A Comprehensive Study on Cross-View Gait Based Human Identification with Deep CNNs

Zifeng Wu, Yongzhen Huang, *Member, IEEE*, Liang Wang, *Senior Member, IEEE*,
Xiaogang Wang, *Member, IEEE*, and Tieniu Tan, *Fellow, IEEE*

Abstract—This paper studies an approach to gait based human identification via similarity learning by deep convolutional neural networks (CNNs). With a pretty small group of labeled multi-view human walking videos, we can train deep networks to recognize the most discriminative changes of gait patterns which suggest the change of human identity. To the best of our knowledge, this is the first work based on deep CNNs for gait recognition in the literature. Here, we provide an extensive empirical evaluation in terms of various scenarios, namely, cross-view and cross-walking-condition, with different preprocessing approaches and network architectures. The method is first evaluated on the challenging CASIA-B dataset in terms of cross-view gait recognition. Experimental results show that it outperforms the previous state-of-the-art methods by a significant margin. In particular, our method shows advantages when the cross-view angle is large, i.e., no less than 36 degree. And the average recognition rate can reach 94 percent, much better than the previous best result (less than 65 percent). The method is further evaluated on the OU-ISIR gait dataset to test its generalization ability to larger data. OU-ISIR is currently the largest dataset available in the literature for gait recognition, with 4,007 subjects. On this dataset, the average accuracy of our method under identical view conditions is above 98 percent, and the one for cross-view scenarios is above 91 percent. Finally, the method also performs the best on the USF gait dataset, whose gait sequences are imaged in a real outdoor scene. These results show great potential of this method for practical applications.

Index Terms—Deep learning, CNN, human identification, gait, cross-view

1 Introduction

Gait is a kind of behavioral biometric feature, whose raw data are video sequences presenting walking people. It is particularly suitable for long-distance human identification, and requires no explicit co-operation by subjects, compared with other kinds of biometric features such as fingerprint and iris. Consequently, the gait feature can be captured more easily in long-distance and uncontrolled scenarios. Nowadays, a large quantity of cameras for video surveillance are installed in various places such as streets, stations, airports, shopping malls, office buildings and even private houses, which enables the gait recognition technology to be one of the useful tools for crime prevention and forensic identification. Gait analysis has already contributed to evidence collection for convicting criminals in Denmark [1] and UK [2].

For vision-based gait recognition, one of the biggest challenges is to disentangle the identity-unrelated factors which yet alter gait appearances drastically. These factors can be grouped into subject-related ones such as walking speed, dressing and carrying conditions, device-related ones such as different frame rates and filming resolutions, and environment-related ones such as illumination conditions and camera viewpoints. Among these, the change of viewpoints would be one of the most tricky factors. Recently, cross-view variance has become one of the key problems in many video-related tasks. It is usually the case that the performance of an approach ignoring cross-view variations would drop drastically when the viewpoint changes. For example, action recognition rates on the INRIA Xmas Motion Acquisition Sequences dataset [3] can drop from 80.0 percent [3] to lower than 20.0 percent [4], when evaluated under cross-view conditions. The cause is that the appearances of objects can be substantially altered, leading to intra-class variations larger than inter-class variations. This is also true for most state-of-the-art gait recognition approaches, which are based on gait energy images (GEI) [5]. These are obtained by averaging properly aligned human silhouettes in gait sequences. Example GEIs belonging to two subjects are shown in Fig. 1. Given a probe in Column b, it is not hard to tell the more similar gallery GEI in Column a, according to the shapes and poses of the average silhouettes. However, it will not be the case for those probes in Columns c-j, with view angle and/or walking condition variations. Global changes in appearances are presented in Fig. 1, when the cross-view angle is large, e.g., up to 54 degree (from 36 degree in Column a to 90 degree in Column d). As a result, there will be a drastic negative impact on gait recognition, as reported by Yu et al. [6].

Approaches to cross-view gait recognition in the recent literature can be roughly grouped into three categories. The

- *Z. Wu is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and the University of Adelaide. E-mail: zifeng.wu@adelaide.edu.au.*
- *Y. Huang, L. Wang, and T. Tan are with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China. E-mail: {yzhuang, wangliang, tnt}@nlpr.ia.ac.cn.*
- *X. Wang is with the Chinese University of Hong Kong, China. E-mail: xgwang@ee.cuhk.edu.hk.*

Manuscript received 27 Sept. 2014; revised 17 Jan. 2016; accepted 9 Feb. 2016. Date of publication 22 Mar. 2016; date of current version 16 Jan. 2017.
Recommended for acceptance by A. Vedaldi.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2016.2545669

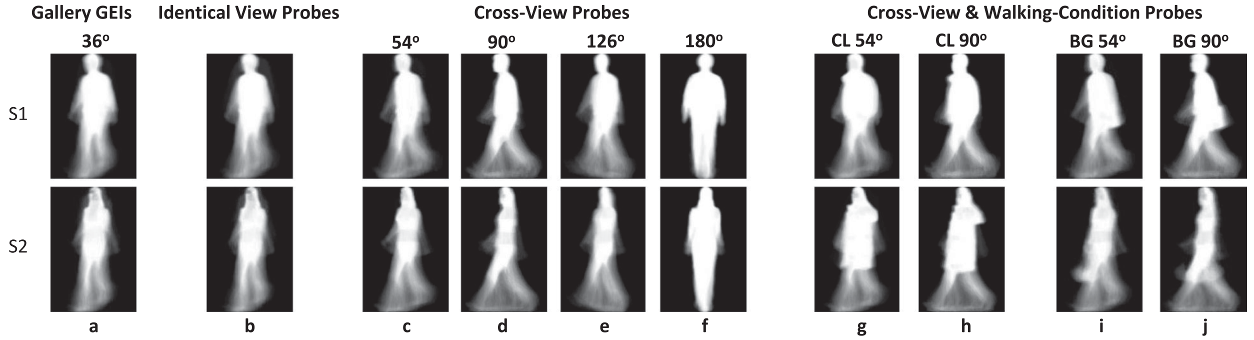


Fig. 1. Example GEIs of two subjects (S1-S2) in the CASIA-B gait dataset [6]. Column a: GEIs in the gallery, with view angle 36 degree . Column b: Probes with the same view angle. Columns c-f: Probes with view angle variations. Columns g&h: Probes with view angle and clothing variations. Columns i&j: Probes with view angle and carrying condition variations. Gait recognition amounts to identifying the most similar gallery GEI for each probe.

first category is to reconstruct the 3D structure of a human body, so that it can generate arbitrary 2D views by projecting the 3D model [7], [8]. Approaches of this kind can achieve promising performance, but they usually require multiple calibrated cameras under fully-controlled and cooperative environments. The recent development of 3D technology has brought us economic depth sensors, e.g., the Kinect. However, there are many more scenarios where only 2D cameras are available, which limits the application of these approaches in practice. The second category is to use handcrafted view-invariant features for gait recognition [9], [10]. These approaches can perform well for their specific scenarios, but usually it is hard to generalize for other cases. The last category amounts to learning the cross-view projections [11], [12], with which one can normalize gait features from one view to another, or to one or more common canonical views, e.g., the side view. In this way, one can compare the normalized gait features extracted from any two videos in order to compute their similarity. Empirical results reported by Kusakunniran et al. [12] showed that these approaches can obtain good results when the cross-view angle is relatively small, i.e., 18 degree. However, when the angle increases up to 36 degree, the performance will drop drastically.

