# SIAMESE NEURAL NETWORK BASED GAIT RECOGNITION FOR HUMAN IDENTIFICATION

*Cheng Zhang, Wu Liu, Huadong Ma, Huiyuan Fu*

Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia,
Beijing University of Posts and Telecommunications, Beijing, China, 100876
{zhangcheng,liuwu,mhd,fhy}@bupt.edu.cn

## ABSTRACT

As the remarkable characteristics of remote accessed, robust and security, gait recognition has gained significant attention in the biometrics based human identification task. However, the existed methods mainly employ the hand-crafted gait features, which cannot well handle the indistinctive inter-class differences and large intra-class variations of human gait in real-world situation. In this paper, we have developed a Siamese neural network based gait recognition framework to automatically extract robust and discriminative gait features for human identification. Different from conventional deep neural network, the Siamese network can employ distance metric learning to drive the similarity metric to be small for pairs of gait from the same person, and large for pairs from different persons. In particular, to further learn effective model with limited training data, we composite the gait energy images instead of raw sequence of gaits. Consequently, the experiments on the world's largest gait database show our framework impressively outperforms state-of-the-arts.

***Index Terms***— Gait recognition, Siamese neural network, gait energy image, feature learning

## 1. INTRODUCTION

Biometrics based automatic human identification is one of the most fundamental and significant research topic in computer vision field. Among the massive biometric authentication traits, the discrimination of human gait is strongly supported in the research of biomechanics, physical medicine studies, and psychological studies [1]. Furthermore, compared with other biometrics (*e.g.,* facial, iris, fingerprint, and voice), human gait gives more attractive characteristics: 1) *remote accessed*—it can identify subjects from a distance without interrupting the subjects; 2) *robust*—even in low resolution videos, the gait still works well; 3) *security*—it is difficult to imitate or camouflage human gait. Therefore, gait recognition, which aims essentially to discriminate individuals by the way they walk, has gained significant attention.

However, accurate gait recognition is still a challenging work as 1) the unconspicuous inter-class differences from different people; and 2) the large intra-class variations from the same person as the different walking speeds, viewpoints, clothing, and belongings. To solve these challenges, two kinds of gait recognition methods are studied: model-based and appearance-based. Model-based approaches [2, 3] directly extract human body structure from the images with higher resolution images of a subject as well as higher computational cost. Differently, appearance-based methods [4, 5, 6, 7, 8] mainly fucus on extracting gait features from captured image sequences regardless of the underlying structure. Therefore, this kind of methods can perform recognition at lower resolutions, which makes them suitable for outdoor applications when the parameters of the body structure are difficult to precisely estimate. Nonetheless, the human-crafted gait features in the existed methods can extremely hard to break through feature representation bottleneck when facing with the gait and appearance changes of a walking person with massive kinds of walking speed, viewpoint, clothing, and carrying.

Recently, for extracting the robust feature, deep learning method is well known for its superiority than traditional methods in plenty of fields [9, 10, 11, 12, 13]. For example, the Convolution Neural Networks (CNN) can automatically learn commendable features from the given training images, which significantly improve the image classification accuracy. However, to learn sufficient features, the CNN requires a mass of training data for all the categories. Conversely, in the area of gait recognition, the number of subjects is very large (*e.g.*, hundreds or thousands), with only a few examples per subject [5, 8, 14, 15]. Besides, gait recognition for human identification is not a typical classification problem [16, 17]. Therefore, we cannot directly use CNN on gait recognition as the huge domain gap between them.

Motivated by the above observations, and aiming to address the aforementioned gait recognition challenges, we proposed a Siamese neural network based gait recognition for human identification. First of all, to solve the data limitation problem, we use Gait Energy Image (GEI) [4] instead of raw sequence of human gait. As removing most noisy information while keeping the major human shapes and body

