Pattern Recognition 98 (2020) 107069

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

A model-based gait recognition method with body pose and human prior knowledge

Rijun Liao[a,c], Shiqi Yu[b,c,*], Weizhi An[a,c], Yongzhen Huang[d,e]

[a] College of Computer Science and Software Engineering, Shenzhen University, China
[b] Department of Computer Science and Engineering, Southern University of Science and Technology, China
[c] Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen, China
[d] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China
[e] Watrix technology limited co. ltd, China

ARTICLE INFO

Article history:
Received 27 April 2019
Revised 7 August 2019
Accepted 26 September 2019
Available online 30 September 2019

Keywords:
Gait recognition
Human body pose
Spatio-temporal feature,

ABSTRACT

We propose in this paper a novel model-based gait recognition method, *PoseGait*. Gait recognition is a challenging and attractive task in biometrics. Early approaches to gait recognition were mainly appearance-based. The appearance-based features are usually extracted from human body silhouettes, which are easy to compute and have shown to be efficient for recognition tasks. Nevertheless silhouettes shape is not invariant to changes in clothing, and can be subject to drastic variations, due to illumination changes or other external factors. An alternative to silhouette-based features are model-based features. However, they are very challenging to acquire especially for low image resolution. In contrast to previous approaches, our model *PoseGait* exploits human 3D pose estimated from images by Convolutional Neural Network as the input feature for gait recognition. The 3D pose, defined by the 3D coordinates of joints of the human body, is invariant to view changes and other external factors of variation. We design spatio-temporal features from the 3D pose to improve the recognition rate. Our method is evaluated on two large datasets, CASIA B and CASIA E. The experimental results show that the proposed method can achieve state-of-the-art performance and is robust to view and clothing variations.

1. Introduction

The gait characterizes is the walking style of a person. It can be used as a biometric feature to identify a person. Compared with other biometric features such as fingerprint, face, iris and palmprint, gait has its unique advantages such as non-contact during acquisition, hard to fake and particularly suitable for long-distance human identification. Gait recognition algorithms have become more and more robust during the past decades, and can nowadays be used in various 'real world' applications such as video surveillance, crime prevention and forensic identification.

Gait is a behavioral biometric; it is not as robust as physical biometric features as fingerprint, iris and face. When there are some variations, such as view, carrying, clothing, illumination, the gait feature may change drastically. In order to improve the stability of the extracted features, some earlier work tried to model a human body and to capture difference of motion patterns among different subjects [1–3]. The ideas of using the body part motion are straightforward and reasonable. But it is very challenging to locate and track each body part accurately.

In the pass two decades, the appearance-based gait recognition methods [4,5] are more popular than the model-based ones. The appearance-based methods usually use the human silhouettes as raw input data. These methods can achieve very high recognition rates when there are not obvious variations. However, the variations in real applications can change human shape greatly and decrease performance severely. In contrast, model-based features are based on human body structure and movements. So they are not so sensitive to human shape and human appearance relatively. Recently the progress in human body pose estimation [6] is bring more hope to the model-based methods. Model-based features are normally extracted from human body structures and local movement patterns, so they can handle many variations, especially view variations.

We propose in this paper a novel model-based gait recognition method, *PoseGait*, which exploits human pose as feature. We demonstrate experimentally that the pose feature, defined in a low dimensional space, can achieve recognition rates in par with the

* Corresponding author.
E-mail address: yusq@sustech.edu.cn (S. Yu).

https://doi.org/10.1016/j.patcog.2019.107069
0031-3203/

appearance-based features, while being invariant to external factors changes. The contributions of our work are as follows:

- We propose a novel model-based gait recognition method, *PoseGait*, which exploits human pose as feature. The method can achieve high recognition rate despite the low dimensional feature (only 14 body joints).
- We design dedicated features based on 3D pose information. We demonstrate experimentally the advantage of these features.
- CNN nor RNN/LSTM can successfully extract spatio-temporal gait feature with the help of fusing two losses.

The rest part of the paper is organized as follows. In the next section, we introduce the related work. Section 3 describes the proposed method. Experiments and evaluation are presented in Section 4. The last section gives the conclusions and future work.

2. Related work

In this section, we will briefly review existing gait recognition methods. Two categories of methods, appearance-based and model-based ones, are presented. We also introduce pose estimation methods that extract human body parameters.

2.1. Appearance-based methods

Appearance-based methods usually use the human silhouettes as raw input data. Gait energy image (GEI) [4] is one of the most popular feature which has a low computational cost and can achieve a relative high recognition rate. One common pipeline of GEI-based methods are: 1) Extract the human silhouettes from videos; 2) compute a gait energy image GEI through aligning the silhouettes and averaging them; 3) and then compute the similarities between each of two GEIs. GEI+PCA [4] is one simple approach of GEI-based methods, it can achieve good accuracy when there are no obvious variations. But it is hard to handle view angles, clothing and other variations.

In order to increase the robustness of GEI, some researchers have focused on reducing the disturbance of the shape changes and occlusion. For instance, Huang et al. [7] increase robustness to some classes of structural variations by fusing shifted energy image and the gait structural profile. For tackling with view condition variations, Yu et al. [8] employ the Stacked Progressive Auto-Encoders (SPAE) trying to transform gait images from arbitrary angles to a specific view. These methods can handle with view, clothing and carrying conditions to a certain extent. But the performance of these methods is still not good enough, since GEI will lead to some temporal information missing.

Recently, some researchers directly use human silhouettes as input data instead of using the average of them. The method in [9] is the first one using a deep CNN model to extract feature from human silhouette sequence, and its performance outperforms the previous state-of-the-art approaches by a outstanding gap. In addition, Chao et al. [10] regard gait as a set consisting of independent silhouettes rather than a continuous silhouettes (Wu et al. [9]). They propose a new network named GaitSet to extract invariant feature from that set. In [11] Zhang et al. also demostrated experimentally that temporal feature between frame can achieve better performance than GEI. A similar conclution can also be found in [5]. These methods can achieve high accuracy in terms of cross-view condition. But it is still challenging to handle with cross-carrying and cross-clothing conditions very well. The main reason is that human appearance and shape can be changed greatly when there are some variations.

2.2. Model-based methods

The model-based methods extract features by modeling human body structure and local movement patterns of different body parts. Compared with appearance-based methods, model-based methods can be robust to many variations if human bodies are correctly and high accurately modeled. But it is not a easy task. Some earlier model-based methods [2] even mark different body parts manually, or use some specific devices to obtain the human joint positions. In addition, body modeling from images normally has a heavy computational cost. For these reasons, model-based methods are not as popular as appearance-based one.

In some early work such as in [1] by Nixon et al. it is shown that human body motion has the capacity to identify different types of gait. They use a simple stick model to simulate legs, and then use an articulate pendulum movement to simulate the leg movement during walking. Finally, frequency components are extracted as gait features for human identification. In addition, Wang et al. [3] argue that the changes of the angle of each joint in temporal domain can be beneficial to recognition. They proposed a multi-connected rigid body model, and the body is divided into 14 parts and each part connected through a joint. To extract the temporal feature from body joints, Feng et al. use the human body joint heatmap extracted from a RGB image instead of using a binary silhouette to describe the human body pose [12]. The heatmaps are sent into a Long Short Term Memory (LSTM) recurrent neural network to extract temporal feature.

