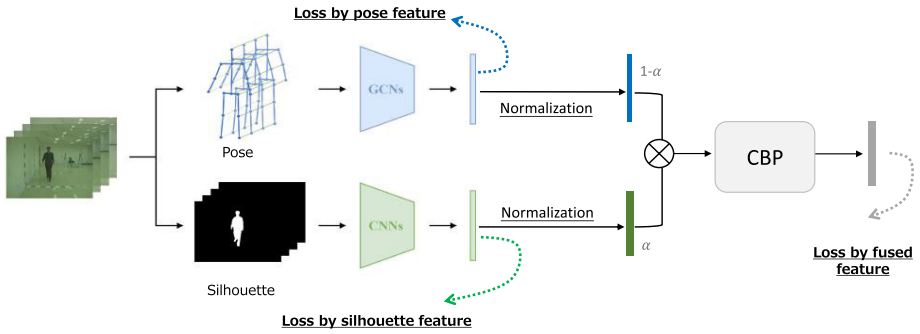


FIGURE 2. The algorithm of our fusion model.

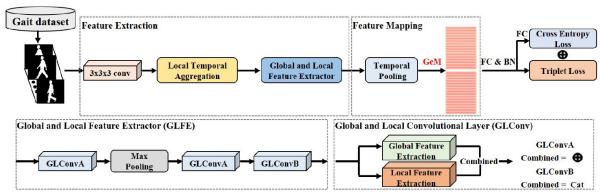


FIGURE 3. The algorithm of GaitGL, silhouette based model [8].

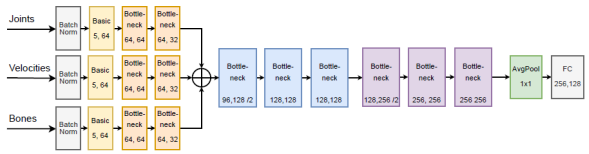


FIGURE 4. The algorithm of GaitGraph, pose based model [13].

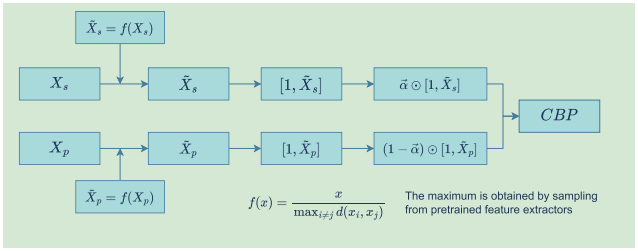


FIGURE 5. The detailed procedure for feature fusion.

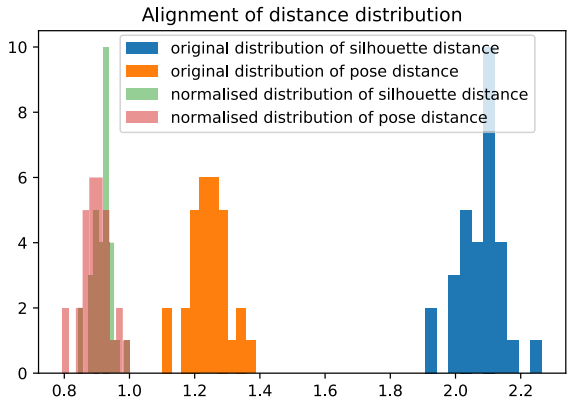


FIGURE 6. The distribution example.

d_2-dimensional vector. For this issue, each of the m local parts of the silhouette feature is concatenated with the same pose feature, and this can be thought of as the combination of global and local information, i.e. the pose feature represents holistic information while silhouette stride feature is local. Therefore the final dimension of the fused feature is $[m, d_3]$.

The distance of different features are samples from different distributions which might not be comparable (e.g. one is $\mathcal{O}(1)$ and the other one is $\mathcal{O}(100)$). The difference between the magnitude of the distances of different features would make the fusion algorithm hard to train. To make the distance of each feature comparable, we divide the feature by the maximum of the distance of the corresponding feature. This is more effective than the L_2 normalisation of the feature (a more common way of normalisation), since the normalised feature does not guarantee comparable distances between different features which affect the performance directly. The maximum of the distance can be approximated by the sample maximum of the distance from the pre-trained feature extractors. Our idea can be considered as distribution alignment which is a common technique applied in multi-modal models. Usually, these features can be coordinated through structure constraint [52], such as WSABIE [53] and canonical correlation analysis [54].

After normalisation, the features are also weighted by trainable parameter vector $\boldsymbol{\alpha}$ as $\boldsymbol{\alpha}^T \odot [1, X_s]$ and $(1 - \boldsymbol{\alpha})^T \odot [1, X_p]$ to adjust the importance of different features where $\odot$ refers to the pointwise multiplication and each element of $\boldsymbol{\alpha}$ is within [0, 1].

IV. EXPERIMENT

Three widely used benchmark gait datasets are adopted for experiments, i.e., CASIA-B [5], OUMVLP [55], and SOTON-small [56]. With these datasets, the proposed fusion method is compared with the SOTA methods, besides, cross data experiments are conducted to verify the robustness of our method. At last, ablation studies are presented to illustrate the contribution of different techniques.

