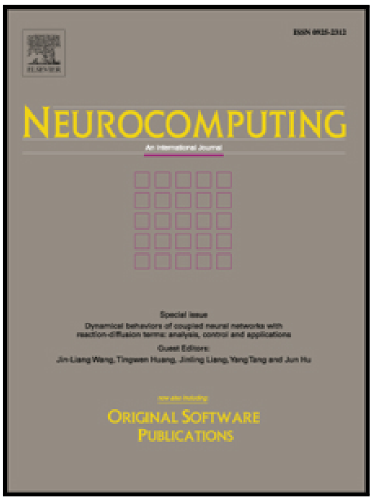


Invariant Feature Extraction for Gait Recognition Using Only One Uniform Model

Shiqi Yu, Haifeng Chen, Qing Wang, Linlin Shen, Yongzhen Huang

PII: S0925-2312(17)30276-X
DOI: 10.1016/j.neucom.2017.02.006
Reference: NEUCOM 18072

To appear in: *Neurocomputing*

Received date: 21 November 2016
Revised date: 26 January 2017
Accepted date: 3 February 2017

Please cite this article as: Shiqi Yu, Haifeng Chen, Qing Wang, Linlin Shen, Yongzhen Huang, Invariant Feature Extraction for Gait Recognition Using Only One Uniform Model, *Neurocomputing* (2017), doi: 10.1016/j.neucom.2017.02.006

Invariant Feature Extraction for Gait Recognition Using Only One Uniform Model

Shiqi Yu[a,*], Haifeng Chen[a], Qing Wang[a], Linlin Shen[a], Yongzhen Huang[b]

[a] *Computer Vision Institute, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, 518060, China*
[b] *National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China*

Abstract

Gait recognition has been proved useful in human identification at a distance. But many variations such as view, clothing, carrying condition make gait recognition is still challenging in real applications. The variations make it is hard to extract invariant feature to distinguish different subjects. For view variation, one view transformation model can be employed to convert the gait feature from one view to another. Most existing models need to estimate the view angle first, and can work for only one view pair. They can not convert multi-view data to one specific view efficiently. Other variations also need some specific models to handle. We employed one deep model based on auto-encoder for invariant gait extraction. The model can synthesize gait feature in a progressive way by stacked multi-layer auto-encoders. The unique advantage is that it can extract invariant gait feature using only one model, and the extracted feature is robust to view, clothing and carrying condition variation. The proposed method is evaluated on two large gait datasets, CASIA Gait Dataset B and SZU RGB-D Gait Dataset. The experimental results show that the proposed method can achieve state-of-the-art performance by only one uniform model.

Keywords: Gait recognition, deep learning, invariant feature

[*] Corresponding author
Email addresses: shiqi.yu@szu.edu.cn (Shiqi Yu), chenhaifeng@email.szu.edu.cn (Haifeng Chen), 2009150166@email.szu.edu.cn (Qing Wang), llshen@szu.edu.cn (Linlin Shen), yzhuang@nlpr.ia.ac.cn (Yongzhen Huang)

1. Introduction

Gait, known as human walking style, is a kind biometric feature for human identification at a distance. Compared with other biometric features, such as face, iris, palmprint and fingerprint, gait has great potential in human identification because of its unique advantages such as non-contact, hard to fake and obtainable at a distance. Therefore gait recognition in surveillance attracted increasing attention in computer vision community.

There are many pioneer works on gait recognition. Some of them are model-based methods [1, 2, 3], and some are appearance-based ones [4, 5, 6, 7]. These works show that gait recognition is feasible in human identification at a distance. But gait recognition is still a challenging task because of view, clothing, occlusion and other variations. These challenges can affect the recognition accuracy greatly. Among these challenges, view variation is one of the most commons because we can not control the walking direction of subjects in real applications. Many existing view invariant gait recognition methods [6, 8, 9, 10, 11, 12] heavily depend on the accuracy of view angle estimation. For each gallery and probe angle pair, a model is need to be trained, and the model can only transform specific view. Besides of view variation, clothing can also change the human body appearance and shape greatly. Some clothes, such as long overcoats, can occluded the leg motion. Carrying condition is another factor which can effect feature extraction because it is not easy to segment the carried object from a human body in images.

The unique advantage of our work is that only one uniform model is trained which can handle gait data with view, clothing and carrying condition variations. The gait data captured with multiple variations can be transformed into the side view without knowing the specific view angles, clothing type and the object carried. So this method has great potential in real scenes.

The rest of the paper is organized as follows. Section 2 discusses related works. Section 3 describes the proposed invariant feature extraction model. Experiments and evaluation are presented in Section 4. The last section, Section

5, gives the conclusions.

2. Related Work

In the following part of this section, we will briefly review gait recognition methods which are invariant to changes.

Some researchers paid close attention to view invariant gait recognition more than a decade ago. Some early methods, such as that in [13], use static body parameters measured from gait images as a kind of view invariant feature. Kale *et al.* [14] used the perspective projection model to generated side view feature from any other arbitrary view. Actually the relation between two views can not be modeled by a simple linear model, such as the perspective projection model.

Some other researcher employed more complex models to handle this problem. Makihara *et al.* [8] designed a view transformation model (VTM) in the frequency-domain features nor the spatial domain. The method RSVD-VTM proposed in [9] is in spatial domain. It uses reduced SVD to construct a VTM and optimized Gait Energy Image(GEI) feature vectors based on linear discriminant analysis (LDA), and achieves relative good improvements. According to the great capability of robust principal component analysis (RPRC) in feature extraction, Zheng *et al.* [6] established a robust VTM via RPCA for view invariant feature extraction. Kusakunniran *et al.* [10] took the view transformation as a regression problem, and used the sparse regression based on the elastic net as the regression function. Bashir *et al.* [15] formulated a gaussian process classification framework to estimate view angle in probe set, then uses canonical correlation analysis(CCA) to model the correlation of gait sequences from different views. Luo *et al.* proposed a gait recognition method based on partitioning and CCA [16]. They separated GEI image into 5 non-overlapping parts, and for each part they used CCA to model the correlation. In [17] Xing *et al.* also used CCA. But they reformulated the traditional CCA to deal with high-dimensional matrix, and reduce the computational burden in view invariant feature extraction. Lu *et al.* [18] proposed one method which can handle arbitrary walking directions by cluster-based averaged gait images. But if there is not similar walking direction in the gallery set, the recognition rate will decrease.

Some other researchers also tried to solve view variance using only one model. Such as Hu *et al.* [19] proposed a method named as ViDP which extracts view invariant features using a linear transform. Wu *et al.* [20] trained deep convolution neural networks using supervised information and achieved high accuracies.

The clothing invariant gait recognition methods are not as many as view invariant ones in the literature. In [21] clothing invariant gait recognition is implemented by dividing the human body into 8 parts and analyzing the discrimination capability of different parts. In [22] Guan *et al.* proposed a random subspace method (RSM) for clothing-invariant gait recognition by combining multiple inductive biases for classification.