In recent years, some researchers use human body skeleton and body joints to recognize different persons. For example, Kastaniotis et al. [13] use skeleton data from the low-cost Kinect sensor instead of a specific equipment in [2]. It shows that the body joints from Kinect contains enough information for human identification. But in video surveillance, the commonly used cameras are mostly RGB ones, not Kinect sensors. From the previous work we can conclude that the accuracy of the human body model is crucial for gait recognition.

2.3. Pose estimation methods

Human body pose estimation has achieved great progress in recent years. One of the earlier works can be found in [14]. The method can recover 3D human poses from silhouettes by adopting multiview locality-sensitive sparse coding in the retrieving process. These last years, most recent work are deep learning based methods [15,16]. In [17] Hong et al. learnt a non-linear mapping from 2D images to 3D poses using a deep autoencoder network. Recently, a bottom-up pose estimation method [6] using deep CNN can create accurate human models. The method can handle multiple persons in an image. It can predict vector fields, named Part Affinity Fields (PAFs), that directly estimate the association between anatomical parts in an image. They have designed an architecture to jointly learn part locations and its association. Their method has achieved the state-of-the-art performance on the MSCOCO 2016 key points challenge and the MPII Multi-Person benchmark. The system in [6], OpenPose, can jointly detect human body joints including hands, feet, elbows, etc.

The 2D poses extracted from images are not invariant to viewpoints, but the 3D ones are. To estimate a 3D pose from one image is an ill-posed problem. But with constrains of the human body, Chen et al. [18] proposed a 3D human pose estimation method from one single RGB image. The idea is simple, but it outperforms many state-of-the-art 3D pose estimation system. It also does not need multiple cameras and is suitable to video surveillance applications.

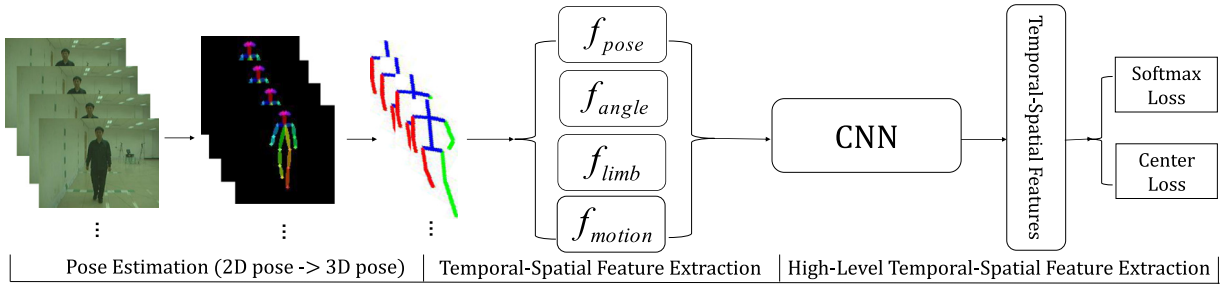

Fig. 1. The framework of the proposed method. The 3D pose are estimated from the 2D one. The 3D pose f_{pose} and three kinds of additional spatio-temporal features (f_{angle}, f_{limb} and f_{motion}) are concatenated as the input of CNN.

In summary, we think that there will be increasing model-based gait recognition method in future with the progress of human body modeling.

3. The proposed method PoseGait

In our proposed method, *PoseGait*, 3D human body joints are taken as feature for gait recognition. We use 3D pose information because it is robust to view variation. Compared with some appearance-based features, such as GEI [4], the feature used in the proposed method is low dimensional and far more compact because there are only some joint positions and not high dimensional images. In order to extract the temporal feature, we employ the poses extracted from a sequence of frames. According to pioneer work in [3], the motion patterns and angles are important for human identification. In this work, we design some handcrafted features based on human prior knowledge to improve the efficiency of feature extraction. Four kinds of features are concatenated together as described in Section 3.3. It can be regarded as that the features are fused at the input level, nor at the learning or output level as in [19,20]. During the training phase, two losses are combined to reduce the intra-class variation and improve the inter-class variation. The framework of the whole method is shown in Fig. 1. The implementation details are described in the following part of this section.

3.1. Human body pose features

3.1.1. 2D pose feature

To reduce the effect of carrying and clothing variation on gait recognition, we introduce a pose feature for gait recognition. In some early works [2,3], it is shown that joint motion has sufficient capacity to identify different subjects. But automatic accurate pose estimation was challenging in the years before deep learning. In the proposed method the joints are estimated using OpenPose [6]. The estimated pose consists of 18 body joints: Nose, Neck, RShoulder (right shoulder, the following names are in the similar manner), RElbow, RWrist, LShoulder, LElbow, LWrist, RHip, RKnee, RAnkle, LHip, LKnee, LAnkle, Reye, LEye, REar and LEar.

In images the size of a human body is changing according to the distance between the subject and the camera. The human bodies of different subjects all are normalized to a fixed size. The distance between the neck and the hip is regarded as the unit length. The position of the hip is at the center of RHip and LHip. The neck is placed at the origin of the plane coordinate system. So the body joints are normalized as follows:

$$J_i' = \frac{J_i - J_{neck}}{H_{nh}} \tag{1}$$

where $J_i \in \mathbb{R}^2$ is the position of body joint i, J_i' is the normalized position of J_i, J_{neck} is the neck position, the H_{nh} is the distance between the neck and the hip. The normalized poses from three different subjects are shown in Fig. 2. From the figure we can find step lengths and leg styles at some instance are different among the three different subjects.

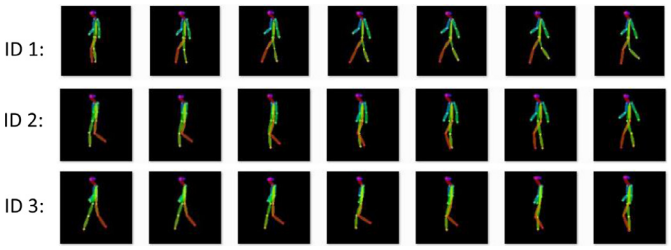

Fig. 2. 2D poses from three different subjects. Step lengths and leg styles at some instance are different among the three different subjects.

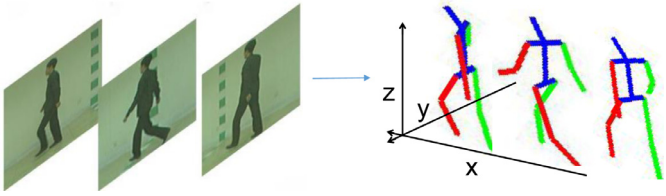

Fig. 3. The estimated 3D poses (right ones) from RGB images (left ones).

3.1.2. 3D pose feature

The pose extracted from images is 2D. When the view is changed, the 2D pose will also be changed drastically. So it is not robust to the view variation. The solution is to estimate the 3D pose from the 2D pose [18]. The unique advantage of [18] is that it can estimate 3D pose from one single image, specifically from the 2D pose by human body constrains. It makes the method feasible in real applications.

In [18], the input data should be the positions of 14 joints. The 2D joints are Head, Neck, RShoulder, RElbow, RWrist, LShoulder, LElbow, LWrist, RHip, RKnee, Rankle, LHip, LKnee and LAnkle. But there are 18 joints can be estimated using OpenPose method [6]. So we averaged the position of Nose, Reye, LEye, Rear and Lear as the position of Head. The 3D pose feature f_{pose} of frame c is defined as:

$$f_{pose} = \{J_0, J_1, \ldots, J_N\} \tag{2}$$

where $J_i = \{x_i, y_i, z_i\}$, $i \in \{0, 1, 2, \ldots, N\}$ and $N = 13$.