A. DATASET

CASIA-B [5] is composed by walking sequences of 124 subjects, with 10 sequences per subject, i.e., 6 of normal walking status (NM), 2 of bag carrying status (BG), and 2 of clothes

changing status (CL). Each walking sequence is captured from 11 different angles from 0° to 180° with the uniform interval of 18°. Following the standard of the community [39], [45], we take the first 74 subjects as the training set and the remaining 24 subjects as the test set. In the test, the first four NM sequences are taken as the gallery while the remaining ones are used as the probe.

OUMVLP [55] is a large public gait dataset released by Osaka University. The dataset consists 10,307 subjects with a wide coverage of ages (from 2 to 87). There are in total 14 angles, ranging from 0° to 90°, and 180° to 270°, with uniform interval of 15°. Each view includes 2 walking sequences (#00-01). Following the test protocol of OUMVLP [55], the dataset are devided into training set with 5153 subjects and test set with 5154 subjects. sequences #01 are kept in the gallery and sequences #00 are regarded as the probe.

SOTON-small [56], as part of the DARPA funded programme, is released by the University of Southampton. It contains 10 subjects of 14 versatile walking scenarios, including walking in different clothes(CL), bag carrying(BG), different speed(SP) and normal walking scenario(NM). Each walking scenario has four different views: **a** is Normal track, **d** is Oblique track, **e** is Normal, elevated view of track, **f** is Frontal view of track. The first 6 subjects are taken as the training set, and the remaining four subjects are the test set. Only normal walking scenario is the gallery and the remaining scenarios are the probe.

B. EXPERIMENTS SETTING

Silhouette gait feature is obtained by GaitGL model, where the silhouette images are resized into 64 × 64 [8], and the length of the gait sequence is dataset dependant, i.e. 30 frames for OUMVLP and 60 for CASIA-B and SOTON-small. Since the silhouette images of Sonton-small are not provided officially, we extracted them by using DeepLabV3 [7]. GaitGraph model is adopted to get the pose based gait feature. For CASIA-B and SOTON-small datasets, the poses are estimated by HRNet [11], while for OUMVLP dataset the officially provided pose information, which is claimed to be extracted by AlphaPose [57], is directly adopted. After pose estimation, each joint is described by two attributes, i.e., 2D coordinate and the estimation confidence. Besides, several data augmentation schemes are applied to pose such as random noise and flip. The final dimension of the fused feature is 128 for CASIA-B and SOTON-small, and 256 for OUMVLP.

1) LOSS

To build the model, different loss functions are imposed separately on the silhouette feature, pose feature, as well as the fused feature. The triplet loss (TL) is employed to supervise the fused feature, with the margin being set as 0.2. The loss for the silhouette branch follows the setting of GaitGL [8], which includes cross-entropy loss (CE) and triplet loss. Also, the margin threshold for triplet loss is set to 0.2. For the pose feature, supervised contrastive loss(SC) is adopted with the temperature being 0.01. The final loss are the weighted sum of the three losses, and the weight values are empirically set to be 1, 0.2, and 0.1 in our implementation.

$$Loss = \alpha TL_{fuse} + \beta(CE_{sil} + TL_{sil}) + \gamma SC_{pose}.$$

TABLE 1. Parameter numbers of gait models.

	CASIA-B	OUMVLP	Sonton-small
GaitGL	3.09667M	95.62045M	2.54781M
OpenGait	0.31352M	0.76528M	0.31352M
Our method	3.41032M	96.38573M	2.86146M

2) OPTIMIZER

The optimization strategy for fusion model applies a multi learning rate strategy. In our implementation, both the GaitGL and GaitGraph model are firstly pre-trained as in [8] and [13], and are then fine tuned in our fusion framework, where Adam scheme is adopted for optimization with the learning rate being $1e^{-6}$. The learning rate for the fusion parameters are $1e^{-3}$.

3) CALCULATING RESOURCES

The experiments are conducted on two GPUs of NVIDIA RTX A6000. Since pretrained models for the feature extractors are applied, the new parameters needed to be trained from the beginning are very few. Therefore the fusion model converges fast. We also provided analysis on the number of parameters in Table 1, and we can see that the magnitude of our fusion model is of the same order of the GaitGL. Since the fusion part is kept as simple as possible, the parameters does not see a significant increase.

C. EXPERIMENTS RESULT

We compare the accuracy of the proposed fusion method with single modality models, including pose based models: GaitGraph [13] and GaitMixer [42], and silhouette based models: GaitSet [9], GaitPart [10], GaitGL [8], CSTL [26], and Gait-D [41]. Two fusion methods BiFusion [39] and TwoBranches [45] are also included as benchmarks. Firstly, the experimental results within single dataset are presented, that is, both the training and test data come from the same dataset. Then, cross-dataset experiments are conducted to verify the robustness of our method, that is, the model is trained by one dataset while tested on another dataset. In all experiments, the top one accuracy after excluding the identical view is used as the evaluation metric.