On the other hand, deep learning has become flourishing in the computer vision community recently. Particularly, the deep convolutional neural networks (CNNs) [13] were used to tackle with various tasks, updating the record scores one after another. To name a few image-based milestone works, Krizhevsky et al. [14] trained a very deep 1,000-way image classifier, which is a network with five convolution and three fully-connected layers. Sermanet et al. [15] established an integrated framework for image classification and object detection based on multi-scale deep CNNs. Farabet et al. [16] dealt with the scene labeling task using hierarchical features learned via CNNs. On these tasks, CNN has shown overwhelming advantages over classic approaches due to its deep and highly non-linear model, which can learn rich features in a discriminative manner given sufficient labeled training data [14]. As for video-related works, Karpathy et al. [17] applied CNNs to large-scale video classification. Although the dataset with one million videos [17] has greatly pushed forward the research on CNNs for videos, there was still a notable gap between dense-trajectory-based [18] and CNN-based [17] approaches in terms of action recognition. However, Simonyan and Zisserman [19] have recently proposed a kind of two-stream convolutional networks, which matches the state-of-the-art performance. They used an extra column of CNN to capture the motion between frames from multi-frame dense optical-flow features.

Turn back to the task of human identification. There are various kinds of biometric features, e.g., iris [20], fingerprint [21], face [22] and gait [23], which bear various characteristics. Most recently, Taigman et al. [24] achieved human-level performance in face verification using CNNs. A large number of training samples were required to combat over-fitting. They learned features out from their large-scale social face classification dataset (SFC), with 4.4 million labeled faces from 4,030 people. The features were learned for directly classifying (identifying) the subjects in SFC, and were subsequently used to initialize a Siamese network [25], which would verify pairs of faces. Besides, Sun et al. [26] obtained comparable results with a joint Bayesian model learned over hidden features of deep CNNs. One of the most notable points is that they explicitly considered 60 local patches instead of only one patch covering a whole face image. They extracted features with different CNNs from these patches respectively. But note that gait is based on temporal sequences in videos, which is its biggest difference compared with other kinds of biometric features. As mentioned above, the research on CNNs for video-based tasks is still open. Here, instead of the generic topic, we consider a specific one, i.e., human identification via cross-view gait video sequences. Our method is similar to the third category of classic approaches mentioned before, in the sense that it takes advantage of labeled cross-view pairs (identical or not) as well. However, our method amounts to directly predicting the similarity given a pair of samples, in an end-to-end manner. The networks will automatically learn discriminative changes of gait features which suggest the change of human identity. Considering the drastic cross-view variations of gait sequences, our method should benefit from its highly non-linear deep model, compared with the previous linear and/or shallow ones [12], [27]. We are working on gait sequences instead of still face images [24], [26], and our method needs no extra data and pre-training by human classification as required in DeepFace [24], yet still greatly outperforms the current best gait recognition approaches.

We summarize this paper's contributions as follows.

- We present a CNN-based method for gait recognition. The networks will automatically learn to recognize the most discriminative changes of gait features, so as to predict the similarity given a pair of them.
- We provide an extensive empirical evaluation in terms of various tasks (cross-view and cross-walking-condition) with different preprocessing approaches and network architectures.
- We greatly advance the record scores on the CASIA-B [6], OU-ISIR [28] and USF [29] datasets, showing that our method works pretty well and can generalize for thousands of categories and complex backgrounds. Our method shows great potential for practical use in terms of its high average hit rates under cross-view conditions, i.e., about 98 percent for CASIA-B and 91 percent for OU-ISIR.

In the remaining part of this paper, after introducing more related works on gait recognition in Section 2, we will present a brief introduction to CNNs in Section 3, and then demonstrate the details of our method in Section 4. Experimental results are given in Section 5, and conclusion of this paper is given in Section 7.

2 Related Work

Approaches to gait recognition in the recent literature can generally be grouped into two typical categories, i.e., model-based [7], [8], [30] and appearance-based [9], [10], [11], [12]. The first one amounts to modeling the underlying structure of human body, while the latter extracts gait representations directly from videos, without explicitly considering the underlying structure. The work in this paper falls in the latter category. In many scenarios, it would be difficult to precisely restore body structures from videos taken under uncontrolled conditions. In contrast, appearance-based approaches can still work well given these videos. As one common pipeline, appearance-based approaches first obtain human silhouettes from all frames in a video; compute a gait energy image [5] by aligning the silhouettes in the spatial space and averaging them along the temporal dimension; then evaluate the similarities between pairs of GEIs, e.g., by the Euclidean distance; and finally assign the label to the video by a nearest neighbor classifier. There are many alternatives for GEIs, e.g., chrono-gait images (CGIs) [31] and gait flow images [32]. However, a recent empirical study by Iwama et al. [28] shows that GEI, despite of its simplicity, is the most stable and effective kind of features for gait recognition on their proposed dataset with 4,007 subjects.

As mentioned before in Section 1, cross-view gait recognition methods can be roughly divided into three categories. The first category is based on 3D model of human body, while the second category is based on handcrafted view-invariant features. The methods, most related to this paper, belong to the third category, which amounts to learning the projections across different viewpoints. These methods rely on the training data to cover the views which appear in the gallery and probe samples. With learned mapping matrices, gait features in different views can be projected into certain common subspace for better matching. Compared with the first two categories of cross-view gait recognition methods, the third category can be applied for scenarios with no explicit action by subjects, and can also be directly applied to views which are significantly different from the side view, e.g., frontal or back view.

To name a few methods in the third category, Makihara et al. [11] proposed an SVD-based view transformation model (VTM) to project gait features from one view into another. Kusakunniran et al. [33] used truncated SVD to avoid the oversizing and over-fitting problem of VTMs. After pointing out the limitations of SVD-based VTMs, they reformulated the VTM reconstruction problem as a support vector regression (SVR) problem [34]. They selected local regions of interests based on local motion relationships, instead of global features [11], [33], to build VTMs through support vector regression. They further improved the performance by introducing sparsity to the regression [35]. Instead of projecting gait features into one common space, Bashir et al. [36] used canonical correlation analysis (CCA) to project each pair of gait features into two subspaces with maximal correlation. Kusakunniran et al. [12] claimed that there may exist some weakly-correlated or non-correlated information in global gait features across views and carried out motion co-clustering to partition the global gait features into multiple groups of segments. They applied CCA on these segments, instead of using the global gait features as Bashir et al. did in [36]. Most of the above mentioned methods trained multiple mapping matrices, one for each pair of viewpoints. Recently, Hu et al. [27] proposed to apply a unitary linear projection, named as view-invariant discriminative projection (ViDP). The unitary nature of ViDP enabled cross-view gait recognition to be conducted without knowing the query gait views. On the other hand, Hu [37] designed a kind of gait feature named as enhanced Gabor gait (EGG), which encodes both statistical and structural characteristics with a non-linear mapping. The regularized local tensor discriminant analysis (RLTDA) was applied for dimensionality reduction. RLTDA was supposed to be able to capture the nonlinear manifolds which are robust against view variations, but it is sensitive to initialization. For that reason, a number of RLTDA learners were accordingly fused for obtaining better performance.

3 Background: CNN

A CNN [13], [38] is usually composed of a series of stacked stages, each of which can be further decomposed into several stacked layers, including a filter bank layer (a convolution layer), a non-linear activation function, a spatial pooling layer and sometimes a normalization layer. A convolution layer is derived from a fully-connected layer via two steps of simplifications in order to reduce the number of parameters. In a fully-connected layer, as the name tells, its neurons are fully connected to those of its previous layer. The first simplification is to impose the spatial locality so that the neurons are only connected to local regions of the previous layer. This kind of layers are also known as the locally-connected layers. The next step is to share the weights across all spatial locations. After that, we will obtain a convolution layer. The classic non-linear activation functions include the hyperbolic tangent function $\tanh(x)$

and the logistic function $f(x) = 1/(1 + e^{-x})$. However, Krizhevsky et al. [14] pointed out that networks with the rectified linear unit (ReLU) [39] $f(x) = max(0, x)$ can be trained several times faster due to ReLU's non-saturating characteristic. ReLU might not be the optimal choice, but it is favorable for the sake of efficiency. The spatial pooling amounts to down-sampling by preserving only one activity for each local region of a feature map. The preserved value can either be the maximum or the average activity within that region. Empirical results show that max pooling performs better in most cases.

