Neurocomputing 395 (2020) 86–94

Contents lists available at ScienceDirect

Neurocomputing

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

Siamese denoising autoencoders for joints trajectories reconstruction and robust gait recognition

Weijie Sheng[a], Xinde Li[b,*]

[a] School of Automation, Key Laboratory of Measurement and Control of CSE Ministry of Education, Southeast University, Nanjing, China
[b] School of Cyber Science and Engineering, Southeast University, Nanjing, China

ARTICLE INFO

Article history:
Received 23 May 2019
Revised 16 January 2020
Accepted 27 January 2020
Available online 5 February 2020

Communicated by Dr Zhang Zhaoxiang

Keywords:
Gait recognition
Siamese denoising autoencoder
Joints trajectories reconstruction
Autoencoder with LSTM
Skeleton-based gait recognition

ABSTRACT

Dynamics of body skeletons convey significant information for human gait recognition. However, it is inevitable that missing points, overlapping error, or confusion of left and right error will frequently occur during the process of skeleton estimation. Existing skeleton-based methods have difficulty in achieving satisfactory performance in gait recognition since they treat the noisy data and the normal data equally to the recognition process. In this paper, we propose a novel skeleton-based model called Siamese Denoising Autoencoder networks (Siamese DAE), which can automatically learn to remove position noise, recover missing skeleton points and correct outliers in joint trajectories. More precisely, we construct an encoder that compresses the characteristics of input trajectories into a latent space and a decoder that attempts to reconstruct more accurate skeleton trajectories from the latent feature. The corrected joint trajectories not only lead to higher discriminative power but also stronger generalization capability. Moreover, we design a Siamese structure to reduce intra-class variations and increase inter-class variations of the encoded features. Experiments demonstrate that our method enhances the robustness against inaccurate skeleton estimation and achieves substantial improvements over mainstream skeleton-based methods for gait recognition.

1. Introduction

As a unique biometric feature that can be obtained from a distance, and without subjects' attention or cooperation [2], human gait has a vast application prospect for crime prevention and forensic identification. Previous research [9] has shown that human gait, precisely the walking pattern, cannot be easily imitated, or intentionally faked. To make walking pattern analysis more quantitative, researchers attempted to use various input data types such as human's silhouette [28], skeleton [17], depth map [30], optical flow [24], and dense trajectories [6]. These inputs usually require specific equipment and different pre-processing operations to obtain discriminative features. In our work, we select the skeleton as the input of our system since it is confirmed to be more robust to covariate conditions such as clothing, carrying, and occlusion.

The dynamic skeleton modality can be naturally represented by the kinematics of human body parts, in the form of joint trajectory information. We can acquire this information through sensors such as wearable accelerometers [25], motion capture [34], RGB-D cameras [23], and industrial-level lidar systems [11]. These systems provide accurate and rich information for gait analysis accompanied by high-cost burden of equipment. Some previous works use a low-cost phone camera to acquire acceleration data, but this method requires sensors to be held at the subject's feet which abdicates the non-invasion advantage of gait. In recent research, Kinect is widely used for skeleton-based gait recognition. Such RGB-D cameras provide not only a standard color image, but also the stereo information of a scene. However, the Kinect-based method is restricted to measurement distance. According to experimental investigation [23], the detection range of the depth camera is 0.5m~4.5m. As the distance increases, there will be more position errors compared to ground truth which is generated by a time-of-flight sensor [8]. Besides, heavy noise could appear on several frames in the process if a large object blocks a few joints or initial estimation fails.

The recent success of deep learning has lead to the surge of deep learning based pose estimation algorithms [5,10], by which we can also obtain skeleton and joint trajectories of human bodies without costly burden. Nevertheless, existing skeleton-based methods, especially using the pose estimation algorithm [41], have difficulty in achieving satisfactory performance. Because, in the estimation process, it is easy to introduce missing points and

* Corresponding author.
E-mail address: xindeli@seu.edu.cn (X. Li).

https://doi.org/10.1016/j.neucom.2020.01.098
0925-2312/

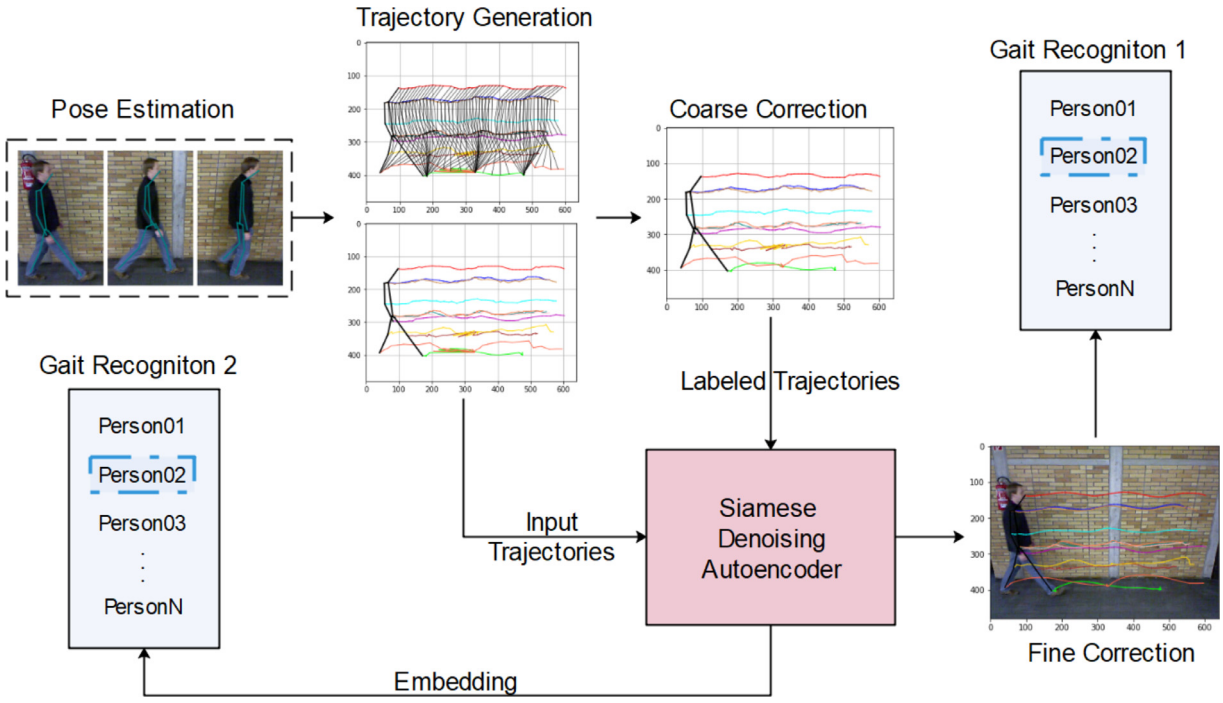

Fig. 1. The overall framework of the proposed gait recognition method. The red box indicates the Siamese DAE model. Gait Recognition 1 and Gait Recognition 2 respectively represent the trajectories based method and embedding vector based method described in Section 3.

outliers to the trajectories. These noisy frames are equally treated as characteristic frames, which seriously affect the expression of gait features. In this paper, we propose a generic representation by extending siamese networks to a dynamic sequence Denoising Autoencoder model (see Fig. 1), called Siamese Denoising Autoencoders Networks (Siamese DAE), to learn how gait features correlate temporally and spatially and reconstruct missing and error joints automatically through this model.