1) NORMAL WALKING

The experimental results on CASIA-B, OUMVLP and SOTON-small is tabulated in Table 2, Table 3, and Table 4. It can be seen that our result is significantly better than that

TABLE 2. Experiment results on CASIA-B.

Prob	Models	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
NM	Gaitgraph	85.3	88.5	91.0	92.5	87.2	86.5	88.4	89.2	87.9	85.9	81.9	87.7
	GaitGL	96.0	98.3	99.0	97.9	96.9	95.4	97.0	98.9	99.3	98.8	94.0	97.4
	GaitSet	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	94.1	98.6	99.3	98.5	94.0	92.3	95.9	98.4	99.2	97.8	90.4	96.2
	GaitMixer	94.4	94.9	94.6	96.3	95.3	96.3	95.3	94.7	95.3	94.7	92.2	94.9
	CSTL	97.2	99.0	99.2	98.1	96.2	95.5	97.7	98.7	99.2	98.9	96.5	97.8
	Bifusion	98.0	99.1	**99.5**	**99.3**	**98.7**	**97.5**	**98.5**	**99.1**	**99.6**	99.5	96.8	98.7
	TwoBranch	97	97.9	98.4	98.3	97.2	97.3	98.2	98.4	98.3	98.1	96	97.7
	Ours	**98.9**	**99.3**	99.2	98.4	98.3	97.3	98	99	99.4	**99.5**	**98.1**	**98.8**
BG	Gaitgraph	75.8	76.7	75.9	76.1	71.4	73.9	78.0	74.7	75.4	75.4	69.2	74.8
	GaitGL	92.6	96.6	96.8	95.5	93.5	89.3	92.2	96.5	98.2	96.9	91.5	94.5
	GaitSet	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	89.1	94.8	96.7	95.1	88.3	84.9	89.0	93.5	96.1	93.8	85.8	91.5
	GaitMixer	83.5	85.6	88.1	89.7	85.2	87.4	84.0	84.7	84.6	87.0	81.4	85.6
	CSTL	91.7	96.5	97.0	95.4	90.9	88.0	91.5	95.8	97.0	95.5	90.3	93.6
	Bifusion	95.8	97.9	**98.2**	97.6	94.4	91.6	93.9	96.6	**98.5**	98.3	93.1	96.0
	TwoBranch	91.9	94.6	96.4	94.3	94.4	91.6	94.1	95.4	95.5	93.9	89.5	93.8
	Ours	**97.6**	**98**	97.8	**98.4**	**97.5**	**96.3**	**97.1**	**98**	98.4	**98.7**	**96.3**	**97.8**
CL	Gaitgraph	69.6	66.1	68.8	67.2	64.5	62.0	69.5	65.6	65.7	66.1	64.3	66.3
	GaitGL	76.6	90.0	90.3	87.1	84.5	79.0	84.1	87.0	87.3	84.4	69.5	83.6
	GaitSet	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	70.7	85.5	86.9	83.3	77.1	72.5	76.9	82.2	83.8	80.2	66.5	78.7
	GaitMixer	81.2	83.6	82.3	83.5	84.5	84.8	86.9	88.9	87.0	85.7	81.6	84.5
	CSTL	78.1	89.4	91.6	86.6	82.1	79.9	81.8	86.3	88.7	86.6	75.3	84.2
	Bifusion	88.7	93.9	95.6	93.8	91.4	89.4	92.3	93.8	94.2	93.7	86.2	92.1
	TwoBranch	87.4	**96**	97	94.6	94	90.1	91.5	94.1	93.8	92.6	88.5	92.7
	Ours	**91.4**	95.9	**97**	**95.8**	**94.3**	**93.4**	**94.3**	**95.9**	**94.9**	**95.3**	**90.6**	**94.8**

TABLE 3. Experiment results on OUMVLP.

Prob	Models	0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	Mean
NM	Gaitgraph	58.3	71.2	74.2	75.8	74.8	72.6	68	53.3	61.1	56.2	72	71.1	68.2	63.3	67.1
	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
	GaitSet	79.3	87.9	90.0	90.1	88.0	88.7	87.7	81.8	86.5	89.0	89.2	87.2	87.6	86.2	87.1
	GaitPart	82.6	88.9	90.8	91.0	89.7	89.9	89.5	85.2	88.1	90.0	90.1	89.0	89.1	88.2	88.7
	CSTL	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
	Bifusion	86.2	90.6	91.3	91.56	91	90.8	90.5	87.8	89.5	90.4	90.7	90	89.8	89.3	89.9
	Gait-D [59]	84.3	**92.6**	90.6	**92.1**	90.5	91.3	**92.1**	87.6	**90.4**	**92.6**	**91.3**	**92.2**	**94.5**	**92.3**	**91**
	Ours	**88.7**	91.6	**91.7**	91.8	**91.9**	**91.3**	91.1	**90.5**	90.1	90.7	91	91	90.4	90	90.8

of the single modal methods on normal case (NM) which is considered as an easy one. Comparing to other single modal models, our fusion method achieved significant increase for CASIA-B and OUMVLP. Even comparing to other fusion methods, we still achieved 1% increase on CASIA-B. The SOTON-small dataset has seen a even larger increase for the NM case of our fusion method, with around 6% increase.

