Improving Gait Recognition with 3D Pose Estimation

Weizhi An[1], Rijun Liao[1], Shiqi Yu[1], Yongzhen Huang[2], and Pong C Yuen[3]

1. College of Computer Science and Software Engineering,
Shenzhen University, Shenzhen, 518060, China.
2. National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, 100190, China
3. Department of Computer Science, Hong Kong Baptist University,
Hong Kong SAR, China
anweizhi2016@email.szu.edu.cn, 2150230306@email.szu.edu.cn,
shiqi.yu@szu.edu.cn, yongzhen.huang@nlpr.ia.ac.cn, pcyuen@comp.hkbu.edu.hk

Abstract. Gait is a kind of attractive biometric feature for human identification in recent decades. The view, clothing, carrying and other variations are always the challenges for gait recognition. One of the possible solutions is the model based methods. In this paper, 3D pose is estimated from 2D images are used as the feature for gait recognition. So gait can be described by the motion of human body joints. Besides, the 3D pose has better capacity for view variation than the 2D pose. Experimental results also prove that in the paper. To improve the recognition rates, LSTM and CNNs are employed to extract temporal and spatial features. Compared with other model-based methods, the proposed one has achieved much better performance and is comparable with appearance-based ones. The experimental results show the proposed 3D pose based method has unique advantages in large view variation. It will have great potential with the development of pose estimation in future.

Keywords: Gait recognition, 3D pose, LSTM, CNNs

1 Introduction

Gait as a kind of biometric feature has a great potential for human identification at a distance. Compared with other kinds of biometric features such as fingerprint, iris, palmprint and face, gait has unique advantages like non-contact, hard to fake. Therefore, gait recognition has attracted more and more attention in the computer vision field. Although many creative works have been proposed on gait recognition, it is still a challenge task due to view variation, clothing occlusion, carrying bags which could reduce the recognition rate drastically.

There are mainly two kinds of methods for gait recognition: the appearance-based methods and the model-based ones. The appearance based methods [20, 23, 19, 16] usually extract appearance features from human silhouettes. The appearance based methods were popular in the pass decades for the efficiency of feature extraction. However, this kind of methods are easily affected by shape changes like clothing and carrying bags. The recognition accuracy could also drop rapidly when evaluated under clothing, carrying conditions. Another category of methods is based on human models which employ modelling human body and local movement patterns of different body parts. Many model-based methods [14, 13, 11, 22] employ static structures of body and motion. It is evident that the model-based methods can be insensitive to occlusions, clothing changing and some other variations. But it was challenging to build an accurate human model in the past. It mainly relies on markers attached on human bodies or using special sensors to track body joints.

With the development of the pose estimation which can directly extract human pose from images, gait recognition also benefited from that. There are some pose-based gait recognition methods in the literature [8, 3, 9]. It can be easily understood that human joints are insensitive under the carrying bags and clothing conditions if the joints can be estimated accurately. Some pioneer researchers have worked on gait recognition based one human pose. Liang *et al.* [8] use skeleton data acquired from the Kinect sensors. Feng *et al.* [3] use the human body joint heatmap as the feature for gait recognition. They feed the joint heatmap of consecutive frames to Long Short Term Memory (LSTM) to extract the gait features. Our prior work [9] proposed a 2D pose-based gait recognition method and used the temporal-spatial network (PTSN) to extract the gait feature. Different from the method in [9], 3D pose feature is used in the proposed method. Experimental results also show that the 3D pose feature is superior to 2D feature.

Our contributions in this paper are: (1) The 3D pose is estimated directly from 2D images from one camera only, and camera calibration and special senors and markers are not needed. (2) LSTM and CNNs are combined to capture both temporal and spatial information from consecutive 3D pose. (3) Only one uniform 3D pose model is needed which can handle view, carrying and clothing variations.

The rest of the paper is organized as follows. Section 2 describes the proposed 3D pose model. Experiments and evaluation are presented in Section 3. The section 4 conclude the conclusions.