Given the above settings, the lth convolution stage's activities $\boldsymbol{H}_l$ can be concisely formulated as

$$\boldsymbol{H}_l = \text{pool}(\max(0, \boldsymbol{W}_l \otimes \boldsymbol{H}_{l-1} + \boldsymbol{b}_l)), \tag{1}$$

where $\otimes$ denotes the convolution operation, $\boldsymbol{H}_{l-1}$ is the input rendered by the previous layer, $\boldsymbol{H}_0$ is the original input data, $\boldsymbol{W}_l$ contains a number of filters, and $\boldsymbol{b}_l$ contains a number of biases shared across different spatial locations. For each feature map in $\boldsymbol{H}_l$, accordingly, there will be one filter in $\boldsymbol{W}_l$ and one entry in $\boldsymbol{b}_l$ (the bias). Note that the ReLU is also integrated in Eq. (1), as well as the spatial max pooling. Considering that ReLU does never saturate in $[0, +\infty)$, it is safe to feed data into networks with no local contrast normalization, as long as there are some examples producing positive activities [14]. However, Krizhevsky et al. [14] also reported that their proposed cross-map local response normalization can aid generalization. It implements a form of lateral inhibition, introducing competition among the big activities on adjacent feature maps. For an activity a_i at certain spatial location on the ith feature map, the cross-map normalized activity b_i can be computed as [14]

$$b_i = a_i \bigg/ \left(\gamma + \alpha \sum_{j \in \text{nb}(k,i)} (a_j) \right)^{\beta}, \tag{2}$$

where α, β, γ and k are all configurable parameters. Once a network gets initialized, its feature maps will be arranged in certain order. Let there be N feature maps, and the k neighbors of the ith feature map $\text{nb}(k, i)$ will be $\{j | j = \max(0, i - n/2), \cdots \min(N - 1, i + k/2)\}$. Notably, only the activities at the same spatial location participate in this kind of normalization.

There are millions of parameters in a deep CNN. Usually, for a specific task, the given data cannot afford to train a good model due to severe over-fitting. To this end, one can apply data augmentation to increase the size of training data, by transforming the original examples in various ways, e.g., rescaling, rotating and cropping. Besides, Krizhevsky et al. [14] reported that the dropout technique [40] is often helpful for combating over-fitting. It amounts to dropping neurons with a rate of 50 percent during training. Dropped neurons will neither contribute to the forward nor the backward propagation. Accordingly, the activities should be multiplied by 0.5 during testing. Dropout has been explained as an efficient way for combining many different networks [40], which reduces co-adaptations of neurons and forces networks to learn more robust features.

Finally, it should be noted that, in the recent literature, CNNs are usually built up with the above layers, but the network architecture will vary according to specific tasks and purposes.

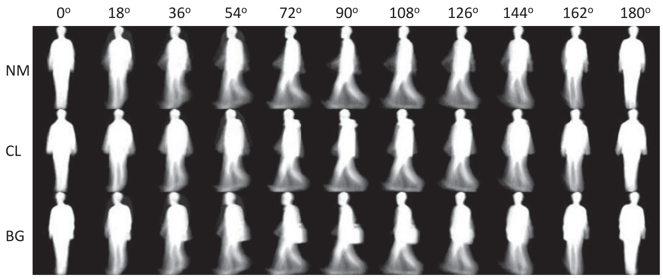


Fig. 2. Example GEIs of one subject in the CASIA-B gait dataset [6] from view 0 degree (left) to 180 degree (right) with an interval of 18 degree. Top row: normal walking (NM). Mid row: with a coat (CL). Bottom row: with a bag (BG).

4 Method

4.1 Gait Recognition

To accomplish gait recognition, one has to predict the identity of a probe sample, given a gallery which is composed of registered gait samples. Or with formulation, suppose there is one probe sample $\boldsymbol{x}$ and N_{g} samples in the gallery $\{(\boldsymbol{x}_i, y_i = i) | i = 1, \ldots, N_{\text{g}}\}$, where y_i denotes the identity of sample $\boldsymbol{x}_i$. Given the above data, the identity of probe $y(\boldsymbol{x})$ is to be predicted.

Most of the widely-used gait recognition datasets provide gait energy images [5], which are the average silhouettes along the temporal dimension. For example, some GEIs of one subject in the CASIA-B gait dataset [6] are shown in Fig. 2. The considered cases include 11 views and three different conditions, i.e., normal walking, with a coat and with a bag. In the easiest case, probe GEIs and those in the gallery are in an identical viewpoint, and at the same time they are all in the normal walking condition. In that case, computing the similarities based on the Euclidean distance achieved pretty good results [6]. This paper will consider two cases which are much harder to deal with.

- *Cross-view*. Probe GEIs and those in the gallery are in different viewpoints.
- *Cross-walking-condition*. Besides the cross-view configuration, probe GEIs are either with a coat or with a bag, while GEIs in the gallery are under the normal walking condition.

The pipeline of a typical GEI-based gait recognition method is illustrated in Fig. 3. First, extract human silhouettes from a raw video sequence using an off-the-shelf approach such as background subtraction based on the Gaussian mixture model [29]. Then, align and average the silhouettes along the temporal dimension to get a GEI. Third, given a probe GEIs and those in the gallery, evaluate the similarities between each pair of probe and gallery GEIs. As a simple implementation, one can calculate the Euclidean distance between two GEIs directly [6]. And finally, assign the identity of the probe GEI, usually with the nearest neighbor classifier, based on the computed similarities in the third step. Different from previous methods, the third step above is realized with deep convolutional neural networks in this paper.

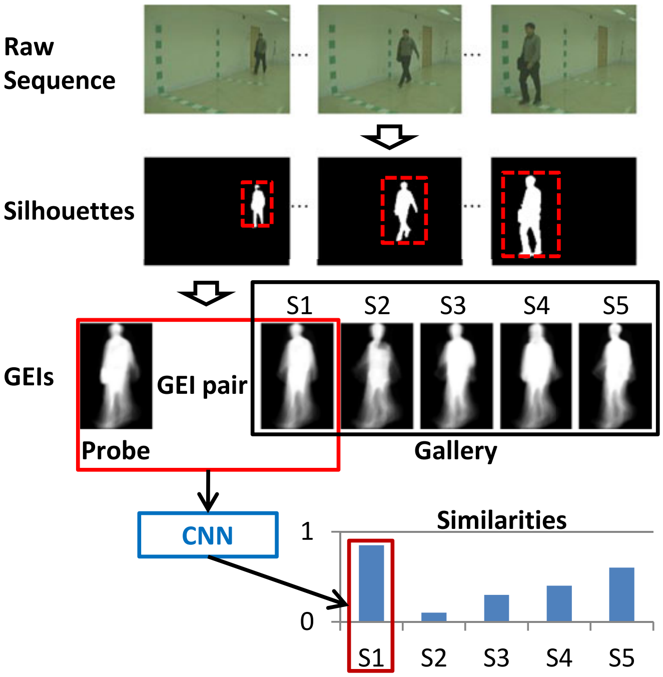


Fig. 3. Pipeline of a typical GEI-based gait recognition method. The probe sample is assigned to S1, which is evaluated as the most similar one.

4.2 Network Architectures

A network to this end should predict the similarity given a pair of gait GEIs, as illustrated in Fig. 3. In the plainest way, similarities are computed as $s_i = \exp(-\|x - x_i\|^2)$. In the context of neural networks, one layer taking in two inputs can simulate the *subtraction* to compute the difference between a pair of features. Another layer can subsequently summarize the entries of the difference to predict a similarity. To train the network in a discriminative manner, the latter layer requires two nodes instead of one in order to constitute a two-way classifier. Putting a soft-max layer on its top, we can train the whole network with the logistic regression loss. In a word, we need at least two layers with trainable parameters to constitute an effective predictor, calculating similarities from pairs of gait features.