The variations on gait data can cause the recognition rate decrease greatly. Some methods in the literature can only solve a specific variation, such as view and clothing. A general method which can extract variant gait feature using only one model should be attractive.

3. Proposed Method

In gait recognition, when the angle between the walking direction of and the camera is 90° (the side view), it is the best view for gait recognition because of more dynamic information. We would try to transform the gait data from any views, clothing and carrying condition to the side view with normal clothing condition and not carrying objects using one uniform non-linear model, and then extract invariant feature. The proposed model is inspired by the one in [23] where a model based on auto-encoder which is named as Stacked Progressive Auto-Encoders(SPAE). The model in [23] is proposed to deal with multi-view face recognition. We adapt it to deal with the view, clothing and carrying condition challenges. The framework is illustrated in Figure 1. we will describe the framework in the following subsections.

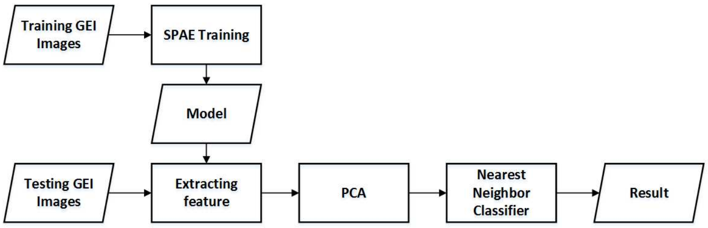


Figure 1: The flowchart of the proposed view invariant gait recognition.

3.1. Gait Energy Image

Gait energy image [4], an appearance-based recognition method, which is produced by averaging the silhouettes in one gait cycle in a gait sequence as illustrated in Figure 2, is well known for its robustness to image noise and reduction on computation. The pixel values in a GEI are the probabilities of the positions are occluded by a human body. According to the success of GEI in gait recognition, we take GEI as the input raw data of our method. The silhouettes and energy images used in the experiments are produced as those in [24].

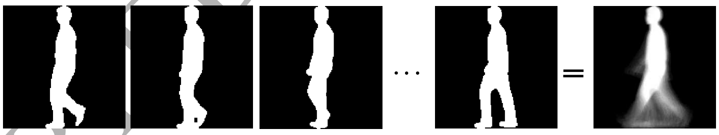


Figure 2: Gait energy image (the right one) is produced by averaging the silhouette in one gait cycle.

3.2. Auto-Encoder for Image Transformation

Auto-encoder [25] is one of the popular models in recent years. It can be used to extract compact features. As shown in Figure 3, an auto-encoder usually contains three layers: one input layer, one hidden layer and one output layer. There are two parts in an auto-encoder, encoder and decoder. The encoder can transform the input data into a new representation in the hidden layer. It

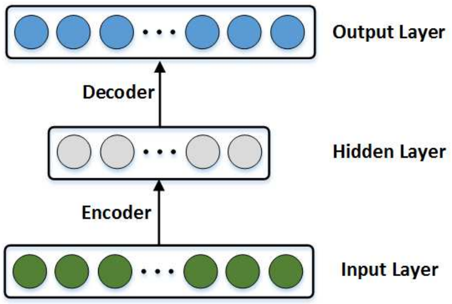


Figure 3: schematic diagram of auto-encoder.

usually consists of a linear and a nonlinear transformation as follows:

$$y = f(x) = s(Wx + b) \tag{1}$$

where $f(\cdot)$ denotes the encoder, W denotes the linear transformation, b denotes the basis and $s(\cdot)$ is the nonlinear transformation, also called activation function, such as:

$$s(x) = \frac{1}{1 + e^{-x}} \tag{2}$$

or

$$s(x) = \ln(1 + e^{-x}) \tag{3}$$

The decoder can transform the hidden layer representation back to input data as follows:

$$x' = g(y) = s(W'y + b') \tag{4}$$

where $g(\cdot)$ denotes the decoder, W' and b' denote the linear transformation and basis in decoder and x' is the output data. 100

We usually use the least square error as the cost function to optimize the parameters in W, b, W' and b'.

$$\begin{aligned}[W, b, W', b'] &= min\Sigma_{i=1}^{N} \parallel x_i - x_i' \parallel^2 \\ &= min\Sigma_{i=1}^{N} \parallel x_i - g(f(x_i)) \parallel^2\end{aligned} \tag{5}$$

where x_i denotes the i_{th} one of the N training samples and x_i' means the correspond output of x_i. In our experiments, we train auto-encoder use Caffe [26, 27] with Euclidean loss and Stochastic Gradient Descent (SGD).

The traditional auto-encoder can reconstruct the input. If we replace the output with a different data what distinguishes with the input data, the whole auto-encoder could be regarded as a regression function. But it would be really hard for just one auto-encoder to deal with large angle change, clothing and carrying variations. As shown in Figure 4(a), the difference between 54° images and 90° ones is much larger than that between 72° images and 90° ones, especially in the leg part. It would be very difficult for just one auto-encoder to transform 54° images to 90° ones. But if we use one auto-encoder to transform 54° images to the 72° ones, and then use another auto-encoder to transform 72° images to 90° ones, it would be much easier. So multiple auto-encoders are needed to deal with gait variations. Some more auto-encoders are needed to handle clothing and carrying condition variations as shown in Figure 4(b).

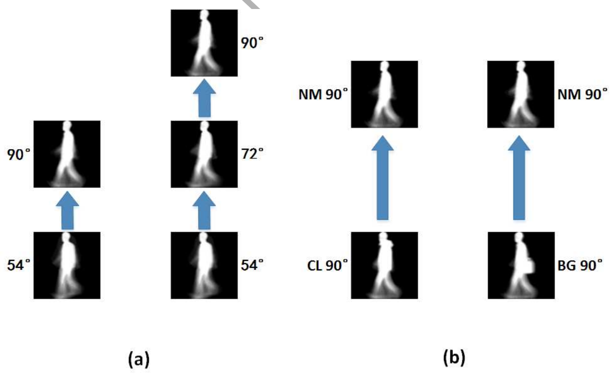


Figure 4: Auto-encoders for gait image transformation. (a)It is more difficult for one auto-encoder to transform 54° images to 90° one than transform 72° image to 90° ones. We could gradually transform 54° images to 72° ones with one auto-encoder and then 72° images to 90° one with another auto-encoder(b)Two auto-encoders are employed to handle the clothing and carry condition variations respectively.

3.3. SPAE for Gait Variations

The main idea of the proposed method is stacked some auto-encoders together to deal with the view, clothing and carrying condition variations. In model training, the output is synthesized in a progressive way.

Side view contains more dynamic information about the gait in gait recognition. So we would try to convert all the gait energy images to side view. But it is difficult for one auto-encoder to deal with all the variations, so we set each auto-encoder to solve a small variation. The first layer of auto-encoders is employed to handle the clothing variation by fitting GEIs with coats to ones of normal clothing. The second layer fits GEIs with bags to ones without bags to reduce carrying condition variation. The view angles are kept unchanged after the transformation of the first two layers as shown in Figure 5.

