PoseMapGait: A model-based Gait Recognition Method with Pose Estimation Maps and Graph Convolutional Networks

Rijun Liao[a], Zhu Li[a,*], Shuvra S. Bhattacharyya[b], and George York[c]

[a]*Department of Computer Science and Electrical Engineering, University of Missouri-Kansas City, MO, USA.*
[b]*Department of Electrical and Computer Engineering and UMIACS, University of Maryland, College Park, MD, USA.*
[c]*Department of Electrical and Computer Engineering and UAS Research Center, US Air Force Academy, Colorado Springs, CO, USA.*

Abstract

Gait recognition is a particularly effective way to avoid the spread of COVID-19 while people are under surveillance. Because of its advantages of non-contact and long-distance identification. One category of gait recognition methods is appearance-based, which usually extracts human silhouettes as the initial input feature and achieves high recognition rates. However, the silhouette-based feature is easily affected by the view, clothing, bag, and other external variations. Another category is based on model-based, one popular model-based feature is extracted from human skeletons. The skeleton-based feature is robust to many variations because it is less sensitive to human shape. However, the performance of skeleton-based methods suffers from recognition accuracy loss due to limited input information. In this paper, instead of relying on coordinates from skeletons, we exploit that *pose estimation maps*, the byproduct of pose estimation. It not only preserves richer cues of the human body compared

⋆

*Corresponding author

Email addresses: rlyfv@mail.umkc.edu (Rijun Liao), lizhu@umkc.edu (Zhu Li), ssb@umd.edu (Shuvra S. Bhattacharyya), george.york@usafa.edu (and George York)

The views expressed in this article, book, or presentation are those of the author and do not necessarily reflect the official policy or position of the United States Air Force Academy, the Air Force, the Department of Defense, or the U.S. Government. Approved for public release: distribution unlimited. PA#: USAFA-DF-2022-568

with the skeleton-based feature, but also keeps the advantage of being less sensitive to human shape compared with the silhouette-based feature. Specifically, the evolution of pose estimation maps is decomposed as one heatmaps evolution feature (extracted by *gaitMap-CNN*) and one pose evolution feature (extracted by *gaitPose-GCN*), which denote the invariant features of whole body structure and body pose joints for gait recognition, respectively. Our method is evaluated on two large datasets, CASIA-B and the CMU Motion of Body (MoBo) dataset. The proposed method achieves the new state-of-the-art performance compared with recent advanced model-based methods.

Keywords: COVID-19, Gait Recognition, Pose Estimation Maps, Heatmaps Evolution Feature, Poses Evolution Feature, Graph Convolutional Networks (GCN)

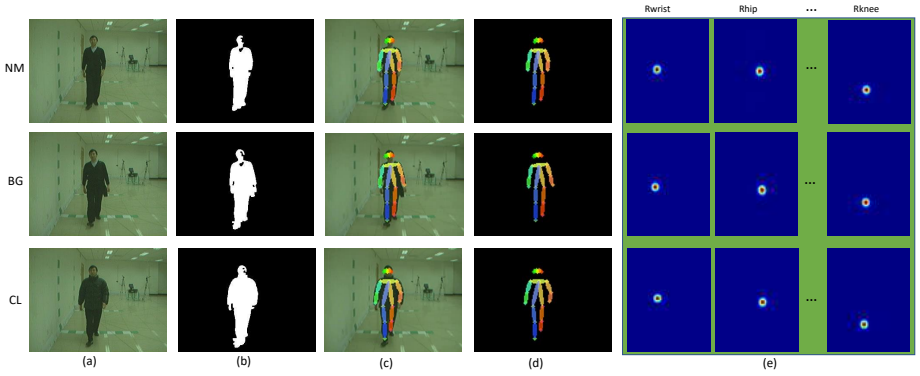


Figure 1: Comparison of three types of raw input feature in three walking conditions on CASIA-B dataset. NM: normal walking, BG: walking with a bag, CL: walking with a coat. a) Original video frames. b) Human silhouette-based input feature. c) Human pose estimation results. d) Human skeleton-based input feature. e) Human pose estimation maps based input feature.

1. Introduction

1.1. Motivation

With the outbreak of novel coronavirus 2019 (COVID-19), the development of biometric recognition technologies to address the various concerns induced by the rapid spread of COVID-19 has become urgent. For contact biometrics such as fingerprint and palm-print, it is clear that they would accelerate the spread of the virus. For non-contact biometrics, face recognition is one of the mature biometric recognition technologies. But it is very challenging to identify subjects when people wear facial masks. The iris recognition also faces challenges when people wear virus protection glasses, and also brings some risks that people may touch the devices due to the iris data collection at close range.

Gait, is a walking style of a person, which also can be used as a biometric feature to identify a person. Compared with the above biometric features, gait has its unique advantages such as being non-contact and hard to fake. More importantly, gait is still available at a long-distance human identification, which is particularly suitable for monitoring people during the period of COVID-19. Since non-contact and long-distance are two important factors to avoid the rapid spread of COVID-19. Gait recognition technology also has a great potential application in other areas, such as video surveillance, crime prevention, and forensic identification.

Gait is a behavioral biometric, it would change drastically when there are some variations, such as view, carrying, clothing, and occlusion. In order to improve the robustness of extracted features, some earlier model-based approaches [1, 2] tried to capture motion patterns by modeling the human body for each subject. However, it is very challenging to locate and track the human body accurately at that earlier time because of technical reasons.

The appearance-based gait recognition approaches [3, 4] usually extract the human silhouettes (Fig. 1 (b)) from RGB images as raw input data. These approaches are more popular than the model-based ones in the past two decades because human silhouettes are easy to obtain and can achieve high recognition

rates. However, there are many drastic variations in real applications, such as changes in clothing or carrying. These variations would change the human silhouette shape greatly and give rise to reducing performance dramatically. By contrast, model-based approaches are not so sensitive to human shape and human appearance because they focus on human body structure and movement modeling.

Recently, with the development of deep learning and human body pose estimation. The performance of locating and tracking human body parts becomes more and more accurate, which brings hope to the model-based approaches. Some works [5, 6, 7, 8, 9] extracted accurate human skeleton feature (Fig. 1 (d)) by using the human body pose estimation algorithm (Fig. 1 (c)). These works have achieved good performance and made a great contribution to the development of model-based approaches. But these works suffer from recognition accuracy loss compared with appearance-based approaches. One main reason for this is that the skeleton usually consists of several body joint coordinates, which is a low dimensional feature and the contained information is very limited compared with the human silhouette.

In this paper, instead of relying on the coordinates from human joints, we exploit pose estimation maps (Fig. 1 (e)), the byproduct of pose estimation. We find that pose estimation maps not only preserve richer cues of the human body to benefit gait recognition compared with the skeleton-based feature, but also are less sensitive to human appearance compared with the silhouette-based feature. Inspired by the popular work [10] of human action recognition, we, therefore, propose a novel model-based gait recognition method, *PoseMapGait*, which exploits human pose estimation maps as the raw input data. Different from [10] which created two handcrafted images from heatmaps and poses data before feeding into CNN, our invariant gait feature is learned automatically from heatmaps and poses data by set pooling and gait graph construction. Simulation results demonstrated that the pose estimation maps feature can bring significant performance improvement compared with recent advanced model-based approaches.

1.2. Method Overview and Contributions

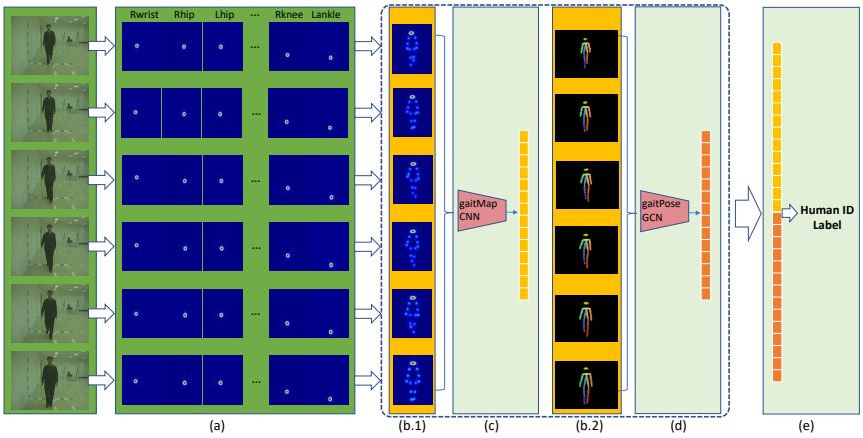


Figure 2: The overview of the proposed method. a) Pose estimation maps of each body part are predicted by extracting the byproduct of pose estimation. b) For each frame, pose estimation maps are aggregated to generate a heatmap (b.1) and a pose (b.2). c) Heatmaps evolution feature extraction, which denotes the invariant feature of body structure. d) Poses evolution feature extraction, which denotes the invariant feature of the body pose joints. e) In the inference stage, two types of evolution features are concatenated to measure the similarity between the gallery and probe videos, and then predict the human ID label.