The predictor can be concisely formulated as

$$s_i = \eta(\varphi(\phi(x), \phi(x_i))), \tag{3}$$

where x is a probe GEI, x_i is a gallery GEI, ϕ projects x and x_i into a common space, φ computes the weighted difference between its two inputs, and η predicts the final similarity. Here, ϕ can be composed of one or more convolution stages and fully-connected layers; φ must take two inputs, and can be a convolution stage or a fully-connected layer. As for the predictor η, the most compact version can be a linear two-way classifier, consisting of a fully-connected layer and a soft-max layer. However, we can stack one or more convolution stages and fully-connected layers below this compact η to compose deeper networks. Note that all bias terms are omitted in Eq. (3) for conciseness.

Below, we highlight three different network architectures investigated in this paper, as illustrated in Fig. 4. Note that all of the three finally involve deep and non-linear matching, despite of their names. The key difference is when and where to start matching the features of GEI pairs, i.e., at the

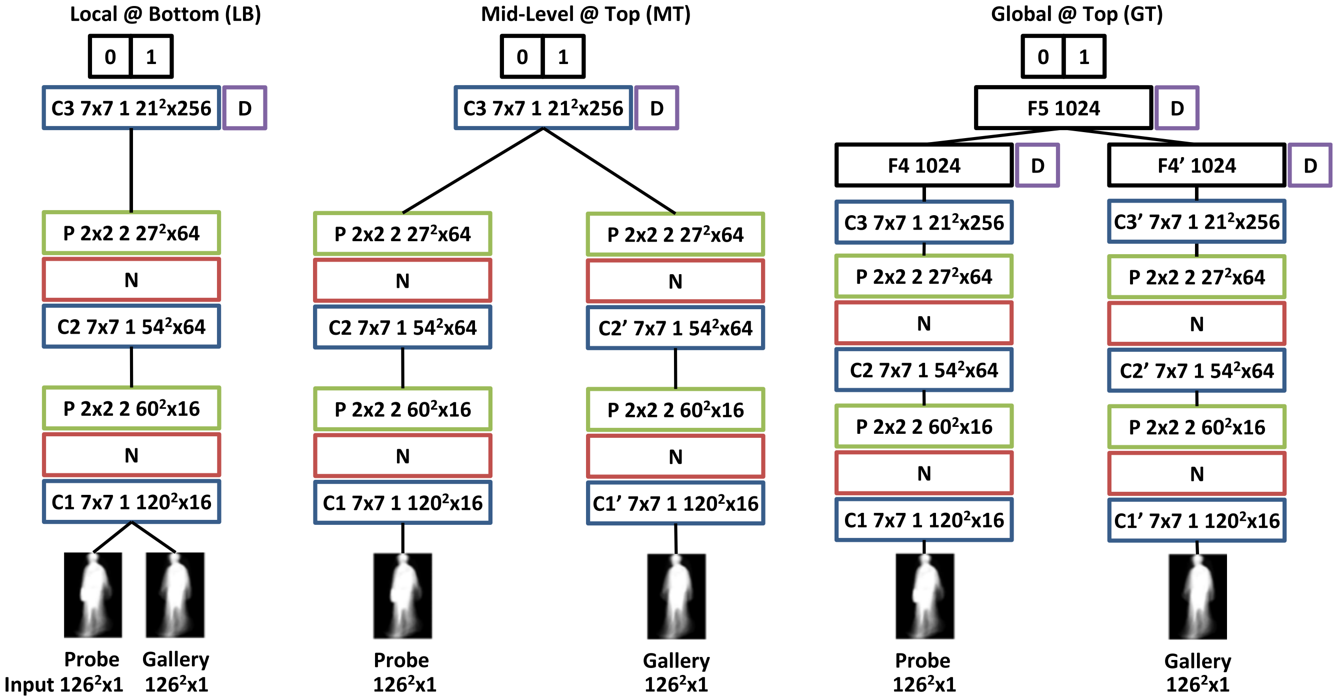


Fig. 4. Three network architectures to be investigated. Rectangles with capital letters denote different kinds of layers. Blue with C: Convolution layer. Red with N: Normalization layer. Green with P: Spatial pooling layer. Black with F: Fully-connected layer. Two adjacent squares filled with zero and one together denote a linear two-way classifier. A purple square to the right of a rectangular with D indicates that this layer applies the dropout technique. Directly next to each of the C and F letters, there is a number denoting the index of that corresponding layer in the whole network. The strings following each of them (C1, C2, C3, C1', C2' and C3') are formatted as the size of the filters, the stride and the dimensions of the feature maps. Likewise, those following the P letters are formatted as the size of the pooling cells, the stride and the dimensions of the down-sampled feature maps. An integer following F4 or F5 denotes the number of neurons. Better viewed in color.

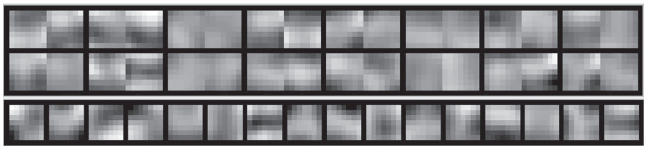


Fig. 5. Top: Learned pair-filters in the first network, i.e., LB in Fig. 4. Every two horizontally adjacent filters constitute a pair-filter. Bottom: Filters of the first convolution stage in the second network, i.e., MT in Fig. 4. The weights are normalized into gray-scale values for illustration. Lighter pixels denote positive weights, while dark ones denote negative weights.

bottom layer or the top layer, over local features or global features. In the networks, if not mentioned explicitly, all filters are of size 7 × 7 and applied with a stride of one; all spatial pooling operations are applied in 2 × 2 neighborhoods with a stride of two; and all neurons are ReLUs except those of the last layer before the soft-max. More details will be given in Section 5.2.

4.2.1 Matching Local Features at the Bottom Layer (LB)

In this network, pairs of GEIs are compared with each other within local regions, and only linear projection is applied before computing the differences between pairs of GEIs, which is realized by the 16 pair-filters in the bottom-most convolution stage. A pair-filter takes two inputs and can be seen as a weighted comparator. At each spatial location, it will first re-weight the local regions of its two inputs respectively, and then render the sum of these weighted entries to simulate the *subtraction*. As shown in Fig. 5, although in a weighted manner, some of the learned pair-filters are indeed subtracting gallery GEIs from probe GEIs. The idea of pair-filters here is similar to the one proposed by Sun et al. [41] in terms of face verification. Here, the motivation is to simulate the linear normalization-based approaches to cross-view gait recognition [11]. That amounts to projecting GEIs of different views into a common space where the GEIs become more comparable. However, the different point is that there are two more convolution stages above the matching layer. The subsequent deep and non-linear part of the network is supposed to be beneficial to learning complex patterns from the differences between GEI pairs. Given GEIs in size 88 × 128, a neuron on Layer C3 has a receptive field in size 46 × 46. And feature maps on Layer C3 are in size 11 × 21. From these features, the top-most two-way classifier will mine the most discriminative ones, which can tell the probability of a GEI pair having the same identity, i.e., the similarity. The network can be concisely formulated as

$$s_i = \boldsymbol{W}_4 f(\boldsymbol{W}_3 f(\boldsymbol{W}_2 (f(\boldsymbol{W}_1 \boldsymbol{x}) + f(\boldsymbol{W}'_1 \boldsymbol{x}_i)))), \tag{4}$$

where $\boldsymbol{W}_l$ denotes the weights on the lth layer (or filters for convolution layers), and f is a non-linear activation function. Note that the bias terms, the spatial pooling and cross-map normalization operations are omitted here for clarity, and that $\boldsymbol{W}_1$ and $\boldsymbol{W}'_1$ are different from each other. The formulations of the next two networks are similar. The key point is that there is always one matching layer with a pair of weight matrices, i.e., $\boldsymbol{W}_l$ and $\boldsymbol{W}'_l$, applying the weighted subtraction.