The major contributions of this work lie in three aspects:

- To our knowledge, this is the first work that emphasizes the importance of joints recovery to the skeleton-based gait recognition. We integrate the Siamese structure and the Denoising Autoencoder network into joints trajectories correction and gait recognition.
- We propose a novel idea for the Denoising Autoencoder network training, which allows our models to automatically correct the missing points and outliers in joints trajectories.
- We design two stages for trajectories correction: coarse correction and fine correction. The reconstructed trajectories and encoded features embeddings obtain superior discrimination power and achieve higher performance as compared to previous approaches on the challenging task of TUM GAID dataset.

The remaining of this paper is organized as follows: Section 2 reviews the related works and points out the severe disadvantages of conventional skeleton-based methods. Section 3 explains the implementation details of the proposed methodology and illustrate the Siamese Denoising Autoencoders network structure. Section 4 describes gait recognition datasets and represents the experimental results. Section 5 concludes the paper.

2. Related work

Current gait recognition studies can be classified into two categories: model-based methods and appearance-based methods. The model-based methods generally aim to fit a walking model by utilizing information of skeleton joints and then compute features based on the model. The appearance-based methods, however, extract features directly from the video imagery or silhouette images derived therefrom.

Conventional appearance-based approaches for gait recognition usually rely on silhouette feature such as Gait energy image (GEI) [16], which is computed by averaging of silhouette images of a walking person. This compact representation makes it easier to be discriminated. However, the major drawback of the appearance-based approach is that silhouettes are vulnerable to covariates such as occlusion, clothing, and carrying condition. There are mainly two kinds of strategies for alleviating the effects of covariate conditions. The first strategy is to discover more powerful hand-crafted features of gait, including frame difference energy image (FDEI) [7], active energy image (AEI) [39], depth gradient histogram energy image (DGHEI) [12] and Enhanced Gabor representation of the GEI [14]. The latest works extend these features to across-views gait recognition problems. Ben et al. [3] present a method called Coupled bilinear discriminant projection (CBDP) to overcome the problem of aligning gait energy images (GEIs) across views. Ben et al. [4] propose a general tensor representation framework with three criteria of tensorial coupled mappings to improve the representation of GEI features. The second strategy is utilizing deep learning methods to capture discriminant information. Notably, the deep convolutional neural network (CNN) has shown overwhelming advantages over classic hand-crafted approaches due to its deep and highly non-linear attribute. In [29], they feed in GEI as an input to a CNN network for gait recognition called GEINet. To reach the trade-off between spatial displacement caused by subject difference and view difference, Takemura et al. [33] uses the Siamese network with a pair of inputs and contrastive loss for gait recognition. Recently, various generative models have been used on gait recognition. Yu et al. [37] employs the Stacked Progressive Auto-Encoders(SPAE) to transform the clothing and carrying conditions into normal walking. Zha and Fan [36] proposes GaitGAN model based on generative adversarial networks (GAN), in order to generate invariant gain images that

are side-view images with normal clothing and without carrying bags.

Unlike the appearance-based methods, the model-based methods recognize gait by expressing movement as a set of joint position trajectories, rather than as a sequence of images. They are considered insensitive under the carrying and clothing conditions if the joints can be estimated accurately. These joints information can be acquired through additional sensors such as marker-less depth cameras (e.g., Kinect) [31], marker-based motion capture cameras (e.g., MoCap) [1]. Based on the reasons that gait recognition can be transformed from a spatiotemporal problem into the spatial domain, specifically the 2D image domain, Oktem et al. [27] utilize the content-based image retrieval (CBIR) techniques in their works. Inspired by action recognition, Kastaniotis et al. [18] achieves a satisfying identification rate by expressing the dynamic characteristics of human walking sequences efficiently. In [38], they use PCA algorithm to extract gait features and categorize these features to identify humans by SVM, and it achieves quite good performance.

The existing skeleton-based methods are limited in a real-world environment with inaccurate skeleton estimation since the noisy frames decrease the discriminative power of models. The human motion recovery algorithm plays a decisive role in recognition performance that has been deeply studied in previous works for a long history. With the smoothness constraint and the bone-length constraint which take the kinematic information into the recovery process, Xia et al. [35] recovers motions using sparse representations of incomplete frames and a learned dictionary through an optimization model. In [40], they establish a sparse optimization model and apply L_0 optimization to all point sets of different frames by integrating the spatiotemporal constraints.

Similar to our research, Nguyen et al. [26] employ a stack of encoder and decoder formed by LSTMs to estimate a weak gait index for a sequence of skeletons, which can be used to detecting abnormal walking gaits. To deal with clothing and carrying condition variations, Liao et al. [20] propose a pose-based temporal-spatial network (PTSN) to extract the temporal features and spatial features from gait pose. It effectively improves the performance of gait recognition on the CASIA B dataset. In the latest skeleton-based approach [8], skeleton quality for each frame is measured for constructing a quality-adjusted cost matrix between input frames and registered frames to prevent matching with noisy patterns. It enhances the robustness against inaccurate skeleton estimation results. Different from the usual way of applying the LSTM model for skeleton sequence data, Liao et al. [21] choose the Convolutional Neural Network (CNN), Ktena et al. [19] leverage the Graph Convolutional Network (GCN), Hu et al. [15] employ Hidden Markov Model (HMM) to model temporal data. And all these models achieve excellent performance. These works indicate that different basic models have their unique power for handling sequence data in recognition tasks.

3. Approach

3.1. Pipeline overview

The pipeline of our approach is illustrated in Fig. 1. We can obtain the skeleton-based data by pose estimation algorithms. Usually, the data is a sequence of joints coordinates in 2D space, which is extracted from consecutive frames. Based on the skeleton sequences, we concatenate joints coordinates to form continuous trajectories for corresponding time intervals. These original joints trajectories generally contain a lot of noises and missing points caused by inaccurate estimation or partial occlusion. We construct a Siamese Denoising Autoencoders network to reconstruct missing and error joints. The implementation of this network can be divided into two steps: coarse correction and fine correction. In the first step, Remarkable outlier such as missing points, overlapping points and confusion error of leg joints, can be revised by traditional motion recovery methods based on the temporal consistency knowledge. Linear interpolation [22] and Kalman filters [32] can effectively predict the position of outliers and correct error points with the available spatial-temporal information. Some inconspicuous outliers have to be modified with manual labeling before model training. In the second step, the coarsely-corrected trajectories will be treated as the label of the Siamese DAE model. Then, the labeled trajectories can automatically be further optimized by the network to obtain finely-corrected trajectories. Finally, not only the finely optimized trajectories can be used as the gait feature for recognition, but also the embedding vectors from the encoder can be discriminated for gait recognition.

