Pose-based temporal-spatial network (PTSN) for gait recognition with carrying and clothing variations

Rijun Liao[1], Chunshui Cao[2], Edel B. Garcia[3], Shiqi Yu[1], and Yongzhen Huang[2]

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. Advanced Technologies Application Center (CENATAV),
7ma A 21406, Playa, Havana, Cuba.
2150230306@email.szu.edu.cn
ccs@mail.ustc.edu.cn
egarcia@cenatav.co.cu
shiqi.yu@szu.edu.cn
yongzhen.huang@nlpr.ia.ac.cn

Abstract. One of the most attractive biometric techniques is gait recognition, since its potential for human identification at a distance. But gait recognition is still challenging in real applications due to the effect of many variations on the appearance and shape. Usually, appearance-based methods need to compute gait energy image (GEI) which is extracted from the human silhouettes. GEI is an image that is obtained by averaging the silhouettes and as result the temporal information is removed. The body joints are invariant to changing clothing and carrying conditions. We propose a novel pose-based gait recognition approach that is more robust to the clothing and carrying variations. At the same time, a pose-based temporal-spatial network (PTSN) is proposed to extract the temporal-spatial features, which effectively improve the performance of gait recognition. Experiments evaluated on the challenging CASIA B dataset, show that our method achieves state-of-the-art performance in both carrying and clothing conditions.

Keywords: Gait recognition, Pose-based, PTSN network

1 Introduction

Gait is a kind of behavioral biometric feature, that is suitable for human identification at a distance. In consequence, gait recognition technology has attracted increasing attention in video surveillance.

There have been mainly two categories of gait approaches with different highlights. The first one is model-based methods [3] which employ modelling of human body structure and local movement patterns of different body parts.

The second category of gait approach is appearance-based methods [4, 5] which directly extract gait representations from videos. Gait energy image (GEI) is the feature most applied, because of its good compromise between recognition rate and simplicity of computation. However, there are different cons concerning the use of human silhouettes, first, in wild conditions the extraction is affected by illumination changes and many silhouettes appear incomplete. Second, even when the extraction step is performed correctly, the shape depends on the view angles, clothes variations and the carring conditions.

Some authors faced this problem removing the parts of silhuoettes affected by variations and retain only those uninfluenced parts to eliminate the effects of clothing and carried objects. But, recognition rates are not good enough. In order to handle with the clothing and carrying variations, Huang et al. [13] increase robustness to some classes of structural variations by fusing Shifted Energy Image and the Gait Structural Profile. In [12], Hossain et al. analyze the discrimination capability of different parts through dividing the human body into eights parts. Yu et al. [21] employ the Stacked Progressive Auto-Encoders (SPAE) trying to transform the clothing and carrying conditions into normal walking. In [1] the authors propose a novel covariate cognizant framework to deal with the presence of such clothes and carring covariates. They describe gait motion by forming a single 2D spatio-temporal template from video sequence. Guan et al. [10] proposed a random subspace method (RSM) framework for clothing-invariant gait recognition by combining multiple inductive biases for classification. In Liang et al. [16] the golden ratio takes the characteristics of clothing into consideration, enabling all the clothing parts to be discarded and the unaffected parts of the gait to be retained. Das et al. [7] introduced the use of rotation forest ensemble classifier in gait recognition, and experimentally demonstrates its superiority to random subspace method in this field.

Some researchers have studied the problem as a pose-based gait recognition, for example [15] uses skeleton data provided by the low-cost Kinect sensors. In [9] instead of using binary silhouette to describe each frame, they use the human body joint heatmap. They feed the joint heatmap of consecutive frames to Long Short Term Memory (LSTM). The hidden activation values at the last timestep is used as their gait feature.

Our approach is based on early studies on gait perception that showed that joints' motion over time is sufficient for humans to identify familiar persons. Until now, only structural feature was not enough to human identification problem in gait analysis, since pose estimation requires accurate tracking of body parts, which is known to be a very challenging problem considering the nonrigidness and self-occlusion of the human body. However, a recently bottom-up method [6] for pose estimation using deep learning opens the door to retake approaches based on dynamic parameters. We believe the body joint is invariant to changing clothing and changing carrying conditions. Our contribution in this paper is a pose based temporal-spatial network that combines a LSTM and Convolutional Neural Network (CNN) to capture the dynamic and static information of a gait sequence. Our method is robust to the clothing variations and carring conditions.