4.2.2 Matching Mid-Level Features at the Top Layer (MT)

In this network, two extra non-linear projections are applied before computing the differences between pairs of GEIs on Layer C3, as shown in the middle part of Fig. 4. The motivation is to apply deep non-linear normalization to GEIs instead of the shallow linear one in LB. This kind of normalization, composed of three convolution stages, is non-linear and significantly different from previous normalization-based works [11]. The non-linearity enables the network to learn a more complex projection if needed. The projected features are matched at Layer C3 so as to obtain the differences between pairs of GEIs, in size 11 × 21, which are the very features for subsequent classification. Note that these are not strictly global features which directly describe a whole GEI. Instead, they are concatenations of mid-level features at many spatial locations. As same as in LB, the receptive field of a neuron on Layer C3 is in size 46 × 46, which is also the receptive field of these mid-level features. The sizes of local regions for matching are actually the same in Networks LB and MT. The key point is the order. LB directly computes the weighted differences at the bottom layer (with local features), and then learns to recognize the patterns in the obtained differences with the rest two convolution layers. In contrast, MT learns mid-level features first, and then computes the weighted differences. Note that these configurations also keep the model complexity of LB and MT consistent with each other. Network MT is related to Siamese networks [25] in the sense that the filters are shared between its two columns. However, the learned weighted difference is used instead of the direct absolute difference.

4.2.3 Matching Global Features at the Top Layer (GT)

In this network, pairs of GEIs are compared with each other by learned global features. As shown in the right part of Fig. 4, it has two more fully-connected layers compared with Network MT, i.e., Layers F4, F4′ and F5. Similarly, Layers F4 and F4′ will share their weights. However, being different from Networks LB and MT, the weighted differences are not computed within local regions, but from global features at Layers F4 and F4′. Each of them is the descriptions of a whole GEI, with only 1,024 entries, which is much more compact than those of Networks LB and MT, i.e., 59k (11 × 21× 256). The model complexity of Network GT is higher than the previous two due to the use of fully-connected layers, which can lead to over-fitting depending on the size of training data. However, the advantage of this network is its compactness, which can lead to computational efficiency. In Network LB, given one probe and $N_{\rm g}$ gallery GEIs, $N_{\rm g}$ pairs of GEIs have to be passed through the whole network. However, there is a much more efficient implementation based on this network. First, we can store in advance the output of Layer F4′ for all gallery GEIs. Second, feed a probe GEI to the network once and obtain the output of Layer F4. Finally, compare the two 1,024-dimensional features using Layer F5 and the two-way classifier. Note that, for Network MT, it is not so attractive to store the output of Layer C2, since it is of size 29k (17 × 27 × 64) even after down-sampling by the spatial pooling layer.

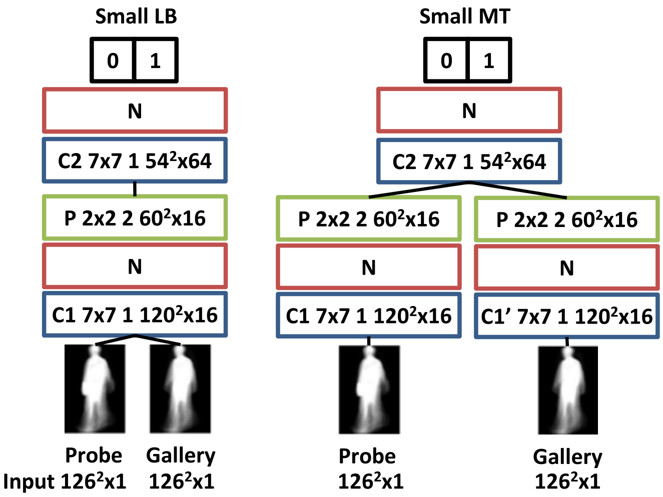


Fig. 6. The small LB and MT networks. Refer to Fig. 4 for more details.

In addition to the variation in terms of architecture, we also consider networks with different depths. There are four of them, i.e., small LB and MT, as illustrated in Fig. 6, and large LB and MT, as illustrated in Fig. 7. The small networks are derived from the standard ones by removing the third stage of convolutions. To make sure that the representations fed into the two-way classifiers are almost of the same dimension, the pooling layers in the second stage are also removed. These two networks are supposed to apply simpler and more local matching, considering less number of layers and the smaller receptive fields of the neurons in the second stage (than in the third stage of a standard network). On the other hand, the large networks are derived by stacking two more convolution stages on top of the standard ones. They are supposed to apply more complex and global matching.

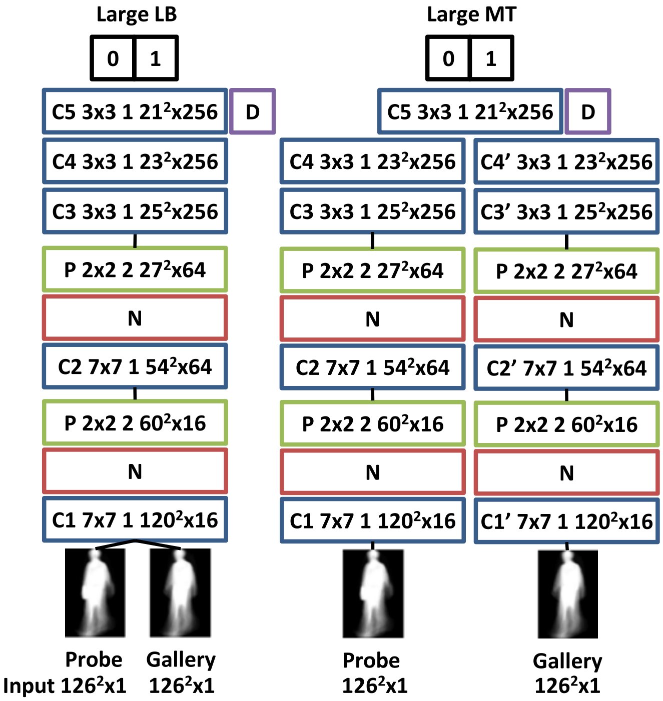


Fig. 7. The large LB and MT networks. Refer to Fig. 4 for more details.

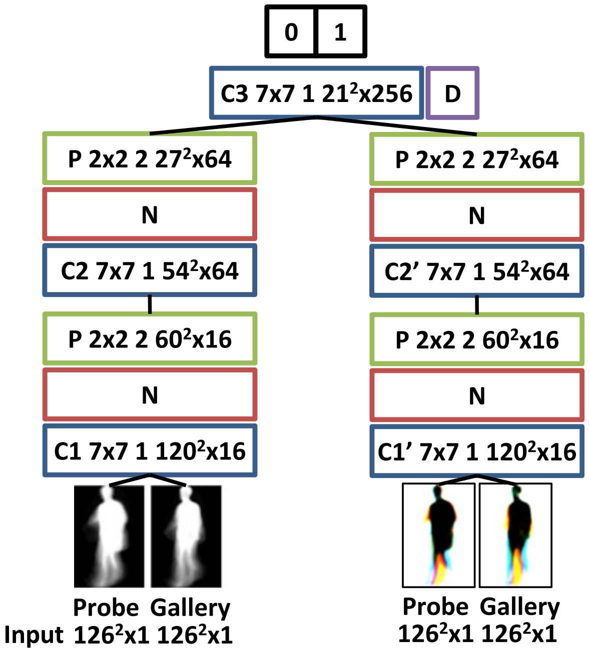


Fig. 8. The two-stream network. Refer to Fig. 4 for more details.

4.3 Temporal Information

It is natural to consider temporal information in gait recognition. For example, Bissacco and Soatto proposed a hybrid dynamical model of human motion for gait analysis [42]. To capture temporal information under the framework of CNNs, Simonyan and Zisserman trained an additional network on top of temporal features [19], i.e., optical flow. They built up a two-stream network to capture appearance from still frames and motion between frames spontaneously. Besides, another kind of methods introduced 3D structures into CNNs [17], [43]. The idea was to treat adjacent frames as different channels. In both of the above mentioned two kinds of methods, no obvious advantages of CNNs were shown in terms of recognition accuracies. Although better CNN features learned on the ImageNet classification dataset [44] can greatly boost the performances on video-based tasks [45], CNNs have not yet been shown to be overwhelmingly effective in capturing temporal information. Recurrent neural networks [46], especially long short-term memory models [47], are supposed to better deal with temporal sequences [48], [49], but it is out of the scope of this paper. Based on the above considerations, we here inspect two methods described as follows.