The overview of the proposed method is shown in Fig. 2. Given each frame of a video, we predict a pose estimation map for each body part by extracting the byproduct of pose estimation. These pose estimation maps not only can preserve global information, which reflects whole shapes that suffer less from the appearance noise, but also preserve local information, which reflects the location movement of body parts.

To reduce the redundancy of pose estimation maps, we average pose estimation maps of all body parts to form an averaged pose estimation map (heatmap, Fig. 2 (b.1)) for each frame. More importantly, the heatmap can better represent global human body structure information compared with separate pose estimation maps (Fig. 2 (a)). According to the study of Liu *et al.* [10], the averaged pose estimation map (heatmap) provides richer information to reflect human body structure and is beneficial to object recognition. In order to ex-

tract high-level spatial-temporal information of body structure from a sequence of heatmap images, the heatmap convolutional neural networks (*gaitMap-CNN*) is designed to extract heatmaps evolution feature.

Since the heatmap image has no explicit differentiation of body parts, that is, there is no connectivity relationship between body part joints. We further predict joint location from the pose estimation map of each body part, generating a pose (Fig. 2 (b.2)) for each frame to extract the body pose invariant feature. Unlike the above skeleton-based methods [5, 6], which merely considered a sequence of human joint coordinates modeling, we construct a gait graph that not only considers inter-frame connection with the same joints, but also considers intra-body connection based on naturally connected human body joints. The pose graph convolutional networks (*gaitPose-GCN*) is designed to extract high-level poses evolution feature from the gait spatial-temporal graph.

Intuitively, the heatmaps evolution feature (Fig. 2 (c) yellow vector) and poses evolution feature (Fig. 2 (d) orange vector) benefit the recognition of general movements of global body structure and elaborate movements of body parts. Thereby, both features are fused to generate the discriminative feature and predict the human ID label.

Compared with appearance-based methods [3, 4, 11], the proposed method has more robustness and enhances the utility of gait recognition in real applications. Unlike most appearance-based works [3, 4, 11], they usually use human silhouettes as the initial input data, the silhouette would be changed greatly when some big variations exist in the real world, as shown in the variation of walking with a coat in Fig. 1 (b). We exploit the pose estimation maps, which are not so sensitive to human shape. In addition, they [3, 4] would ignore the human body part modeling because the human silhouette is a kind of image which is combining hands, feet, and other human parts together. Fan *et al.* [12] divide silhouette equally into four parts in order to model the body part movement. However, it can not strictly divide the human body structure. In contrast, pose estimation maps can model the local movement more accurately as it consists of separate human body joints.

Compared with model-based methods [6, 7, 9], the proposed method provides a new way to exploit a model-based algorithm, which inspires researchers to rethink model-based approaches and promotes the development of gait recognition. Unlike previous works [5, 6, 7, 9], they usually extract discriminative features from human body skeletons (Fig. 1 (c)), which suffer from accuracy loss due to limited input information. We abandon the human pose and use a more informative feature, *pose estimation maps*, the byproduct of pose estimation, to solve this challenge. In addition, PTSN [5], PoseGait [9] and other works [6, 7] use CNN or LSTM to analyze coordinates from skeletons, they partly ignore the human topological structure such as the connectivity relationship between body part joints. While we construct a gait graph to analyze both inter-frame connections with the same joints and intra-body connections based on naturally connected human body joints.

To summarize, our contributions are three-fold.

- **Flexible:** We propose a novel model-based gait recognition method as the evolution of pose estimation maps, called *PoseMapGait*, which exploits human pose estimation maps as the initial input data. Compared with appearance-based methods that use human silhouettes as input data, the pose estimation maps are less sensitive to human shape. In addition, they have richer information compared with the model-based methods that use human skeletons as input data. The visualization of three types of input data is shown in Fig 1.

- **Interpretable:** Instead of using pose estimation maps directly, the evolution of pose estimation maps is decomposed as an evolution feature (Fig. 2 (c) yellow vector) of heatmaps and an evolution feature (Fig. 2 (d) orange vector) of estimated 2D human poses in a biologically interpretable way, which denote the invariant features of whole body structure and body pose joints for gait recognition, respectively.

- **Effective:** Some experiments are performed on popular gait dataset CASIA-B [13] and the CMU Motion of Body (MoBo) dataset [14]. Compared with

previous model-based methods using skeleton pose information or with the assistance of hand-crafted features, our models achieve a state-of-the-art recognition rate. Experiment results show that our proposed method is more robust to various variations, which enhances the utility of gait recognition in real applications.

2. Related Work

In this section, we will briefly review existing gait recognition methods. Approaches in the recent gait recognition literature can be roughly grouped into two categories, appearance-based and model-based approaches. We also briefly introduce graph convolutional neural networks in this section.

2.1. Appearance-Based Approaches

Appearance-based methods usually use the human silhouettes as raw input data, and these methods can be also roughly divided into two categories, namely template-based approaches and sequence-based approaches.

Template-based approaches would create a gait template by rendering pixel-level operators on the human silhouette images. Template creation and template matching are common pipelines of template-based approaches. Gait Energy Image (GEI) template [15] and Chrono-Gait Image (CGI) template [16] are two very popular gait template features. In the template matching step, the most common solutions are to reduce the effect of view variation by using View Transformation Model (VTM). VTM can transform gait template features from one view to another view for improving the discriminative capability of the template feature. Like Yu *et al.* [17] proposed Stacked Progressive Auto-Encoders (SPAE) can transform GEI with arbitrary views to a specific angle GEI. GaitGANv2 [18] was proposed to directly deal with the view, bag, and clothing variation by using a generative adversarial network model, while SPAE [17] requires 7 stacks to deal with small view variation one by one. To lighten the burden of view-invariant feature extraction for CNNs, DV-GEIs [4, 3] was proposed

to provide a much denser view sampling to deal with the cross-view problem. These template-based methods have made a great contribution to the development of gait, however, the performance is not good enough because it would reduce some temporal information during the process of template generation.

Sequence-based approaches directly employ a sequence of human gait features like human silhouettes as input data. Wu *et al.* [19] proposed the first work based on deep CNNs for gait recognition to extract gait features from a sequence of human silhouettes. Different from Wu *et al.* [19] which uses continuous human silhouettes, Chao *et al.* [11] introduced the GaitSet network to further improve performance based on unordered silhouettes set. Rather than dealing with human silhouettes for gait recognition, GaitNet [20] was proposed to explicitly disentangle pose and appearance features from RGB images. Sequence-based approaches can achieve high performance in terms of cross-view condition. This is because a sequence of human gait features contains rich temporal information compared with template-based methods. However, it can not deal with cross-carrying and cross-clothing variations very well. The main reason is that human appearance and shape can be changed greatly when these variations exist in the real world, and lead to a decrease in performance.

2.2. Model-Based Approaches

The model-based approaches extract features through modeling human body structure and analyzing movement patterns of different human body parts. These methods are robust to many variations because they are not so sensitive to human appearance compared with appearance-based approaches.

In the early works, model-based approaches are not an easy task because it requires human bodies are correctly and high accurately modeled. To obtain human joint positions, some earlier methods [1] even mark human body parts manually or with the assistance of some specific devices. Nixon *et al.* [21] argue that human body movement has the ability to recognize different subjects' gait patterns. They simulate legs and leg movement by using a simple stick model and an articulate pendulum movement. A multi-connected rigid body model

was proposed by Wang *et al.* [2]. They divide into 14 parts and each part is connected through a joint. This work shows that the changes in the angle of each joint are beneficial to extracting the temporal information of gait.

In recent years, with the development of pose estimation algorithms. Some researchers [22] extract accurate human body joints information from an RGB image or a video by using pose estimation models. Feng *et al.* [23] used a Long Short Term Memory recurrent neural network to extract temporal feature from human joints. The body structure spatial information was lost because the authors just considered temporal information from each human joint heatmap separately. Liao *et al.* [5] proposed a pose-based temporal-spatial network (PTSN) to extract static and dynamic information from the human body skeleton. PTSN-3D [6] was proposed to future improve its robustness to view variation by estimating 3D pose from a single image. In the following years, Liao *et al.* [9] introduced PoseGait based on the human body pose and human prior knowledge. This method can achieve a high recognition rate despite the low dimensional feature with only 14 body joints. In order to promote the study of model-based approaches, OU-ISIR provides a multi-view large population dataset with pose sequence [7], this dataset is opened to the public for research. These works boost greatly the development of model-based approaches, but it still needs to further improve the recognition rate due to limited input information compared with appearance-based methods.

2.3. Graph Convolutional Neural Networks

Traditional neural networks CNN or LSTM usually process data with grid attributes (such as images), but many data have a topological structure in daily life and scientific research. Recently, graph convolutional networks GCN appeared and developed quickly. There are two types of convolution operations according to the high-dimensional domain. The first one is based on the spectral domain and the second one is based on the spatial domain. The first one uses the eigenvalues and eigenvectors of the Laplacian matrix of the graph into spectrum [24]. The second one processes the nodes in the graph and their neigh-

boring nodes based on some rules. The spatial-temporal graph convolutional networks (ST-GCN) [25] is based on the second one. ST-GCN has achieved representative performance by applying graph convolution to the human action recognition field. Our paper is also based on the spatial domain.