3.2. Pose trajectory generation

Not all of the joints can effectively contribute to the performance of gait recognition, and some even do the opposite owing to inaccurately estimation. So a significant error will be brought if we select the ineffective joints. After comparing different 2D pose estimation methods, we choose the OpenPose estimation algorithm as our body detector. OpenPose is a real-time multi-person keypoint detection library for body, face, hands, and foot estimation. The most significant advantage to us is that it is invariant of running time to the number of detected people. Each skeleton is represented by a collection of 25 joint positions in a 2D space. We should select the practical joints to extract more robust feature. Concretely, many body joints such as eyes joints and ears joints are discarded from each input skeleton due to their lack of significant contribution in describing the posture for gait analysis.

In most side-view captured walking videos, the arm away from the camera is mostly occluded by the body. On the other hand, the arm near the camera can be accurately estimated. An accurate arm trajectory can fully express the feature of two moving arms during walking, since left and right arm trajectories often have fixed phase difference and share the same curved paths. In this way, we use the pre-trained OpenPose model to acquire the coordinates of 12 joints illustrated in Fig. 2, including the nose, neck, right or left shoulder, right or left elbow, right or left wrist, right hip, middle hip, left hip, right knee, left knee, right ankle, left ankle. Then, we concatenate joints coordinates to form continuous trajectories by the motion over time. Finally, we normalize the coordinates concerning the distance between a subject's hip and neck to discard the influence of variations in the distance of different subjects from the camera.

3.3. Autoencoder with LSTM

An autoencoder is a particular type of neural networks, where the input and output are pair-wise defined. It is part of the so-called unsupervised learning. Unlike feedforward neural networks, recurrent neural network (RNN), such as the LSTM network, are capable of learning the complex dynamics within the temporal ordering of input sequences as well as use internal memory to remember or use information across long input sequences. This LSTM structure is successfully applied in many applications such as speech recognition, natural language processing, and video representation. Since the LSTM autoencoder structure is designed in a mixture of their respective advantages, it allows us to use the model to both support variable-length input sequences and to reconstruct variable-length sequences. Our system is formed by LSTM autoencoder structure whose input is a sequence and output is also a sequence. It can be separated into an encoder and a decoder that share the same latent space. The encoder converts

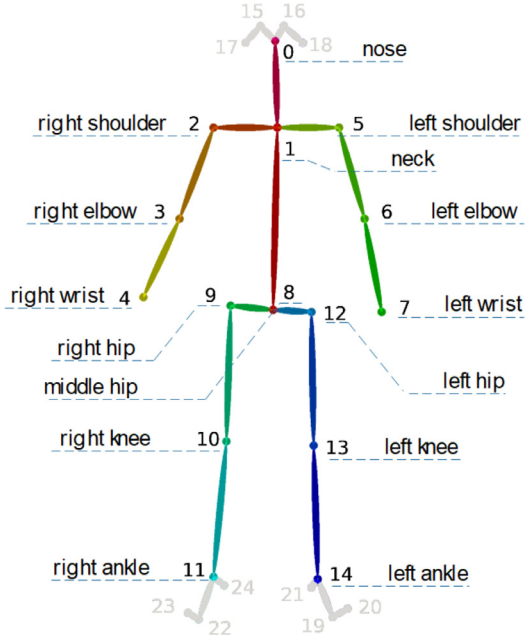

Fig. 2. An illustration of the 25 joints estimated by the OpenPose algorithm. The black marked points indicate the selected joints and the grey numbers correspond to the discarded joints.

a variable-length signal sequence into fixed-length embedding vectors as input to other models, while the decoder turns this fixed-length vector into a variable-length target sequence.

Fig. 3 presents the LSTM autoencoder structure. The left part of the figure is the encoding stage, using LSTM as the basic module, and the middle red module represents the final coding vector. In the encoding steps, the encoder reads the input joints trajectories $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \ldots, \boldsymbol{x}_{n-1}, \boldsymbol{x}_n]$ sequentially and the hidden state of the LSTM is updated accordingly. The hidden state vectors from the last step then are transformed to a D-dimensional embedding vector. The embedding vector then is added with original inputs as the input of the decoding steps to reconstruct the output sequences of n joints trajectories $\boldsymbol{X}^{'}$. Finally, the D-dimensional embedding vector is extracted as the feature representation for the input joints trajectories.

3.4. Siamese DAE structure

Siamese network is commonly used in metric learning systems on pairwise identity verification, including pairwise face verification and pairwise speaker verification and so on. It is essential that not only the architecture of the subnetworks is identical, but the weights have to be shared among them as well. The main goal of the network structure is to learn optimal feature representations of the input pairs where matched pairs are pulled closer, and unmatched pairs are pushed far away.

We propose to learning a discriminative embedding vector using a Siamese architecture of the autoencoder network in a supervised setting. The siamese DAE network, presented in Fig. 4, consists of two identical DAE networks sharing the same weights; each takes joints trajectories as input and outputs the reconstructed trajectories. After the encoding stage, these two encodings act as latent feature representation, which is demonstrated in the figure with green blocks and orange blocks. By learning a similarity metric, the siamese networks can learn useful data encodings that can be further used to compare the inputs of the respective subnetworks. The subsequent networks are usually formed by a single fully connected layer with one output, that corresponds to the similarity estimate. Using this, we compare and tell if the two input trajectories are from the same identity or not. Finally, the Siamese architecture can reduce intra-class variations and increase inter-class variations of the encoded embedding vectors. In the results, we demonstrate the performance of the method for the tasks of classification between using siamese and non-siamese structure, showing that the siamese structure leads to significantly improved results compared to non-siamese structure.

3.5. Loss function

Reconstruction loss. Different to the common autoencoder, the loss of which is defined between an input sequence and its output sequence, our reconstruction loss is defined between an reference sequence after coarse correction and its reconstructed sequence outputted from the model. The main idea is quite similar to the traditional denoising autoencoder, which takes a partially corrupted input whilst training to recover the original undistorted input. We replace the corrupted input with inaccurate trajectories and treat the modified trajectories as undistorted input to be reconstructed. Hereby, the model have an ability to automatically correct the missing points and outliers in joints trajectories. We use the Mean Square Error (MSE) as the loss function to evaluate the loss between reference inputs and reconstructed outputs, which is defined as:

$$\mathcal{L}_{reconstruction\ loss} = \frac{1}{2n}\sum_{i=1}^{n}\left\|\boldsymbol{x}_i^{'} - \boldsymbol{x}_i^{ref}\right\|^2$$

where n represents the number of training frames, $\boldsymbol{x}_i^{'}$ represents a frame from the reconstructed sequence, and the $\boldsymbol{x}_i^{ref}$ represents a frame from the reference sequence.