First, we train a two-stream network as illustrated in Fig. 8. Note that this is not an MT network. Instead, it is composed of two LB networks. The left stream takes a pair of GEIs as the input, which is the counter part of the stream processing still images in Simonyan and Zisserman's network [19]. The right stream takes a pair of chrono-gait images [31] as the input, which is the counter part of the stream processing optical flow features [19]. Wang et al. [31] carefully designed the CGI to carry temporal information by color mapping. They started from the extraction of contours in each gait image, and utilized a color mapping function to encode each of the gait contour images in the same gait sequence and aggregated them into a single CGI. Two computed CGIs are shown in Fig. 8 for example.

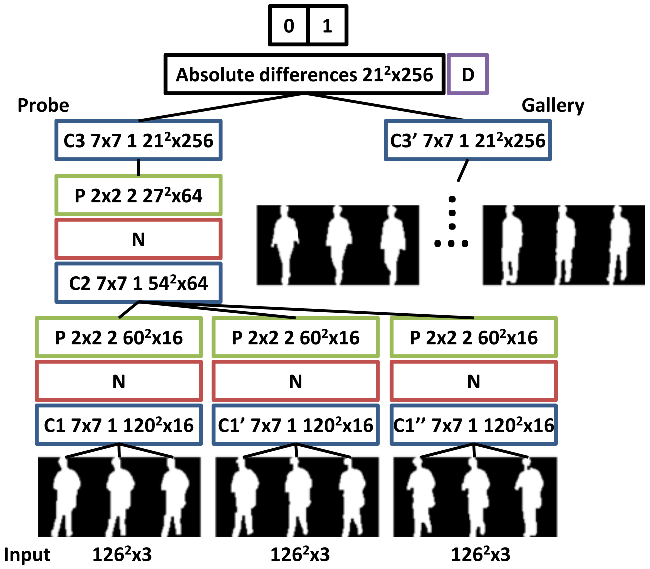


Fig. 9. The 3D CNN network. Refer to Fig. 4 for more details.

Second, we train a network with 3D convolutions in its first and second layers, as illustrated in Fig. 9. To this end, the network should take raw frames as the input. During training, each time we feed it with a pair of sequence slices, each of which is composed of nine adjacent frames sampled from a gait sequence. During testing, we feed it with all frames of a sequence (nine by nine to fit the network input), and average the output. One limitation of this method is that the network only takes nine adjacent frames as its input during training, while a sequence may have dozens of frames. To enlarge the temporal window of a training sample, we randomly pick multiple (16 in this paper) groups of adjacent frames, feed them within the same mini-batch and average the computed feature maps in the network after Layers C3 and C3′ in Fig. 9. We empirically find it to be important for high recognition rates. In addition to the above described network, we will also investigate another two networks, which share a similar structure but respectively have no or only one 3D convolution layer.

5 Experiments

5.1 Datasets

The first one is the CASIA-B gait dataset [6]. There are 124 subjects in total, and 110 sequences per subject. Namely, there are 11 views (0, 18, . . . , 180 degree) and 10 sequences per subject for each view. Among the 10, six are taken under normal walking conditions (NM). Four of them are in the gallery (NM #1-4) and the rest two are kept as probes (NM #5-6). Another two are taken when the subjects are in their coats (CL), kept as probes (CL #1-2), and the remaining two are taken with bags (BG), kept as probes (BG #1-2). Example GEIs extracted from this dataset can be found in Figs. 2 and 11. Cross-view gait recognition on this dataset is challenging, especially when the cross-view angle is larger than 36 degree [6], [12]. It becomes even harder when probe and gallery samples are under different walking conditions. There are few reported results under this configuration [27], though this is a widely-used dataset for gait recognition.

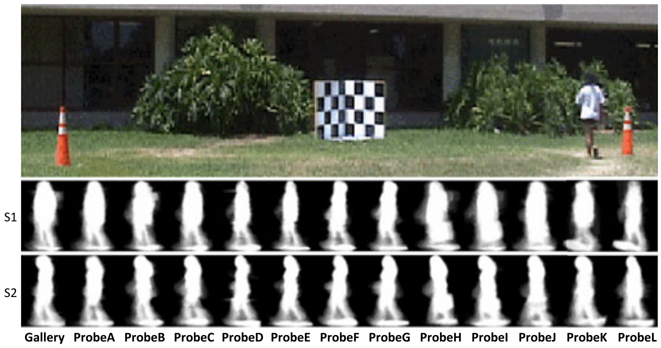


Fig. 10. Top: The stage where the USF gait sequences are imaged. Bottom: The GEIs of two subjects under different conditions. See the text for details.

The second one is the OU-ISIR gait dataset [28]. There are 4,007 subjects (2,135 males and 1,872 females) with ages ranging from one to 94 years old. For each subject, there are two sequences available, one in the gallery and the other as a probe sample. Four view angles are considered (55, 65, 75, 85 degree). The cross-view angles are smaller than those in CASIA-B, and there are no variations in walking conditions. However, this dataset allows us to determine statistically significant performance differences between gait recognition approaches due to its big number of subjects.

The third one is the USF gait dataset [29], which is also one widely-used benchmark in the gait recognition community. There are 122 subjects in total, for each of whom there are five covariates, leading to 32 possible conditions under which gait sequences can be imaged. These include two different shoe types (A and B), two carrying conditions (with or without a briefcase), two surface types (grass and concrete), two viewpoints (left and right) and two time instants. The sequences in this dataset are recorded in an outdoor scene, with more complex backgrounds, which is more close to a practical scenario. As a result, the obtained GEIs are more noisy and of lower quality. The outdoor scene and some example GEIs are shown in Fig. 10.

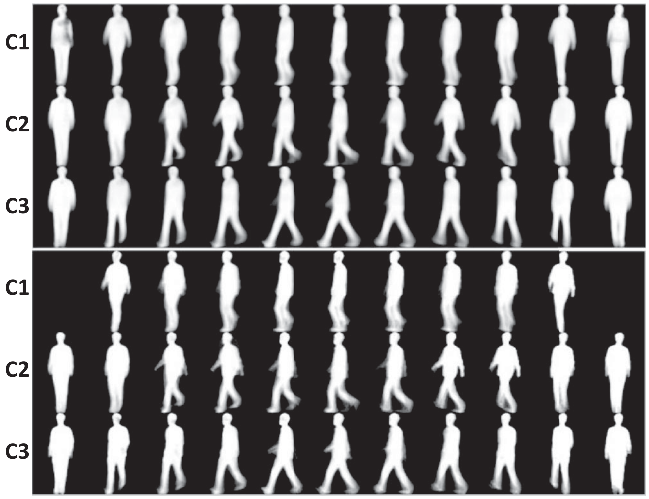


Fig. 11. Clustered GEI centers and example subGEIs of one subject in the CASIA-B gait dataset [6] from view 0 degree (left) to 180 degree (right) with an interval of 18 degree. Top row: The three clustered centers, each sub-row denotes a center. Second row: The sub-GEIs under the normal walking condition.

5.2 Implementation Details

5.2.1 GEI

After being obtained from raw frames by off-the-shelf background subtraction approaches, the silhouettes have to be cropped, rescaled and aligned, as shown in the second row of Fig. 3. When people are walking under normal conditions, we observe that the height of silhouettes in a sequence usually varies stably, and that the gravity center of silhouettes usually moves stably. Considering that, we here adopt a heuristic approach as follows. For each frame of a gait sequence, first, locate the top and bottom pixels of the silhouette and record the distance between them along the vertical dimension (the height). Second, compute the gravity center of the silhouette, which will be used to locate it along the horizontal dimension next. Third, with the gravity center, the height obtained in the first step and a given aspect ratio, i.e., 11/16, draw a rectangle box circling the silhouette, as shown in the second row of Fig. 3. Finally, crop off the region within the rectangle and resize it into a given size (88 × 128). This preprocessing returns a series of roughly aligned silhouette sequences. Classically, one has to segment a sequence so as to compute its GEI within a gait circle [5]. However, we empirically find that using the whole sequence works comparably well for our method. We choose this version of GEIs only for the sake of convenience in implementing mixture of GEIs as presented below.