2) VIEW CHANGE

The view changes caused much interference for the practical application of gait recognition. Our fusion model shows potential to make gait recognition more robust to view changes. Firstly, regarding to the CASIA-B and OUMVLP datasets, the average accuracy of the fusion model varies around 2 % at different angles from Table 2 and Table 3. However, for other models, the performance at different angles fluctuates significantly. For the fusion models, taking the BG case of CASIA-B as an example, the performance of Bifusion ranges from 93.1 to 98.2, while the performance of TwoBranch ranges from 89.5 to 96.4. For OUMVLP, the fluctuation of Bifusion ranges from 86.2 to 91.56, and even for Gait-D, the performance fluctuates up to about 10 %. For the Sonton-small Table 4, due to the relatively coarse angle classification, the performance at each angle varies significantly. However, the performance of our fusion model is relatively stable which is about 10 % for all the cases. However, the performance gap for single modal models can vary by up to 20 %.

TABLE 4. Experiment results on Sonton-small.

Prob	Models	a^1	d^2	e^3	f^4	Mean
NM	Gaitgraph	90.7	91.6	90.1	83.3	88.9
	GaitGL	**98.85**	**99.54**	98.85	60.82	89.5
	Ours	98.24	99.52	**98.97**	**86.76**	**95.87**
BG	Gaitgraph	89.1	87.8	85.6	82.2	86.2
	GaitGL	99.07	99.79	97.52	71.52	91.977
	Ours	**99.69**	**99.90**	**98.59**	**91.90**	**97.52**
CL	Gaitgraph	41.1	35.6	34.4	49.6	40.2
	GaitGL	**91.77**	**87.34**	**92.61**	66.24	84.493
	Ours	86.88	84.37	90.89	**79.96**	**85.52**
SP	Gaitgraph	**93.9**	91.1	**91.7**	85.4	90.6
	GaitGL	90.11	93.98	88.60	72.47	86.29
	Ours	93.33	**96.54**	91.29	**86.88**	**92.01**

TABLE 5. Cross data experiment: trained by OUMVLP, tested on CASIA-B.

((a)) Cross data experiment: trained by CASIA-B, tested on SOTON-small

Model of CASIA-B	NM	BG	CL	SP
GaitGL	71.41	64.71	40.05	51.23
GaitGraph	53.42	52.54	35.58	52.58
Ours	75.14	74.96	44.15	69.36

((b)) Cross data experiment: trained by OUMVLP, tested on CASIA-B

Model of OUMVLP	NM	BG	CL
GaitGL	71.66	61.36	25.83
GaitGraph	35.32	22.87	13.08
Ours	73.7	64	27.8

3) BAG CARRYING AND CLOTHES CHANGES

In practical scenarios of gait recognition, subjects often carry backpacks or change clothes, which can cause interference to the accuracy of recognition. For CASIA-B, our fusion model exhibits only a 1 % decrease in performance in the BG case and demonstrates stable performance compared to single modal models Table 2. The reason behind this improvement is the incorporation of a pose-based model that directly models the human structure and provides robustness against bag carrying interference. Similarly, our fusion model provides relatively stable performance in scenarios where people change clothes, with only a 4 % decrease in performance in the CL case compared to the NM case. In contrast, single modal models exhibit inferior performance in the CL case, such as a 20 % decrease in performance for GaitGL compared to the NM case and a 14 % decrease in performance. On the SOTON-small dataset Table 3, the challenging nature of the CL case is fully exemplified, where Gaitgraph exhibits only 40 % average performance. However, our fusion model can still provide a 1 % performance improvement compared to GaitGL.

4) SPEED CHANGES

In practical scenarios, subjects may walk at different speeds, which can also affect the performance of our models. Regarding to the SOTON-small dataset from Table 4, we made an interesting observation that GaitGraph exhibits higher average performance than GaitGL. In this scenario, the pose-based model demonstrates robustness advantages, especially in the F (frontal) view, where the performance difference between the two models is significant, with 72.47 vs 85.4. Our fusion model shows a significant advantage in the F view, improving the performance to 92.

From the above analysis, it is evident that gait recognition in real-world applications can be influenced by various factors. Single modal models expose significant performance fluctuations. In addition, there is no one feature which has overwhelming dominance in all scenarios. Therefore, it is necessary to use fusion models to deal with complex scenarios in practical applications.

5) CROSS-DATA

To further verify the robustness and the generalization ability for our fusion method, two cross-data experiments are conducted. In these experiments, the training and test data are from different datasets, more specifically, the model is trained by OUMVLP \ CASIA-B data while tested on CASIA-B \ SOTON-small. From Table 5, steady accuracy improvement is exhibited by our fusion method. In the first experiment, the accuracy on CASIA-B is improved by around 2% for all the cases. For SOTON-small, significant improvement is obtained by our fusion method on the hard cases, e.g. 10% boost for bag carrying (BG) and speed changing (SP) conditions.