Recently, graph convolutional neural networks (GCN) has been successfully applied in many works for human action recognition. Yan *et al.* [25] proposed an ST-GCN network to extract spatial-temporal feature from the human skeleton. Graph network is also widely used for other fields, such as point cloud compression [26] and sparse feature extraction [27]. In contrast, GCN is not used often for gait recognition. This is because gait recognition usually uses human silhouettes as gait raw input feature, and a human silhouette is an image that lacks a topological structure. In this paper, we exploit pose estimation maps in gait recognition and extract their spatial-temporal information by constructing a gait graph.

3. Generation of Pose Estimation Maps

In this section, we will describe the generation of robust gait input features. Given a video of people's walking sequence, we extract the pose estimation maps (Fig. 2 (a)) from each frame, and then generate a heatmap (Fig. 2 (b.1)) and a pose (Fig. 2 (b.2)) to represent human body characteristic of each frame.

Pose Estimation Maps: The goal of human pose estimation can be modeled as a structure prediction problem. Fang *et al.* [21] proposed AlphaPose which is an accurate multi-person pose estimator. Instead of directly using AlphaPose [21] to evaluate the coordinates of each human pose joints, we exploit the hidden layer of AlphaPose to extract pose estimation maps for body joints. AlphaPose [21] mainly includes three steps to evaluate human pose. 1) detect the bounding box of human, 2) predict estimation maps for body joints 3) predict coordinates of each body joint based on predicted estimation maps. In fact, the pose estimation map is a byproduct of pose estimation.

Let m_i denote the pose estimation map from body joint i. The whole pose

estimation maps output can be formulated as $M = \{m_1, m_2, ..., m_N\}$, where N is the total number of body joints. There are 17 pose estimation maps, including one joint of the nose, and two joints (right and left one) of the eyes, ears, shoulders, elbows, wrists, hips, knees, and ankles, as shown in Fig. 2 (a). These pose estimation maps will be normalized in a fixed human bounding box during the extraction. Consequently, the human bodies of different subjects will be normalized to a fixed size, which removes the variation of a human body size changes due to the different distances between the subject and the camera.

Heatmaps & **Poses:** For a RGB frame of a video, N types of pose estimation maps are predicted, namely $\{m_1, m_2, ..., m_N\}$. Since the pose estimation map of each body part is separate, we average them and as describe a heatmap h (Fig. 2 (b.1)) to better represent global human body structure information. The averaging operation can also reduce the redundancy of pose estimation maps. The heatmap h can be expressed as follows:

$$h = \frac{1}{N} \sum_{i=1}^{N} m_i \tag{1}$$

We further predict joint location from pose estimation map of each body part, generating a pose (Fig. 2 (b.2)) for each frame to extract body pose invariant feature. The pose consists of N coordinates of joints, that is $\{v_1, v_2, ..., v_N\}$. Each v_i has 2D coordinates (x, y) and one confidence score c. v_i is often estimated via Maximum A Posterior (MAP) criterion [28]. For each joint's coordinates and confidence score can be expressed as:

$$v_i\{x, y\} = \arg\max_{v \in Z}(m_i) \tag{2}$$

$$v_i\{c\} = \max_{v \in Z}\{m_i\} \tag{3}$$

where $Z \in \mathbb{R}^2$ denote all coordinates on the image m_i. The confidence score c is the maximum value of pose estimation map. In the end, each frame of a video is described as a heatmap and a pose. Therefore, the video is converted to the evolution of heatmaps and the evolution of poses.

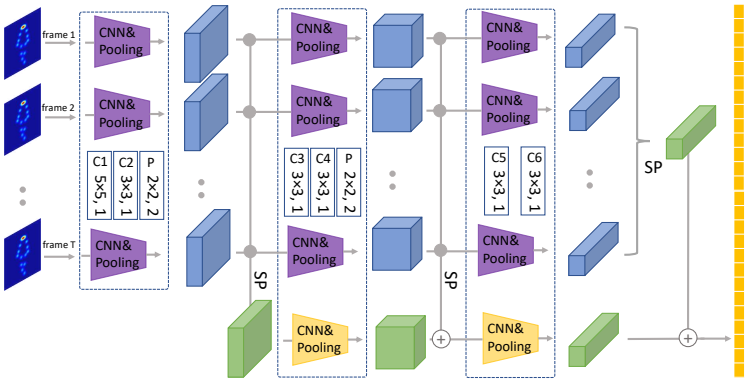


Figure 3: The structure of *gaitMap-CNN*. "SP" represents set pooling, it aims to aggregate feature maps from every frame of heatmap into a set feature map.

4. Evolution of Pose Estimation Maps

This section describes the high-level evolution feature extraction from pose estimation maps by *gaitMapPose-Net*, which consists of two streams, namely, heatmap convolutional neural networks (*gaitMap-CNN*) and pose graph convolutional networks (*gaitPose-GCN*). *gaitMap-CNN* is used to extract the heatmaps evolution feature, while *gaitPose-GCN* is developed to extract the poses evolution feature.

4.1. Heatmaps Evolution Feature

Given a dataset with T frames heatmaps. A set of n heatmaps $H = \{h_t | t = 1, 2, ..., T\}$ are put into heatmap convolutional neural networks (*gaitMap-CNN*). The structure of *gaitMap-CNN* is inspired by the network framework of Chao *et al.* [11] and [29], as shown in Fig. 3. H is a tensor with four dimensions, that is, set dimension, image channel dimension, image height dimension, and image width dimension. We use 3 steps to deal with the gait recognition, formulated as:

$$f_{map} = G(F(H)) \tag{4}$$

where F is a convolutional network aims to extract the frame-level features ($\{f^t|t=1,2,...,T\}$) from each gait heatmap. G is a function which used to map a set of frame-level features ($\{f^t|t=1,2,...,T\}$) to a set-level feature f_{map}. G takes set frame-level features as an input, it is a permutation invariant function which is formulated as:

$$G(\{f^t|t=1,2,...,T\}) = G(\{f^{\pi(t)}|t=1,2,...,T\}) \tag{5}$$

where π is any permutation [30], this operation makes gait immune to the permutation of frames based on the set perspective. And can naturally integrate frames from different videos under different scenarios. G is implemented by an operation called set pooling $G(\cdot) = max(\cdot) + mean(\cdot) + median(\cdot)$, which aims to aggregate gait information of elements in a set. Compared with typical convolutional neural networks which miss the temporal information extraction, set pooling extracts the set-level feature from high-level feature maps, it not only preserves temporal information well, but also processes spatial information sufficiently. The diagram of set pooling can be shown in Fig. 3. f_{map} is the output heatmaps evolution set-level feature.

4.2. Poses Evolution Feature

Gait Graph Construction: Heatmap captures more global body structure information of the gait sequence, while for better recognition performance, body pose information captured by the skeleton key points is also important. Inspired by ST-GCN [25], we compute a body pose information embedding from the skeleton keypoints spatial-temporal graph by employing GCN based framework. A representation of the pose sequences is generated by using pose graph convolutional networks (*gaitPose-GCN*). Specifically, given a dataset with N joints and T frames, we create an undirected spatial-temporal graph as the following formula:

$$G = (V, E) \tag{6}$$

$$V = \{v_{ti} | t = 1, ..., T, i = 1, ..., N\} \tag{7}$$

$$E_S = \{v_{ti} v_{tj} | (i, j) \in \Phi\} \tag{8}$$

$$E_F = \{v_{ti} v_{(t+1)j}\} \tag{9}$$

where the node-set V consists of all the joints in a T frames skeleton sequence. The number of input joints is 18 in the framework of ST-GCN [25], while the number of output heatmaps in AlphaPose [21] is 17, without neck joint. In order to make the joints apply to the ST-GCN network, the mean of the left shoulder and right shoulder is the neck joint.

E is called edge set, composed of two subsets. The first subset $E_S = \{v_{ti} v_{tj} | (i, j) \in \Phi\}$ represents the intra-joint connection at each frame, it represents the spatial gait information. Where Φ is the set of naturally connected human body joints, $\Phi = \{(1,0), (1,2), (2,3), (3,4), (1,5), (5,6), (6,7), (1,8), (8,9), (9,10), (1,11), (11,12), (12,13), (0,14), (0,15), (16,14), (15,17)\}$, as shown in Fig. 4 gray edges. The second subset depicts the inter-frame connection with same joints in consecutive frames, denoted as $E_F = \{v_{ti} v_{(t+1)j}\}$, as shown in Fig. 4 blue edges, it represents the temporal gait information.

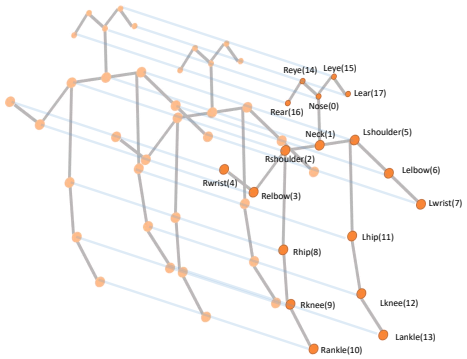


Figure 4: Gait Graph Construction. Orange dots denote the body joints. Gray edges denote the intra-body edges set E_S which represents spatial graph of poses. Blue edges denote the inter-frame edges set E_F which represents temporal graph of poses.