5.2.2 Mixture of GEIs

In a GEI, the noise due to failed pre-processing can be effectively suppressed by averaging silhouettes within a long temporal range. However, this average strategy can lead to the loss of details. The motivation is that the distribution of gait can be seen as a mixture of Gaussian distributions instead of a single Gaussian. The idea is similar to spatial pyramid matching for image classification [50], which slices an image into several regions where classic bag-of-words representations can be computed respectively. Here, we slice a gait sequence and treat these slices separately. We let the silhouettes cluster into several groups, and compute one GEI for each group of silhouettes in a sequence. Here, we highlight two points below.

- Treat silhouettes in different viewpoints separately. The target is to separate a sequence, which can only be in one viewpoint.[1]
- Compare corresponding subGEIs only. It has no merits to compare two silhouettes if they are definitely in different poses.

This extension to GEIs will be evaluated on the CASIA-B dataset [6]. To this end, we randomly pick out 10 silhouettes from each normal walking sequences belonging to the first 24 subjects. For each view angle, we let the silhouettes cluster into three groups. Then, we sort the groups with a kind of handcrafted feature, i.e., the width of the lower part. To compute it, crop a GEI and keep the lower 30 percent of it, find the left-most and right-most pixels whose gray-scale value is bigger than 32, and return the horizontal distance between the two pixels. The clustered centers and some examples are shown in Fig. 11. For 0 and 180 degree, the differences between silhouettes are so subtle that our method might fail to cluster them properly. However, the results are reasonable for the rest nine view angles.

5.2.3 Network Architecture

For some of the networks illustrated in Fig. 4, we also apply group sparsity to their convolution layers. Namely, the filters in the third stage are only applied on 16 of the second stage's 64 channels, and those in the forth stage (if exists) are only connected to 64 of the third stage's 256 channels. As for the cross-map normalization in Eq. (2), following the suggestion by Krizhevsky et al. [14], we set the four configurable parameters as $\alpha = 10^{-4}$, $\beta = 0.75$, $\gamma = 2$ and $k = 5$.

5.2.4 Training

We train the networks using back-propagation with the logistic regression loss, and update the weights with a mini-batch of size 128. We initialize the weights of each layer using a Gaussian distribution with a mean of zero and a standard deviation of 0.01. All the bias terms are initialized with the constant zero. For all layers, the momentums for weights and bias terms are 0.9, and the weight decay is 0.0005. We start with a learning rate of 0.01, and divide it by 10 for two times. In some cases, we have to start from 0.001 otherwise the training will not converge. We decide the iterations according to empirical results on the validation set of the CASIA-B dataset [6]. Usually, we cost about 1.2 million iterations to train one network.

5.2.5 Sampling

We feed the networks with a balanced training set. Namely, the positive and negative samples respectively constitute half of the training set. To obtain a positive sample, we randomly pick a subject, then two of his present view angles, and finally his two sequences in these angles respectively. To obtain a negative one, we do the same except picking out the sequences from two different subjects.

5.2.6 Evaluation

Note that the task is not multi-view but cross-view gait recognition. We do have access to all view angles during training the networks. However, in each test case, the gait sequences registered in the gallery should all be in the same view angle. For the sake of better comparison and analysis, we sometimes iterate the possible probe and gallery view angles so as to cover all the cross-view combinations, as in Table 5; And sometimes report results on a subset of the cross-view combinations, as in Tables 2 and 4. In other experiments whose results are shown in Tables 1 and 3, and those in Figs. 12, 15 and 16, we fix the probe view angle and report the average recognition rates while the gallery view angle varies.

5.3 Impact of Network Architectures

We compare the performances of three networks. The first two are illustrated in Fig. 4, i.e., Local @ Bottom (LB) and Mid-Level @ Top (MT). We do not compare Global @ Top (GT) in

[1] It is true for our used datasets.

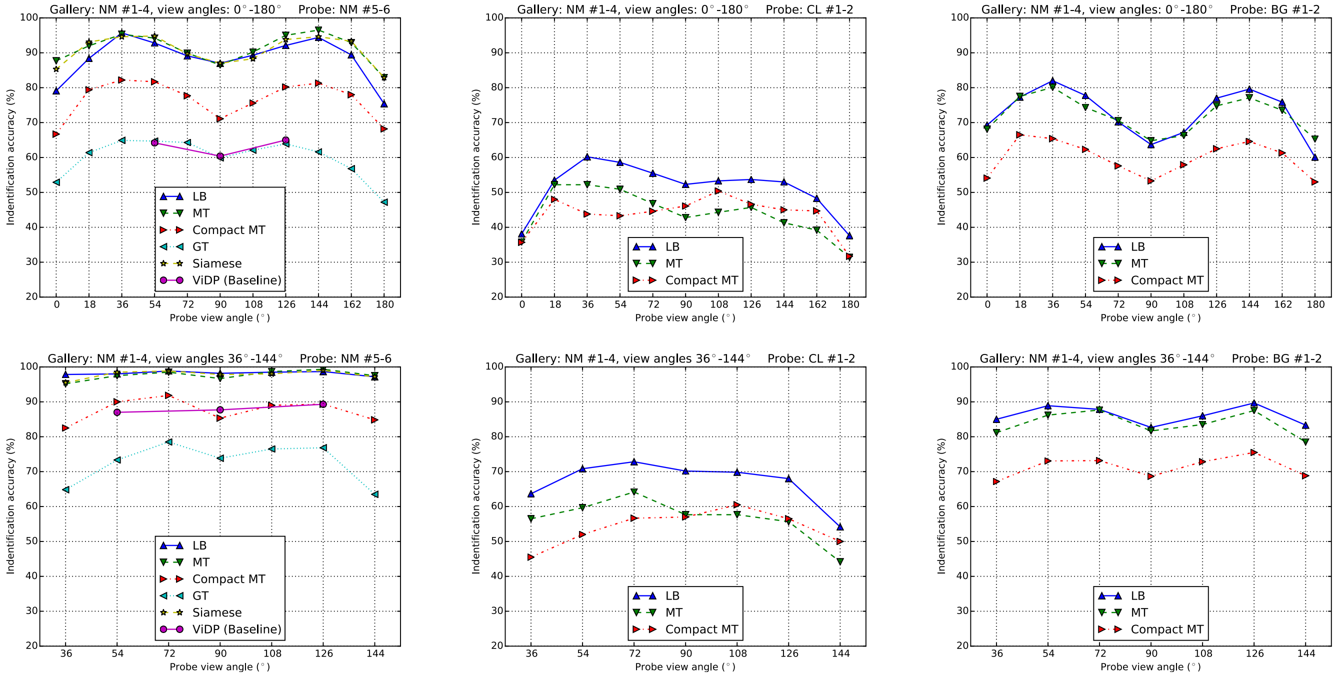


Fig. 12. Impact of network architectures evaluated on CASIA-B. Given a probe angle, we test all allowed gallery angles and report the average accuracies. ViDP [27] is the previous best performer. NM: Normal walking. CL: With a coat. BG: With a bag.

detail considering its less satisfactory performance. In our experiments, Network GT suffers from severe over-fitting, probably due to the small training dataset. However, we sometimes do favor its computational efficiency. As a compromise, we modify Network MT to obtain more compact features, which is the very third network compared here, i.e., Compact Mid-Level & Top (CMT). It amounts to use a larger stride in the third convolution stage. For example, when we use a stride of five, the resulting feature map will be in size $3 \times 5 \times 256$, with only 3,840 entries.