2 Our method

In this paper, a novel pose-based gait recognition approach is proposed to deal with clothing and carrying condition variations. The work-flow of proposed method is illustrated in Fig. 1. The first step is to estimate the pose information from the given consecutive frames. Then, the pose coordinate sequences is extracted and preprocessed. Finally, a pose-based temporal-spatial network (PTSN) is proposed to extract the temporal features and spatial features from gait pose rather than image, which effectively improve the performance. In this section, we will illustrate our method in detail.

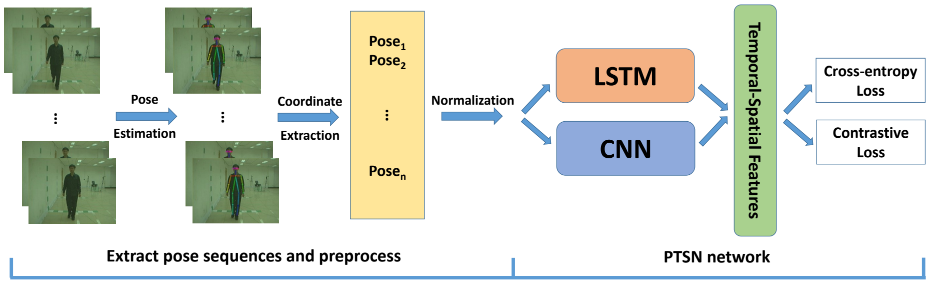

Fig. 1. Work-flow of our pose-based gait recognition approach.

2.1 Pose information

The proposed method employs the pose information to extract the invariant feature for clothing and carrying conditions. For the appearance-based methods, one common pipeline is to evaluate the similarities between pairs of gait energy image (GEI). However, the GEI would be greatly changed by the clothing and carrying condition variations which directly lead to decrease the recognition rate. Besides, the GEI is computed by averaging the silhouettes, which will eliminate the temporal information in the process of walking. In contrast, the human pose is less affected by these variations due to it does not depend on human body appearance and shape. In addition, gait is a process of movements, the pose sequences has powerful representation capacity to capture the invariant features from consecutive frames. Consequently, the invariant features that are robust to clothing and carrying conditions, are extracted from the pose sequences rather than human shape.

We use a pre-trained model of multi-person 2D pose estimation [6] to acquire the human pose. Cao et at. propose the Part Affinity Fields which directly estimate the association between anatomical parts. The pre-trained model can estimate 18 joints, namely Nose, Neck, RShoulder, RElbow, RWrist, LShoulder, LElbow, LWrist, RHip, RKnee, Rankle, LHip, LKnee, LAnkle, Reye, LEye, Rear and Lear, as shown in Fig. 2. Before we use the pose information, we should normalize and select the effective joints in order to extract more robust feature.

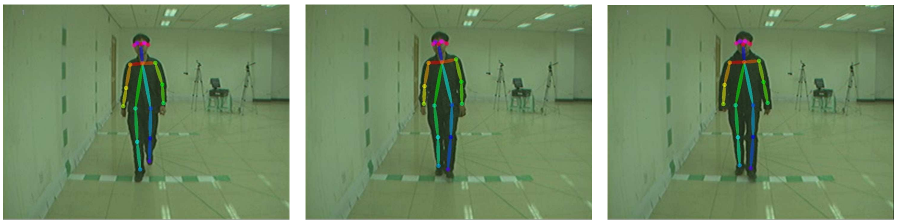

Fig. 2. Normal walking, walking with a bag, and walking with a coat sequences from CASIA B dataset: 18 human joints are shown.

Normalization: The distance between people and camera will change at all the time when people walk through the fixed camera. In order to avoid the influence of this distance change, each joint needs to be normalized. In the process of people walking, the Neck and the center of Hip are two relatively more stable joints than others. As a result, the normalization should be based on that two joints. The equation of normalization is defined as follows:

$$P_i' = \frac{P_i - P_{neck}}{H_{nh}} \tag{1}$$

where $P_i \in \mathbb{R}^2$ be the coordinate of body joint i, P_i' be the normalized coordinate of P_i, P_{neck} is the Neck coordinate, the H_{nh} is the height between the Neck position and the center of Hip position.

Selection of Effective Joints: One of the most important features is the change of human leg movement. Cunado et al. [8] used the legs as a model, as they found harmonics from the motion of legs. In addition, from the Fig. 2, we can find the width of shoulder in the walking with a coat is little bigger than the normal walking and the walking with a bag. Therefore, not all of the joints can effectively boost the performance of gait recognition, and even some joints will perform worse. As the Neck already was used as a base point for normalization, we do not have to choose Neck as an effective joint. Consequently, we choose the RHip, RKnee, Rankle, LHip, LKnee, LAnkle as the effective joints of gait feature. These six effective joints not only have rich representation capacity for gait recognition, but also more robust to the clothing and carrying condition variations than other joints.