Graph Convolution Neural Networks for the poses evolution feature extraction is formulated as:

$$f_{pose}(v_{ti}) = \sum_{v_{tj} \in B(v_{ti})} \frac{1}{Z_{ti}(v_{tj})} f_{in}(v_{tj}) \cdot \mathbf{w}(l_{ti}(v_{tj})) \tag{10}$$

where f_{pose} is output poses evolution feature, $\mathbf{w}$ is the weight function to provide a weight vector for computing the inner product with input feature l_{ti}. Mapping function $l_{ti}(v_{tj}) = d(v_{tj}, v_{ti})$ is the label map for the single frame case at vertex v_{tj}. Here $d(v_{tj}, v_{ti})$ denotes the minimum length of any path from v_{tj} to v_{ti}, for example, $d = 0$ refers to the root node itself and $d = 1$ refers to the remaining neighbor nodes. The convolution operation on graphs is defined to the cases where the input features map f_{in} resides on a spatial graph $G(V, E)$. The normalizing term is defined as $Z_{ti}(v_{tj}) = \{v_{tk} | l_{ti}(v_{tk}) = l_{ti}(v_{tj})\}$. $B(v_{ti}) = \{v_{qj} | d(v_{tj}, v_{ti}) \leq K, |q - t| \leq \Gamma/2\}$, where K is the spatial range, Γ is the temporal range, called the temporal kernel size.

4.3. Feature Fusion and Loss Function

In order to make the gait feature more discriminative. The heatmaps evolution feature f_{map} and poses evolution feature f_{pose} are concatenated to a final invariant gait feature f_{gait}, which benefit the recognition of general movements of global body structure and elaborate movements of body parts, formulated as:

$$f_{gait} = cat(f_{map}, f_{pose}) \tag{11}$$

where *cat* means concatenate operation. The corresponding features f_{gait} among different subjects will be used to compute the loss value by triplet loss function [31], as shown in equation 12. Where d means the distance between two features, f^a_{gait}, f^p_{gait} and f^n_{gait} denote the anchor sample, positive sample and the negative sample, respectively.

$$L_{triplet} = max(d(f^a_{gait}, f^p_{gait}) - d(f^a_{gait}, f^n_{gait}) + margin, 0) \tag{12}$$

5. Experimental Results and Analysis

5.1. Datasets

To evaluate our proposed method, RGB color video datasets are needed because the human pose estimation maps are extracted from RGB color images rather than from silhouettes. One popular gait dataset, CASIA-B Gait Dataset [13], not only provides human silhouettes, but also provides the original color video to the public for research. Therefore, CASIA-B Gait Dataset [13] is selected to evaluate our method. The Institute of Scientific and Industrial Research (ISIR), Osaka University (OU) also has provided many large population datasets such as OU-MVLP [32] and OU-ISIR [33]. However, the original RGB video is not currently available to the public due to privacy issues. Recently, ISIR provides multi-view large population dataset with pose sequence [7]. However, the input data of our proposed method is based on the byproduct of pose estimation rather than the pose estimation final result. Then, we choose the CMU Motion of Body (MoBo) dataset [14] as the second dataset to evaluate our proposed method because it provides original RGB frames and has multiple variations to evaluate the proposed method.

CASIA-B dataset [13], is one of the popular public gait datasets, it was created at the Institute of Automation, Chinese Academy of Sciences (CASIA) in 2005. It consists of 31 females and 93 males, and the total number of subjects is 124. Each subject has 10 sequences, including 6 sequences of normal walking (NM), 2 sequences of walking with a bag (BG), and 2 sequences of walking with a coat (CL), as shown in Fig. 1 (a). In addition, there are 11 cameras to capture the subjects at the same time, the view angles are $\{0°,18°,\cdots, 180°\}$.

The CMU Motion of Body (MoBo) dataset [14] was collected at Carnegie Mellon University in March 2001. It contains 25 individuals walking on a treadmill in the CMU 3D room. Each subject has four different walk patterns with one sequence, including slow walking, fast walking, incline walking, and walking with a ball, as can be shown in Fig. 5. The average walking speeds of slow walking, fast walking, incline walking, and walking with a ball are 2.06,

2.82, 1.96, and 2.04 mph, respectively. In terms of incline walking, the treadmill was set to the maximum incline of 15°. In addition, each subject is captured using 6 high-resolution color cameras distributed evenly around the treadmill, cameras labels are $vr03\_7, vr05\_7, vr07\_7, vr13\_7, vr16\_7$, and $vr17\_7$. According to the angle definition of the CASIA-B dataset, we define the angle set of MoBo dataset is {0°,45°,90°, 180°, 225°, 315°}, as shown in Fig. 6

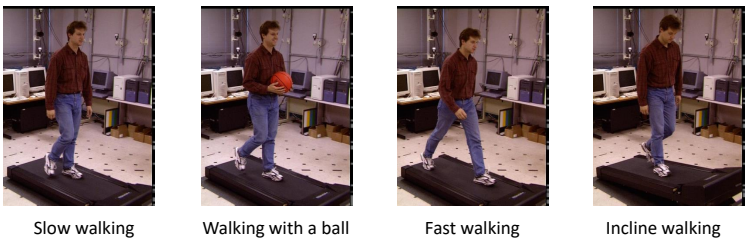


Figure 5: Four walking conditions on MoBo dataset.

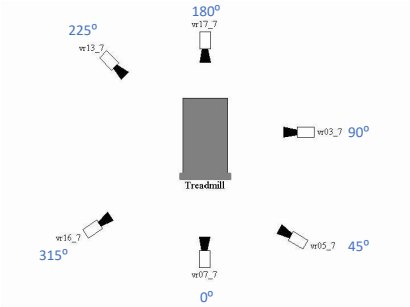


Figure 6: View angle definition on MoBo [14] dataset.

5.2. Experimental settings

CASIA-B dataset [13]: In order to compare with latest model-based methods, our experimental setting is the same as PTSN [5], PTSN-3D [6] and PoseGait [9]. That is, the first 62 subjects are put into the training set and the rest of the subjects are put into the test set. In the test set, the gallery set consists of the first 4 normal walking sequences of each subject, and the probe

set consists of the rest of 2 normal walking sequences, 2 sequences of walking with a bag, and 2 sequences of walking with a coat, as shown in Table 1.

Table 1: Experimental setting on CASIA-B dataset. NM: normal walking, BG: walking with a bag, CL: walking with a coat.

Training	Test	
	Gallery Set	Probe Set
ID: 001-062 Seqs: NM01-NM06 BG01-BG02, CL01-CL02	ID: 063-124 Seqs: NM01-NM04	ID: 063-124 Seqs: NM05-NM06 BG01-BG02, CL01-CL02

The CMU Motion of Body (MoBo) dataset [14]: Following the experimental setting of the above method, the first 13 subjects are put into the training set and the remaining 12 subjects are put into the test set. In the test set, the gallery set consists of slow walking condition, because it is closer to natural walking compared with other walking patterns. For the probe set, it consists of several conditions, that is, fast walking, incline walking, and walking with a ball, as shown in Table 2. Each condition only has one walking sequence.

Table 2: Experimental setting on the CMU Motion of Body (MoBo) dataset.

Training	Test	
	Gallery Set	Probe Set
ID: 01-13 slow walking, fast walking, incline walking, walking with a ball	ID: 14-25 slow walking	ID: 14-25 fast walking, incline walking, walking with a ball

Implementing details: The *gaitMapPose-Net* network consists of two streams, namely, *gaitMap-CNN* and *gaitMapPose-Net*. In terms of *gaitMap-CNN* network, the input size of pose estimation maps is 64×44. The total number T of set frames is to be 30. For the *gaitMapPose-Net* network, the number N of human joints is set as 18. The spatial range K and temporal range Γ are set to be 2 and 9. Adam is selected as an optimizer [34]. The learning rate is $1e-4$. The margin in triplet loss is set to be 0.2. The models are trained with 2 NVIDIA 1080TI 12GB. The implementation is based on PyTorch with CUDA 9.0.

5.3. Experimental results and discussions on CASIA-B dataset

The experimental results of the proposed method *PoseMapGait* on the CASIA-B dataset, as shown in Fig 7, including normal walking, carrying a bag, and clothing three conditions. The gallery set consists of the first 4 normal walking sequences with 11 views. The probe set includes three sets, that is, the rest 2 normal sequences, 2 walking with bag sequences, and 2 walking with a coat sequences, each set also has 11 views, as shown in Fig. 7. For each probe set of evaluation, there are 121 recognition rates in each figure. From Fig. 7, it can be found that *PoseMapGait* can achieve a high recognition rate when the gallery angle is equal to the probe angel, and the overall performance is the best when the probe set under normal walking sequences, following by walking with a bag sequence.