In order to reduce the effect of view variation, we set the x direction to the subject's frontal direction, and the y direction to the side of the body defined by the left shoulder and the right one, and the z is the vertical one to the ground. The 3D pose is rotated and normalized in this 3D space as shown in Fig. 3.

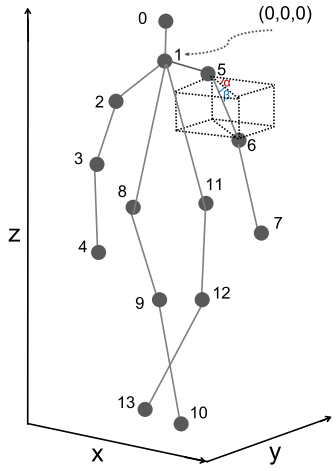

Fig. 4. The angle information α, β between left shoulder joint J_5 and left elbow joint J_6. The local movement pattern would be captured through the change of α and β during walking process. The human joints should be confined human 3D skeleton model according to the human prior knowledge. Thus there are 13 pairs of angle motion on each frame pose.

3.2. Designing spatio-temporal features

It should be helpful to design some handcrafted features based on 3D pose (ie., joints position in a 3D Euclidean space), such as joint angles, motions for gait recognition. The features based on prior knowledge will ease the task of the deep neural network. There is a similar approach in [21], coined *EigenJoints*, for human action recognition: it can combine the features of static pose, motion and offset to improve action recognition. Inspired by *EigenJoints*, we design three kinds of additional spatio-temporal pose features: 1) joint angle, the angle changing at some joints; 2) limb length, the human body static measures; 3) joint motion, the dynamic feature to describe the motion patterns.

3.2.1. Joint angle

In [3] Wang et al. proposed a model-based method to employ joint angle and joint trajectories of lower limbs to capture the dynamic feature of gait. The authors demonstrate experimentally that the joint angle changing in lower limbs can benefit gait recognition. Compared with the method in [3], the joint positions are more accurate in the proposed method. Besides, the joint positions are in a 3D space not a 2D one, and not only the lower limbs but all the joints can be captured for the model-based feature extraction.

The angle feature f_{angle} of frame c is defined in the following equations. As described in Fig. 4, two angles α and β are defined.

$$f_{angle} = \{(\alpha_{ij}, \beta_{ij}) | (i, j) \in \Phi\} \tag{3}$$

$$\alpha_{ij} = \begin{cases} arctan\frac{y_i - y_j}{x_i - x_j}, & x_i \neq x_j \\ \frac{\pi}{2}, & x_i = x_j \end{cases} \tag{4}$$

$$\beta_{ij} = \begin{cases} arctan\frac{z_i - z_j}{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}}, & (x_i - x_j)^2 + (y_i - y_j)^2 \neq 0 \\ \frac{\pi}{2}, & (x_i - x_j)^2 + (y_i - y_j)^2 = 0 \end{cases} \tag{5}$$

where $J_i = (x_i, y_i, z_i), J_j = (x_j, y_j, z_j)$. The (i, j) is in the set of Φ, and Φ = {(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)}. The angles are defined between two adjacent joints. The selection of the two adjacent joints is constrained by the human 3D skeleton model. The angles α and β are defined between left shoulder joint J_5 and left elbow joint J_6 as shown in Fig. 4.

3.2.2. Limb length

The limb lengths are the distances between two adjacent joints. The limb lengths can be regarded as a model-based spatial feature. The work in [2,3] demonstrates that this feature is feasible for gait recognition. So we involve the spatial feature into *PoseGait* method. The feature is also robust to view, carrying condition and clothing variations. The definition of the limb length of frame c is:

$$f_{limb} = \|J_i - J_j\|_2 \tag{6}$$

where $(i, j) \in \Phi$ and Φ is the same with that defined previously. There are 13 static limb lengths for each human skeleton.

3.2.3. Joint motion

The walking style can be described by the motion of joints. It is reasonable and straightforward to use joint motions as the feature for gait recognition. In [22] Chen et al. proposed a kind of feature FDEI which uses the difference between frames to capture the dynamic information for gait recognition. FDEI is the difference between human body silhouettes, and here we use the difference between body joints. The joint motion is defined by the difference between two adjacent frames, frame c and frame $c+1$ is as follows.

$$f_{motion} = P_{c+1} - P_c \tag{7}$$

where $P = \{J_0, J_1, \ldots, J_N\}, J_i = \{x_i, y_i, z_i\}$, and $i \in \{0, 1, 2, \ldots, N\}, N = 13$.

3.3. Fusion of features

For each frame, we can get the 4 kinds of features as described previously, f_{pose}, f_{angle}, f_{limb} and f_{motion}. The four vectors can be concatenated to a long vector to present the pose, motion and static body measures. The feature vector from different frames can be put together to form a feature matrix as show in Fig. 5. Because f_{motion} is computed from two frames, the number of f_{motion} vectors is less one than other kinds of features. We put a zero vector to make the matrix complete. The number of frames is fixed. The size of the feature matrix has a fixed size. It can be taken as the input of a CNN model.

3.4. The network design

Since the feature is extracted frame by frame and is sequence data, it is reasonable to employ a method for sequential data such as RNN [23] and LSTM [24]. In our previous works [25,26] we proposed one method named as PTSN which can combine CNN and LSTM for gait recognition. But some researchers [27,28] argued that CNN is better than RNN on recognition tasks. Compared with CNN, RNN is computationally expensive and sometimes difficult to train. In addition, Zhang et al. [29] showed that CNN has enough capability to model temporal data. So we choose CNN nor LSTM for the proposed method.

For feature extraction in gait recognition, it is crucial to reduce the intra-class variation and enlarge the inter-class one. As suggested in [30,31], a multi-loss strategy is employed to optimize the network. There are two losses, the Softmax loss (Eq. (8)) and the center loss (Eq. (9)). The softmax loss can classify the input into different classes. That means the softmax loss can enlarge the inter-class variation. The center loss can keep the features of different classes separable by minimizing the intra-class variation. The two losses are combined as defined in Eq. (10).

$$L_{softmax} = -\sum_{i=1}^{m} \log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}} \tag{8}$$

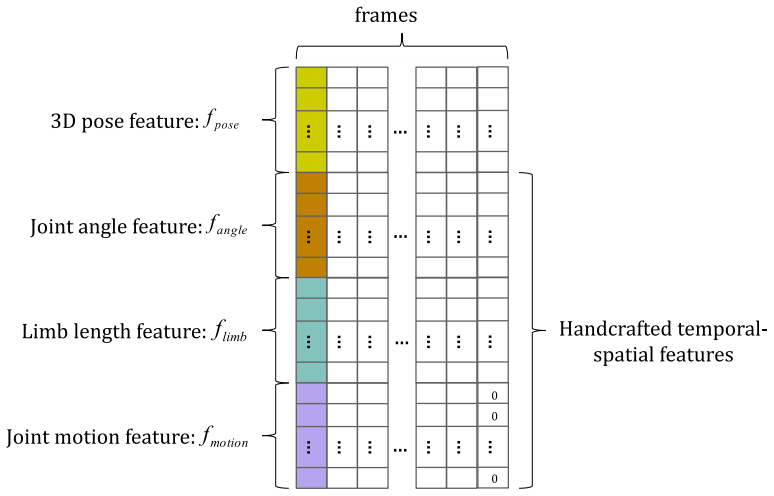

Fig. 5. The four kinds of features from a sequence are concatenated to form a feature matrix.

$$L_{center} = \frac{1}{2}\sum_{i=1}^{m} ||x_i - c_{y_i}||_2^2 \tag{9}$$

$$L = L_{softmax} + \lambda \cdot L_{center} \tag{10}$$

where $x_i \in \mathbb{R}^d$ is the ith feature that belongs to the y_ith class, d, $W \in \mathbb{R}^{d\times n}$ and $b \in \mathbb{R}^d$ denote the feature dimension, last connected layer and bias term, respectively, and $c_{y_i} \in \mathbb{R}^d$ is the y_ith class center of deep features. In our experiments λ was set to 0.008. It is the best value in our experiments.