Even comparing to other fusion methods, our methods show advantages over other fusion methods. For CASIA-B, our methods are higher than the other methods by around 2% and 1% increment over OUMVLP dataset. The two methods focus on designing more representative single modal models with little focus on the fusion step, where they took simple summation or concatenation as the final representations. But, our method focuses on aligning two spaces so that the fusion method can take effect. As the results shown, the fusion methods by combining two complementary features can achieve high recognition accuracy, and also shows robustness against view changes and clothes changes. The robustness are very good features when making applications.

Our approach has significant improvements compared to other fusion methods. For instance, in CASIA-B's BG case, our method achieved 2% improvement, while on OUMVLP, there was a 1% increase. Different from other fusion methods focusing on designing complex single feature, our method works towards ensuring comparability between different features through feature normalisation technique, that is normalised each feature by its maximum of the distance. This allows both features to effectively exploit their advantages.

D. FUSION APPROACH COMPARISON

This article adopts the CBP fusion method, and to verify the effectiveness of our fusion method, we also conducted

experiments on other fusion methods. Table 8 shows that using the concatenation method has advantages compared to directly adding, with an improvement of 0.1% to 1% on the average level of the three datasets. This is probably because concatenation increases the dimensions of the embedding, thereby enhancing the expressive ability of the representation. However, compared with CBP, the CBP method demonstrates superior fusion capabilities. On the OUMVLP dataset, we achieved a 0.4 % improvement. Especially in the CASIA-B CL case, we achieved a 2 % improvement. In the SOTON-small dataset, we also achieved an improvement in average performance. CBP not only include the first order information but also the second order interaction of the two features, and thus has superiority in extracting more detailed interaction of two features.

E. HYPERPARAMETER

Our fusion method contains a hyperparameter d_3 as the final dimension of the CBP, and d_3 is positive related to the accuracy theoretically. We conduct experiments to test the effects of the different choice of d_3 on CASIA-B. From Table 6, it can be seen that smaller d_3, e.g. 64, leads to lower accuracy whereas the degrade is limited to 0.3%. As d_3 increases, the accuracy becomes insensitive, that is, the accuracy is very similar when d_3 being 128 and 256.

F. ABLATION STUDY

Our method contains two steps, one is normalisation and the other is fusion by CBP. When we fuse the two features, an important parameter is α where the importance of the two features are weighted. In order to study the effect of each step, we design ablation experiments regarding the effects of normalisation, and parameter α using CASIA-B as an example.

1) NORMALISATION

We compare the accuracy with and without normalisation in order to study the effect of normalisation in a quantitative way. To eliminate the influence of the specific fusion method, the experiments are conducted on both concatenation and CBP fusion methods. From Table 7, we can see that the accuracy for CL case drops significantly without normalisation for both CBP and concatenation fusion approach. The performance degrade for the other two cases is smaller and the average accuracy drop is around 1%. With this result, it can be concluded that proper normalization is a key scheme toward better fusion result.

2) WEIGHT PARAMETER

$\boldsymbol{\alpha}$, controlling the importance of different features, is also important to obtain high accuracy for our fusion method. Thus, we conduct experiments to compare the accuracy difference with or without parameter $\boldsymbol{\alpha}$. The experiment is also conducted for concatenation and CBP fusion methods for a solid verification. As suggested by Table 7, for both fusion approach, the average drop is around 2% by deleting the wright parameter $\boldsymbol{\alpha}$. For CBP fusion approach, the CL case sees a significant accuracy drop of around 3%, and for concatenation approach, the BG case is affected most with an accuracy decline of 2%.

In summary, both normalisation and weight parameter $\boldsymbol{\alpha}$ contribute to the high accuracy of the fusion method. In addition, it is worth noticing that, in both cases, the CBP fusion approach outperforms the concatenation fusion approach which also exhibits the advantages of the CBP method as a fusion approach.

TABLE 6. Hyperparameter d_3 experiments on CASIA-B.

d_3	NM	BG	CL	Avg
64	98.54	96.9	94.27	96.6
128	98.8	97.8	94.8	96.89
256	98.55	97.52	94.55	96.87

TABLE 7. Ablation experiments on CASIA-B.

Fusion approach	N[1]	W[2]	NM	BG	CL	Avg
Concat	✓	✓	98.69	97.11	92.76	96.1
	✗	✓	97.91	96.21	91.56	95.22
	✓	✗	97.34	94.85	91.28	94.49
CBP	✓	✓	98.8	97.8	94.8	96.89
	✗	✓	98.42	96.39	93.1	95.62
	✓	✗	97.26	94.89	91.66	94.61

[1] **N**: normalisation module
[2] **W**: weight parameter module

TABLE 8. Experiment results on different fusion methods.