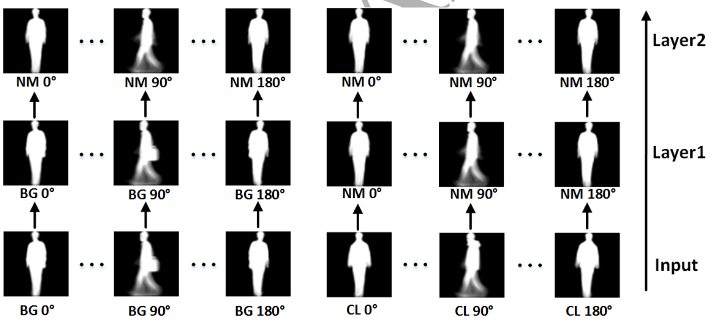


Figure 5: The first two layers employed to handle the clothing and carrying condition variations.

For the view variation, auto-encoders will convert the GEIs at a larger view angle to an adjacent smaller one. At the same time those gait energy images at smaller view angles are kept unchanged. Then after some auto-encoders, all the images would gradually become side view images as shown in Figure 6, it would be very helpful for improving the accuracy of gait recognition.

It is assumed that there are $2 \times L + 1$ views in the dataset. The difference between the adjacent angles is $\Delta = 18^\circ$ and $L = 5$. The view angles of the gait

135 data are $\{0^\circ, 18^\circ, \cdots, 180^\circ\}$. The auto-encoder in first layer would map the gait images at 0° to 18°, and the gait images at 180° to 162°. Meanwhile it keeps the gait images from 18° to 162° unchanged. Then auto-encoder in second layer would map the gait image which is smaller than 36° to 36°, and larger than 144° to 144°. The last layer would map all the images to 90° but maintain images
140 at 90° unchanged. Figure 6 shows a schematic view of the training phase in a progressive way.

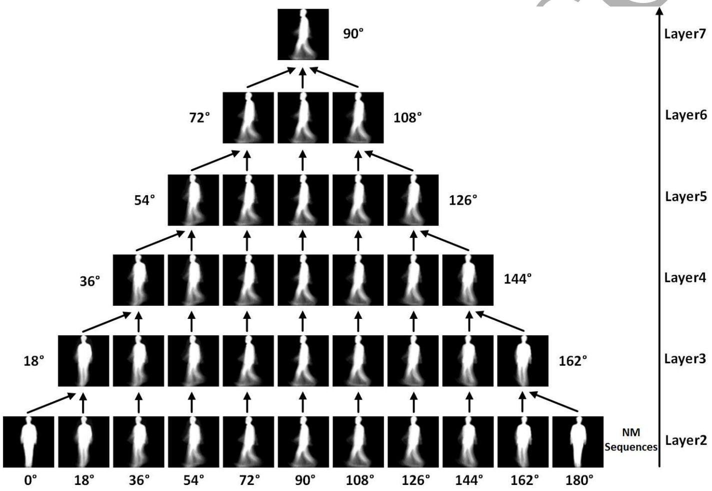


Figure 6: The stacked AEs are employed to deal with view variation. Each layer deals with one small view variation. After some layers, all the images would be transformed to the side view images.

We train each layer individually and the output of a hidden layer is the input of the next layer. After training all the auto-encoders, the whole network is fine tuned by optimizing all layers together as bellow.

$$\begin{aligned} &[W_j \mid_{j=1}^{L}, b_j \mid_{j=1}^{L}, W'_L, b'_L] \\ &= \arg\min \Sigma_{i=1}^{N} \parallel x'_i - g_L(f_L(f_{L-1}(\cdots(f_1(x_i))))) \parallel^2 \end{aligned} \tag{6}$$

145 j means the j_{th} layer in all L layers.

3.4. Invariant Feature Extraction

As the GEIs with clothing and carrying condition variations are transformed to normal ones and the view variations become smaller layer by layer, the output of topmost layer f_L should be the synthesized normal side view feature and is robust to the variations. But lower layers should also contain some beneficial information. So we cumulate the representation in multiple hidden layers at descending order as follows:

$$F = [f_{L-i}, f_{L-i+1}, \cdots, , f_L] \tag{7}$$

where $0 \leq i \leq L-1$. We then use Principal Component Analysis (PCA) to extract more compact feature. In [23], LDA is employed for discriminant feature extraction. LDA needs relatively a large amount samples in each class to model the intra-class variance. The number of samples in our dataset is limited, and experimental results show that LDA can not improve the recognition rate obviously. Considering the computational cost of LDA, we did not use LDA in our experiments as in [23]. The structure of final model is shown in Figure 7.

4. Experiments and Analysis

4.1. Datasets

Two datasets, CASIA B and SZU RGB-D, are involved in our experiments to evaluate the proposed method. CASIA B gait dataset [24] is one of the largest public gait databases, which was created by the Institute of Automation, Chinese Academy of Sciences in January 2005. It consists of 124 subjects (31 females and 93 males) captured from 11 views. The view range is from 0° to 180° with 18° interval between two nearest views. There are 10 sequences for each subject. There are 6 sequences for normal walking (”nm”), 2 sequences for walking with a bag (”bg”) and 2 sequences for walking in a coat (”cl”). Figure 8 shows the samples at 11 views from a subject of normal walking.

SZU RGB-D [28] is a large RGB-D gait dataset created by our group using ASUS Xtion PRO LIVE which is a kind of RGB-D sensor. The sensor can

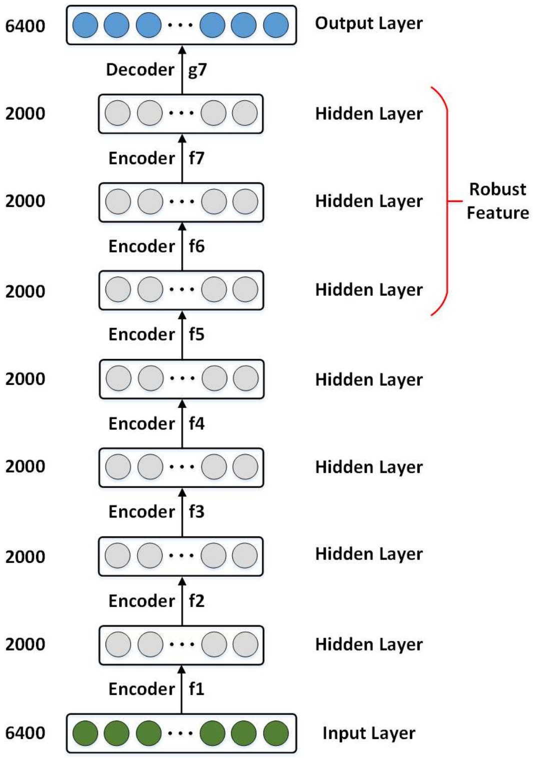


Figure 7: The structure of final model. The numbers on the left side are the number of nodes in each layer.