2.2 PTSN for Gait Temporal-Spatial Features

We borrow the idea of Deep Evolutional Spatial-Temporal Networks [23], and propose a pose-based temporal-spatial network (PTSN) to capture the dynamic and static information of gait pose. The proposed PSTN mainly consists of two kinds of networks, as shown in Fig. 1. Firstly, we use the Long Short-Term Memory (LSTM) [11] to extract the temporal features from gait pose sequences. Secondly, the Convolutional Neural Network (CNN) is used to extract the spatial features from static gait pose frames. Finally, the two types of features are combined to capture the dynamic-static information of gait pose, which has powerful representation capacity to extract invariant features from different gaits.

LSTM for Temporal Features: As gait is a process with a series of movements, it is natural to consider to extract the dynamic information from the walking sequence. Simonyan et al. [19] trained an additional network on top of optical flow in order to capture temporal information under the framework of CNN. Although CNN can achieve state-of-the-art performance on image classification tasks, it has not yet been shown to be effective in capturing dynamic information. In contrast, the Long Short-Term Memory is supposed to better handle with temporal sequences. Therefore, we employ the LSTM to extract the temporal features from consecutive pose frames.

CNN for Spatial Features: The LSTM can effectively extract the dynamic information, but it has not enough capacity to extract the static information of gait, such as the length between Ankle and Knee. In order to complement the information of static appearance, Zhang et al. [23] proposed a multi-signal convolutional neural network (MSCNN) to extract spatial features from static frames. Unlike the MSCNN, we fuse CNN with LSTM in the top fully convolutional layer, which effectively boost the performance of gait recognition.

2.3 Definition of Loss Function

In order to extract the temporal-spatial features with large between-gait variations and reduce the within-gait variations, we adopt a multi-loss strategy to optimize the PTSN network. The Cross-entropy Loss classifies each gait sequence into different gaits, and the Contrastive Loss constrains the relationship between the temporal-spatial features.

Cross-entropy Loss: In the task of recognition, many researchers [23] use the recognition signal as supervision. Because of features have to be classified into different classes, so the Cross-entropy Loss is useful to pull apart the temporal-spatial features of different gaits. The Cross-entropy Loss can promote the temporal-spatial features with large between-gait variations, it is defined as:

$$CELoss = -\sum_{i} y_i \log(p_i) \tag{2}$$

where y_i is the true distribution of sample i, and p_i is the predicted probability of gaits.

Contrastive Loss: The Cross-entropy Loss can push temporal-spatial features apart, but it has not a strong capacity to reduce the variations of identical human gaits. Many researchers employ another loss function to constrain the feature, such as Zhang et al. [23] use the VeLoss and Wen et al. [20] adopt the center loss. In order to extract powerful features, we adopt an additional Contrastive Loss, which is not only helpful to enlarge the between-gait variations, but also can reduce the within-gait variations. The Contrastive Loss is defined as:

$$CTLoss = \frac{1}{2}y\|f_i - f_j\|_2^2 + \frac{1}{2}(1-y)max(\lambda - \|f_i - f_j\|_2^2, 0) \tag{3}$$

where f_i and f_j are features of two input sequences. $y = 1$ when the two input sequences are from the same human gait, then the f_i and f_j will to be close. $y = 0$ means that the two input sequences are from different human gaits. In this case, the distance of f_i and f_j is limited to be larger than margin λ.

3 Experiments and Analysis

3.1 Experimental setting

To evaluate the performance of the proposed pose-based gait recognition approach, several experiments are performed on the challenging CASIA B gait dataset [22]. CASIA B dataset is one of the largest public gait databases. It has 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). The three kinds of sequences as shown in Fig. 2. In these 10 sequences, each sequence has 11 views which were captured from 11 cameras, the view angle set of camera is $\{0°,18°,\cdots, 180°\}$. Like the experimental setting of SPAE [21] and GaitGAN [18], we also set the first 62 subjects as the training set and the rest of subjects as the test set. 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 be shown in Table 1.

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

3.2 Experimental results on CASIA B dataset

Our experimental results on test set of CASIA B dataset are shown in Table 2. The gallery set of Table 2 is the first 4 normal walking sequences at a specific view, the probe sets has three types which are the last 2 normal sequences, 2 carrying bags sequences and 2 with coats sequences, respectively. In the tables, each column represents a view of gallery set and probe set.