Dataset	Status	Add	Concat	CBP
CASIA-B	NM	98.59	98.69	98.8
	BG	97.21	97.11	97.8
	CL	92.47	92.76	94.8
OUMVLP	NM	90.22	90.41	90.8
SOTON-small	NM	95.63	95.58	95.87
	CL	85.56	85.96	85.52
	BG	94.81	95.76	97.52
	SP	91.27	91.94	92.91

V. CONCLUSION

In this paper, a fusion model is proposed to combine the strength of both the silhouette and pose for gait recognition. Our method focuses on constructing a reasonable way to combine the two features, instead of designing complicated but cumbersome single feature models as most fusion methods do [39], [45]. Although GaitGL and GaitGraph are the chosen feature extractors in our paper, generally, the feature extractor can be any model. Hence, in comparison to other fusion methods [39], [45], our method is very flexible. Extensive experiments are conducted on CASIA-B,

OUMVLP and SOTON-small to show the effectiveness of our method. In addition, our method is more robust to view changes, bag carrying, clothes changes and speed changes. These factors affect the real world gait recognition significantly, and thus, our method shows great potential in real world application. Besides, cross-data experiments are also conducted to further prove the robustness. Despite the promising performance of our model, we intend to further fit more effective single modal models into our fusion models, such as 3DCNN [58] and multi scale pose model [39].

REFERENCES

[1] I. Bouchrika, M. Goffredo, J. Carter, and M. Nixon, "On using gait in forensic biometrics," *J. Forensic Sci.*, vol. 56, no. 4, pp. 882–889, Jul. 2011.
[2] A. Badiye, N. Kapoor, P. Kathane, and K. Krishan, *Forensic Gait Analysis*. Treasure Island, FL, USA: StatPearls Publishing, May 2020.
[3] S. An, Y. Tuncel, T. Basaklar, G. K. Krishnakumar, G. Bhat, and U. Y. Ogras, "MGait: Model-based gait analysis using wearable bend and inertial sensors," *ACM Trans. Internet Things*, vol. 3, no. 1, pp. 1–24, Feb. 2022.
[4] C. Shen, S. Yu, J. Wang, G. Q. Huang, and L. Wang, "A comprehensive survey on deep gait recognition: Algorithms, datasets and challenges," 2022, *arXiv:2206.13732*.
[5] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition," *IPSJ Trans. Comput. Vis. Appl.*, vol. 10, no. 1, pp. 1–14, Dec. 2018.
[6] B. Garcia-Garcia, T. Bouwmans, and A. J. R. Silva, "Background subtraction in real applications: Challenges, current models and future directions," *Comput. Sci. Rev.*, vol. 35, Feb. 2020, Art. no. 100204.
[7] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[8] B. Lin, S. Zhang, and X. Yu, "Gait recognition via effective global-local feature representation and local temporal aggregation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14628–14636.
[9] H. Chao, K. Wang, Y. He, J. Zhang, and J. Feng, "GaitSet: Cross-view gait recognition through utilizing gait as a deep set," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3467–3478, Jul. 2022.
[10] C. Fan, Y. Peng, C. Cao, X. Liu, S. Hou, J. Chi, Y. Huang, Q. Li, and Z. He, "GaitPart: Temporal part-based model for gait recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14213–14221.
[11] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5686–5696.
[12] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.
[13] T. Teepe, A. Khan, J. Gilg, F. Herzog, S. Hörmann, and G. Rigoll, "Gait-graph: Graph convolutional network for skeleton-based gait recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 2314–2318.
[14] M. Jiang, J. Dong, D. Ma, J. Sun, J. He, and L. Lang, "Inception spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. Int. Symp. Control Eng. Robot. (ISCER)*, Feb. 2022, pp. 208–213.
[15] R. Cadene, H. Ben-younes, M. Cord, and N. Thome, "MUREL: Multimodal relational reasoning for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1989–1998.
[16] H. Ben-Younes, R. Cadene, M. Cord, and N. Thome, "MUTAN: Multimodal tucker fusion for visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2631–2639.
[17] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 317–326.
[18] Niyogi and Adelson, "Analyzing and recognizing walking figures in XYT," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 1994, pp. 469–474.
[19] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer, "The humanID gait challenge problem: Data sets, performance, and analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 162–177, Feb. 2005.
[20] G. V. Veres, M. S. Nixon, and J. N. Carter, "Model-based approaches for predicting gait changes over time," in *Proc. Int. Conf. Intell. Sensors, Sensor Netw. Inf. Process.*, Dec. 2005, pp. 325–330.
[21] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, Feb. 2006.
[22] K. Bashir, T. Xiang, and S. Gong, "Gait recognition using gait entropy image," in *Proc. 3rd Int. Conf. Imag. Crime Detection Prevention (ICDP)*, Dec. 2009, pp. 1–6.
[23] Q. Ma, S. Wang, D. Nie, and J. Qiu, "Recognizing humans based on gait moment image," in *Proc. 8th ACIS Int. Conf. Softw. Eng., Artif. Intell., Netw., Parallel/Distrib. Comput. (SNPD)*, vol. 2, Aug. 2007, pp. 606–610.
[24] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001.
[25] S. Hou, X. Liu, C. Cao, and Y. Huang, "Set residual network for silhouette-based gait recognition," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 3, no. 3, pp. 384–393, Jul. 2021.
[26] X. Huang, D. Zhu, H. Wang, X. Wang, B. Yang, B. He, W. Liu, and B. Feng, "Context-sensitive temporal feature learning for gait recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12889–12898.
[27] A. Sepas-Moghaddam and A. Etemad, "View-invariant gait recognition with attentive recurrent learning of partial representations," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 3, no. 1, pp. 124–137, Jan. 2021.
[28] B. Lin, S. Zhang, Y. Liu, and S. Qin, "Multi-scale temporal information extractor for gait recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 2998–3002.
[29] J. Zheng, X. Liu, X. Gu, Y. Sun, C. Gan, J. Zhang, W. Liu, and C. Yan, "Gait recognition in the wild with multi-hop temporal switch," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 6136–6145.
[30] J. Liang, C. Fan, S. Hou, C. Shen, Y. Huang, and S. Yu, "Gaitedge: Beyond plain end-to-end gait recognition for better practicality," in *Proc. Eur. Conf. Comput. Vis.* Tel Aviv, Israel: Springer, Oct. 2022, pp. 375–390.
[31] T. Chai, X. Mei, A. Li, and Y. Wang, "Silhouette-based view-embeddings for gait recognition under multiple views," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 2319–2323.
[32] A. F. Bobick and A. Y. Johnson, "Gait recognition using static, activity-specific parameters," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Dec. 2001, p. I.
[33] M. S. Nixon and J. N. Carter, "Automatic recognition by gait," *Proc. IEEE*, vol. 94, no. 11, pp. 2013–2024, Nov. 2006.
[34] A. Y. Johnson and A. F. Bobick, "A multi-view method for gait recognition using static body parameters," in *Proc. Int. Conf. Audio- Video-Based Biometric Person Authentication*, 2001, pp. 301–311.
[35] C. Yam, M. S. Nixon, and J. N. Carter, "Automated person recognition by walking and running via model-based approaches," *Pattern Recognit.*, vol. 37, no. 5, pp. 1057–1072, May 2004.
[36] D. K. Wagg and M. S. Nixon, "On automated model-based extraction and analysis of gait," in *Proc. 6th IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2004, pp. 11–16.
[37] D. K. Wagg and M. S. Nixon, "Automated markerless extraction of walking people using deformable contour models," *Comput. Animation Virtual Worlds*, vol. 15, nos. 3–4, pp. 399–406, Jul. 2004.
[38] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12018–12027.
[39] Y. Peng, K. Ma, Y. Zhang, and Z. He, "Learning rich features for gait recognition by integrating skeletons and silhouettes," 2021, *arXiv:2110.13408*.
[40] Md. B. Hasan, T. Ahmed, and M. H. Kabir, "HEATGait: Hop-extracted adjacency technique in graph convolution based gait recognition," in *Proc. 4th Int. Conf. Adv. Comput. Technol., Inf. Sci. Commun. (CTISC)*, Apr. 2022, pp. 1–6.
[41] S. Gao, J. Yun, Y. Zhao, and L. Liu, "Gait-D: Skeleton-based gait feature decomposition for gait recognition," *IET Comput. Vis.*, vol. 16, no. 2, pp. 111–125, Mar. 2022.