Siamese loss. In terms of the cost function, one can divide the Siamese structure methods into distance metric learning and similarity metric learning, where the cost function is defined based on a distance metric and a similarity measurement, respectively. Here we use the similarity metric loss function, which is a combination of similarity-loss and dissimilarity-loss. Generally, the contrastive loss can be used as the similarity metric loss function, which can reduce the variations of identical gait features and push apart the gait features from different identities. The contrastive loss is defined as:

$$\mathcal{L}_{contrastive\ loss} = \frac{1}{2N}\sum_{i=1}^{N} yd^2 + (1-y)\max(\lambda - d, 0)^2$$

where N denotes the number of training pairs, $d = \|a_i - b_i\|_2$, represents the Euclidean distance of two embedding vectors. $y = 1$ when the two input samples are from the same identity. $y = 0$ means that the two input samples are from different personalities. The λ is set to guarantee the distance of two sample features to be larger than a fixed margin. This loss function can be well expressed as the matching level of input pairs, and can also be used to extract the features for recognition.

Smoothness constraint loss. To reduce the effect of noise that may affect the coordinates of skeleton joins, we utilize the smoothing filter to each sequence. The smoothness constraint loss is defined as:

$$\mathcal{L}_{smoothness\ loss} = \beta\|\boldsymbol{X}^{'}\mathbf{C}\|^2$$

$\boldsymbol{X}^{'}$ is the reconstructed sequence.

$$\boldsymbol{X}^{'} = \left[\boldsymbol{x}_1^{'}, \boldsymbol{x}_2^{'}, \boldsymbol{x}_3^{'}, \ldots, \boldsymbol{x}_{n-1}^{'}, \boldsymbol{x}_n^{'}\right]$$

where $\boldsymbol{x}_i^{'}$represents a frame from the reconstructed sequences. β is the tuning parameter that is usually determined by experience. $\mathbf{C}$

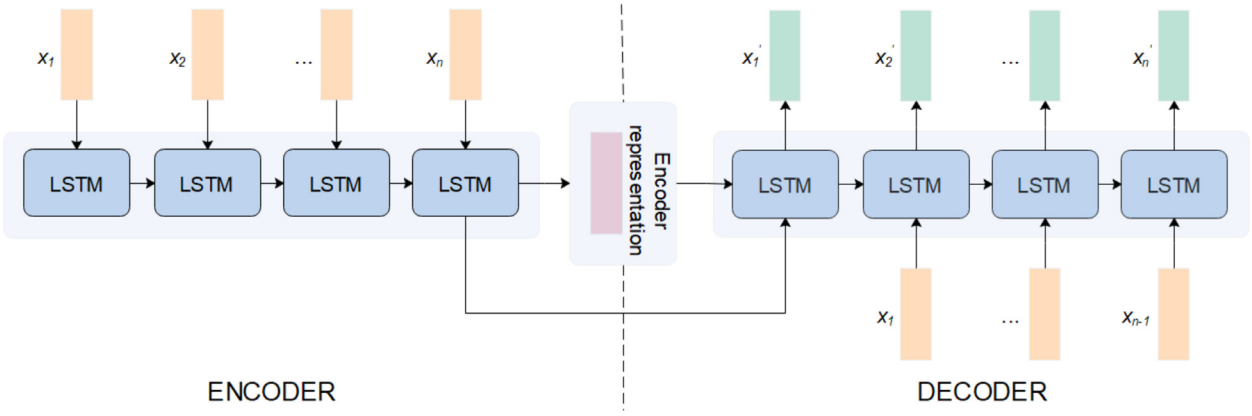

Fig. 3. Our LSTM autoencoder structure that uses two LSTMs for the encoding and decoding stages. $x_1, x_2, \ldots, x_n$ indicate the input sequence, and $x_1', x_2', \ldots, x_n'$ indicate the reconstructed sequence.

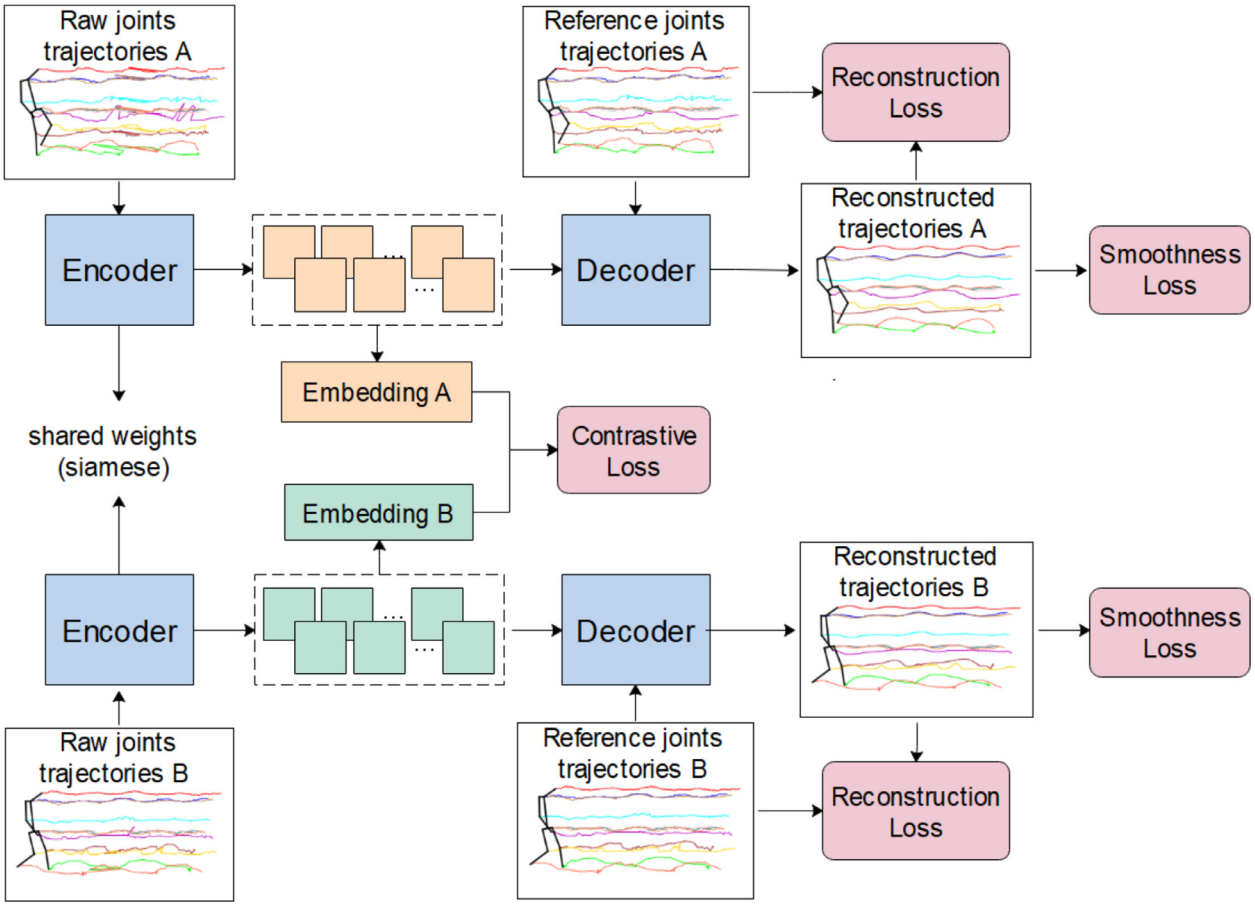

Fig. 4. The siamese DAE structure. The green blocks and orange blocks indicate the encoded embeddings. The red boxes represent all kinds of loss functions of the model.