4. Experimental results and analysis

4.1. Datasets

To evaluate the proposed gait recognition method, RGB color video frames are needed because the human poses should be estimated from color images and cannot from silhouettes. We chose CASIA B Gait Dataset [32] since it contains the original color video frames. The OU-ISIR research group in Osaka University also provided several large gait datasets [33]. But the large datasets from OU-ISIR cannot provide the original frames because of the privacy issue. We chose CASIA E Dataset as a second large dataset.

CASIA B dataset is one of the popular public gait datasets widely used in research community. It was created at the Institute of Automation, Chinese Academy of Sciences (CASIA). It contains 124 subjects in total (31 females and 93 males). There are 10 sequences for each subject, 6 sequences of normal walking (NM), 2 sequences of walking with bag (BG) and 2 sequences of walking with coat (CL). There are 11 views which were captured from 11 cameras at the same time, the view angles are {0°, 18°, ⋯, 180°}. Fig. 6 illustrates the samples at 11 views from a subject of normal walking.

CASIA E is a newly created gait dataset by the Institute of Automation, Chinese Academy of Sciences and the Watrix company. The dataset contains 1014 subjects and is much larger than CASIA B. Different from other gait datasets with more than one thousand subjects, the gait data was collected from at 13 different views. The views is from 0° to 180° with a 15° interval between two adjent views in the horizontal direction. There are 6 sequences for each subject. They are 2 sequences for normal walking (NM), 2 for walking with a bag (BG) and 2 for walking in a coat (CL). We are working on the privacy issue and will release the dataset later.

Fig. 6. Normal walking sequences at 11 views from CASIA B dataset.

Table 1
Experimental setting on CASIA B dataset.

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

4.2. Experimental settings

The first set of experiments is carried on CASIA B dataset. We put the first 62 subjects into the training set and the rest of subjects into the test set as the experimental setting of SPAE [8] and GaitGAN [34]. In the test set, the gallery set consists of the first 4 normal walking sequences of each subjects, and the probe set consists of the rest of sequences, as shown in Table 1.

For the experiments on CASIA E dataset, the experimental setting is similar with that on CASIA B. The first 507 subjects of CASIA E dataset are put into the training set, and the other 507 subjects are put into the test set. In the test set, there are two types of settings. For the normal walking with the identical view condition, we put the first normal walking sequence into the gallery and the second normal walking sequence into the probe set because each subject has only two normal walking sequences at the same view. In the second setting, the first 2 normal walking sequences are put into the gallery and the others into probe set. The experimental setting is shown in Table 2.

Table 2
Experimental setting on CASIA E dataset.

Training	Test	
	Gallery set	Probe set
ID: 001–0507 Seqs: NM01-NM02 BG01-BG02, CL01-CL02	ID: 508–1014 Seqs: NM01-NM02	ID: 508–1014 Seqs: NM01-NM02 BG01-BG02, CL01-CL02

* *Note:* In the test set, for the normal walking with the identical view condition, gallery set contains NM01 and probe set contains NM02.

Table 3
Implementation of the CNN on CASIA B dataset.

Layers	Number of filters	Filter size	Stride	Activation function
Conv.1	32	3 × 3	1	P-ReLU
Conv.2	64	3 × 3	1	P-ReLU
Pooling.1	–	2 × 2	2	–
Conv.3	64	3 × 3	1	P-ReLU
Conv.4	64	3 × 3	1	P-ReLU
Eltwise.1	Sum operation between Pooling.1 and Conv.4			
Conv.5	128	3 × 3	1	P-ReLU
Pooling.2	–	2 × 2	2	–
Conv.6	128	3 × 3	1	P-ReLU
Conv.7	128	3 × 3	1	P-ReLU
Eltwise.2	Sum operation between Pooling.2 and Conv.7			
FC.1	512	–	–	–

In order to make the readers follow the method easily, we also listed the CNN implementation details on CASIA B dataset in Table 3. It is a light weighted network with 7 convolutional layers. For the experiments on CASIA E there is about 10 times data for training. We found that a deeper network can achieve better results. The implementation of the deeper network is shown in Table 4. We also tried to use the deeper network for CASIA E to train a model for CASIA B. But the model is tend to be overfitted because the data is not enough for the deeper network.

4.3. Experimental results on CASIA B dataset

The complete experimental results on CASIA B dataset are listed in Tables 5–7. The evaluation of view, carrying condition and clothing variations are shown in the three tables. In experiments, the first 4 normal walking sequences at a specific view are put into the gallery set, and the last 2 normal sequences, 2 walking with a bag sequences and 2 walking with a coat sequences are put into the probe set of the three sets of experiment respectively. For each set of experiments, there are 121 combinations. That means there are 121 recognition rates in each table.

4.4. Effectiveness of the handcrafted features by prior knowledge

To light the burden of feature extraction for CNN and make the feature more discriminative, handcrafted features by human prior knowledge are involved in the proposed method. To prove the effectiveness of the handcrafted features, we designed 5 sets of experiments. For the first set of experiments, only the 3D pose feature f_{pose} was used. For the second to the fouth ones, the feature f_{angle}, f_{limb} and f_{motion} were evaluated respectively. In the fifth one, four kinds of features were evaluated by concatenating them to be a feature vector $[f_{pose}, f_{angle}, f_{limb}, f_{motion}]$. The handcrafted features by prior knowledge include joint angles, limb lengths and joint motions.

There are also 121 combinations in each group of experiments. To show the results clearly, we just show the average of the 121 recognition rates in Table 8. From the results it can be found

Table 4
Implementation of the CNN on CASIA E dataset.