2 3D Pose Feature Extraction

3D human pose contains more information than 2D [9]. Compared with 2D pose [9], the pose information of our proposed method is in the three dimension, which is definitely beneficial to dealing with view problem. In addition, we use the center loss rather than contrastive loss to constrain the gait feature, which can reduce the complexity in the training process and improve the performance of gait recognition. It is inherently view invariant because it is in a 3D space. Given the 3D human model, the feature at any view can be synthesized from the 3D model. The proposed method employs the 3D pose information estimated from 2D images by CNN. It is inspired by the facial expression recognition

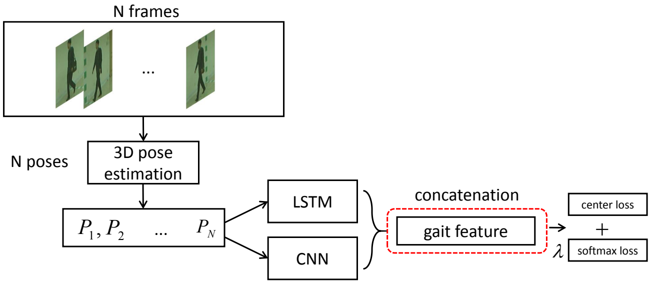


Fig. 1: The framework of the proposed method.

method in [21]. We extract temporal features based on 3D pose from consecutive frames by LSTM, and spatial features by CNN. A multiple loss strategy is employed to enhance the gait feature extraction and improve recognition rates. The framework of the proposed method is shown in Fig. 1.

2.1 3D Pose Estimation

Estimating a high accuracy 3D pose is a challenge because it can be cast as a non-linear optimization problem [7]. Recently, Chen *et al.* [2] explore 3D human pose estimation from single RGB image and it is straightforward to implement with off-the-shelf 2D pose estimation systems and 3D mocap libraries. It outperforms almost all state-of-the-art 3D pose estimation system, so we use it to obtain the gait pose which contains 14 joints. The 14 joints are Nose, Neck, Right Shoulder, Right Elbow, Right Wrist, Left Shoulder, Left Elbow, Left Wrist, Right Hip, Right Knee, Right Ankle, Left Hip, Left Knee, Left Ankle, Right Eye, Left Eye, Right Ear and Left Ear. Some gait RGB images and the correspondent 3D pose are show in Fig. 2.

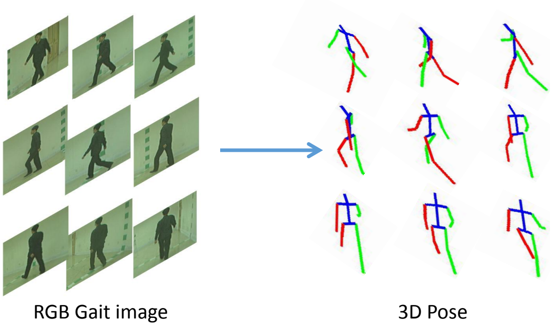


Fig. 2: Some gait RGB images and the correspondent 3D pose estimated from the RGB images.

2.2 The Feature Learning

In the proposed architecture there are two networks. They are LSTM and CNNs respectively. The size of the input data is $N \times 42$ where N stands for N consecutive frames selected from a video and 17 body joints (each joint has its position with (x, y, z). The consecutive poses can be considered as dynamic variation so we implement LSTM as a temporal network to extract dynamic features from the consecutive poses. Another network based on CNNs is constructed to extract features from still poses. The dynamic features extracted by LSTM and the spatial features by the CNNs are finally concatenated as our gait feature to improve the recognition.

LSTM for temporal feature: Since that gait is the walking style and different person has different gait. Gait can also be regarded as the dynamic motion of different body joints in the temporal space. LSTM is a network which is good at extracting features in the temporal domain. It contains self-connected memory units. It can improve long range contextual information of in the temporal domain. So it is effective in capturing dynamic information. We put the joint positions into a LSTM network to extract the temporal feature.

CNNs for spatial feature: A CNNs model is also designed in the proposed method to extract the spatial information. we want to emphasis here that the input of CNNs is the pose data which is the same with the one to LSTM. For most CNNs based gait recognition methods, the input is 2D images which is the appearance data. The input is a global representation of a gait sequence. As illustrated in Table 2, we implemented ResNet [4] which add shortcut and can help to extract deep global information. Then, the LSTM and CNNs are fused to extract temporal and spatial gait features in a gait sequence. At last, the temporal gait feature and the spatial one are concatenated as the feature.

2.3 Loss Functions

After the gait feature extraction by LSTM and CNNs, a multi-loss strategy is involved to improve the recognition rate. For gait features, it is a great challenge that intra-class is larger than inter-class sometimes. The softmax loss [12] can help to enlarge the inter-class distance, and the center loss [15] is good at reducing the intra-class distance. So the softmax loss and the center loss are fused to boost our network.

Softmax loss Our networks can learn discriminant features under the supervision of the gait labels. The softmax loss could classify each gait pose into the correspondent subject, and it also effectively to enlarge the inter-class distance. It is defined as:

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

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.