is a tridiagonal square matrix that is used to measure the distance between neighbor frames in $\boldsymbol{X}'$, and

$$\boldsymbol{C} = \begin{bmatrix} -1 & 1 & & & & \\ 1 & -2 & 1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & & 1 & -2 & 1 \\ & & & & 1 & -1 \end{bmatrix}_{n \times n}$$

Total loss. Finally, the objective function of Siamese DAE is the weighted sum of the three losses and will be back propagated through the whole model. The total loss is deployed for the entire model training and is described by the following equation:

$$\mathcal{L}_{total\ loss} = (1 - \lambda_1 - \lambda_2)\mathcal{L}_{reconstruction\ loss} + \lambda_1 \mathcal{L}_{contrastive\ loss} + \lambda_2 \mathcal{L}_{smoothness\ loss}$$

where λ_1 and λ_2 are hyperparameters to balance the weight of three loss functions. In our experiment, the λ_1 is set to value 0.1, and the λ_2 is set to value 0.02.

3.6. Gait recognition

Through Siamese denoising model, we will get fixed-length embedding vectors with excellent discriminatory properties and corrected skeleton sequences. In this case, we have two methods for gait recognition. The first is to construct a time series model to recognize the reconstructed skeleton sequence. The second is to treat the embedding vectors as identification features, which are linearly separable with each other and can be easily classified with classification methods.

Trajectories based method. From the difference between the original joints trajectories and the reconstructed joints trajectories, the later is more precise and more explicit. Generally, joints trajectories can be discriminated by an end-to-end standalone deep neural network. We use multilayer LSTM cells recurrent network to extract temporal features and two layers fully connected perceptron networks with Softmax activation function to classify different identifications.

Embedding vector based method. The embedding vectors from the encoding stage are linearly separable with each other. Meanwhile, the Siamese structure reduces their intra-class variations and increase their inter-class variations. So the encoded vectors have powerful discrimination and can therefore be classified with e.g. linear SVM or perceptron networks.

4. Experiments

To evaluate our approach, we conduct several experiments for verification tasks on the TUM GAID dataset. In Section 3, we have explained the proposed methodology and illustrated the Siamese Denoising Autoencoders network structure. To measure the quality of the learned feature, we report experiments and evaluated results in comparison to state-of-art methods.

4.1. Dataset and training details

The TUM GAID database is one of the most challenging gait databases, comprising 305 different subjects and offering two main experiments. The goal in the first one is to identify 305 different subjects using 10 gait sequences for each person. These sequences are recorded in three different covariate conditions: normal walking (N), walking with a backpack (B), walking with coating shoes (S). To further investigate the challenges of time variation, a subset of 32 subjects is recorded a second time for another experiment. The goal of the second experiment is to identify 32 subjects using 20 gait sequences for each person, while ten of them were taken in January and the other 10 in April. A total of 32 subjects participated in both experiments, thus have 10 more sequences in three conditions: normal walking (TN), walking with a backpack (TB), walking with coating shoes (TS). To be consistent with the recommended experiments in [13], we also split the database into three partitions: 100 subjects for training, 50 subjects for validation, and the rest 155 subjects for testing. In the test set, the recordings N1, N2, N3, N4 are used as the gallery set, and the recordings N5, N6, B1, B2, S1, S2 are used as probe set.

In all the experiments, the input is a set of image sequences in size of 640×480, without any background subtraction or resizing operation. Since each original image sequence has a different temporal length, we extract subsequences of 32 frames from the full-length sequences with a three frames interval. By pose estimation algorithm, we extract the joints coordinates from each continuous 32 image sequences to obtain $32 \times 12 \times 2$ trajectories vector as the input of our model. As for the TUM GAID dataset, we can get a set of 45,073 examples. To increase the number of samples available for training, we compute the corresponding mirror sequences with the left side and the right side exchanged.

Table 1
State-of-the-art on TAUM GAID. Percentage of correct recognition to TUM GAID for diverse methods. Each column corresponds to a different covariate condition. Best results are marked in bold.

	N	B	S	TN	TB	TS	AVG
GEI	99.4	27.1	52.6	44.0	6.0	9.0	39.7
GVI	99.0	47.7	94.5	62.5	15.6	62.5	63.6
SVIM	98.4	64.2	91.6	65.6	31.3	50.0	66.9
SEIM	99.0	18.4	96.1	15.6	3.1	28.1	43.4
RSM	**100.0**	79.0	97.0	58.0	38.0	57.0	71.5
CNN-SVM	99.7	97.1	97.1	59.4	50.0	62.5	77.6
PFM	99.7	**99.0**	**99.0**	78.1	56.3	46.9	79.8
Ours	98.7	93.6	98.0	**81.4**	**76.2**	**78.1**	**87.7**

We use Bidirectional LSTM cells for the Autoencoder structure to eliminate the variate condition of walking direction. After some prior testing, we adopt two layers LSTM for encoding stage with the corresponding number of hidden units 128 and 64 respectively and two layers LSTM for decoding stage with 64 and 128 hidden units in order. All LSTM layers use the rectification (ReLU) activation function. To reduce overfitting in LSTMs, we use the Dropout technique after each LSTM layer by the approach expressed in [42], which can regularize the RNN neural network without sacrificing its valuable memorization ability. The dropout rate is set to 0.5 during training. In this work, we empirically used 256 hidden units for each fully connected perceptron networks.

We adopt the Adam optimizer to adjust the weight of neurons by calculating the gradient of the loss function expressed in Section 3. We use the early stopping criterion and a batch size of 256 samples. The initial learning rate is $\eta = 0.001$, which is kept unchanged until the validation error stops improving for more than the patience epochs. The maximum number of epochs is set to 10,000, although the training may be stopped earlier if the validation error stops improving, and the training stage is shown in Fig. 6. It is obvious that the loss quickly converged after a few training epochs. It also shows that after about 8000 epochs, the training stage stopped earlier.