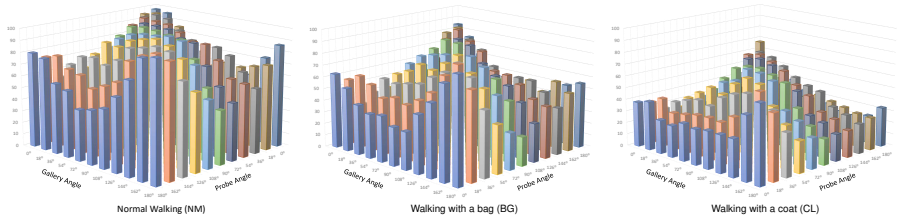
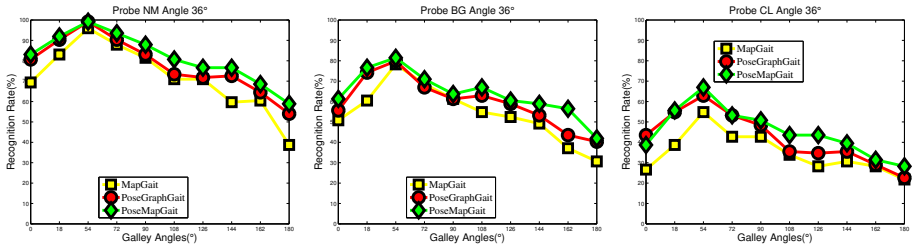
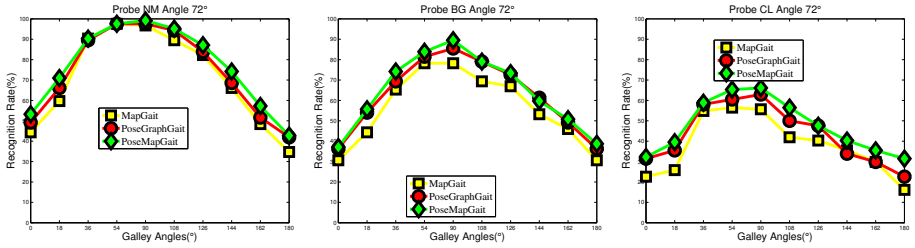


Figure 7: The experimental results when probe under three conditions on CASIA-B dataset.

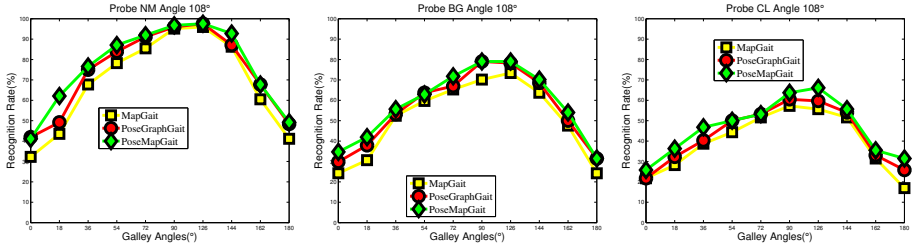
In order to study the complementary of between heatmaps evolution feature (extracted from heatmaps Fig. 2 (b.1)) and poses evolution feature (extracted from poses Fig. 2 (b.2)) for gait recognition, three different models are trained on CASIA-B dataset. *MapGait* is the model trained with heatmaps by using *gaitMap-CNN* network, *PoseGraphGait* is the model trained with poses by using *gaitPose-GCN* network, while *PoseMapGait* is the model trained with the fusion of heatmaps and poses by using *gaitMapPose-Net* network. Because of limited space, we only list 4 probe angles with a $36°$ interval, that is, $36°, 72°, 108°$ and $144°$. The first column of Fig. 8 compares the recognition rates at different probe angles in normal walking sequences, the second column is for the comparison in walking with a bag sequence, and the third column is in walking with a



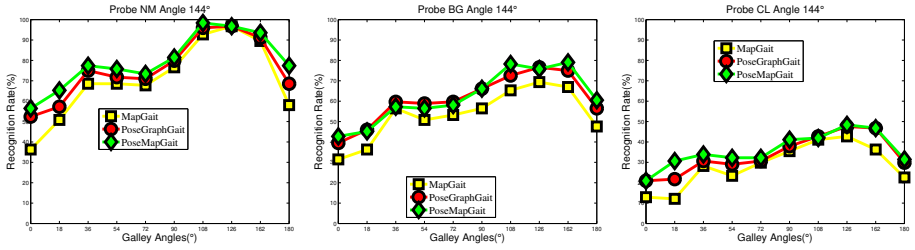
(a) (b) (c)



(d) (e) (f)



(g) (h) (i)



(j) (k) (l)

Figure 8: The complementary between body structure and body pose joints . From left column to right column are NM, BG and CL condition respectively. *MapGait*, *PoseGraphGait* and *PoseMapGait* are models trained with heatmaps data (Fig. 2 (b.1)), poses data (Fig. 2 (b.2)), and the fusion of two data, respectively.

coat sequences. From Fig. 8, we can see the performance of *MapGait* is better than that of *PoseGraphGait* at many points. One reason for this is that the input data of *MapGait* contains more body structure information than that of *PoseGraphGait*. In addition, the *PoseMapGait* model achieves the best performance among them, which shows that the poses evolution feature can bring performance improvement to the heatmaps evolution feature despite its limited input information (only 18 joints).

5.4. Comparisons with model-based approaches on CASIA-B dataset

To analyse the performance of our proposed methods *PoseGraphGait*, *MapGait*, and *PoseMapGait*, we compare them with recent state-of-the-art model-based methods on CASIA-B dataset, including PTSN [5], PTSN-3D [6] and PoseGait [9], where the input data of PTSN [5] is based on the 2D human joints, PTSN-3D [6] and PoseGait [9] are based on the 3D human joints. The comparison is shown in Table 3, results are the mean accuracies on rest 10 views excepting the identical-view cases, we can get the mean accuracies by averaging the 10 accuracies in Fig. 8.

From Table 3, it is clear that *PoseMapGait* can achieve the best accuracy on mean accuracy of 11 gallery views in all three walking conditions, that is, 75.7% (NM), 58.1% (BG), and 41.2% (CL), respectively. The mean accuracy gap between the *PoseMapGait* (58.1%) and the state-of-the-art method PoseGait (39.6%) can even reach 18.5% under the carrying a bag condition. The second-best performance is *MapGait*, which takes human structure heatmaps as input data. The high performances of *MapGait* and *PoseMapGait* show that the pose estimation maps feature is more able to promote the development of model-based approaches compared with the human skeleton-based feature.

In addition, the mean performance of *PoseGraphGait* is also better than those of recent advanced model-based methods [5, 6, 9] whether under the condition of normal walking, or under the conditions of walking with a bag and walking with a coat. The input data between these methods and ours are all based on human skeletons. Unlike these methods which merely considered a se-

Table 3: Average recognition rate (%) comparisons with model-based approaches on CASIA-B dataset. Excluding Identical-view Cases. (NM: normal walking, BG: walking with a bag, CL: walking with a coat)

Gallery angle NM #1-4	0°-180°											
Probe angle NM #5-6	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
PTSN [5]	34.5	45.6	49.6	51.3	52.7	52.3	53	50.8	52.2	48.3	31.4	47.4
PTSN-3D [6]	38.7	50.2	55.9	56	56.7	54.6	54.8	56	54.1	52.4	40.2	51.9
PoseGait [9]	48.5	62.7	66.6	66.2	61.9	59.8	63.6	65.7	66	58	46.5	60.5
PoseGraphGait (ours)	46.5	66.3	71.9	74.9	71.0	70.4	68.6	71.9	70.6	65.2	47.3	65.9
MapGait (ours)	56.5	71.5	78.0	79.6	74.0	74.4	73.8	77.8	76.0	73.5	58.6	72.2
PoseMapGait (ours)	**59.9**	**76.2**	**81.7**	**83.1**	**76.8**	**76.1**	**76.3**	**81.1**	**79.6**	**75.4**	**66.1**	**75.7**
Probe angle BG #1-2	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
PTSN [5]	22.4	29.8	29.6	29.2	32.5	31.5	32.1	31	27.3	28.1	18.2	28.3
PTSN-3D [6]	27.7	32.7	37.4	35	37.1	37.5	37.7	36.9	33.8	31.8	27	34.1
PoseGait [9]	29.1	39.8	46.5	46.8	42.7	42.2	42.7	42.2	42.3	35.2	26.7	39.6
PoseGraphGait (ours)	37.9	47.3	54.4	55.1	56.3	51.5	51.1	53.6	53.4	48.8	35.0	49.5
MapGait (ours)	43.5	51.1	59.7	60.7	62.5	56.9	55.9	58.6	61.1	55.2	41.9	55.2
PoseMapGait (ours)	**47.7**	**56.1**	**63.9**	**63.3**	**64.2**	**59.5**	**58.1**	**61.5**	**61.9**	**58.2**	**44.3**	**58.1**
Probe angle CL #1-2	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
PTSN [5]	14.2	17.1	17.6	19.3	19.5	20	20.1	17.3	16.5	18.1	14	17.6
PTSN-3D [6]	15.8	17.2	19.9	20	22.3	24.3	28.1	23.8	20.9	23	17	21.1
PoseGait [9]	21.3	28.2	34.7	33.8	33.8	34.9	31	31	32.7	26.3	19.7	29.8
PoseGraphGait (ours)	24.6	32.8	34.8	38.6	37.9	39.6	39.8	37.8	28.5	27.1	24.1	33.2
MapGait (ours)	27.7	35.3	42.0	45.2	43.2	44.7	43.1	41.9	33.8	30.1	26.5	37.6
PoseMapGait (ours)	**30.4**	**41.9**	**45.2**	**48.9**	**47.3**	**48.1**	**46.5**	**44.9**	**36.0**	**34.5**	**29.6**	**41.2**

quence of human joint coordinates modeling, we construct a gait graph that not only considers inter-frame connections with the same joints, but also considers intra-body connections based on naturally connected human body joints. The comparison shows that the pose graph gait can further boost the development of the pose skeleton for gait recognition.