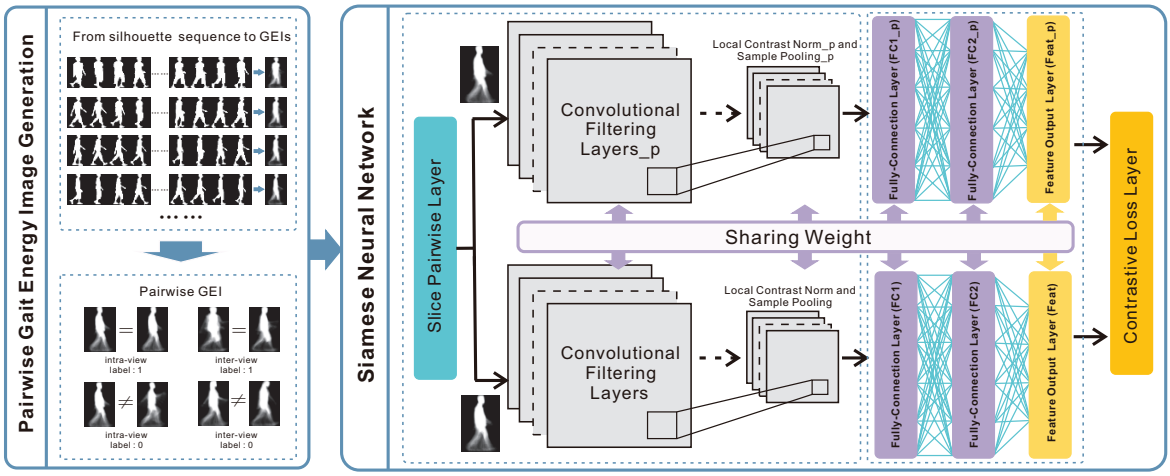


**Fig. 1**. The framework of our proposed Siamese neural network based gait recognition for human identification.

changes during walking, the GEI representation can help deep neural network quickly capture the discriminative biometrics information in human gait. In particular, aiming to further learn sufficient feature representations to tackle gait recognition for human identification, we exploit the siamese neural network [18, 19], which can simultaneously minimize the distance between similar subjects and maximize the distance between dissimilar pairs with a distance metric learning architecture. With the well-learned gait features, the K-Nearest Neighbor (KNN) method [20] is used to identify the same person in surveillance environment. Finally, the evaluations on the world's largest and most comprehensive gait benchmark dataset demonstrate that the proposed method can impressively beyond state-of-the-art with nearly 5% improvement in intra-degree human identification.

In summary, this paper makes the following contributions:

- To the best of our knowledge, we present one of the first attempts to study the deep neural network based gait recognition framework for human identification with distance metric learning.
- In the end-to-end framework, we leverage the competitive GEI presentation as the input of network while holistically exploit the Siamese neural network to learn effective feature representations for human identification.
- The comprehensive evaluations show that we impressively outperform the state-of-the-arts on the world's largest challenge gait benchmark dataset.

## 2. THE PROPOSED METHOD

Fig. 1 gives an overview of the proposed framework by applying Siamese network based gait recognition to identify person. Firstly, we combine the raw sequence of surveillance images into GEIs, which are used as the input of the deep neural network. Next, instead of the conventional CNN, we use Siamese network to learn sufficient feature representations of gait for human identification. The Siamese neural network contains two parallel CNN architectures sharing the same parameters. In the training stage, the two images in the similar or dissimilar pairs separately entrance the two parallel CNNs. Then the output of the CNNs are combined by the contrastive layers to compute the contrastive loss. After that, the back-propagating with contrastive loss is used to fine-tune the model. In the testing stage, for one query gait sequence, the GEI is composted and only one CNN is used to extract the gait feature. Finally, we use the KNN method to identify the person with the similar human gaits. Next, we will describe each component in detail.

### 2.1. Gait Energy Image

To represent one sequence of human gait recorded by surveillance camera, we can extract the silhouette of human [21] and average them into the GEI representation. Here, we use GEI as it represents a human motion sequence in a single image while preserving temporal information. Besides, this averaging operation cannot only well maintain the original information of the gait sequences, but also be robust to incidental silhouette errors in individual image. Consequently, GEI is popular applied in various gait analysis tasks [6, 8]. Some examples of the GEI extracting process can be found in Fig.1. Next, the computed GEIs are feeded into our deep architecture to further learn the gait features.

### 2.2. Conventional CNN based Gait Recognition

To begin with, we attempt to fine-tune the conventional CNN on the gait dataset for gait recognition based human identification task as 1) CNN is able to learn discriminative features automatically by exploring deep architecture at multiple level of abstracts from raw data, without any domain knowledge; and 2) fine-tuning from a pre-trained model is a good solution to solve the data limitation problem and speed up the convergence of new model.

Here, we employ the CNN architecture as discussed in [10] and only change the 1,000 label output to the number

of subjects (i.e., 3835) in the gait dataset. Next, we take Caffe-BLVC model[1] to initialize the network as its good performance on ImageNet dataset. In the training process, we fix all convolutional layers and only fine-tune the fully connected layers. After training, we take the activations of three fully connected layers (CNN.FC1, CNN.FC2, and CNN.FC3) as the feature representations and employ the KNN method to identify the same person in surveillance environment. According to our comprehensive experiments, we find that the "CNN.FC1" give the best performance for gait recognition.