Layers	Number of filters	Filter size	Stride	Activation function
Conv.1	32	3 × 3	1	P-ReLU
Conv.2	64	3 × 3	1	P-ReLU
Pooling.1	–	2 × 2	2	–
Conv.3	64	3 × 3	1	P-ReLU
Conv.4	64	3 × 3	1	P-ReLU
Eltwise.1	Sum operation between Pooling.1 and Conv.4			
Conv.5	128	3 × 3	1	P-ReLU
Pooling.2	–	2 × 2	2	–
Conv.6	128	3 × 3	1	P-ReLU
Conv.7	128	3 × 3	1	P-ReLU
Eltwise.2	Sum operation between Pooling.2 and Conv.7			
Conv.8	128	3 × 3	1	P-ReLU
Conv.9	128	3 × 3	1	P-ReLU
Eltwise.3	Sum operation between Eltwise.2 and Conv.9			
Conv.10	256	3 × 3	1	P-ReLU
Pooling.3	–	2 × 2	2	–
Conv.11	256	3 × 3	1	P-ReLU
Conv.12	256	3 × 3	1	P-ReLU
Eltwise.4	Sum operation between Pooling.3 and Conv.12			
Conv.13	256	3 × 3	1	P-ReLU
Conv.14	256	3 × 3	1	P-ReLU
Eltwise.5	Sum operation between Eltwise.4 and Conv.14			
Conv.15	256	3 × 3	1	P-ReLU
Conv.16	256	3 × 3	1	P-ReLU
Eltwise.6	Sum operation between Eltwise.5 and Conv.16			
Conv.17	256	3 × 3	1	P-ReLU
Conv.18	256	3 × 3	1	P-ReLU
Eltwise.7	Sum operation between Eltwise.6 and Conv.18			
Conv.19	256	3 × 3	1	P-ReLU
Conv.20	256	3 × 3	1	P-ReLU
Eltwise.8	Sum operation between Eltwise.7 and Conv.20			
FC.1	512	–	–	–

some interesting conclusions. Firstly, if there is not any variation, pose feature f_{pose} can achieve a best recognition rate 60.92%. The pose feature is also the best one among these kinds of features. Secondly the motion feature f_{motion} achieves a recognition rate of 30.38% where there is clothing variation, and it is the best among the four individual features. That means motion is robust to the clothing variation. At last the performance can be improved obviously by combining all these features. It also means that the handcrafted features can improve the recognition rate obviously.

4.5. Comparisons with appearance-based methods

As stated in the previous part of the paper, the model-based feature used in the proposed method is compact and has less redundant information as some appearance-based features. It means the feature extraction is more challenging. To show effectiveness of the model-based features, we make comparisons with some appearance-based methods. There are two groups of comparisons according to different experimental settings.

The first group of comparisons is made between the proposed method and four state-of-the-arts which have the same experimental settings in Section 4.2, namely GEI+PCA [4], SPAE [8], GaitGANv1 [34] and GaitGANv2 [35]. The average recognition rates for the probe data being NM, BG and CL with the view variation are shown in Fig. 7.

From Fig. 7, it can be found that the proposed method achieves much higher recognition rates than those of GEI+PCA, SPAE and

Table 5
Recognition rates when the probe data is normal walking on CASIA B dataset.

		Probe angle (normal #4-5)										
		0°	**18°**	**36°**	**54°**	**72°**	**90°**	**108°**	**126°**	**144°**	**162°**	**180°**
Gallery angle (normal #1-4)	**0°**	95.97	88.71	61.29	45.97	27.42	26.61	27.42	30.65	47.58	65.32	63.71
	18°	89.52	96.77	95.16	83.06	54.84	35.48	41.13	37.10	60.48	65.32	64.52
	36°	69.35	95.16	95.97	93.55	76.61	66.13	59.68	51.61	58.87	50.00	45.16
	54°	38.71	81.45	95.16	96.77	91.94	83.06	75.00	59.68	54.03	44.35	38.71
	72°	32.26	54.03	75.81	88.71	95.97	96.77	79.84	70.97	59.68	37.10	24.19
	90°	27.42	39.52	66.94	81.45	91.13	97.58	90.32	70.97	66.13	41.13	22.58
	108°	27.42	37.90	62.90	71.77	83.06	89.52	97.58	91.13	81.45	59.68	31.45
	126°	38.71	44.35	53.23	66.13	70.16	75.81	92.74	94.35	91.94	79.84	43.55
	144°	42.74	50.81	58.87	62.90	53.23	64.52	83.87	94.35	96.77	87.90	60.48
	162°	62.10	60.48	51.61	41.94	37.10	31.45	53.23	71.77	89.52	97.58	80.65
	180°	68.55	63.71	49.19	31.45	22.58	20.16	21.77	35.48	62.90	89.52	97.58

Table 6
Recognition rates when the probe data is with carrying variation on CASIA B dataset.

	Probe angle (walking with a bag #1-2)										
	0°	**18°**	**36°**	**54°**	**72°**	**90°**	**108°**	**126°**	**144°**	**162°**	**180°**
0°	74.19	56.45	41.13	27.42	16.13	14.52	16.94	20.97	25.00	37.90	34.68
18°	56.45	75.81	72.58	58.06	34.68	29.03	19.35	28.23	32.26	33.87	33.87
36°	41.94	70.97	77.42	74.19	56.45	42.74	40.32	37.90	35.48	35.48	29.03
54°	30.65	60.48	73.39	76.61	65.32	56.45	49.19	42.74	37.10	30.65	21.77
72°	22.58	42.74	54.84	65.32	69.35	62.90	52.42	45.16	38.71	23.39	18.55
90°	20.97	28.23	51.61	59.68	66.94	70.16	58.06	54.84	44.35	22.58	14.52
108°	19.35	26.61	41.94	43.55	59.68	66.13	70.97	65.32	57.26	31.45	16.13
126°	21.77	28.23	33.87	42.74	45.97	50.81	65.32	69.35	65.32	43.55	24.19
144°	33.06	30.65	33.06	35.48	37.10	48.39	50.00	60.48	74.19	61.29	33.06
162°	36.29	41.94	29.03	23.39	22.58	28.23	29.03	32.26	58.87	65.32	50.00
180°	42.74	35.48	21.77	16.13	14.52	15.32	16.13	21.77	30.65	52.42	60.48

Table 7
Recognition rates when the probe data is with clothing variation on CASIA B dataset.

		Probe angle (walking with a coat #1-2)										
		0°	**18°**	**36°**	**54°**	**72°**	**90°**	**108°**	**126°**	**144°**	**162°**	**180°**
Gallery angle (normal #1-4)	**0°**	46.77	33.87	21.77	13.71	14.52	15.32	12.90	23.39	25.00	28.23	24.19
	18°	33.87	48.39	52.42	32.26	26.61	16.94	20.97	21.77	29.84	25.00	22.58
	36°	26.61	42.74	57.26	50.81	45.97	31.45	32.26	35.48	31.45	28.23	21.77
	54°	17.74	23.39	56.45	61.29	51.61	47.58	37.90	37.90	30.65	21.77	12.90
	72°	14.52	20.16	47.58	51.61	58.06	50.81	45.16	42.74	31.45	21.77	12.10
	90°	9.68	19.35	37.10	55.65	54.03	56.45	58.06	50.00	34.68	20.97	9.68
	108°	8.87	13.71	33.06	40.32	48.39	44.35	59.68	47.58	41.94	20.97	10.48
	126°	13.71	15.32	23.39	29.03	33.06	40.32	61.29	54.84	46.77	29.03	17.74
	144°	21.77	25.81	28.23	29.84	26.61	30.65	48.39	51.61	55.65	43.55	20.97
	162°	29.03	22.58	22.58	21.77	13.71	13.71	28.23	30.65	42.74	58.06	37.90
	180°	29.84	23.39	12.90	9.68	12.90	13.71	14.52	20.16	23.39	36.29	39.52

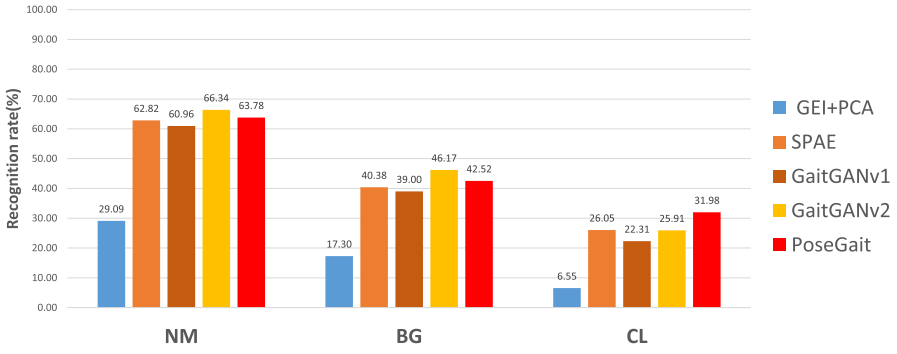

Fig. 7. The average recognition rates for the probe data being NM, BG and CL with the view variation.