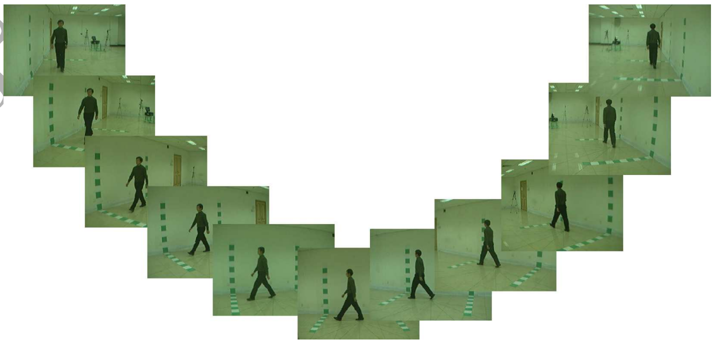


Figure 8: Walking sequences at 11 views from CASIA B dataset.

capture color and depth images. The sensor is fixed to a tripod, about 80cm high from the ground. Subjects walk in the scene, and are demanded to walk in two directions. So gait data can be captured from two views. The first one is the side view (90°), the second is about 30° away from the side view (60°). For each view, there were 4 video sequences captured. Two sequences are right walking ones, and two are left walking. So there are 8 different sequences for each subject. The dataset contains 99 subjects. When subjects walk, synthesized color images (RGB image) and depth images are captured. In this experiment, we just use the color image to generate GEI image for each walking sequence. Figure 9 shows some samples from SZU RGB-D Gait Dataset.

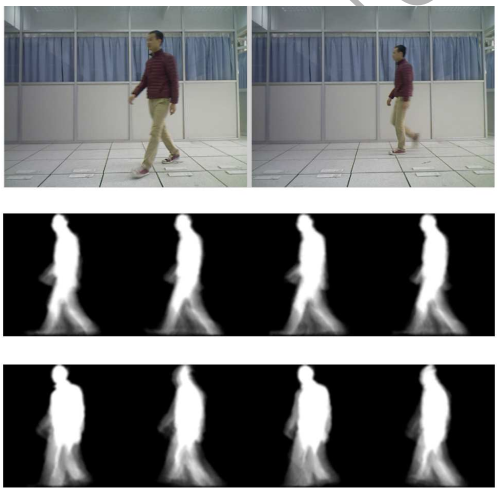


Figure 9: Image samples from SZU RGB-D Gait Dataset. The first row contains color images, and the second and third rows contains the GEI images from 8 different sequences.

4.2. Experimental Design

The first experiment to evaluate the proposed method is carried on CASIA B dataset. The experiment mainly focus on view, clothing and carrying condition variations in gait recognition. We put the first two normal, wearing coats and carrying bags sequences of the first 62 subjects into the training set and the remaining 62 subjects 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. The experiment design is listed in Table 1.

Table 1: Experimental design on CASIA B dataset

Training	Test	
	Gallery Set	Probe Set
ID: 001-062 Seqs: nm01,nm02 bg01,bg02,cl01,cl02	ID: 063-124 Seqs: nm01-nm04	ID: 063-124 Seqs: nm05,nm06 bg01,bg02,cl01,cl02

We also evaluate the proposed method SZU RGB-D dataset, and the model exactly the same as that in CASIA B dataset is used in this experiment. The gait sequences from the first 49 subjects are put into the training set, and the sequences from the remaining 50 subjects are put into the test set. In the test set, the first sequences (No.01-02) are put into gallery set and the others into probe set as those shown in Table 2.

Table 2: Experimental design on SZU RGB-D dataset

Training	Test	
	Gallery Set	Probe Set
ID: 01-49 Seqs: 01-08	ID: 50-99 Seqs: 01-02	ID: 50-99 Seqs: 03-08

4.3. Model Parameters

In the experiments, the model with 7 layers as shown in Figure 7 are used. In the training phase, we use the caffe software [26] to training model which is very popular in deep learning field. First of all, each auto-encoder needs to be trained independently. The initial weights in the layers are set to random values in Gaussian distribution, and the initial bias values are set to zeros before training. We set the base learning rate to 0.1, the maximum number of iteration is 60,000 and the activation function is *sigmoid*. After that, the 7 trained layers are combined in stacked way, and fine tuned as a whole model. In the fine tuning, the base learning rate is set to 0.01 and the maximum number of iteration is 30,000.

One important parameter to the proposed model is the numbers of hidden layer neurons. We set all the number of different layers to the same, and find the optimal one by experiments. The experimental results on CASIA B dataset with view variations are shown in Figure 10. The model with different numbers of neurons from 500 to 6,000 is evaluated. From the results, it can be found that when the number is 2,000 the model achieves the best recognition rate. So in the following experiments, we set the number to 2,000.

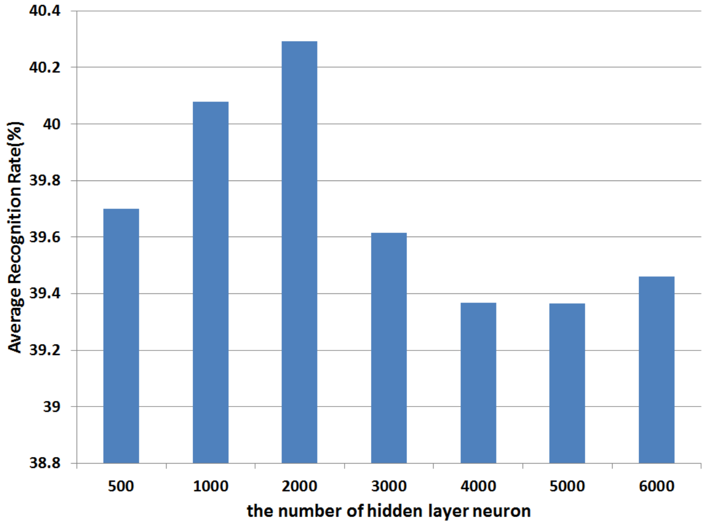


Figure 10: Average recognition rate of experiments with view variation. When the number of hidden layer neurons is 2,000, the highest recognition rate is achieved.

Beside the feature from the topmost layer, we can also select the features from lower layers as the invariant gait feature for recognition. The results of different combinations of layers are shown in Figure 11. From the results we can find that the feature consists of the last three layers, the 5-th and 6-th 7-th ones, achieves the highest recognition rate. So we concatenate the outputs from the last three layers as a long vector (size $2,000 \times 3 = 6,000$) as the invariant gait feature.