However, as we discussed before, the conventional CNN-based method treats gait recognition as a classification problem with 3835 categories, which neglects the huge domain gap between classification and recognition. Furthermore, as the conflict between large categories number and small samples per category, the CNN-based method cannot effectively solve the gait recognition based human identification task. In order to solve this problem, we propose the Siamese network based framework in the next subsection.

### 2.3. Siamese Network based Gait Recognition

The Siamese network was first introduced in [18, 19] to be applied to face and signature verification tasks. The main idea of the network is to learn a function that maps input patterns into a latent space where similarity metric to be small for pairs of the same objects, and large for pairs from different ones. Therefore, the network is best suited for verification scenarios where the number of classes is very large, and/or examples of all the classes are not available at the time of training. Definitely, gait recognition is one of such verification scenarios.

As shown in Fig. 1, the Siamese neural network designed for gait recognition contains two parallel CNN architectures, which of them consists of two parts: 1) the two convolution layers and max-pooling layers, and 2) three fully connection layers. This network accepts inputs of size 128×88×3 pixels. Using shorthand notation, the full architecture of each branch is $C(20,5,1)$-$N$-$P$-$C(50,5,1)$-$N$-$P$-$FC(500)$-$FC(10)$-$FC(2)$, where $C(d,f,s)$ indicates a convolutional layer with $d$ filters of spatial size $f\times f$, applied to the input with stride $s$. $FC(n)$ is a fully connected layer with $n$ nodes. All max-pooling layers $P$ pool spatially in non-overlapping 2×2 regions and all normalization layers $N$ use the same parameters: $n = 5$, $alpha = 10^{-4}$, $beta = 0.75$. The final feature output layers are connected to a contrastive loss layer. In the training stage, the two branches of the network will be optimized simultaneously with the weight sharing mechanism. Pairwise images with similar or dissimilar labels separately entrance the two CNNs. Then the output of the CNNs are combined by the contrastive layers to compute the contrastive loss. After that, the back-propagating with contrastive loss is used to fine-tune the model.

Specifically, consider a pair of GEIs $x_1$ and $x_2$, let $y$ be a binary label of the pair, $y = 1$ if the images $x_1$ and $x_2$ belong to the same subject (*i.e.*, "genuine pair") and $y = 0$ otherwise (*i.e.*, "impostor pair"). $W$ is the shared parameter matrix throughout the Siamese architecture which needs to be learned. We can use $W$ to map $x_1$ and $x_2$ into $S_W(x_1)$ and $S_W(x_2)$, which are the two points in the latent low-dimensional space. Then the distance $E_W(x_1,x_2)$ between $x_1$ and $x_2$ can be measured by:

$$E_W(x_1,x_2) = ||S_W(x_1) - S_W(x_2)||_2^2. \quad (1)$$

We can define the contrastive loss function as follows:

$$\mathcal{L}(W) = \sum_{i=1}^{P} L(W,(y,x_1,x_2)^i), \quad (2)$$

$$\begin{aligned} L(W,(y,x_1,x_2)^i) =& (1-y)\cdot max(m - E_W(x_1,x_2)^i, 0) \\ &+ y\cdot E_W(x_1,x_2)^i \end{aligned} \quad (3)$$

where $(y,x_1,x_2)^i$ is the *i*-th pair, which is composed of a pair of GEIs with corresponding label $y$, $P$ is the number of the training pairs. The positive number $m$ can be interpreted as margin.

In the implementation, the training set is selected from OULP-C1V1-A-Gallery [8] dataset, with 20,000 similar GEI pairs and randomly selected 20,000 dissimilar pairs. In the testing stage, we send the query GEI into one of the CNNs, then compute the feedforward network based on the matrix multiplication for one time to extract features, the whole scheme will be very efficient. In the following experiments, we denote the feature representation as "SiaNet.FC".

## 3. EXPERIMENTS

### 3.1. Data Preparation

In our experiment, we evaluate the proposed approach on the OULP-C1V1-A dataset from the OU-ISIR LP gait benchmark [8]. Compared with other database [14, 5], OU-ISIR LP contains the world's largest number of subjects with a wide age and an almost balanced gender ratio. The dataset records two sequences for each subject: probe (*i.e.*, query) sequence and gallery (*i.e.*, source) sequence. The sequences are constituted by silhouette images, which are normalized into 128×88 pixels. Some examples of the dataset can be found in Fig. 1. In addition, each sequence is further divided into 4 slices based on the observation angles (55°, 65°, 75°, 85°) and 1 slice including all four angles (named All). Moreover, a standard directory structure of gallery and probe offers fair comparison test bed. In the experiment, we use gallery for training. Specifically, there consists multiple gait cycles for each subject in the gallery set. Certainly, no probe images are used in any training stage. In the following experiments, we test both intra-degree and inter-degree with the testing protocol in [8].