Center Loss In gait recognition many challenges such as the view variation can cause the recognition rate to drop drastically because the intra-class distance is mostly greater than the inter one. The center loss is effective to reduce the intra-class distance. It is defined as:

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

where $c_{y_i} \in \mathbb{R}^d$ is the y_ith class center of pose features. When the distance between the pose and its correspondent center is large, it adds penalty so that the intra-class can be reduced.

Fusion of loss functions To enlarge the inter-class distance and reduce the intra-class one, the softmax lass and the center loss are fused. They are fused as follows.

$$L = L_S + \gamma L_c \tag{3}$$

where γ is to balance the weight of two loss functions, and in our experiment the γ is set to value 0.005.

3 Experimental Results and Analysis

3.1 Dataset

CASIA-B gait dataset [18] is one of the largest public gait databases in this world, and it contains 124 subjects captured from 11 views with the view range from 0° to 180° with 18° interval between two nearest views. The set of view angles are $\{0°, 18°, \cdots, 180°\}$. 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). The CASIA-B dataset consists 13640 video sequences and with 2 or 3 gait cycles in each sequence.

3.2 Implementation Details

The experimental setting of the proposed method is the same with those in [9]. All the gait data including "nm", "bg" and "cl" are all involved. The first 62 subjects are put into the training set and the remaining 62 ones into the test set. In the test set, the first 4 normal walking sequences of each subjects are put into the gallery set and the others into the probe set as shown in Table 1.

According to our framework in Fig 1, the 3D pose is estimated from images using the method in [2]. The 3D pose contains 14 joints. The height of the

Table 1: Experimental setting on CASIA-B dataset.

Training	Test	
	Gallery Set	Probe Set
ID: 001-062	ID: 063-124	ID: 063-124
NM01-NM06	SNM01-NM04	NM05-NM06
BG01-BG02, CL01-CL02		BG01-BG02, CL01-CL02

subjects in the images is not fixed because the distance between the subjects and the camera is not fixed. So the human pose is normalized to a fixed size. To be specifically, it is that the distance between the neck and the hip is normalized to a fixed size.

The 3D pose joint data of train set are fed into the networks. The details of our networks involving CNNs and LSTM are shown in Table 2 and Table 3 respectively.

Table 2: Implementation details of the CNN.

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	N	2× 2	2	N
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	N	2× 2	2	N
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	128	3× 3	1	P-ReLU
FC.1	512	N	N	N

3.3 Impact of Temporal Network

The proposed method combines the LSTM and CNNs to extract temporal and spatial features respectively. To evaluate the efficiency of the LSTM, experiments are carried out with only CNNs with exactly the same train and set settings. We compute the average recognition rates under the view variation, carrying and clothing conditions. From the results shown in Fig. 3, we can find that the proposed method outperforms the method with CNNs only. It shows the efficiency of the temporal information by LSTM.

Table 3: Implementation details of LSTM

Layers	Activation function
FC	ReLU
FC	ReLU
FC	ReLU
FC	ReLU
LSTM	N

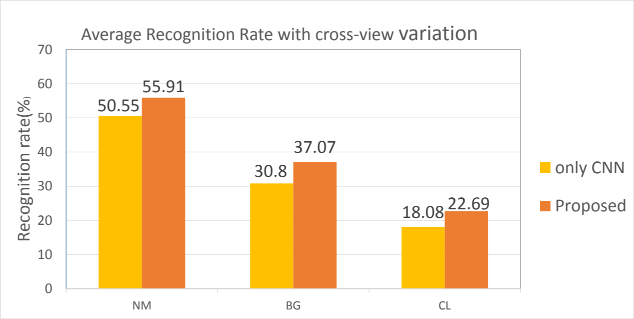


Fig. 3: The cross-view average recognition comparison between proposed and only using CNNs model on CASIA-B dataset.