Table 8
The recognition rates of different features.

Feature	Probe set		
	NM	BG	CL
$[f_{pose}]$	60.92%	39.16%	29.71%
$[f_{angle}]$	46.97%	26.60%	25.61%
$[f_{limb}]$	42.40%	25.55%	10.63%
$[f_{motion}]$	48.95%	30.31%	30.38%
$[f_{pose}, f_{angle}, f_{limb}, f_{motion}]$	**63.78%**	**42.52%**	**31.98%**

GaitGANv1. The recognition rates are comparable with those of GaitGANv2. It should be noticed that the proposed method achieves a much better recognition rate when there is a clothing variation. That means the proposed method is more robust to the clothing variation. It is the advantage of the model-based features. The raw feature is body joints and robust to clothing while the appearance-based features tend to be changed by clothing.

In [9], Wu et al. proposes a method using CNN and achieves very high recognition rates. It is also an appearance-based method using human body silhouettes. But the experimental setting in [9] is different from that of the proposed method. The models in [9] are trained using the first 74 subjects in CASIA B dataset. To compare fairly, we also did experiments using the same experimental setting with Wu's method. The experimental results are listed in Table 9.

The experimental results of DeepCNNs [9] and the proposed method are also listed in Table 9. Both models are trained with gait sequences of the first 74 subjects. The method of DeepCNNs [9] has achieved a very high performance. There are two reasons for this high accuracy. Firstly, the feature they used is a kind of appearance-based one which is a high dimension one. We only used 14 body joints as gait feature. Secondly, they train CNN with pairs in a verification manner, so the number of combinations for training could be more than a million. By contrast, our models were trained in a classification manner nor a verification manner. The number of samples were much less than that in [9].

4.6. Effectiveness on view variation

From the previous experiments, it can be found the proposed method can achieve comparable recognition rates with state-of-the-art, and it is also more robust to clothing variation. Here we also want to compare the proposed method with some cross-view gait recognition methods to show the effectiveness on view variation. Some cross-view gait recognition methods are FD-VTM [36], RSVD-VTM [37], RPCA-VTM [38], R-VTM [39], GP+CCA [40] and C3A [41]. The probe angles chosen are 54°, 90° and 126°, and it is the same experimental setting with these methods in the original papers by their authors. The experimental results are shown in Fig. 8.

It can be clearly found that the proposed method achieves much high recognition rates when the difference between the gallery angle and the probe one is large. The greater is the difference, the more obvious improvements. The improvements can be easily understood because the poses used for gait recognition are in 3D spaces and normalized to the same view angle. That is the reason the proposed model-based method is more robust to view variation.

4.7. Experimental results on CASIA E dataset

For further evaluation on the performance of the proposed method, we carried out experiments on the CASIA E dataset. Since this dataset is still not public available, we implemented some

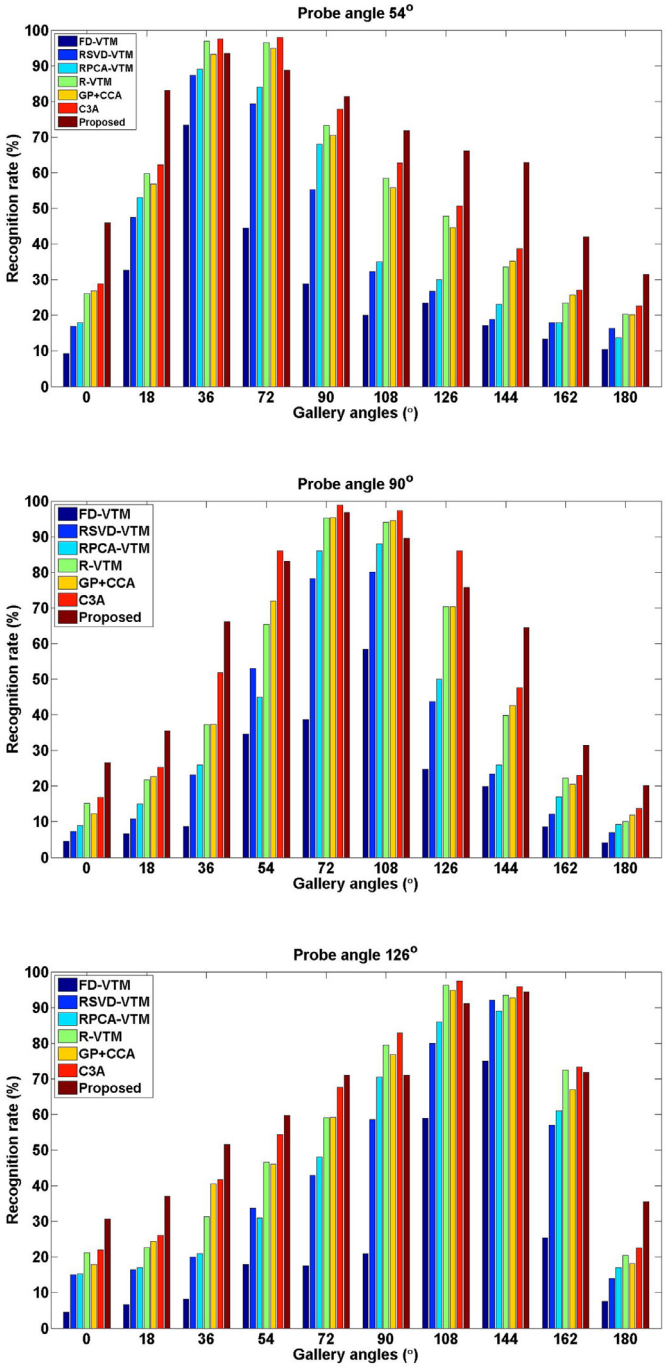

Fig. 8. Comparing with some cross-view methods at probe angle 54°, 90° and 126°. The gallery angles are the remaining 10 angles except the corresponding probe angle. The proposed method achieves much higher recognition rates when the difference between the gallery angle and the probe one is large.

methods by ourselves and can not cite the results from the original paper. In the experiments there are three methods, and they are GEI+PCA [4], GaitGANv2 [35] and the proposed *PoseGait*. The experiment settings for the three methods are that shown in Table 2. The experimental results rates are shown in Fig. 9. We only list 4 probe angles with a 60° interval as the limited space. Each row represents a probe angle, the compared angles are 0°, 60°, 120° and 180°. Three columns in Fig. 9 shows the comparison with normal walking (NM), carrying a bag (BG) and clothing (CL) condition, respectively.