5.5. Comparisons with appearance-based approaches on CASIA-B dataset

As mentioned in the previous part of the paper, the model-based feature (pose estimation maps) used in the proposed method is compact and has less redundant information compared with the appearance-based feature. This means that the joint maps feature extraction is more challenging for model-based algorithms considering the prediction accuracy of joint maps' location. In order to show the effectiveness of the pose estimation maps feature, we compare it with recent state-of-the-art appearance-based approaches. Including SPAE [17], GaitGAN [35], GaitGANv2 [18] and DV-GEIs-pre [4]. The experimental results of these methods can be shown in Table 4.

Table 4: Comparisons with appearance-based approaches at average accuracy (%) on CASIA-B dataset. Excluding identical-view cases. (NM: normal walking, BG: walking with a bag, CL: walking with a coat)

Gallery angle NM #1-4	0°-180°											
Probe angle NM #5-6	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
SPAE [17]	50.0	58.1	61.0	63.3	64.0	62.1	62.3	66.3	64.4	54.5	46.7	59.3
GaitGAN [35]	41.9	53.5	63.0	64.5	63.1	58.1	61.7	65.7	62.7	54.1	40.6	57.2
GaitGANv2 [18]	48.1	61.9	68.7	71.7	66.7	64.8	66.0	70.2	71.6	58.9	46.1	63.1
DV-GEIs-pre [4]	64.5	76.2	81.3	80.8	77.1	72.6	74.4	78.9	80.6	75.6	63.7	75.1
PoseMapGait (ours)	59.9	76.2	81.7	83.1	76.8	76.1	76.3	81.1	79.6	75.4	66.1	**75.7**
Probe angle BG #1-2	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
SPAE [17]	34.0	38.6	42.1	42.7	39.0	32.8	31.3	39.9	41.0	35.7	32.3	37.2
GaitGAN [35]	28.5	35.2	42.7	34.4	38.0	33.5	36.2	44.8	41.8	33.3	23.6	35.6
GaitGANv2 [18]	37.2	43.4	46.6	46.0	47.6	41.5	41.2	48.5	48.8	42.2	31.6	43.1
PoseMapGait (ours)	47.7	56.1	63.9	63.3	64.2	59.5	58.1	61.5	61.9	58.2	44.3	**58.1**
Probe angle CL #1-2	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
SPAE [17]	21.5	25.4	27.3	28.1	26.9	22.2	22.3	26.3	24.8	21.5	19.6	24.2
GaitGAN [35]	9.8	15.2	24.8	25.0	24.7	19.9	22.7	24.5	27.7	18.0	11.9	20.4
GaitGANv2 [18]	20.7	23.1	26.6	30.8	28.2	23.0	24.4	27.4	24.2	21.9	16.0	24.2
PoseMapGait (ours)	30.4	41.9	45.2	48.9	47.3	48.1	46.5	44.9	36.0	34.5	29.6	**41.2**

From Table 4, we can see that the proposed method not only achieves the highest recognition rates of each gallery view than SPAE, GaitGAN, and GaitGANv2, but also obtains much higher recognition rates in all three walking conditions. It should be noticed that the mean accuracy gap between the *PoseMapGait* (41.2%) and GaitGANv2 [18] (24.2%) can even reach 17.0% under the carrying a coat condition. That means that the proposed method is more robust to the view, carrying a bag, and clothing variations. This is the advantage of the pose estimation maps. The raw feature is robust to human shape while the appearance-based features tend to be changed greatly.

We also compare with GaitSet [11] method, which has achieved very high performance in gait recognition. There are two experimental settings on GaitSet [11], one is based on Table 1, that is, the first 62 subjects are put into the training set and the rest of the 62 subjects are put into the test set. In the second experimental setting, we set the first 74 subjects as the training set and the rest of the 50 subjects as the test set. In order to show the potential of our proposed in a larger dataset, we also implement another experiment that uses the first 100 subjects as the training set. The experimental results are listed in Table 5, the evaluation of calculating mean accuracy is the same as the mean

accuracy of the above Table 4.

Table 5: Mean accuracy (%) comparison with GaitSet [11] approach on CASIA-B dataset. Excluding Identical-view Cases.

Conditions	Methods	Training Subjects			Growth Rate	
		62	74	100	62 to 74	62 to 100
normal walking (NM)	GaitSet [11]	92.0	95.0	95.8	3.3 %	4.0 %
	PoseMapGait (ours)	75.7	79.3	89.3	**4.7 %**	**15.3 %**
	Gap	16.3	15.7	**6.5**	-1.4%	-11.3%
walking with a bag (BG)	GaitSet [11]	84.3	87.2	91.8	3.4 %	8.1 %
	PoseMapGait (ours)	58.1	61.1	74.2	**5.2 %**	**21.7 %**
	Gap	26.2	26.1	**17.6**	-1.8%	-13.6%
walking with a coat (CL)	GaitSet [11]	62.5	70.4	83.1	12.6 %	24.8 %
	PoseMapGait (ours)	41.2	48.1	63.2	**16.7 %**	**34.9 %**
	Gap	21.3	22.3	**19.9**	-4.1%	-10.1%

From Table 5, it can be seen that the method of GaitSet [11] has achieved very high performance. There are two reasons that our model-based approach *PoseMapGait* is inferior to the appearance-based approach GaitSet [11]. For one reason, they used human silhouettes as input data which is a high dimension feature, while our pose estimation maps consist of only 17 compact joint heatmaps. The semantic information is very limited compared with the human silhouettes. Another reason is that the performance of model-based methods depends heavily on the accuracy of body part locating and tracking, while the accuracy of joint heatmaps extraction is more challenging in such low-resolution gait recognition conditions compared with human silhouette extraction.

To analyze the potential of pose estimation maps on gait recognition, we calculate the growth rates and gaps of the proposed *PoseMapGait* and GaitSet [11] method from 62 training subjects to 74 and 100 training subjects. From Table 5, it is clearly found that the growth rates of the proposed method are better than those of GaitSet [11]. On the condition of walking with a coat, the proposed method can achieve a 16.7% growth rate with only 12 additional training samples. When the number of training subjects increases from 62 to 100, the proposed method shows great growth rates (21.7 % and 34.9 %) under the conditions of walking with a bag and a coat. In addition, with the increase of training subjects, it is obvious that the gaps of mean accuracy between GaitSet [11] and

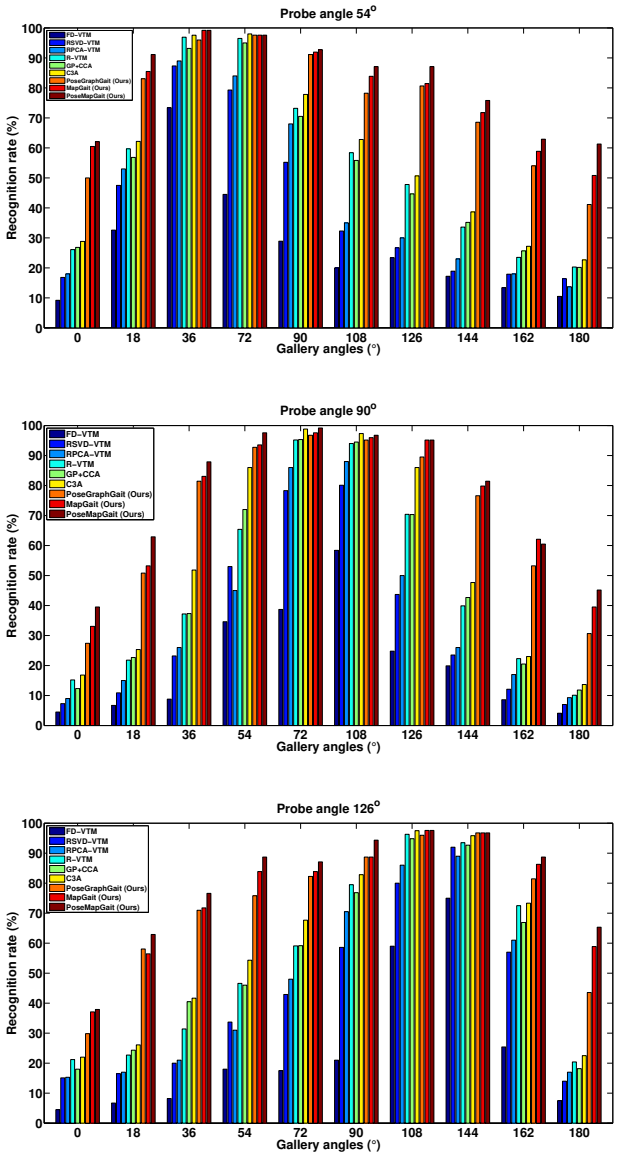


Figure 9: Comparing with some cross-view methods at probe angle 54°, 90° and 126°. The gallery angles are the remaining 10 angles excluding the corresponding probe angle.