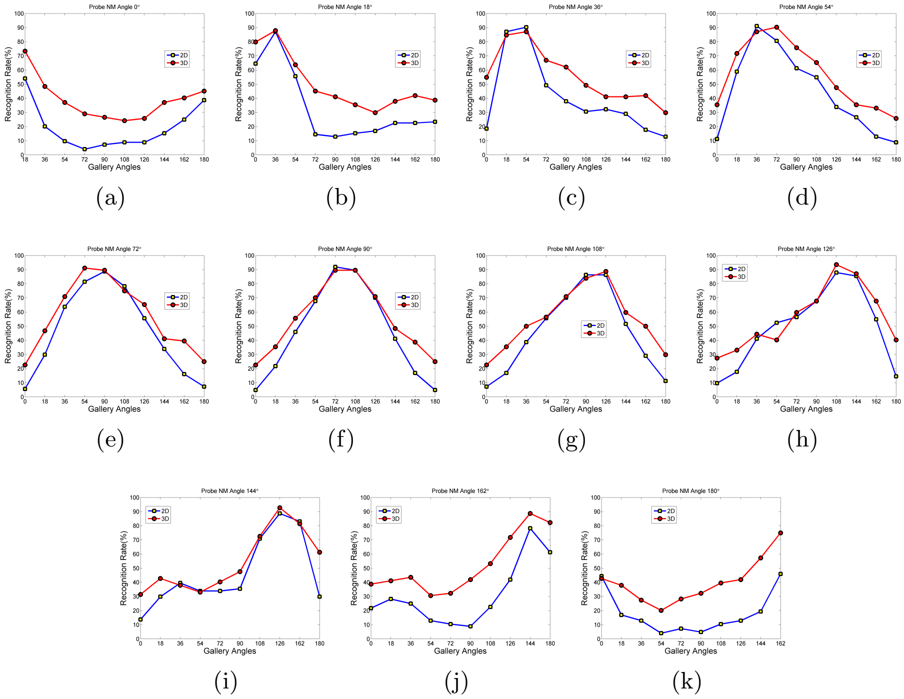
3.4 Comparisons with 2D Pose

Our prior work in [9], named as PTSN, is a 2D pose based gait recognition method. Different from the 3D joint positions extracted from images, it is only the 2D positions used in [9]. The proposed method is compared with PTSN under the cross-view variations to evaluate that the 3D pose is more robust to view variation. The experimental design of the proposed method is the same with that of PTSN as shown in Table 1. Fig. 4 shows the recognition rates of PTSN and the proposed method at each probe angle. It is clearly shown that the proposed can achieve much better results especially when there is a larger view variation. That means the proposed method is more robust to view variation.

3.5 Comparison with other Cross-view Methods

We compared the proposed method with some other sate-of-the-art works. They are FD-VTM [10], RSVD-VTM [5], RPCA-VTM [23], R-VTM [6], GP+CCA [1], C3A [17] and PTSN [9]. For the limitation of space, we only selected the results of 54°, 90°, 126° probe angles. It is the same setting with that in [9]. The experimental results are shown in Fig. 5.

We want to emphasize here that only the positions of 14 joint are taken as the input. No other kinds of appearance based features are sent into the networks. From the results, we can find that the proposed method performs well in large view variation especially. The results also show that the proposed 3D model owns advantages in handling cross-view condition.



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k)

Fig. 4: Comparison between the 2D pose method PTSN [9] (blue lines) and the proposed method (red lines).

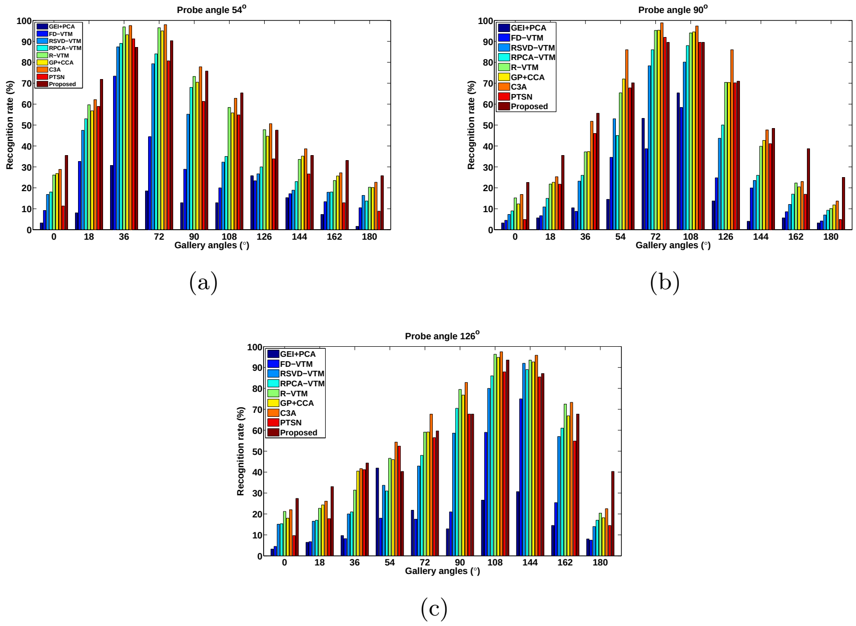
4 Conclusions and Future Work

In this paper, we proposed a gait recognition method based on 3D body pose to handle the cross-view variations. A 3D pose estimation method based on CNN is used to estimate the position of human body joints. Then the positions in a sequence can be sent to neural networks to train the networks. Since 3D pose is used in the proposed method, the proposed method is more robust to view variation and others. Experimental results also prove that. Even only the joint positions are used for recognition, state-of-the-art recognition rates are achieved.