Table 9
Average recognition rate(%) comparisons of the proposed method with Wu's on CASIA-B. Both models are trained with the same experimental setting of 74 training subjects.

Gallery angle NM #1-4	0°-180°											
Probe angle NM #5-6	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
DeepCNNs [9]	88.7	95.1	98.2	96.4	94.1	91.5	93.9	97.5	98.4	95.8	85.6	94.1
PoseGait	55.3	69.6	73.9	75	68	68.2	71.1	72.9	76.1	70.4	55.4	68.72
Probe angle BG #1-2	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
DeepCNNs [9]	64.2	80.6	82.7	76.9	64.8	63.1	68.0	76.9	82.2	75.4	61.3	72.4
PoseGait	35.3	47.2	52.4	46.9	45.5	43.9	46.1	48.1	49.4	43.6	31.1	44.5
Probe angle CL #1-2	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
DeepCNNs [9]	37.7	57.2	66.6	61.1	55.2	54.6	55.2	59.1	58.9	48.8	39.4	53.98
PoseGait	24.3	29.7	41.3	38.8	38.2	38.5	41.6	44.9	42.2	33.4	22.5	35.95

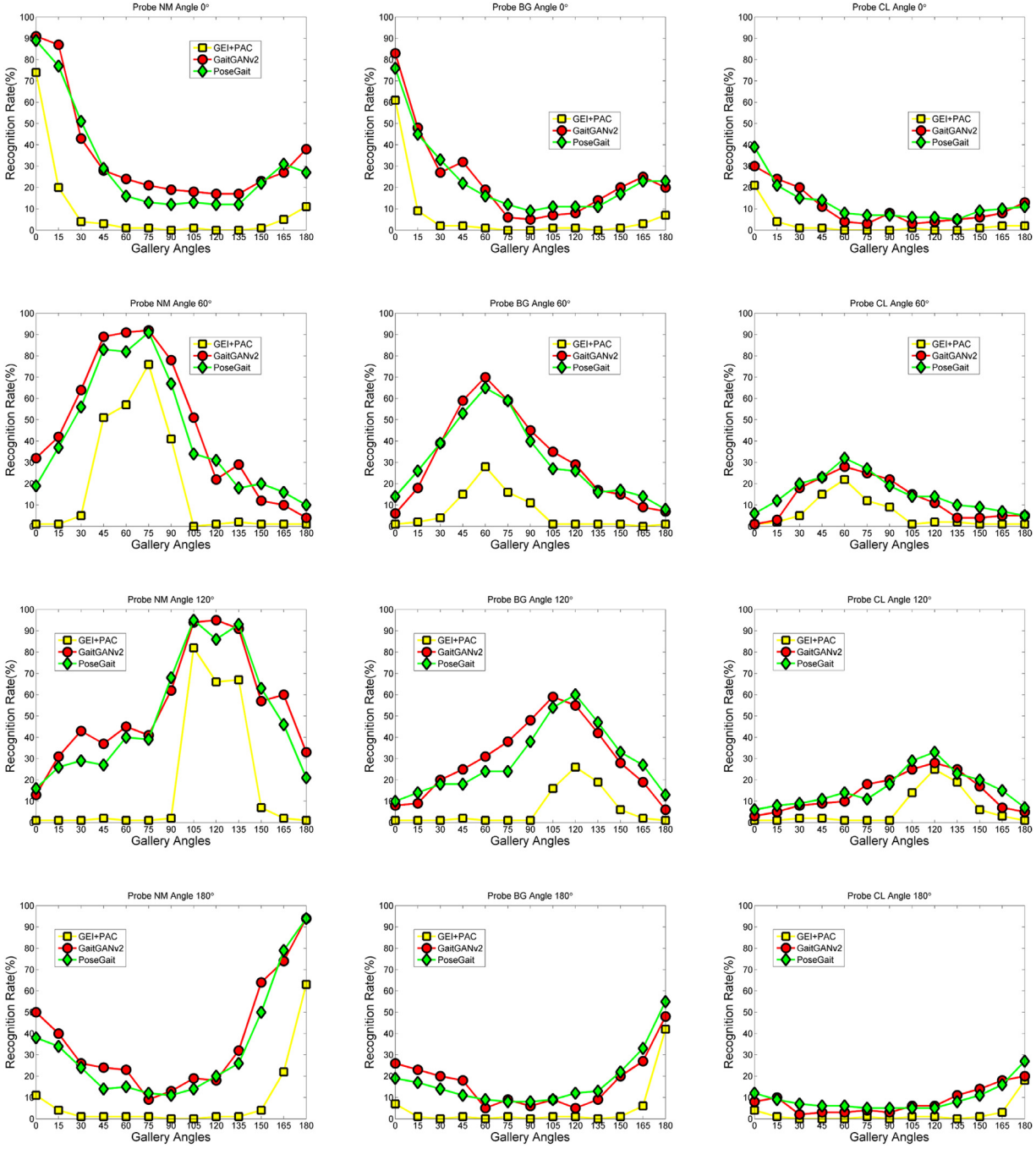

Fig. 9. The experimental results on a large dataset, CASIA E, with over 1000 subjects.

Table 10
The computational cost of different steps in the proposed method.

Step	Time (ms)	Description
2D Pose estimation	91.2	GPU
2D to 3D	204.3	CPU & Matlab code
CNN feature extraction	1.78(CASIA B), 11.8(CASIA E)	GPU

From the results it can be found that the proposed method achieves comparable performance with the state-of-the-art method, GaitGANv2. When there is clothing variation, the proposed method is a little better than GaitGANv2. It is similar with that in Fig. 7. The experiments on the two datasets demonstrates that the proposed method has its own advantage on view variation.

4.8. Computational cost analysis

In the proposed method, most computational cost is from the pre-processing step. That is the body pose estimation from images. We ran the proposed method on a server with a Tesla K80 (12GB) GPU and listed the time consumed of different steps in Table 10. For the 2D pose estimation step, all images were resized to 368×368 as in [6]. It took about 0.2s to convert one 2D pose to a 3D one. The computational cost of 2D to 3D actually is low enough. But because the code is in Matlab language and is not optimized, it took a lot of time. It can be much faster if the code is written in C/C++ or some other languages. For the gait feature extraction part, the CNN model for CASIA B is simpler than that for CASIA E as described in Section 4.2. The model for CASIA only took 1.78ms on GPU. Even the heavy model for CASIA E only took 11.8ms. According to the time consumed for the proposed method. It can be shown that the proposed method is fast and efficient.

5. Conclusions and future work

With the progress on human body modeling based on deep learning, we proposed a model-based gait recognition method, named *PoseGait*, in the paper to invest model-based features for gait recognition. *PoseGait* employs 3D human body poses as feature. The feature is very compact since there are only body joints in it. Experimental results on CASIA B and CASIA E datasets show that proposed method performance is comparable with some state-of-the-art appearance-based methods. In addition, we combine three types of spatio-temporal features based on human prior knowledge with body pose to enrich the feature and improve the recognition rate. The experiments also show that CNN can extract temporal feature efficiently and achieve better results than LSTM or RNN.

It also should be noticed that the proposed model-based method just achieves comparable accuracy with state-of-the-art. Even so, it shows that model-based methods have great potential on gait recognition. Besides OpenPose, there are also some other work which can model human bodies with more details. Such as DensePose in [42] can model the human body surface with a mesh. We ever tried to use the mesh for gait recognition. But DensePose can only model the body surface which faces to the camera and can not estimate the surface occluded. That makes the data is not completed and difficult to be used for gait recognition. In future human body modeling will continue to improve. Surely model-based gait recognition will also be improved with better human body models.