PoseMapGait become smaller. The gap can achieve 6.5 % when the training set under 100 subjects with the normal walking condition. What's more, from the table, we can see the gaps in growth rates enlarge when the number of training subjects increases from 62 to 100. The comparison shows that the pose estimation maps feature has a great potential to deal with external environmental factors. We believe that with the development of pose estimation algorithms and the increase in gait data volume, the performance of the proposed method can ban further improved.

5.6. Effectiveness on View Variation

From the above experiments, it can be found the proposed method can achieve state-of-the-art performance compared with recent model-based methods. In order to show the effectiveness of the view variation of the proposed method, we compared our methods (*PoseGraphGait*, *MapGait* and *PoseMapGait*) with some cross-view gait recognition methods. Including FD-VTM [36], RSVD-VTM [37], RPCA-VTM [38], R-VTM [39], GP+CCA [40] and C3A [41]. We choose three probe angles, that is, 54°, 90°, and 126°, and the experimental setting is the same as these methods in the original papers. The recognition rates are shown in Fig. 9.

It can be clearly found that our methods (*PoseGraphGait*, *MapGait* and *PoseMapGait*) achieve much high performance when the difference between the gallery angle and the probe angle is large. The greater the difference, the more obvious improvements. The greater the difference, the more obvious the improvement. It is the advantage of proposed methods that focuses on human body movement modeling which is more robust to view variation.

5.7. Experimental results on MoBo dataset

The complete experimental results on the MoBo dataset are listed in Fig. 10. The evaluation under the variations of view, fast walking, incline walking, and walking with a ball are shown in these figures. In the experiment, the slow walking sequences at a specific view are put into the gallery set, and the fast

walking, incline walking, and walking with a ball are put into the probe set of the three sets of the experiment, respectively. For each set of experiments, there are 36 combinations. That means there are 36 recognition rates in each figure. It is easy to found the recognition rate will be high when the gallery angle is equal to the probe angle.

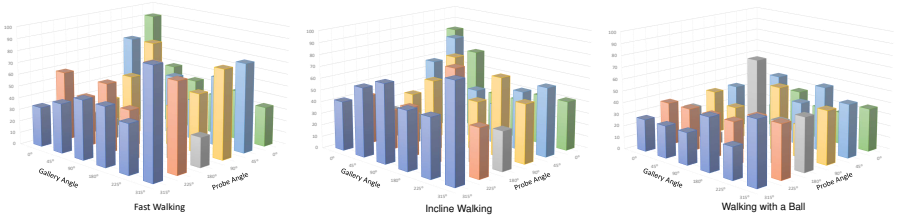


Figure 10: The experimental results when probe under three conditions on MoBo dataset.

5.8. Comparisons on MoBo dataset

As mentioned in previous experimental results on the CASIA-B dataset, the pose estimation maps feature used in the proposed method contains richer gait information compared with some model-based features, and has less redundant information compared with some appearance-based features. To show the effectiveness of the pose estimation maps feature, we make comparisons with some advanced methods on the MoBo dataset. Including recent popular model-based method PoseGait [9], and appearance-based method GaitGANv2 [18], DV-GEIs-pre [4] and DV-GEIs [3]. We implemented these methods by ourselves as they do not cite the experimental results of the MoBo dataset from the original paper. In order to better analyze the comparisons, we analyze two types of comparisons according to variation conditions.

Firstly, we compare with average recognition rates on identical-view cases under three different conditions (gallery data being slow walking), the comparison is shown in Fig. 11. The average recognition rate is by averaging the 6 recognition rates when the gallery angle is equal to the probe angle from the above experimental results (Fig. 10). From Fig. 11, it is clear that our three

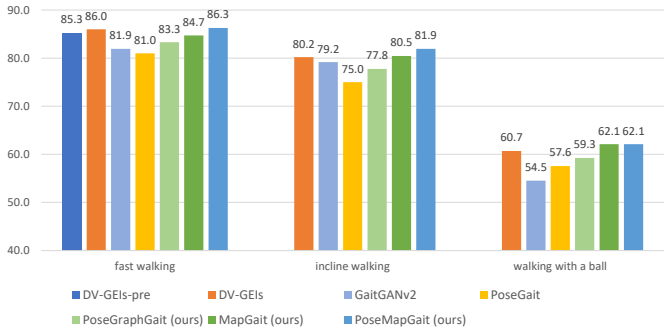


Figure 11: The average recognition rates for the probe data being fast walking, incline walking and walking with a ball on identical-view cases.

types of models can achieve better performance no matter under the condition of fast walking, or the conditions of incline walking and walking with a ball. Additionally, there is a big gap (7.6%) between *PoseMapGait* (62.1%) and GaitGANv2 (54.5%) when probe data is walking with a ball, which shows that the pose estimation maps feature has better robustness to human shape compared with human silhouettes feature.

Secondly, we further compare with average recognition rates under three different conditions on the cross-view case (excluding identical-view cases), as shown in Table 6. The average recognition results are the mean accuracies on the rest of 5 views except the identical-view cases, we can get the mean accuracies by averaging the 5 accuracies in Fig. 10. It should be notice that the number of subjects on CASIA-B (124 subjects) is much more than that of subjects on MoBo (25 subjects), and the types of variation conditions on MoBo (5 variations: view, slow, fast, incline, and with a ball walking) is more than that of CASIA-B (4 variations: view, normal, with a bag, and with a coat walking). Therefore, the overall recognition rates are inferior to that of CASIA-B. But our methods *PoseMapGait* can still achieve the best accuracy on mean accuracy of 6 gallery views in all three walking conditions, that is, 46.7% (fast walking), 42.5% (incline walking), and 37.9% (walking with a ball), respectively. The comparison

Table 6: Average accuracies (%) on MoBo dataset under three different experimental settings, excluding identical-view cases.

Gallery angle (slow walking)	0°, 45°, 90°, 180°, 225°, 315°						
Probe angle (fast walking)	0°	45°	90°	180°	225°	315°	Mean
GaitGANv2 [18]	31.7	46.7	50.0	33.3	36.7	50.0	41.4
DV-GEIs-pre [4]	41.7	41.7	51.7	41.7	40.0	53.3	45.0
DV-GEIs [3]	41.7	48.3	48.3	38.3	51.7	53.3	**46.9**
PoseGait [9]	38.3	45.0	43.3	25.0	36.7	45.0	38.9
PoseGraphGait (ours)	33.3	50.0	45.0	33.3	40.0	55.0	42.8
MapGait (ours)	40.0	41.7	55.0	35.0	43.3	56.7	45.3
PoseMapGait (ours)	48.3	43.3	43.3	40.0	46.7	58.3	46.7
Probe angle (incline walking)	0°	45°	90°	180°	225°	315°	Mean
GaitGANv2 [18]	36.7	51.7	38.3	31.7	43.3	46.7	41.4
DV-GEIs [3]	40.0	51.7	36.7	35.0	38.3	51.7	42.2
PoseGait [9]	36.7	50.0	36.7	30.0	35.0	51.7	40.0
PoseGraphGait (ours)	38.3	48.3	33.3	36.7	45.0	40.0	40.3
MapGait (ours)	35.0	55.0	31.7	35.0	43.3	51.7	41.9
PoseMapGait (ours)	40.0	46.7	40.0	38.3	45.0	45.0	**42.5**
Probe angle (walking with a ball)	0°	45°	90°	180°	225°	315°	Mean
GaitGANv2 [18]	32.7	27.3	30.9	27.3	32.7	27.3	29.7
DV-GEIs [3]	38.2	27.3	32.7	41.8	40.0	36.4	36.1
PoseGait [9]	32.7	21.8	32.7	23.6	38.2	34.5	30.6
PoseGraphGait (ours)	29.1	16.4	34.5	30.9	41.8	36.4	31.5
MapGait (ours)	38.2	21.8	29.1	45.5	38.2	38.2	35.2
PoseMapGait (ours)	36.4	30.9	32.7	45.5	41.8	40.0	**37.9**

of *PoseMapGait* and PoseGait [9] shows that the proposed pose estimation maps feature can further improve the performance of gait recognition compared with the skeleton-based feature. And the comparison between *PoseMapGait* and GaitGANv2 [18], DV-GEIs-pre [4], DV-GEIs [3] shows that the pose estimation maps feature is robust multiple variations compared with the appearance-based feature, which enhances the utility of gait recognition in real applications.