In particular, we use rank-1 and rank-5 identification rates as evaluation measures, which denote the percentages of correct subjects out of all the subjects appearing within the first and fifth ranks respectively.

[1] "Caffe Model Zoo," http://caffe.berkeleyvision.org/model_zoo.html

| | Rank-1 Identification Rate (%) | | | | | Rank-5 Identification Rate (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | 55° | 65° | 75° | 85° | All | 55° | 65° | 75° | 85° | All |
| HWLD [6] | — | — | — | 87.70 | 95.50 | — | — | — | 94.70 | **98.50** |
| GEI [8] | 84.70 | 86.63 | 86.91 | 85.72 | 94.24 | 92.39 | 92.84 | 92.78 | 93.01 | 97.13 |
| FDF [8] | 83.89 | 85.49 | 86.59 | 85.90 | 94.17 | 91.53 | 92.81 | 92.88 | 92.83 | 97.10 |
| CNN.FC1 | 73.96 | 76.71 | 77.87 | 78.82 | 86.09 | 86.64 | 88.67 | 89.39 | 90.09 | 93.56 |
| **SiaNet.FC** | **90.12** | **91.14** | **91.18** | **90.43** | **96.02** | **94.98** | **95.90** | **95.92** | **95.97** | 98.31 |

**Table 1**. Comparison results of different methods in term of the Rank-1 and Rank-5 Identification Rates

### 3.2. Evaluation on Intra-view Human Identification

We first evaluate our approach for intra-view identification task. Table. 1 shows the comparison results of our approach and state-of-the-art gait recognition methods, including GEI and FDF template matching strategy [8], HWLD [6] and CNN-based method. Note, [6] only tests their HWLD method in near-profile view set (85) and all the angles set (All). As we can see, our Siamese network based method is better than the other techniques in all tests cases and achieves the state-of-the-art performance. Compared to hand-crafted feature matching method GEI, FDF and HWLD, our method can obviously improve the identification rate, which demonstrates that the proposed Siamese network based method can automatically learn commendable features from the given GEIs. For example, cascade convolutional architecture of Siamese network has ability to capture the massive complexity structure around several ambiguous regions such as human neck, hip and shoulder. Furthermore, compared with traditional CNN-based method, our Siamese network based method also achieves much better accuracy because its distance metric learning architecture can well solve the verification scenario of gait recognition based human identification.

### 3.3. Evaluation on Inter-view Human Identification

In the other hand, a change in view between query gait sequences and source samples occurs frequently in the real world human identification scenario. Therefore, we verify the robustness of our proposed methods in cross-view gait recognition task. We select 4 types of inter-view tests and compare with recently published works including AVTM_PdVS [7], AVTM [7], woVTM [7] and RankSVM [22]. The first three approaches are particularly designed for gait-based human identification with cross-view matching by applying an extra 3D gait volume for training view transformation model. RankSVM is a typical representative metric learning-based approach aiming at gait recognition task with covariate variations especially for crow-view matching. Note that, the four matchers only selected 1,912 subjects in OU-ISIR LP dataset, whose data were captured for evaluation by calibrated cameras for testing. Differently, we evaluate our method on the whole 3,835 persons set, which is more difficult. As shown in Fig. 2, the performance improvements of our method (SiaNet) are consistent and stable, *i.e.,* all the cross-view test cases improved compared to other methods. It demonstrates that

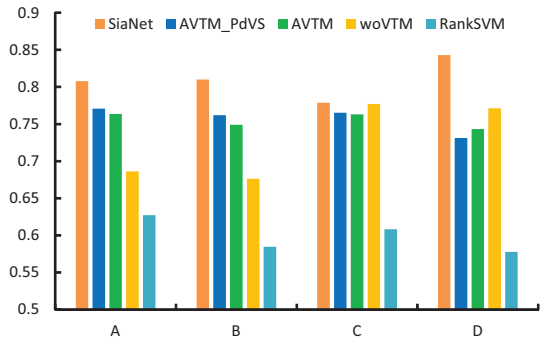


**Fig. 2**. Comparison of the cross-view matching approaches on different types of inter-view test (in terms of rank-1 identification rate). Group A∼D stand for (65,75), (75,65), (75,85) and (85,75).

proposed method is quite robust to view-change variations in this four testing groups. The reason is that in the training stage, we send both intra- and inter-view pairwise GEIs into the Siamese neural network. In this way, we can train the similarity metric to be small for pairs from same subjects, and large for pairs from different subjects, which enhance the robustness of the gait-based human identification method under the view-change condition.