After the training process, the model possesses a strong ability to reconstruct joints trajectories with outliers automatically corrected. To prove this point, we choose two representative samples and compare their raw joint trajectories, manually corrected trajectories and reconstructed trajectories in Fig. 5. Each row with three figures represents a set of sample. In the left two figures, the original data contains serious noise caused by incorrect estimation, which includes missing points and overlapping errors. The middle two figures show the manually labeled data, and the right two show the reconstructed data by our model. Notably, the reconstructed trajectories almost coincide with the labeled trajectories, while in the details the reconstructed curves are smoother than the labeled ones with the help of smoothness constraints. Denoising is commonly used to regularized autoencoder network. Experiments proved it to be beneficial in our application as well. For the whole, the Siamese DAE model can learn to reconstruct joints trajectories and correct outliers or position deviation of joints.

4.2. Experiment results

In this section, we verify the performance of our model. The results of the experiments on the TUM GAID dataset are compared with other state-of-the art algorithms in Table 1, where each row corresponds to the results using various kinds of approaches. Each column presents the recognition results of the different covariate conditions including general subset (N, B, S), elapsed time subset (TN, TB, TS), and the average results (AVG) of the six diverse scenarios. The most comparative method is the Pyramidal Fisher Motion (PFM), proposed in [6]. It is a high-level gait

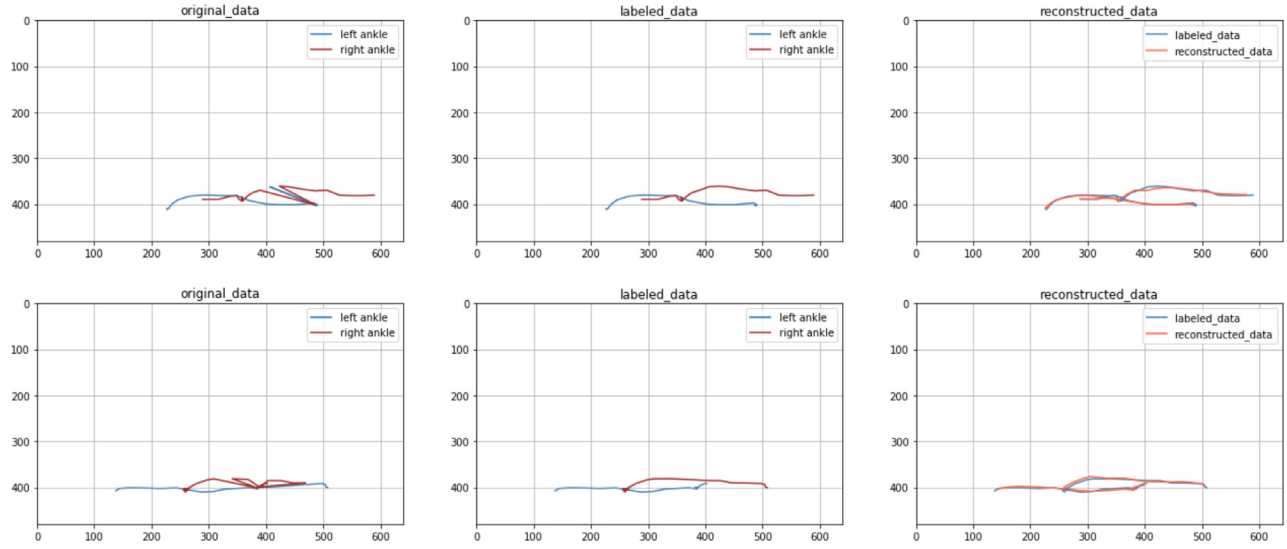

Fig. 5. The plots (640 × 480) of raw joint trajectories and reconstructed trajectories of right ankle and left ankle. The original data contains serious noise caused by incorrect estimation. The reconstructed trajectories is smoother than the labeled trajectories.

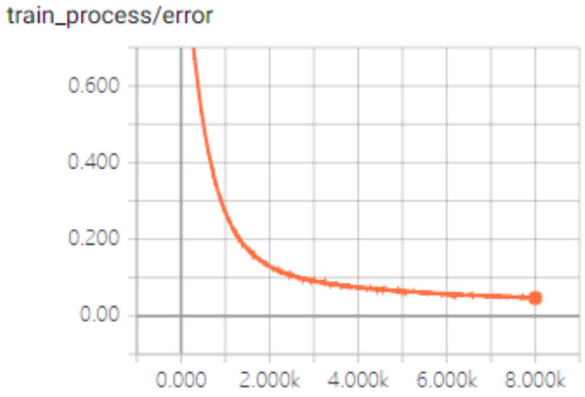

Fig. 6. The training loss curve of the Siamese DAE model, which is visualized by using TensorBoard.

descriptor based on densely sampled short-term trajectories. As Table 1 shows, the performance of our method is not superior to other methods in the first experiment, especially compared to the state-of-the-art algorithm PFM. But it almost obtained the best results of them. While in the second experiment, our approach outperforms PFM with a large margin. It means that our approach is more robust against clothing and time variation than other methods. Using only the joint points dynamic information of the human body, our model will lose some static information like the body shape and clothing texture. Meanwhile, we believe the more accurate positioning of joints, the higher discriminative power of the acquired features. We cannot guarantee the corrected trajectories are 100% accurate, but we believe it is a promising approach and still has potentials for further improvement.

The real-time performance of an algorithm is important for judging the practical applicability, so we test the response time of each module of our method. The most time-consuming module is the pose estimation module. It took about 0.1s to process a sequence on a system with Intel(R) Core(TM) i7-4770 processor and 16GB RAM. After pose estimation of 32 consecutive sequences, it only took 0.03s to get the encoded features and corrected trajectories by our network. The feature recognition module by multilayer LSTM cells or linear SVM spent less than 0.05s. In summary, the real-time performance of our approach can be greatly improved by improving the real-time performance of the pose estimation module. We can choose a lightweight network model or a more efficient pose estimation method.

Table 2
Ablation experiments. Percentage of correct recognition on TUM GAID dataset. Each row corresponding to a different combination of methods from the proposed siamese DAE model. Four kinds of trajectories based methods and two kinds of embedding-based methods are listed.