Table 2. The recognition rate for 11 single views on CASIA B dataset.

View	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
Probe NM 5-6	96.77	99.19	98.39	98.39	94.35	96.77	95.97	95.97	96.77	98.39	95.16	96.92
Probe BG 1-2	89.52	95.16	92.74	87.90	83.87	79.03	84.68	83.06	83.06	90.32	74.19	85.78
Probe CL 1-2	53.23	83.87	87.90	72.58	61.29	61.29	75.00	66.94	70.97	70.16	45.97	68.11

3.3 Comparisons with GEI+PCA, SPAE and GaitGAN

We compare the average recognition rates without view variation with GEI+PCA [17], SPAE [21] and GaitGAN [18], as is shown in Fig. 3. The average recognition rates without view variation are computed by averaging the recognition rates on the 11 single views. The corresponding values for GEI+PCA, SPAE and GaitGAN are obtained in the same way. In normal walking condition, our method achieves comparable performance with GEI+PCA, SPAE and GaitGAN. In walking with carrying condition, the proposed method outperforms these three methods greatly, its recognition rate is higher than the best result by 13%. For walking with clothing condition, our method achieves a high average recognition rate of 68.11%, which exceeds the best result by more than 22%. The comparison shows that our method can effectively handle with carrying and clothing condition variations.

3.4 Comparisons with state-of-the-art

For further illustrate the performance of our method, we also compare the proposed method with state-of-art methods. Including Shanableh et al. [2], Huang et al. [13] and Jeevan et al. [14] which are all appearance-based methods for the 90° view. Since our method does not adopt the fusion scheme, we only compare the nine single-level methods (R1-R9) of Shanableh et al. In addition, we want to emphasize that our method contains only one model to handle with any single view. The 90° view is the best angle for appearance-based methods because of captures more dynamic information, but not for our pose-based method due to many joints are invisible in 90° view. So we use both 90° and 36° views to compare with these methods, the result is listed in Fig. 4. The comparison shows that proposed method achieves comparable performance with state-of-the-art in

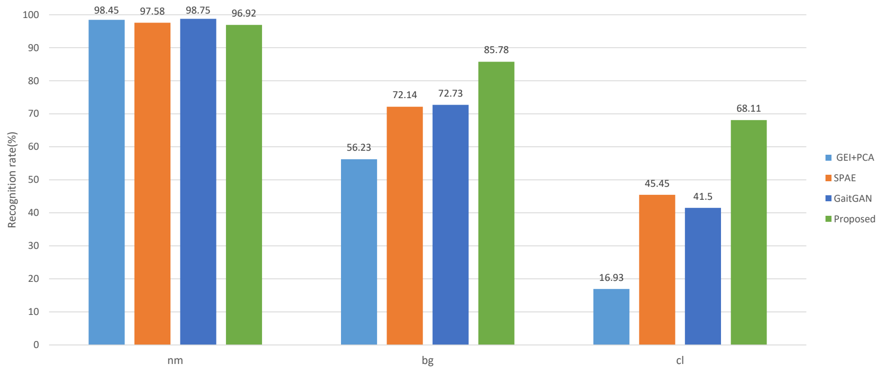

Fig. 3. The average recognition rates compared with GEI+PCA, SPAE and GaitGAN.

normal walking, better than many methods in carrying and clothing conditions. Besides, the comparison of average recognition rate of NM, BG and CL shows that our method achieves good performance, especially for the 36° view.

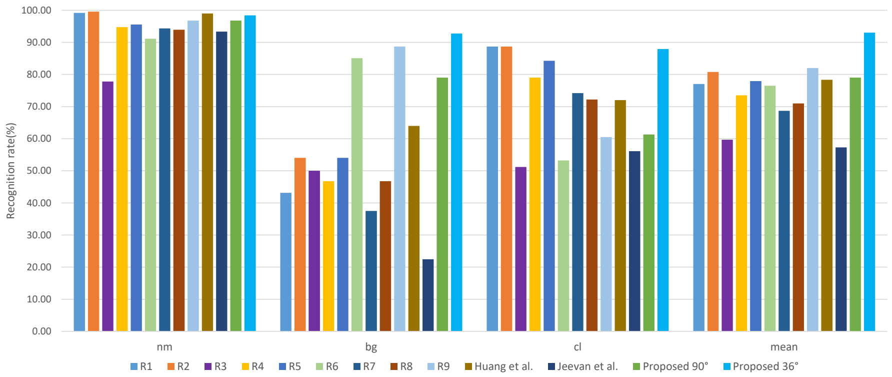

Fig. 4. Comparing with state-of-the-art methods.