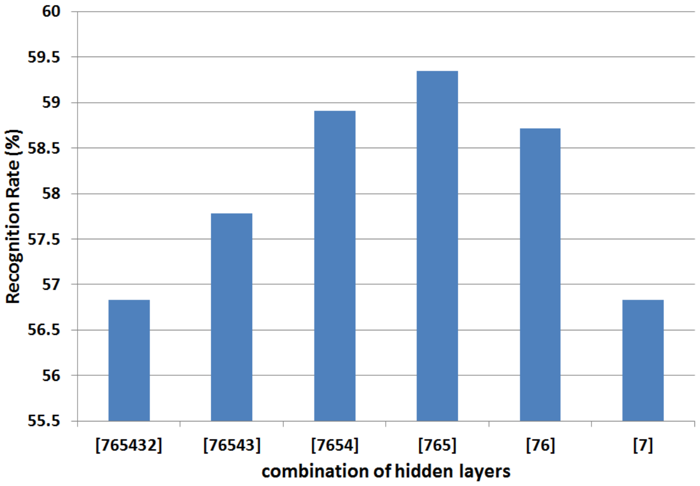


Figure 11: The recognition rates of all combinations for multiple hidden layers. The feature consists of the last three layers $[F_7,\ F_6,\ F_5]$ achieves the highest recognition rate.

The last step of the invariant gait feature extraction is to reduce the dimension using PCA as in [23]. The feature dimension is reduced from 6,000 to 100. The value 100 is chosen according the experiments. After we extract the compact invariant feature, a simple classifier, nearest neighbor (NN), is employed for classification.

4.4. Experimental Results on CASIA B Dataset

To evaluate the performance of the proposed method on variations, the experimental results on CASIA B dataset are given in details in Table 3 - 5. The results in the three tables can evaluate view, clothing and carrying condition variations respectively. For Table 3, the first 4 normal sequences at a specific view are put into the gallery set, and the last 2 normal sequences at another view are put into the probe set. Since there are 11 views in the dataset, there are 121 combinations. All the 121 recognition rates are listed in Table 3. For Table 4, the differences are the probe sets. The probe data is carrying bags data, and the carrying condition is different from the that in the gallery set. The probe sets for Table 5 contain gait data with coats. In the tables, each row correspond to a view angle of the gallery set, whereas the columns correspond to the view angle of the probe set.

Table 3: Recognition rates when the probe data is normal walking data.

		Probe set view(Normal walking, nm05,nm06)										
		0	18	36	54	72	90	108	126	144	162	180
Gallery set view	0	98.39	87.10	58.06	39.52	28.23	33.87	31.45	37.90	46.77	62.10	67.74
	18	85.48	99.19	98.39	75.00	56.45	47.58	41.13	47.58	54.84	52.42	55.65
	36	66.13	96.77	97.58	91.13	67.74	54.03	52.42	54.84	58.87	55.65	46.77
	54	50.00	63.71	83.87	95.97	89.52	82.26	72.58	65.32	57.26	43.55	28.23
	72	37.10	50.81	67.74	83.06	95.97	94.35	91.13	79.84	62.10	37.10	33.87
	90	32.26	35.48	52.42	70.16	95.16	95.97	95.16	80.65	56.45	33.87	29.03
	108	26.61	37.10	47.58	65.32	91.94	95.97	96.77	90.32	70.97	42.74	30.65
	126	33.06	45.16	60.48	72.58	84.68	86.29	93.55	98.39	94.35	59.68	35.48
	144	41.13	51.61	54.03	66.94	60.48	60.48	80.65	96.77	97.58	79.84	56.45
	162	54.03	62.10	53.23	44.35	37.10	38.71	37.90	71.77	87.10	96.77	83.06
	180	74.19	51.61	34.68	25.00	28.23	27.42	27.42	37.90	55.65	78.23	100.0

Table 4: Recognition rates when the probe data is coat wearing data.

		Probe set view(walking wearing a coat, cl01,cl02)										
		0	18	36	54	72	90	108	126	144	162	180
Gallery set view	0	44.35	29.03	22.58	15.32	11.29	10.48	13.71	12.90	17.74	25.00	29.03
	18	34.68	49.19	36.29	29.84	16.13	8.06	11.29	17.74	18.55	16.94	20.16
	36	21.77	46.77	46.77	41.94	29.84	20.16	19.35	24.19	15.32	15.32	15.32
	54	20.16	28.23	39.52	46.77	34.68	29.03	21.77	29.84	20.16	13.71	13.71
	72	13.71	27.42	33.06	37.10	49.19	37.90	29.84	25.00	19.35	14.52	12.10
	90	17.74	16.94	24.19	33.87	46.77	42.74	37.10	33.06	24.19	14.52	14.52
	108	16.94	19.35	27.42	30.65	41.94	40.32	46.77	41.94	32.26	24.19	12.10
	126	20.16	21.77	27.42	29.03	35.48	36.29	39.52	43.55	41.13	28.23	20.97
	144	17.74	19.35	21.77	25.00	23.39	17.74	22.58	34.68	40.32	31.45	22.58
	162	25.81	25.00	23.39	20.97	13.71	12.90	16.13	25.81	34.68	41.13	35.48
	180	26.61	20.16	16.94	16.94	15.32	8.87	12.10	17.74	25.00	30.65	42.74

Table 5: Recognition rates when the probe data is carrying a bag data.

		Probe set view(walking with a bag, bg01,bg02)										
		0	18	36	54	72	90	108	126	144	162	180
Gallery set view	0	79.84	58.87	45.97	23.39	16.13	10.48	12.90	18.55	25.00	40.32	45.97
	18	63.71	81.45	67.74	45.97	31.45	21.77	14.52	20.97	32.26	39.52	39.52
	36	39.52	71.77	70.16	62.10	41.13	28.23	17.74	24.19	37.90	38.71	30.65
	54	25.00	41.13	60.48	66.94	56.45	48.39	40.32	41.94	38.71	25.81	26.61
	72	25.81	28.23	48.39	66.13	74.19	63.71	57.26	59.68	40.32	23.39	25.00
	90	21.77	24.19	33.06	47.58	62.90	65.32	57.26	49.19	35.48	24.19	19.35
	108	23.39	26.61	35.48	52.42	63.71	61.29	62.10	65.32	50.81	25.81	25.00
	126	20.97	33.87	39.52	46.77	51.61	44.35	54.84	75.81	66.94	42.74	25.00
	144	28.23	30.65	36.29	39.52	30.65	24.19	30.65	56.45	72.58	49.19	37.10
	162	37.90	34.68	27.42	24.19	16.94	10.48	14.52	39.52	50.81	68.55	49.19
	180	54.03	36.29	26.61	18.55	18.55	15.32	12.90	23.39	31.45	47.58	74.19

4.5. Comparisons with GEI+PCA

Since GEIs are used as input and try to extract invariant feature, we first compare our method with GEI+PCA [4]. The experimental design about gallery sets and probe sets for GEI+PCA is exactly the same as ours in Table. 1. The first column of Figure 12 shows the comparison of recognition rates with GEI+PCA at different probe angles. For the limitation of space, we only list 5 probe angles with a 36° interval. The second column shows the comparison with different carrying conditions, and the third shows the comparison with different clothing. As illustrated in Figure 12, the proposed method outperforms GEI+PCA at all probe angle and gallery angle pairs. The results show that the proposed method can extract better gait feature which can be robust to view, clothing and carrying condition variations.