Human pose estimation is just improved greatly in these several years with the progress of deep learning. We surely believe that the pose estimation will achieve better performance in future. The work in the paper shows that 3D pose can benefit gait recognition a lot. Gait recognition will be continually benefited by the development of human pose estimation, human body modeling and related topics.

Acknowledgment

The work is supported by the strategic new and future industrial development fund of Shenzhen (Grant No. 20170504160426188).



(a)

(b)

(c)

Fig. 5: Comparing with existing methods at probe angles (a)54°, (b)90° and (c)126° on CASIA-B dataset. The gallery angles are the rest 10 angles except the corresponding probe angle.

References

1. K. Bashir, T. Xiang, and S. Gong. Cross view gait recognition using correlation strength. In *BMVC*, pages 1–11, 2010.
2. C. H. Chen and D. Ramanan. 3D human pose estimation = 2D pose estimation + matching. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7035–7043, 2017.
3. Y. Feng, Y. Li, and J. Luo. Learning effective gait features using lstm. In *International Conference on Pattern Recognition (ICPR)*, pages 325–330, 2017.
4. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.
5. W. Kusakunniran, Q. Wu, H. Li, and J. Zhang. Multiple views gait recognition using view transformation model based on optimized gait energy image. In *IEEE International Conference on Computer Vision Workshops*, pages 1058–1064, 2010.
6. W. Kusakunniran, Q. Wu, J. Zhang, and H. Li. Gait recognition under various viewing angles based on correlated motion regression. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(6):966–980, 2012.
7. B. Kwolek, T. Krzeszowski, A. Michalczuk, and H. Josinski. *3D Gait Recognition Using Spatio-Temporal Motion Descriptors*. Springer International Publishing, 2014.

8. Y. Liang, C. T. Li, Y. Guan, and Y. Hu. Gait recognition based on the golden ratio. *Eurasip Journal on Image and Video Processing*, 2016(1):22, 2016.
9. R. Liao, C. Cao, E. B. Garcia, S. Yu, and Y. Huang. Pose-based temporal-spatial network (ptsn) for gait recognition with carrying and clothing variations. In *the 12th Chinese Conference on Biometric Recognition (CCBR)*, pages 474–483, 2017.
10. Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi. Gait recognition using a view transformation model in the frequency domain. *Computer Vision–ECCV 2006*, pages 151–163, 2006.
11. J. M. Nash, J. N. Carter, and M. S. Nixon. Dynamic feature extraction via the velocity hough transform. *Pattern Recognition Letters*, 18(10):1035–1047, 1997.
12. Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1891–1898, 2014.
13. R. Tanawongsuwan and A. Bobick. Gait recognition from time-normalized joint-angle trajectories in the walking plane. In *IEEE Conference on Computer Vision and Pattern Recognition*, page 726, 2001.
14. L. Wang, T. Tan, W. Hu, and H. Ning. Automatic gait recognition based on statistical shape analysis. *IEEE Transactions on Image Processing*, 12(9):1120–1131, 2003.
15. Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515, 2016.
16. Z. Wu, Y. Huang, L. Wang, X. Wang, and 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):209–226, 2016.
17. X. Xing, K. Wang, T. Yan, and Z. Lv. Complete canonical correlation analysis with application to multi-view gait recognition. *Pattern Recognition*, 50(C):107–117, 2016.
18. S. Yu, D. Tan, and T. Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *International Conference on Pattern Recognition (ICPR)*, pages 441–444, 2006.
19. S. Yu, L. Wang, W. Hu, and T. Tan. Gait analysis for human identification in frequency domain. In *International Conference on Image and Graphics*, pages 282–285, 2004.
20. S. Yu, Q. Wang, L. Shen, and Y. Huang. View invariant gait recognition using only one uniform model. In *23rd International Conference on Pattern Recognition (ICPR2016)*, pages 889–894, 2016.
21. K. Zhang, Y. Huang, Y. Du, et al. Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Transactions on Image Processing*, 26(9):4193–4203, 2017.
22. G. Zhao, G. Liu, H. Li, and M. Pietikainen. 3d gait recognition using multiple cameras. In *International Conference on Automatic Face and Gesture Recognition*, pages 529–534, 2006.
23. S. Zheng, J. Zhang, K. Huang, R. He, and T. Tan. Robust view transformation model for gait recognition. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pages 2073–2076. IEEE, 2011.