## 4. CONCLUSION

In this paper, we have investigated on leveraging Siamese neural network to extract robust and discriminative gait recognition features for human identification. In the framework, the competitive GEI representation is utilized to remove most noise while keeping the major human shapes and body changes during walking. More important, the Siamese neural network is employed to holistically exploit the effective features with directly computing the similarity between two human gaits with parallel Convolution Neural Networks architecture. Experiment results on the benchmark dataset demonstrate the effectiveness and efficient of our proposed method. In the future, we will try to train 3-Dimensional Siamese neural network with more training dataset to further improve the performance of the gait recognition.

**Acknowledgements** This work is supported by the Fund for Creative Research Groups of China (61421061), the National Natural Science Foundation of China (61402048), and the NSFC-Guangdong Joint Fund (U1501254).

## 5. REFERENCES


[1] L. Wang, T. Tan, H. Ning, and W. Hu, "Silhouette analysis-based gait recognition for human identification," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1505–1518, 2003.

[2] C. Yam, M. S. Nixon, and J. N. Carter, "Automated person recognition by walking and running via model-based approaches," *Pattern Recognition*, vol. 37, no. 5, pp. 1057–1072, 2004.

[3] G. Ariyanto and M. S. Nixon, "Model-based 3d gait biometrics," in *Proc. of International Joint Conference on Biometrics*. IEEE, 2011, pp. 1–7.

[4] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 316–322, 2006.

[5] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer, "The humanid gait challenge problem: Data sets, performance, and analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 2, pp. 162–177, 2005.

[6] S. Sivapalan, D. Chen, S. Denman, S. Sridharan, and C. Fookes, "Histogram of weighted local directions for gait recognition," in *Proc. of Computer Vision and Pattern Recognition Workshop*. IEEE, 2013, pp. 125–130.

[7] D. Muramatsu, A. Shiraishi, Y. Makihara, M. Uddin, and Y. Yagi, "Gait-based person recognition using arbitrary view transformation model," *IEEE Trans. on Image Processing*, vol. 24, no. 1, pp. 140–154, 2015.

[8] H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi, "The ou-isir gait database comprising the large population dataset and performance evaluation of gait recognition," *IEEE Trans. on Information Forensics and Security*, vol. 7, no. 5, pp. 1511–1521, 2012.

[9] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[11] W. Liu, T. Mei, Y. Zhang, C. Che, and J. Luo, "Multi-task deep visual-semantic embedding for video thumbnail selection," in *Proc. of Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 3707–3715.

[12] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems*, 2014, pp. 3104–3112.

[13] B. Wang, S. Tang, R. Zhao, W. Liu, and Y. Cen, "Pedestrian detection based on region proposal fusion," in *Proc. of International Workshop on Multimedia Signal Processing*. IEEE, 2015, pp. 1–6.

[14] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *Proc. of International Conference on Pattern Recognition*. IEEE, 2006, vol. 4, pp. 441–444.

[15] H. Ma, C. Zeng, and C. X. Ling, "A reliable people counting system via multiple cameras," *ACM Trans. on Intelligent Systems and Technology*, vol. 3, no. 2, pp. 31, 2012.

[16] Z. Zha, T. Mei, Z. Wang, and X. Hua, "Building a comprehensive ontology to refine video concept detection," in *Proc. of the International Workshop on Multimedia Information Retrieval*. ACM, 2007, pp. 227–236.

[17] X. Yuan, W. Lai, T. Mei, X. Hua, X. Wu, and S. Li, "Automatic video genre categorization using hierarchical svm," in *Proc. of International Conference on Image Processing*. IEEE, 2006, pp. 2905–2908.

[18] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. of Computer Vision and Pattern Recognition*. IEEE, 2005, vol. 1, pp. 539–546.

[19] J. Bromley, J. W Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah, "Signature verification using a siamese time delay neural network," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 7, no. 04, pp. 669–688, 1993.

[20] M. Muja and D. G. Lowe, "Fast matching of binary features," in *Proc. of Computer and Robot Vision*, 2012, pp. 404–410.

[21] Y. Boykov and M. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in nd images," in *Proc. of International Conference on Computer Vision*. IEEE, 2001, vol. 1, pp. 105–112.

[22] R. Martín-Félez and T. Xiang, "Gait recognition by ranking," in *Proc. of European Conference on Computer Vision*, pp. 328–341. Springer, 2012.