We also compared the recognition rates without view variant. By averaging the rates on the diagonal of Table 3, Table 4 and Table 5, the recognition rates without view variant can be computed. The corresponding average rates of GEI+PCA are also obtained in the same manner. The results are shown in Figure 13. When there is no variation, the proposed method achieve a high recognition rate which is almost the same with GEI+PCA. But when variation exists, the proposed method outperforms GEI+PCA greatly.

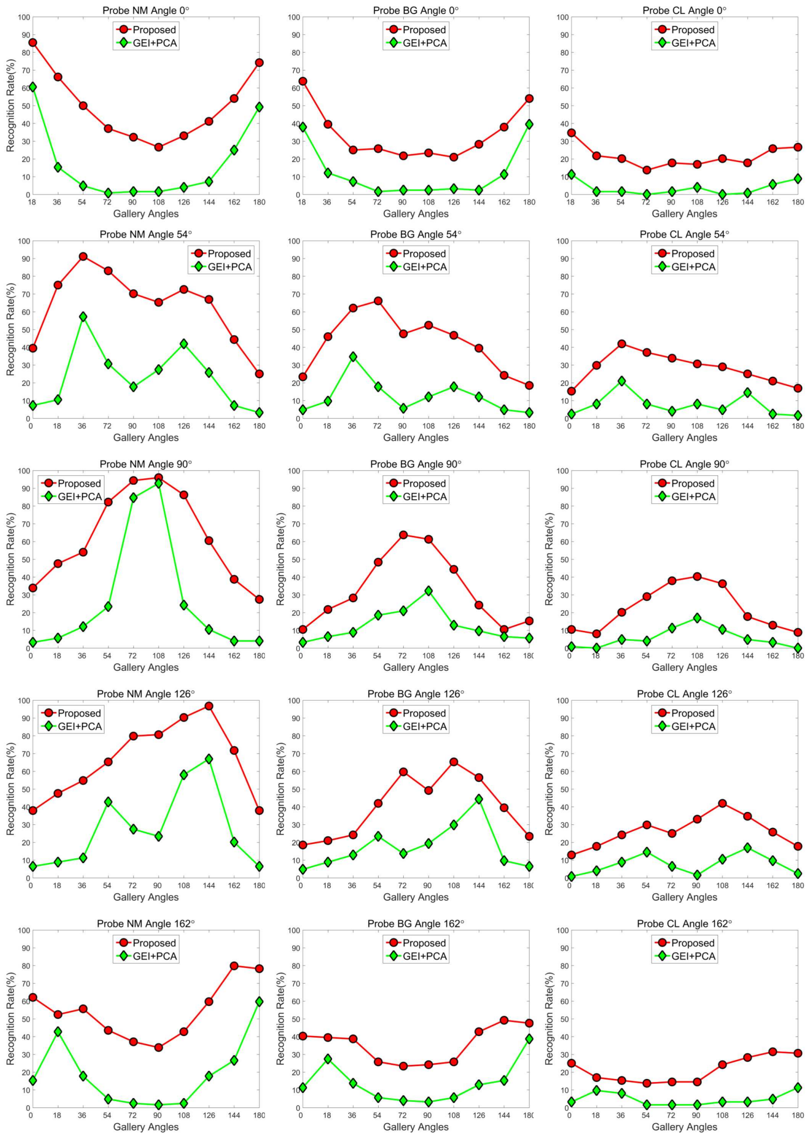


Figure 12: Comparison with GEI+PCA at different probe angles. The red lines are achieved by the proposed method.

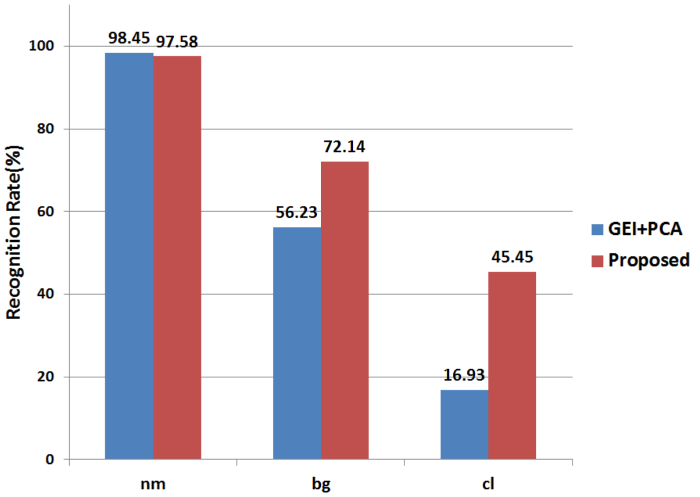


Figure 13: The average recognition rates compare with GEI+PCA. The red bars are achieved by the proposed method.

4.6. Experimental Results on SZU RGB-D Dataset

We also use SZU RGB-D dataset to evaluate the proposed method. Even only two views are in the dataset, we still use the proposed 7-layer model. The dataset does not contain clothing and carrying condition variations. The experiment setup is illustrated in Table 2. Sequence 01-02 are put into the gallery set. But the sequence 03-08 are split into two set. One probe set contains sequence 03-04 which are in side view and the same with the gallery set. The other probe set contains sequence 05-08 which are about 30 degrees from the side view.

The recognition rates of experiments on SZU RGB-D dataset are shown in Figure 14 and Figure 15. As shown in Figure 14, the recognition rate can be very high (over 97%) when there is no view variation. When view variation exists, the recognition rate will drop greatly. GEI+PCA only achieves a recognition rate of 27%. The proposed method is much better than GEI+PCA, and it achieves almost 70%. The experimental results on SZU RGB-D dataset also prove the effectiveness of the prosed method.

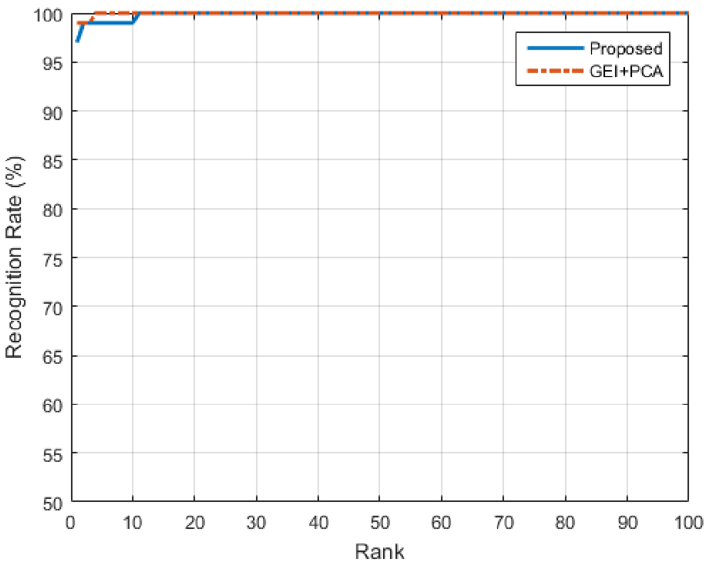


Figure 14: The recognition rates when the probe set contains sequence 03-04.