[42] E. Pinyoanuntapong, A. Ali, P. Wang, M. Lee, and C. Chen, "GaitMixer: Skeleton-based gait representation learning via wide-spectrum multi-axial mixer," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.
[43] L. Wang, J. Chen, Z. Chen, Y. Liu, and H. Yang, "Multi-stream part-fused graph convolutional networks for skeleton-based gait recognition," *Connection Sci.*, vol. 34, no. 1, pp. 652–669, Dec. 2022.
[44] M. Faundez-Zanuy, "Data fusion in biometrics," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 20, no. 1, pp. 34–38, Jan. 2005.
[45] L. Wang, R. Han, and W. Feng, "Combining the silhouette and skeleton data for gait recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.
[46] J. Zheng, X. Liu, W. Liu, L. He, C. Yan, and T. Mei, "Gait recognition in the wild with dense 3D representations and a benchmark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20196–20205.
[47] G. Li, L. Guo, R. Zhang, J. Qian, and S. Gao, "TransGait: Multimodal-based gait recognition with set transformer," *Int. J. Speech Technol.*, vol. 53, no. 2, pp. 1535–1547, Jan. 2023.
[48] L. Zhao, L. Guo, R. Zhang, X. Xie, and X. Ye, "MmGaitSet: Multimodal based gait recognition for countering carrying and clothing changes," *Int. J. Speech Technol.*, vol. 52, no. 2, pp. 2023–2036, Jan. 2022.
[49] R. N. Yousef, A. T. Khalil, A. S. Samra, and M. M. Ata, "Model-based and model-free deep features fusion for high performed human gait recognition," *J. Supercomput.*, pp. 1–38, 2023.
[50] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1449–1457.
[51] N. Pham and R. Pagh, "Fast and scalable polynomial kernels via explicit feature maps," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2013, pp. 239–247.
[52] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.
[53] J. Weston, S. Bengio, and N. Usunier, "Wsabie: Scaling up to large vocabulary image annotation," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, 2011, pp. 2764–2770.
[54] H. Hotelling, "Relations between two sets of variates," in *Breakthroughs in Statistics*. New York, NY, USA: Springer, 1992.
[55] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, vol. 4, Aug. 2006, pp. 441–444.
[56] J. Shutler. (2002). *Small Population, in Depth Database*. [Online]. Available: https://web-archive.southampton.ac.uk/www.gait.ecs.soton.ac.uk/index.html/
[57] W. An, S. Yu, Y. Makihara, X. Wu, C. Xu, Y. Yu, R. Liao, and Y. Yagi, "Performance evaluation of model-based gait on multi-view very large population database with pose sequences," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 2, no. 4, pp. 421–430, Oct. 2020.
[58] Z. Huang, D. Xue, X. Shen, X. Tian, H. Li, J. Huang, and X.-S. Hua, "3D local convolutional neural networks for gait recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14900–14909.