	N	B	S
Raw trajectories	57.2	46.8	52.3
Manually corrected trajectories	84.1	79.8	80.4
DAE reconstructed trajectories	95.6	92	91.2
Reconstructed+Smoothed trajectories	96.4	93.5	93.4
DAE Embedding	93.7	90.4	93.3
Siamese DAE Embedding	98.7	93.6	98.0

4.3. Analysis and discussion

To evaluate the contribution of each component of the proposed framework to the final performance, we design additional ablation experiments to evaluate them respectively. As shown in Table 2, there are mainly six ablation experiments in two groups, which correspond to the two kinds of methods described in Section 3, and each method is evaluated under six covariate conditions as usual. First, we assess the performance of the raw trajectories from the pose estimation algorithm. As expected, the raw trajectories are inferior in discriminative expression and produce a lackluster performance. The manually corrected trajectories alleviate the noise problem and achieve much better performance than the raw trajectories. Thus, the quality of trajectory is crucial for the performance of recognition. In the third row, the reconstructed trajectories by Denoising Autoencoder model achieve significantly higher performance. Meanwhile, it can be confirmed that the performance is improved somewhat depending on the methods of smoothness constraint. Additionally, we evaluate the discrimination power of the embedding vector obtained from the intermediate encoding stage. Obviously, the embedding-based methods also achieve remarkable performance. By comparison, the siamese DAE embedding has stronger discrimination power than DAE Embedding and increases the recognition accuracy by 5 percent approximately. We believe the better performance of the

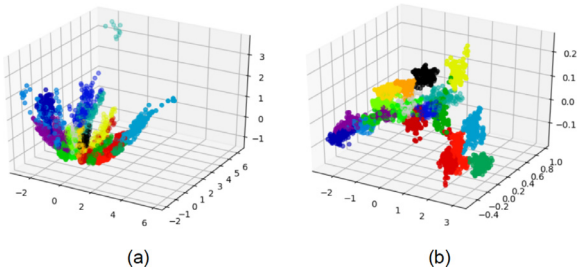

Fig. 7. Scatter plots of two embeddings after PCA. The left figure is from DAE method, and the right one is from Siamese DAE method. Every point in 3D represents a sample, and different identities are symbolized by different colors.

siamese DAE embedding justifies the power of siamese DAE model in skeleton-based gait recognition.

To illustrate more specifically, the encoded DAE embedding vectors are visualized after PCA dimension reduction to see whether there is significant discrimination among different identities. To simplify this, we randomly choose 20 characters for the experiment. As shown in Fig. 7, the two embeddings scatter plots in 3 dimensions are from DAE method and Siamese DAE method. Every point in 3D represents a sample, and different identities are symbolized by different colors. It shows that both of them have discriminative power in some way. Various kinds of classification algorithm can be taken to separate them. However, in the left figure (Fig. 7(a)), samples of all the categories are partially clustered, and some samples in the same category are not gathered together. On the other hand, samples for Siamese DAE in the right figure (Fig. 7(b)) obviously have stronger discrimination. Since the contrastive loss in siamese structure can close the sample distance of identical identities and push apart the samples from different personalities. On the whole, the Siamese DAE model can reduce intra-class variations and increase inter-class variations. Notably, our system only requires a single camera installed at a fixed place, and the method is quite effective on the clothing and time covariates, where the time factor is likely also affected by the change in clothing between recording sessions.

5. Conclusion and future work

In this paper, we proposed the siamese denoising autoencoders network for skeleton-based gait recognition. We have assessed its performance on TUM GAID datasets against some baselines, showing its superiority for discriminative expression of dynamic features, especially when the testing walking conditions are different from the corresponding training conditions. Through our model, the reconstructed joint trajectories have less noise and higher performance for recognition. The embedding layer constrained by contrastive loss from siamese architecture provides comprehensive spatial-temporal features of gait dynamics. Our model is superior in discriminative power and more robust to clothing variation, carrying variation, and even time variation. Currently, the siamese DAE network can reconstruct joint trajectories and realize gait recognition using images taken from side-view. In the future, we will extend this model to deal with cross-view gait recognition condition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Weijie Sheng: Methodology, Software, Investigation, Writing - original draft, Conceptualization. **Xinde Li:** Validation, Resources, Supervision, Writing - review & editing.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 61573097 and 91748106, in part by Key Laboratory of Integrated Automation of Process Industry (PAL-N201704), the Advanced Research Project of the 13th Five-Year Plan (31511040301) and (30601120401), the Fundamental Research Funds for the Central Universities (3208008401), the Qing Lan Project and Six Major Top-talent Plan, and in part by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

[1] M. Balazia, P. Sojka, Learning robust features for gait recognition by maximum margin criterion, in: Proceedings of the 2016 Twenty-third International Conference on Pattern Recognition (ICPR), 2016, pp. 901–906, doi:10.1109/ICPR.2016.7899750.
[2] K. Bashir, T. Xiang, S. Gong, Gait recognition without subject cooperation, Pattern Recognit. Lett. 31 (2010) 2052–2060.
[3] X. Ben, C. Gong, P. Zhang, R. Yan, Q. Wu, W. Meng, Coupled bilinear discriminant projection for cross-view gait recognition, IEEE Trans. Circuits Syst. Video Technol. PP (2019), doi:10.1109/TCSVT.2019.2893736. 1–1
[4] X. Ben, P. Zhang, Z. Lai, R. Yan, X. Zhai, W. Meng, A general tensor representation framework for cross-view gait recognition, Pattern Recognit. 90 (2019), doi:10.1016/j.patcog.2019.01.017.
[5] Z. Cao, T. Simon, S. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, in: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1302–1310, doi:10.1109/CVPR.2017.143.
[6] F. Castro, M. MarÃn-JimÃnez, N. Guil, R. MuÃ±oz-Salinas, Fisher motion descriptor for multiview gait recognition, Int. J. Pattern Recognit Artif Intell. 31 (2016) 1756002, doi:10.1142/S021800141756002X.
[7] C. Changhong, L. Jimin, Z. Heng, H. Haihong, T. Jie, Frame difference energy image for gait recognition with incomplete silhouettes, Pattern Recognit. Lett. 30 (2009) 977–984.
[8] S. Choi, J. Kim, W. Kim, C. Kim, Skeleton-based gait recognition via robust frame-level matching, IEEE Trans. Inf. Forensics Secur. PP (2019). 1–1
[9] J.E. Cutting, L.T. Kozlowski, Recognizing friends by their walk: Gait perception without familiarity cues, Bull Psychon Soc 9 (1977) 353–356.
[10] H. Fang, S. Xie, Y. Tai, C. Lu, RMPE: Regional multi-person pose estimation, in: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2353–2362, doi:10.1109/ICCV.2017.256.
[11] B. Gálai, C. Benedek, Feature selection for lidar-based gait recognition, in: Proceedings of the 2015 International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM), 2015, pp. 1–5, doi:10.1109/IWCIM.2015.7347076.
[12] M. Hofmann, S. Bachmann, G. Rigoll, 2.5d gait biometrics using the depth gradient histogram energy image, in: Proceedings of the 2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS), 2012, pp. 399–403, doi:10.1109/BTAS.2012.6374606.
[13] M. Hofmann, J. Geiger, S. Bachmann, B. Schuller, G. Rigoll, The TUM gait from audio, image and depth (gaid) database: Multimodal recognition of subjects and traits, J. Vis. Commun. Image Represent. 25 (2014) 195–206.
[14] H. Hu, Enhanced Gabor feature based classification using a regularized locally tensor discriminant model for multiview gait recognition, IEEE Trans. Circ. Syst. Video Technol. 23 (2013) 1274–1286.
[15] M. Hu, Y. Wang, Z. Zhang, D. Zhang, J.J. Little, Incremental learning for video-based gait recognition with LBP flow, IEEE Trans. Cybern. 43 (2013) 77–89, doi:10.1109/TSMCB.2012.2199310.
[16] H. Ju, B. Bir, Individual recognition using gait energy image, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2005) 316–322.
[17] D. Kastaniotis, I. Theodorakopoulos, G. Economou, S. Fotopoulos, Gait-based gender recognition using pose information for real time applications, in: Proceedings of the 2013 Eighteenth International Conference on Digital Signal Processing (DSP), 2013, pp. 1–6, doi:10.1109/ICDSP.2013.6622766.
[18] D. Kastaniotis, I. Theodorakopoulos, C. Theoharatos, G. Economou, S. Fotopoulos, A framework for gait-based recognition using kinect, Pattern Recognit. Lett. 68 (2015) 327–335, doi:10.1016/j.patrec.2015.06.020.
[19] S.I. Ktena, S. Parisot, E. Ferrante, M. Rajchl, M. Lee, B. Glocker, D. Rueckert, Metric learning with spectral graph convolutions on brain connectivity networks, Neuroimage 169 (2018) 431–442.
[20] R. Liao, C. Cao, E.B.G. Reyes, S. Yu, Y. Huang, Pose-based temporal-spatial network (PTSN) for gait recognition with carrying and clothing variations, in: Proceedings of the Biometric Recognition – Twelfth Chinese Conference, CCBR