4.7. Comparison with the State-of-the-art

270 In order to better illustrate the performance of the proposed method, we also compare the proposed one with some state-of-the-art methods. To the best of our knowledge, we did not find methods what can extract invariant feature according to different variations. We compared the recognition rates with some view invariant methods. They are FD-VTM [8], RSVD-VTM [9], RPCA-VTM 275 [6], R-VTM [10], GP+CCA [15] and C3A [17].

The probe angles selected are 54°, 90° and 126° as in experiments of those methods. The experimental results are listed in Figure 16. From the results we can find that the proposed method outperforms others when the angle difference between the gallery and the probe is large. It proves that the model can handle 280 large view variation well. When the view variation is not large enough, the proposed method can also improve the recognition rate obviously.

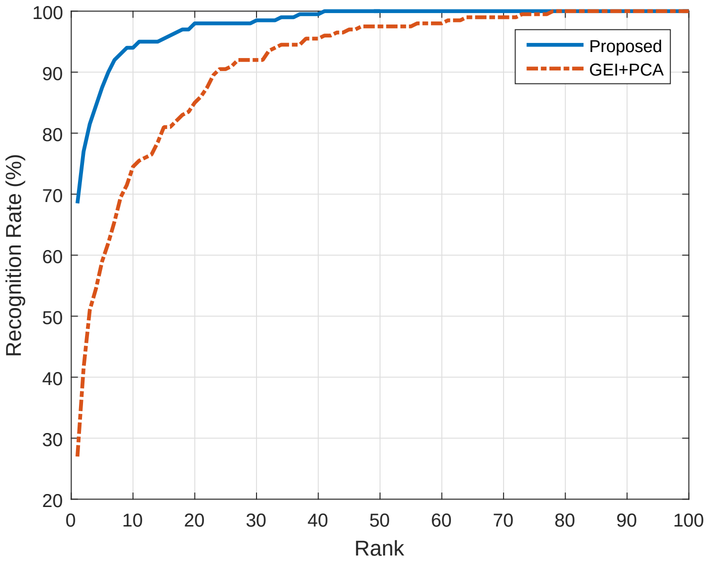
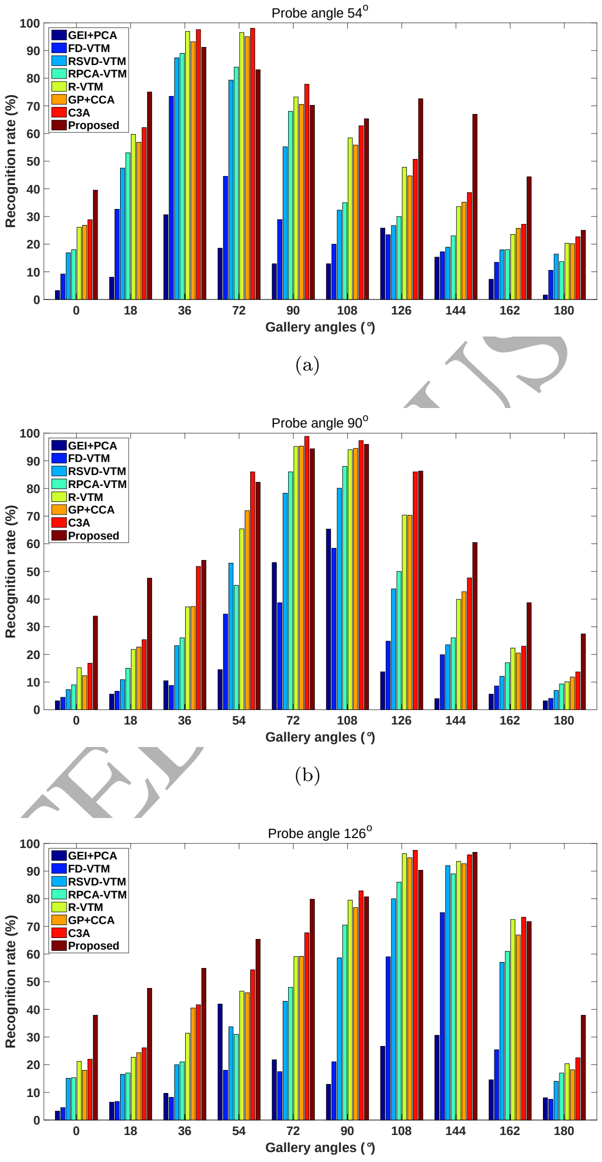


Figure 15: The recognition rates when the probe set contains sequence 05-08.

Table 6: Average recognition rates at probe angles (a)54°, (b)90° and (c)126°. The gallery angles are the rest 10 angles except the corresponding probe angle. The values in the right most column is the average of which at at probe angles (a)54°, (b)90° and (c)126°

	Probe angle			
Method	54°	90°	126°	Average
C3A [17]	56.64%	54.65%	58.38%	56.56%
ViDP [19]	64.2%	60.4%	65.0%	63.2%
CNN [20]	77.8%	64.9%	76.1%	72.9%
Proposed	63.31%	62.10%	66.29%	63.90%



(a)

(b)

(c)

Figure 16: Comparison with the State-of-the-art at probe angles (a)54°, (b)90° and (c)126°. The gallery angles are the rest 10 angles except the corresponding probe angle.

In Table 6 the experimental results of C3A [17], ViDP [19], CNN [20] and the proposed method are listed. Here we want to emphasis that the proposed method contains only one model for any views, and for clothing or carrying condition variations. ViDP can extract view invariant feature using only one linear model, but the recognition rate is not high enough especially when the view variation is large. CNN achieves the highest recognition rate but the supervised information is needed in training. In training step of the proposed method, the human identification labels are not needed. The proposed method can extract robust gait feature by synthesizing the side view data and remove the effect of different variations. Some methods such as C3A need to know the probe angle and gallery angles before to extract gait feature. That means that they have to train many models for all the views pairs (each view pair needs one model). Our method does not need to estimate the view angle, clothing type and carrying condition. It is more feasible in practical applications.

5. Conclusions and Future work

In this paper, we proposed a uniform model based on auto-encoders to extract invariant gait feature for gait recognition. The model could transform gait images at any view to the side view. If the gait images are in different clothing and carrying conditions, they will all be transformed to normal conditions (without coat and carrying objects). So we do not need to know the exact view angles between subjects and camera, and we also do not need to estimate the subjects' clothing types and carrying conditions. Experimental results show that the proposed model can improve recognition rate greatly especially when there is a large view variation and achieves state-of-the-art performance. It is very suitable for practical applications in surveillance.