4 Conclusions and Future Work

In this paper, we proposed a novel pose-based gait recognition approach to handle with clothing and carrying condition variations. In order to extract the dynamic and static information for gait poses from a sequence of frames, a pose-based temporal-spatial network (PTSN) is proposed which can greatly boost the performance. Experimental results show that our method can improve recognition rate greatly especially for the clothing condition, and achieve state-of-the-art performance.

In the future, we will extend this method to handle with other challenging variations, such as view condition. The view variation is an important challenging in gait recognition. The pose-based gait recognition approach has greatly potential to deal with all variations in gait recognition.

Acknowledgments. The work is supported by the Science Foundation of Shenzhen (Grant No. JCYJ20150324141711699).

References

1. H. Aggarwal and D. K. Vishwakarma. Covariate conscious approach for gait recognition based upon zernike moment invariants. *CoRR*, abs/1611.06683, 2016.
2. A. Al-Tayyan, K. Assaleh, and T. Shanableh. Decision-level fusion for single-view gait recognition with various carrying and clothing conditions. *Image and Vision Computing*, 61:54 – 69, 2017.
3. G. Ariyanto and M. S. Nixon. Model-based 3d gait biometrics. In *Biometrics (IJCB), 2011 International Joint Conference on*, pages 1–7. IEEE, 2011.
4. X. Ben, W. Meng, R. Yan, and K. Wang. An improved biometrics technique based on metric learning approach. *Neurocomputing*, 97:44–51, 2012.
5. X. Ben, P. Zhang, W. Meng, R. Yan, M. Yang, W. Liu, and H. Zhang. On the distance metric learning between cross-domain gaits. *Neurocomputing*, 208:153–164, 2016.
6. Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1611.08050*, 2016.
7. S. D. Choudhury and T. Tjahjadi. Clothing and carrying condition invariant gait recognition based on rotation forest. *Pattern Recognition Letters*, 80:1–7, 2016.
8. D. Cunado, M. Nixon, and J. Carter. Using gait as a biometric, via phase-weighted magnitude spectra. In *Audio-and Video-based Biometric Person Authentication*, pages 93–102. Springer, 1997.
9. Y. Feng, Y. Li, and J. Luo. Learning effective gait features using LSTM. In *23rd International Conference on Pattern Recognition, ICPR 2016, Cancún, Mexico, December 4-8, 2016*, pages 325–330, 2016.
10. Y. Guan, C. Li, and Y. Hu. Robust clothing-invariant gait recognition. In *Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IIH-MSP 2012, Piraeus-Athens, Greece, July 18-20, 2012*, pages 321–324, 2012.
11. S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
12. M. A. Hossain, Y. Makihara, J. Wang, and Y. Yagi. Clothing-invariant gait identification using part-based clothing categorization and adaptive weight control. *Pattern Recognition*, 43(6):2281–2291, 2010.
13. X. Huang and N. V. Boulgouris. Gait recognition with shifted energy image and structural feature extraction. *IEEE Transactions on Image Processing*, 21(4):2256–2268, 2012.
14. M. Jeevan, N. Jain, M. Hanmandlu, and G. Chetty. Gait recognition based on gait pal and pal entropy image. In *Image Processing (ICIP), 2013 20th IEEE International Conference on*, pages 4195–4199. IEEE, 2013.

15. D. Kastaniotis, I. Theodorakopoulos, and S. Fotopoulos. Pose-based gait recognition with local gradient descriptors and hierarchically aggregated residuals. *Journal of Electronic Imaging*, 25(6):063019, 2016.
16. Y. Liang, C. Li, Y. Guan, and Y. Hu. Gait recognition based on the golden ratio. *EURASIP J. Image and Video Processing*, 2016:22, 2016.
17. J. Man and B. Bhanu. Individual recognition using gait energy image. *IEEE transactions on pattern analysis and machine intelligence*, 28(2):316–322, 2006.
18. Y. Shiqi, C. Haifeng, G. R. Edel B., and P. Norman. Gaitgan: Invariant gait feature extraction using generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.
19. K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.
20. 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. Springer, 2016.
21. S. Yu, H. Chen, Q. Wang, L. Shen, and Y. Huang. Invariant feature extraction for gait recognition using only one uniform model. *Neurocomputing*, 239:81–93, 2017.
22. S. Yu, D. Tan, and T. Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 4, pages 441–444. IEEE, 2006.
23. K. Zhang, Y. Huang, Y. Du, et al. Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Transactions on Image Processing*, 2017.