2017, 2017, pp. 474–483, doi:10.1007/978-3-319-69923-3_51. Shenzhen, China, October 28–29, 2017
[21] R. Liao, S. Yu, W. An, Y. Huang, A model-based gait recognition method with body pose and human prior knowledge, Pattern Recognit. 98 (2020) 107069.
[22] G. Liu, L. Mcmillan, Estimation of missing markers in human motion capture, Vis. Comput. 22 (2006) 721–728.
[23] X. Ma, J. Peng, Kinect sensor-based long-distance hand gesture recognition and fingertip detection with depth information, J. Sensors 2018 (2018) 1–9.
[24] M.J. Marín-Jiménez, F.M. Castro, N. Guil, F. de la Torre, R. Medina-Carnicer, Deep multi-task learning for gait-based biometrics, in: Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), 2017, pp. 106–110, doi:10.1109/ICIP.2017.8296252.
[25] T.T. Ngo, Y. Makihara, H. Nagahara, Y. Mukaigawa, Y. Yagi, The largest inertial sensor-based gait database and performance evaluation of gait-based personal authentication, Pattern Recognit. 47 (2014) 228–237.
[26] T. Nguyen, H. Huynh, J. Meunier, Skeleton-based gait index estimation with lstms, in: Proceedings of the 2018 IEEE/ACIS Seventeenth International Conference on Computer and Information Science (ICIS), 2018, pp. 468–473, doi:10.1109/ICIS.2018.8466522.
[27] O. Oktem, M. Muftuoglu, F. Senbabaoglu, B. Urman, Walking in colors: Human gait recognition using kinect and CBIR, IEEE Multimedia 20 (2013) 28–36.
[28] A. Roy, S. Sural, J. Mukherjee, Gait recognition using pose kinematics and pose energy image, Signal Process. 92 (2012) 780–792.
[29] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, Y. Yagi, Geinet: View-invariant gait recognition using a convolutional neural network, in: Proceedings of the 2016 International Conference on Biometrics (ICB), 2016, pp. 1–8, doi:10.1109/ICB.2016.7550060.
[30] S. Sivapalan, D. Chen, S. Denman, S. Sridharan, C. Fookes, Gait energy volumes and frontal gait recognition using depth images, in: Proceedings of the 2011 International Joint Conference on Biometrics (IJCB), 2011, pp. 1–6, doi:10.1109/IJCB.2011.6117504.
[31] S. Springer, S.G. Yogev, Validity of the kinect for gait assessment: A focused review, Sensors 16 (2016) 194.
[32] S. Tak, H.S. Ko, A physically-based motion retargeting filter, ACM Trans. Graph. 24 (2005) 98–117, doi:10.1145/1037957.1037963.
[33] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, Y. Yagi, On input/output architectures for convolutional neural network-based cross-view gait recognition, IEEE Trans. Circ. Syst. Video Technol. PP (2017). 1–1
[34] R. Tanawongsuwan, A. Bobick, Gait recognition from time-normalized joint-angle trajectories in the walking plane, in: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, 2001, doi:10.1109/CVPR.2001.991036. II–II
[35] G. Xia, H. Sun, G. Zhang, F. Lei, Human motion recovery jointly utilizing statistical and kinematic information, Inf. Sci. (Ny) 339 (2016). S0020025516000244.
[36] S. Yu, H. Chen, E. Garcia, N. Poh, Gaitgan: Invariant gait feature extraction using generative adversarial networks, in: Proceedings of the Computer Vision & Pattern Recognition Workshops, 2017, pp. 532–539, doi:10.1109/CVPRW.2017.80.
[37] 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.
[38] Y. Zha, Y. Fan, Multi-person gait recognition system based on kinect, in: Proceedings of the 2016 Second IEEE International Conference on Computer and Communications (ICCC), 2016, pp. 353–357, doi:10.1109/CompComm.2016.7924722.
[39] E. Zhang, Y. Zhao, X. Wei, Active energy image plus 2DLPP for gait recognition, Signal Process. 90 (2010) 2295–2302.
[40] Y. Zhang, B. Shen, S. Wang, D. Kong, B. Yin, L_0-regularization-based skeleton optimization from consecutive point sets of kinetic human body, ISPRS J. Photogramm. Remote Sens. 143 (2018) 124–133, doi:10.1016/j.isprsjprs.2018.04.016.
[41] D. Kastaniotis, I. Theodorakopoulos, S. Fotopoulos, Pose-based gait recognition with local gradient descriptors and hierarchically aggregated residuals, Journal of Electronic Imaging 25 (2016) 91–99, doi:10.1117/1.JEI.25.6.063019.
[42] Y. Feng, Y. Li, J. Luo, Learning effective Gait features using LSTM, 2016 23rd International Conference on Pattern Recognition (ICPR) (2016) 325–330.

Weijie Sheng received the B.S. and M.S. degrees in Marine Engineering & Automation from Huazhong University of Science and Technology, Wuhan, China, respectively in 2010 and 2013. Now, he is a Ph.D. student in School of Automation at the Southeast University, Nanjing, China.

Xinde Li received his Ph.D. from the Department of Control, Huazhong University of Science and Technology in June 2007. In December of the same year, he worked in the School of Automation, Southeast University. From January 2012 to January 2013, he visited Georgia Polytechnic University as a national visiting scholar for one year. From January 2016 to the end of August 2016, he worked as a research fellow in the Department of ECE, National University of Singapore. His main research interests include intelligent robots, machine vision perception, machine learning, human-computer interaction, intelligent information fusion and artificial intelligence.