Acknowledgments

The work is partly supported by the National Natural Science Foundation of China (grant no. 61976144) and the Science Foundation of Shenzhen (grant no. 20170504160426188). We also would like to thank Prof. Véronique Prinet for her valuable suggestions and comments.

References

[1] M.S. Nixon, J.N. Carter, J.M. Nash, P.S. Huang, D. Cunado, S.V. Stevenage, Automatic gait recognition, in: Motion Analysis and Tracking, 1999, pp. 3/1–3/6.
[2] 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
[3] L. Wang, H. Ning, T. Tan, W. Hu, Fusion of static and dynamic body biometrics for gait recognition, IEEE Trans. Circuits Syst. Video Technol. 14 (2) (2004) 149–158.
[4] J. Han, B. Bhanu, Individual recognition using gait energy image, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2) (2006) 316–322.
[5] Y. Zhang, Y. Huang, L. Wang, S. Yu, A comprehensive study on gait biometrics using a joint cnn-based method, Pattern Recognit 93 (2019) 228–236.
[6] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, Y. Sheikh, in: OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields, 2018, pp. 1–14. arXiv:1812.08008.
[7] X. Huang, N.V. Boulgouris, Gait recognition with shifted energy image and structural feature extraction, IEEE Trans. Image Process. 21 (4) (2012) 2256–2268.
[8] 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.
[9] Z. Wu, Y. Huang, L. Wang, X. Wang, T. Tan, A comprehensive study on cross-view gait based human identification with deep cnns, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2) (2017) 209–226.
[10] H. Chao, Y. He, J. Zhang, J. Feng, Gaitset: regarding gait as a set for cross-view gait recognition, 2018 arXiv:1811.06186.
[11] Y. Zhang, Y. Huang, S. Yu, L. Wang, Cross-view gait recognition by discriminative feature learning, IEEE Trans. Image Process. (2019). 1–1
[12] Y. Feng, Y. Li, J. Luo, Learning effective gait features using LSTM, in: the 23rd International Conference on Pattern Recognition, 2016, pp. 325–330.
[13] D. Kastaniotis, I. Theodorakopoulos, S. Fotopoulos, Pose-based gait recognition with local gradient descriptors and hierarchically aggregated residuals, J. Electron. Imag. 25 (6) (2016) 063019.
[14] C. Hong, J. Yu, D. Tao, M. Wang, Image-based three-dimensional human pose recovery by multiview locality-sensitive sparse retrieval, IEEE Trans. Ind. Electron. 62 (6) (2015) 3742–3751.
[15] G. Xie, K. Yang, J. Lai, Filter-in-filter: low cost cnn improvement by sub-filter parameter sharing, Pattern Recognit. 91 (2019) 391–403.
[16] C.-L. Zhang, J. Wu, Improving cnn linear layers with power mean non-linearity, Pattern Recognit. 89 (2019) 12–21.
[17] C. Hong, J. Yu, J. Wan, D. Tao, M. Wang, Multimodal deep autoencoder for human pose recovery, IEEE Trans. Image Process. 24 (12) (2015) 5659–5670.
[18] C.-H. Chen, D. Ramanan, 3D human pose estimation = 2D pose estimation + matching, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7035–7043.
[19] J. Yu, X. Yang, F. Gao, D. Tao, Deep multimodal distance metric learning using click constraints for image ranking, IEEE Trans. Cybern. 47 (12) (2017) 4014–4024.
[20] Z. Yu, J. Yu, C. Xiang, J. Fan, D. Tao, Beyond bilinear: generalized multimodal factorized high-order pooling for visual question answering, IEEE Trans. Neural Netw. Learn. Syst. 29 (12) (2018) 5947–5959.
[21] X. Yang, Y.L. Tian, Eigenjoints-based action recognition using nave-bayes-nearest-neighbor, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2012, pp. 14–19.
[22] C. Chen, J. Liang, H. Zhao, H. Hu, J. Tian, Frame difference energy image for gait recognition with incomplete silhouettes, Pattern Recognit. Lett. 30 (11) (2009) 977–984.
[23] M. Schuster, K.K. Paliwal, Bidirectional recurrent neural networks, IEEE Trans. Signal Process. 45 (11) (1997) 2673–2681.
[24] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[25] 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: the 12th Chinese Conference on Biometric Recognition, 2017, pp. 474–483.
[26] W. An, R. Liao, S. Yu, Y. Huang, P.C. Yuen, Improving gait recognition with 3D pose estimation, in: the 13th Chinese Conference on Biometric Recognition, 2018, pp. 137–147.
[27] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014 arXiv:1409.1556.
[28] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[29] Y. Zhang, M. Pezeshki, P. Brakel, S. Zhang, C.L.Y. Bengio, A. Courville, Towards end-to-end speech recognition with deep convolutional neural networks, 2017 arXiv:1701.02720.

[30] K. Zhang, Y. Huang, Y. Du, L. Wang, Facial expression recognition based on deep evolutional spatial-temporal networks, IEEE Trans. Image Process. 26 (9) (2017) 4193–4203.
[31] Y. Wen, K. Zhang, Z. Li, Y. Qiao, A discriminative feature learning approach for deep face recognition, in: European Conference on Computer Vision, 2016, pp. 499–515.
[32] 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.
[33] OU-ISIR Biometric Database. http://www.am.sanken.osaka-u.ac.jp/BiometricDB/index.html.
[34] 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.
[35] 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 Recognit. 87 (2019) 179–189.
[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 Trans. Circuits Syst. Video Technol. 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 Recognit. 50 (2016) 107–117.
[42] R.A. Güler, N. Neverova, I. Kokkinos, Densepose: dense human pose estimation in the wild, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7297–7306.

Rijun Liao received his B.S. degree from the College of Physics and Energy, Shenzhen University, China in 2015. He is currently a master student in the College of Computer Science and Software Engineering, Shenzhen University, China. His research interests include biometrics, computer vision and deep learning.

Shiqi Yu currently is an associate professor in the Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China. He received his B.E. degree in computer science and engineering from the Chu Kochen Honors College, Zhejiang University in 2002, and Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences in 2007. He worked as an assistant professor and an associate professor in Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences from 2007 to 2010, and as an associate professor in Shenzhen University from 2010 to 2019. His research interests include computer vision, pattern recognition and artificial intelligence.

Weizhi An received her B.S. degree from the College of Computer Science and Software Engineering, Shenzhen University, China in 2016. She is currently a master student in the College of Computer Science and Software Engineering, Shenzhen University, China. Her research interests include biometrics, computer vision and deep learning.

Yongzhen Huang received the B.E. degree from the Huazhong University of Science and Technology in 2006 and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences (CASIA) in 2011. In July 2011, he joined the National Laboratory of Pattern Recognition (NLPR), CASIA, where he is currently an associate professor. He has published more than 50 papers in the areas of computer vision and pattern recognition at international journals such as IEEE Transactions on Pattern Analysis and Machine Intelligence, International Journal of Computer Vision, IEEE Transactions on Systems, Man, and Cybernetics, IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions on Multimedia, and conferences such as CVPR, ICCV, NIPS, and BMVC. His current research interests include pattern recognition, computer vision, and machine learning.