In future, we will extend this model to deal with more challenging variations. Currently the view only changing in one dimension. It is that it is changed from the frontal view to the side view and back view. Looking down view as in most surveillance systems should also be involved. Besides, a large dataset (such as contains over 1,000 subjects) collected in a real video surveillance system should also be created to evaluate effectiveness of methods.

Acknowledgment

The authors would like to thank Prof. Meina Kan for her valuable suggestions and Dr. Xianglei Xing for his support on experimental results of some methods. The work is supported by Natural Science Foundation of China (Grant No. 61672357 and 61420106015), the Science Foundation of Guangdong Province (Grant No. 2014A030313556), the Science Foundation of Shenzhen (Grant No. JCYJ20150324141711699 and JCYJ20160422144110140), Strategic Priority Research Program of the CAS (Grant No. XDB02070001) and Youth Innovation Promotion Association CAS (Grant No. 2006121).

References

[1] L. Wang, H. Ning, T. Tan, W. Hu, Fusion of static and dynamic body biometrics for gait recognition, IEEE Transactions on Circuits and Systems for Video Technology 14 (2) (2004) 149–158.

[2] R. Tanawongsuwan, A. Bobick, Gait recognition from time-normalized joint-angle trajectories in the walking plane, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR2001), Vol. 2, 2001.

[3] F. Tafazzoli, R. Safabakhsh, Model-based human gait recognition using leg and arm movements, Engineering Applications of Artificial Intelligence 23 (8) (2010) 1237–1246.

[4] J. Han, B. Bhanu, Individual recognition using gait energy image, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (2) (2006) 316–322.

[5] S. Yu, L. Wang, W. Hu, T. Tan, Gait analysis for human identification in frequency domain, in: the 3rd International Conference on Image and Graphics, Hong Kong, China, 2004, pp. 282–285.

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

[7] M. Hu, Y. Wang, Z. Zhang, D. Zhang, Multi-view multi-stance gait identification, in: Prof. of the 18th IEEE International Conference on Image Processing, 2011, pp. 541–544.

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

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

[10] W. Kusakunniran, Q. Wu, J. Zhang, H. Li, Gait recognition under various viewing angles based on correlated motion regression, IEEE Transactions on Circuits and Systems for Video Technology 22 (6) (2012) 966–980.

[11] X. Ben, W. Meng, R. Yan, K. Wang, An improved biometrics technique based on metric learning approach, Neurocomputing 97 (2012) 44 – 51.

[12] X. Ben, P. Zhang, W. Meng, R. Yan, M. Yang, W. Liu, H. Zhang, On the distance metric learning between cross-domain gaits, Neurocomputing 208 (2016) 153 – 164.

[13] A. Y. Johnson, A. F. Bobick, A multi-view method for gait recognition using static body parameters, in: Proc. of 3rd International Conference

on Audio and Video Based Biometric Person Authentication, Halmstad, Sweden, 2001, pp. 301–311.

[14] A. Kale, A. K. R. Chowdhury, R. Chellappa, Towards a view invariant gait recognition algorithm, in: IEEE Conference on Advanced Video and Signal Based Surveillance, Guildford, UK, 2003, pp. 143–150.

[15] K. Bashir, T. Xiang, S. Gong, Cross-view gait recognition using correlation strength, in: British Machine Vision Conference, BMVC, Aberystwyth, United kingdom, 2010.

[16] C. Luo, W. Xu, C. Zhu, Robust gait recognition based on partitioning and canonical correlation analysis, in: IEEE International Conference on Imaging Systems and Techniques, 2015.

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

[18] J. Lu, G. Wang, P. Moulin, Human identity and gender recognition from gait sequences with arbitrary walking directions, IEEE Transactions on Information Forensics and Security 9 (1) (2014) 51–61.

[19] M. Hu, Y. Wang, Z. Zhang, J. J. Little, D. Huang, View-invariant discriminative projection for multi-view gait-based human identification, IEEE Transactions on Information Forensics and Security 8 (12) (2013) 2034–2045.

[20] Z. Wu, Y. Huang, L. Wang, X. Wang, T. Tan, A comprehensive study on cross-view gait based human identification with deep cnns, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (2) (2017) 209–226.

[21] M. A. Hossain, Y. Makihara, J. Wang, Y. Yagi, Clothing-invariant gait identification using part-based clothing categorization and adaptive weight control, Pattern Recognition 43 (6) (2010) 2281–2291. doi:http://dx.doi.org/10.1016/j.patcog.2009.12.020.

[22] Y. Guan, C. T. Li, Y. Hu, Robust clothing-invariant gait recognition, in: Proc. of the 8th International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2012, pp. 321–324.

[23] M. Kan, S. Shan, H. Chang, X. Chen, Stacked progressive auto-encoders (spae) for face recognition across poses, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1883–1890.

[24] S. Yu, D. Tan, T. Tan, A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition, in: Proc. of the 18'th International Conference on Pattern Recognition (ICPR06), Hong Kong, China, 2006, pp. 441–444.

[25] G. E. Hinton, R. R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.

[26] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast feature embedding, arXiv preprint arXiv:1408.5093.

[27] Caffe Website. http://caffe.berkeleyvision.org/.

[28] S. Yu, Q. Wang, Y. Huang, A large rgb-d gait dataset and the baseline algorithm, in: Proc. of the 8th Chinese Conference on Biometric Recognition(CCBR2013), 2013, pp. 417–424.

Biography

Shiqi Yu received his B.E. degree in computer science and engineering from the Chu Kochen Honors College, Zhejiang University in 2002, and Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences in 2007. He worked as an assistant professor and then as an associate professor in the Shenzhen Institutes of Advanced Technology, Chinese Academy of Science from 2007 to 2010. Currently, he is an associate professor in the College of Computer Science and Software Engineering, Shenzhen University, China. He especially focuses on image classification and related research topics.

Haifeng Chen received his B.S. degree in computer science and engineering from Qufu Normal University, China, in 2013. He is currently a master student in the College of Computer Science and Software Engineering, Shenzhen University, China. His research interests include computer vision and deep learning.

Qing Wang received his B.S. and B.E. degrees in computer science and engineering and pattern recognition from the College of Computer Science and Software Engineering, Shenzhen University, China in 2013 and 2016 respectively. Now He works as a researcher in Tencent Company, China. His research interests include computer vision and deep learning.

Linlin Shen received the B.Sc. degree from Shanghai Jiaotong University, China, and the Ph.D. degree from the University of Nottingham, Nottingham, England. He was a Research Fellow with University of Nottingham, working on MRI brain image processing. He is currently a Professor and the Director of the Computer Vision Institute, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. His research interests include Gabor wavelets, facial recognition, analysis/synthesis and medical image processing. Prof. Shen was listed as the Most Cited Chinese Researcher by Elsevier, he received the Most Cited Paper Award from the journal of Imange & Vision Computing. His cell classificaiton algorithm was the winners of International Contest on Pattern Recognition Techniques for Indirect Immunofluorescence Images held by ICIP 2013 and ICPR 2016.

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