YANG ZHAO was born in Shandong, China. She received the B.S. degree in financial mathematics from Xi'an Jiaotong Liverpool University, China, in 2012, and the M.S. degree in financial mathematics and the Ph.D. degree in statistics from The University of Warwick, U.K., in 2013 and 2019, respectively.

Since 2019, she has been a Researcher on Bayesian neural network, explainable AI, and recommendation system, in various research facilities, such as NEC, and Career Science Laboratory, BOSS ZhiPin. She joined Fujitsu Research and Development Center Company Ltd., in 2022, with a focus on gait recognition.

RUJIE LIU received the M.S. and Ph.D. degrees in signal and information processing from Beijing Jiaotong University, China, in 1998 and 2001, respectively.

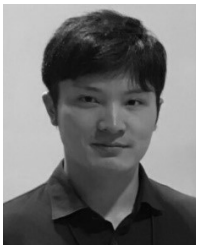
Since 2001, he has been a Researcher with Fujitsu Research and Development Center Company Ltd. He has published more than 40 papers and tens of patents. His research interests include AI, pattern recognition, and image processing.

WENQIAN XUE was born in Shanxi. She received the B.S. and M.S. degrees in communication engineering from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2012 and 2015, respectively.

Since 2015, she has been with Fujitsu Research and Development. She focused on the IoT fault detection with machine learning, from 2015 to 2019. Since 2019, she has been shifts the working interests from machine learning to deep learning, now she is interested in gait recognition and real-scene applications.

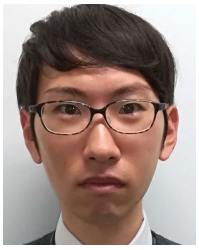
MING YANG was born in Yichang, Hubei, in 1986. He received the M.S. degree in software technology from the School of Computer Science, Wuhan University, in 2012.

Since 2012, he has been a Researcher with Fujitsu Research and Development. His research interests include natural language processing, linked open data, data mining, machine learning, optimization techniques, supply chain forecasting, and gait recognition.



MASAHIRO SHIRAISHI was born in Chiba, Japan, in 1994. He received the B.S. and M.S. degrees in electronics from Meiji University, Tokyo, Japan, in 2018 and 2020, respectively.

Since 2020, he has been a Researcher with Fujitsu, Kawasaki, Japan. His research interest includes computer vision technology in psychophysics. He is currently studying human action recognition technology using camera or radar sensors.



SHUJI AWAI was born in Osaka, Japan, in 1989. He received the B.S. degree from the Faculty of Informatics, in 2012, and the M.S. degree in informatics from the Osaka Institute of Technology, Osaka, in 2014.

Since 2022, he has been a Researcher with Fujitsu, Kawasaki, Japan. His research interests include computer vision, such as action recognition and tracking. He is currently studying gait recognition technology using pose estimation.

YU MARUYAMA was born in Kyoto, Japan, in 1991. He received the B.S. degree in computer engineering from Shinshu University, Nagano, Japan, in 2014, and the M.S. degree in informatics from Kyoto University, Kyoto, in 2016.

From 2016 to 2022, he was a Researcher with Omron, Kyoto. Since 2022, he has been a Researcher with Fujitsu, Kawasaki, Japan. His research interests include vision system for a ping-pong robot, the development of bin-picking algorithm for robot vision, and the development of visual inspection algorithm for deep learning. He is currently studying gait recognition technology using camera.

TAKAHIRO YOSHIOKA was born in Saga, Japan, in 1984. He received the B.S., M.S., and Ph.D. degrees in engineering from Kyushu University, Fukuoka, Japan, in 2007, 2009, and 2012, respectively.

Since 2012, he has been a Researcher with Fujitsu, Kawasaki, Japan. His research interests include computer vision, such as camera calibration, pattern recognition, tracking, and applied computer vision technology in psycho-physics. He is currently studying human action recognition technology using camera or radar sensors.

Dr. Yoshioka is a Committee Member of the Institute of Electronics, Information and Communication Engineers, Japan.

TAKESHI KONNO was born in Fukushima, Japan, in 1977. He received the B.S. and M.S. degrees in engineering from Tokyo Denki University, Saitama, Japan, in 2001 and 2003, respectively.

Since 2010, he has been a Researcher with Fujitsu, Kawasaki, Japan. His research interest includes computer vision technology in psychophysics. He is currently studying human action recognition technology using camera or radar sensors. He is a Committee Member of the Japan Electronics and Information Technology Industries Association.