6. Conclusions and Future Work

To address the various concerns induced by the rapid spread of COVID-19 while people are under surveillance, it is necessary to accelerate the development of gait recognition technology because of its advantages of non-contact and long-distance identification. In this paper, we proposed a novel model-based gait recognition method, called *PoseMapGait*. *PoseMapGait* employs *pose estimation maps* as a gait feature, rather than directly relying on coordinates from skeletons. Compared with the skeleton-based feature, this feature not only has

richer body structure information, but also is more robust to human shape compared to the silhouette-based feature. In addition, we construct a gait graph in order to extract spatial-temporal information from the human skeleton based on the pose graph convolutional networks. The experimental results on CASIA-B and MoBo datasets show that the proposed method achieves the new state-of-the-art performance compared with recent advanced model-based methods, and it is comparable with some state-of-the-art appearance-based methods.

Although the proposed model-based method just achieves comparable accuracy with state-of-the-art appearance-based methods, it shows that model-based methods have a great potential for gait recognition because they are robust for more challenging conditions. In addition to AlphaPose, there are other methods that can model the human body in more detail. For example, DensePose [42] can map all human pixels of an RGB image to the 3D surface of the human body. We believe that with the development of pose estimation algorithms and the quality of the camera, the proposed pose estimation maps feature can make a great contribution to the development of gait recognition, and enhance its utility in real applications.

Acknowledgment

This work is supported in part by the AFOSR DDIP Program and SFFP. The views expressed in this article, book, or presentation are those of the author and do not necessarily reflect the official policy or position of the United States Air Force Academy, the Air Force, the Department of Defense, or the U.S. Government. Approved for public release: distribution unlimited.. PA#: USAFA-DF-2022-568.

References

[1] R. Tanawongsuwan, A. Bobick, Gait recognition from time-normalized joint-angle trajectories in the walking plane, in: Proceedings of the IEEE

Computer Society Conference on Computer Vision and Pattern Recognition, 2001, pp. II–II.

[2] L. Wang, H. Ning, T. Tan, W. Hu, Fusion of static and dynamic body biometrics for gait recognition, IEEE Transactions on circuits and systems for video technology 14 (2) (2004) 149–158.

[3] R. Liao, W. An, Z. Li, S. S. Bhattacharyya, A novel view synthesis approach based on view space covering for gait recognition, Neurocomputing 453 (2021) 13–25.

[4] R. Liao, W. An, S. Yu, Z. Li, Y. Huang, Dense-view geis set: View space covering for gait recognition based on dense-view gan, in: 2020 IEEE International Joint Conference on Biometrics (IJCB), IEEE, 2020, pp. 1–9.

[5] R. Liao, C. Cao, E. B. Garcia, S. Yu, Y. Huang, Pose-based temporal-spatial network (ptsn) for gait recognition with carrying and clothing variations, in: Chinese Conference on Biometric Recognition, Springer, 2017, pp. 474–483.

[6] W. An, R. Liao, S. Yu, Y. Huang, P. C. Yuen, Improving gait recognition with 3d pose estimation, in: Chinese Conference on Biometric Recognition, Springer, 2018, pp. 137–147.

[7] W. An, S. Yu, Y. Makihara, X. Wu, C. Xu, Y. Yu, R. Liao, Y. Yagi, Performance evaluation of model-based gait on multi-view very large population database with pose sequences, IEEE Transactions on Biometrics, Behavior, and Identity Science 2 (4) (2020) 421–430.

[8] N. Li, X. Zhao, C. Ma, Jointsgait: A model-based gait recognition method based on gait graph convolutional networks and joints relationship pyramid mapping, arXiv preprint arXiv:2005.08625.

[9] R. Liao, S. Yu, W. An, Y. Huang, A model-based gait recognition method with body pose and human prior knowledge, Pattern Recognition 98 (2020) 107069.

[10] M. Liu, J. Yuan, Recognizing human actions as the evolution of pose estimation maps, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1159–1168.

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

[12] C. Fan, Y. Peng, C. Cao, X. Liu, S. Hou, J. Chi, Y. Huang, Q. Li, Z. He, Gaitpart: Temporal part-based model for gait recognition, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 14225–14233.

[13] S. Yu, D. Tan, T. Tan, A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition, in: 18th International Conference on Pattern Recognition, 2006, pp. 441–444.

[14] R. Gross, J. Shi, The cmu motion of body (mobo) database, 2001.

[15] J. Han, B. Bhanu, Individual recognition using gait energy image, IEEE transactions on pattern analysis and machine intelligence 28 (2) (2005) 316–322.

[16] C. Wang, J. Zhang, L. Wang, J. Pu, X. Yuan, Human identification using temporal information preserving gait template, IEEE transactions on pattern analysis and machine intelligence 34 (11) (2011) 2164–2176.

[17] S. Yu, H. Chen, Q. Wang, L. Shen, Y. Huang, Invariant feature extraction for gait recognition using only one uniform model, Neurocomputing 239 (2017) 81–93.

[18] S. Yu, R. Liao, W. An, H. Chen, E. B. G. Reyes, Y. Huang, N. Poh, Gaitganv2: Invariant gait feature extraction using generative adversarial networks, Pattern recognition 87 (2019) 179–189.

[19] Z. Wu, Y. Huang, L. Wang, X. Wang, T. Tan, A comprehensive study on cross-view gait based human identification with deep cnns, IEEE transactions on pattern analysis and machine intelligence 39 (2) (2017) 209–226.

[20] Z. Zhang, L. Tran, X. Yin, Y. Atoum, X. Liu, J. Wan, N. Wang, Gait recognition via disentangled representation learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4710–4719.

[21] H.-S. Fang, S. Xie, Y.-W. Tai, C. Lu, RMPE: Regional multi-person pose estimation, in: ICCV, 2017.

[22] Z. Li, S. Yu, E. B. G. Reyes, C. Shan, Y.-r. Li, Static and dynamic features analysis from human skeletons for gait recognition, in: 2021 IEEE International Joint Conference on Biometrics (IJCB), IEEE, 2021, pp. 1–7.

[23] Y. Feng, Y. Li, J. Luo, Learning effective gait features using LSTM, in: the 23rd International Conference on Pattern Recognition, 2016, pp. 325–330.

[24] M. Defferrard, X. Bresson, P. Vandergheynst, Convolutional neural networks on graphs with fast localized spectral filtering, in: Proceedings of the 30th International Conference on Neural Information Processing Systems, 2016, pp. 3844–3852.

[25] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 32, 2018.

[26] Y. Shao, Z. Zhang, Z. Li, K. Fan, G. Li, Attribute compression of 3d point clouds using laplacian sparsity optimized graph transform, in: 2017 IEEE Visual Communications and Image Processing (VCIP), IEEE, 2017, pp. 1–4.

[27] D. Mo, Z. Lai, W. Wong, Locally joint sparse marginal embedding for feature extraction, IEEE Transactions on Multimedia 21 (12) (2019) 3038–3052.

[28] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7291–7299.

[29] L. Pan, R. Liao, Z. Li, S. S. Bhattacharyya, Dynamic, data-driven hyperspectral image classification on resource-constrained platforms, in: International Conference on Dynamic Data Driven Application Systems, Springer, 2020, pp. 320–327.

[30] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Póczos, R. R. Salakhutdinov, A. J. Smola, Deep sets, in: NIPS, 2017.

[31] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815–823.

[32] D. M. T. E. Y. Y. Noriko Takemura, Yasushi Makihara, Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition, IPSJ Trans. on Computer Vision and Applications 10 (4) (2018) 1–14.

[33] H. Iwama, M. Okumura, Y. Makihara, Y. Yagi, The ou-isir gait database comprising the large population dataset and performance evaluation of gait recognition, IEEE Transactions on Information Forensics and Security 7 (5) (2012) 1511–1521.

[34] A. KingaD, A method for stochastic optimization, Anon. International Conferenceon Learning Representations. SanDego: ICLR.

[35] S. Yu, H. Chen, E. B. G. Reyes, N. Poh, Gaitgan: Invariant gait feature extraction using generative adversarial networks, in: Computer Vision and Pattern Recognition Workshops, 2017, pp. 532–539.

[36] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, Y. Yagi, Gait recognition using a view transformation model in the frequency domain, in: European Conference on Computer Vision, 2006, pp. 151–163.

[37] W. Kusakunniran, Q. Wu, H. Li, J. Zhang, Multiple views gait recognition using view transformation model based on optimized gait energy image, in: IEEE 12th International Conference on Computer Vision Workshops, 2009, pp. 1058–1064.

[38] S. Zheng, J. Zhang, K. Huang, R. He, T. Tan, Robust view transformation model for gait recognition, in: 18th IEEE International Conference on Image Processing, 2011, pp. 2073–2076.

[39] W. Kusakunniran, Q. Wu, J. Zhang, H. Li, Gait recognition under various viewing angles based on correlated motion regression, IEEE transactions on circuits and systems for video technology 22 (6) (2012) 966–980.

[40] K. Bashir, T. Xiang, S. Gong, Cross view gait recognition using correlation strength, in: the British Machine Vision Conference, 2010, pp. 1–11.

[41] X. Xing, K. Wang, T. Yan, Z. Lv, Complete canonical correlation analysis with application to multi-view gait recognition, Pattern Recognition 50 (2016) 107–117.

[42] R. A. Güler, N. Neverova, I. Kokkinos, Densepose: Dense human pose estimation in the wild, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7297–7306.