Among the 124 subjects in the dataset, the first 50 are used for training, the next 24 are kept for evaluation, and the rest 50 are preserved for testing. We test our method under the three kinds of walking conditions considered in this dataset, i.e., normal walking (NM), walking wearing a coat (CL) and carrying a bag (BG). The results are reported in Fig. 12. Note that each of these identification accuracies is the average score for a given probe view angle with different gallery view angles. Based on these results, we highlight six points as follows.

5.3.1 LB $\approx$ MT $\gg$ GT

There are no significant gaps between the performances of LB and MT, and they both outperform GT with a clear margin. Considering the number of training data in our task, it is very important to deal with the over-fitting problem. There are a rather large number of trainable parameters on Layers F4 in GT, which is probably one of the reasons why we witness severe over-fitting problems with GT. Besides, pairs of GEIs are matched with each other more locally in LB and MT than in GT. Note that the silhouettes have already been roughly aligned in preprocessing. As a result, the subtle shape and pose changes can be better recognized from local differences obtained in LB and MT.

5.3.2 LB versus MT

The most notable difference between the two is that MT performs better for view angles around 0 or 180 degree. Recall the GEIs given in Fig. 2. These two viewpoints are too different from the other ones. As a result, they would prefer matching at upper layers, where the receptive field of a node is larger and more complex transformation is allowed before the matching.

5.3.3 MT versus CMT

There is a moderate drop in performance for CMT compared with MT. Nevertheless, CMT still outperforms the baseline with a clear margin when we consider all of the 11 view angles. It can be a compromise between the accuracy and efficiency when needed.

5.3.4 MT versus Siamese

The Siamese network can approximately be seen as a special case of MT, which applies the same convolutions in the third stage to its gallery and probe columns (although with different signs). However, MT has more parameters, and may learn more complex differences. As shown in Fig. 12, generally speaking, MT performs slightly better than a Siamese network.

5.3.5 $0^\circ \approx 180^\circ > 90^\circ > \cdots > 36^\circ \approx 144^\circ$

Someone might be surprised about that the 90 degree view, with the richest gait information, seems harder than other view angles such as 36 degree. To explain that, recall the GEIs given in Fig. 2 again. First, it is about the cross-view setting in our experiment. Besides the 0 and 180 degree views, the profile view (90 degree) is the most different one from the other views. Second, considering the BG subset only, a bag usually changes the profile view more than the other views.

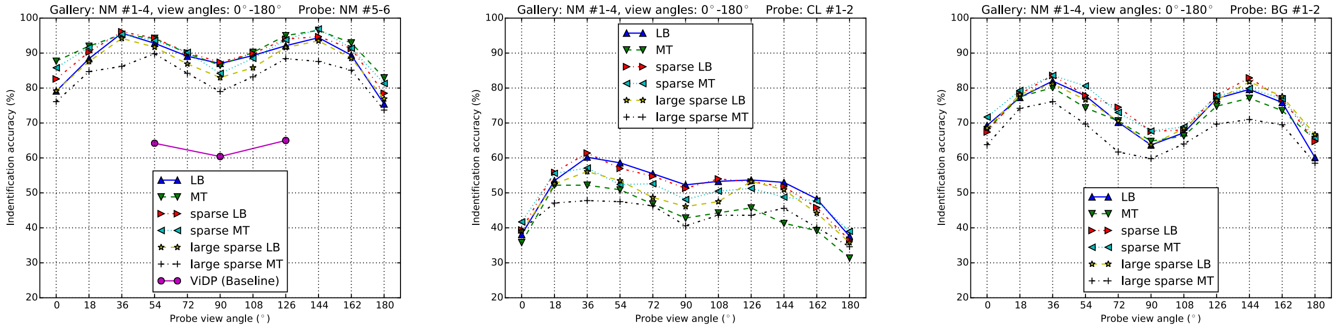


Fig. 13. Impact of network group sparsity evaluated on CASIA-B. ViDP [27] is the previous best performer. NM: Normal walking. CL: With a coat. BG: With a bag.

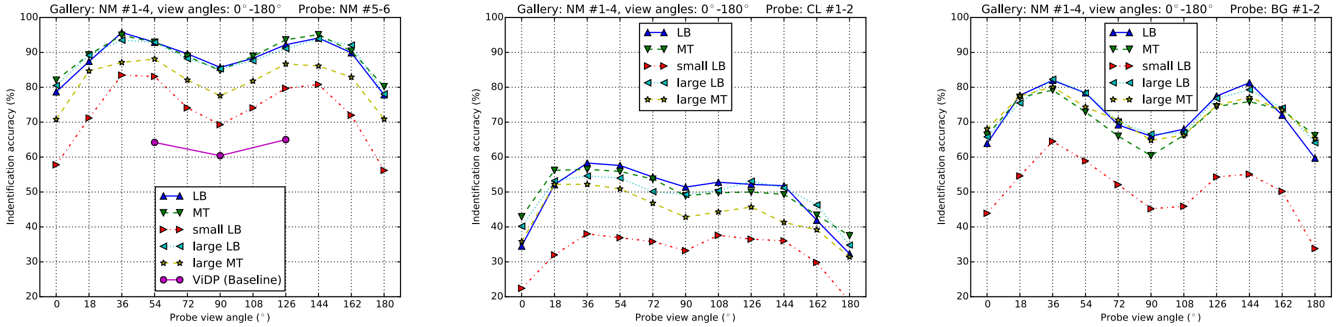


Fig. 14. Impact of network depths evaluated on CASIA-B. ViDP [27] is the previous best performer. NM: Normal walking. CL: With a coat. BG: With a bag.

5.3.6 CL > BG > NM

Cross-view and cross-walking-condition gait recognition is challenging [6], [37]. Our method has obtained very promising results on the NM subset, especially when excluding the four outer view angles, i.e., 0, 18, 162, 180 degree. This is reasonable since silhouettes in these frontal and back viewpoints carry little gait information. As for the BG subset, our method still performs well. However, it needs further improvements for the CL subset, especially when the cross-view angle is larger than 36 degree, as shown in the central part of Fig. 12. The cause behind these results is that carrying a bag only affects a small part of a GEI, while wearing a coat can greatly change the appearance, as shown in Fig. 2. There are little results reported in the literature under these settings. So, we present no baselines for the CL and BG subsets in Fig. 12, and similarly, none in Figs. 15 and 16.

In the remaining part of our experiments, we will use the LB network by default, for two reasons. First, it performs much better than GT. Second, it is generally comparable with MT in accuracy, and meanwhile, it is almost two times as fast and uses half of the GPU memories.

5.4 Impact of Network Group Sparsity

Networks with group sparsity can generate as large feature maps with less trainable parameters, which is sometimes important to suppress over-fitting. Besides, sparse networks are more efficient both during training and testing. We report the performances of the sparse LB and MT, and the large sparse LB and MT networks in Fig. 13. The results show that our strategy to introduce sparsity generally does not affect the performance of an LB network, but can improve the one of an MT network under the CL and BG conditions. There are more trainable parameters in the third stage of an MT network (than in an LB network), so this is probably achieved by suppressing over-fitting.

5.5 Impact of Network Depths

The inspected networks include those with two, three or five stages of convolutions, as illustrated in Figs. 4, 6 and 7. The results obtained on the CASIA-B dataset are given in Fig. 14. Under all of the three conditions, there is an obvious drop in performance for small LB compared with LB. It is too shallow to learn an effective model for gait matching. On the other hand, large LB shows no advantage over LB, and large MT may even perform worse than MT. According to our experiments, this has something to do with over-fitting. Early-stopping or more training data might be needed to train such large networks. Also note that small MT is missing here because it never converges in our experiments with different initializing strategies and learning rate schedules.

5.6 Impact of Input Resolutions

As shown in Fig. 15, down-sampling the GEIs (into 32 × 32 or 64 × 64) leads to a slight drop in performance for the NM and BG subsets. Note that all the curves are far above the baseline method, so down-sampling the GEIs can be a better choice if there is a demand for computational efficiency. There is an exception that the network with half resolution outperforms the one with full resolution on the CL subset. This should have a connection to the comparison between LB and MT. Greater variations would prefer larger

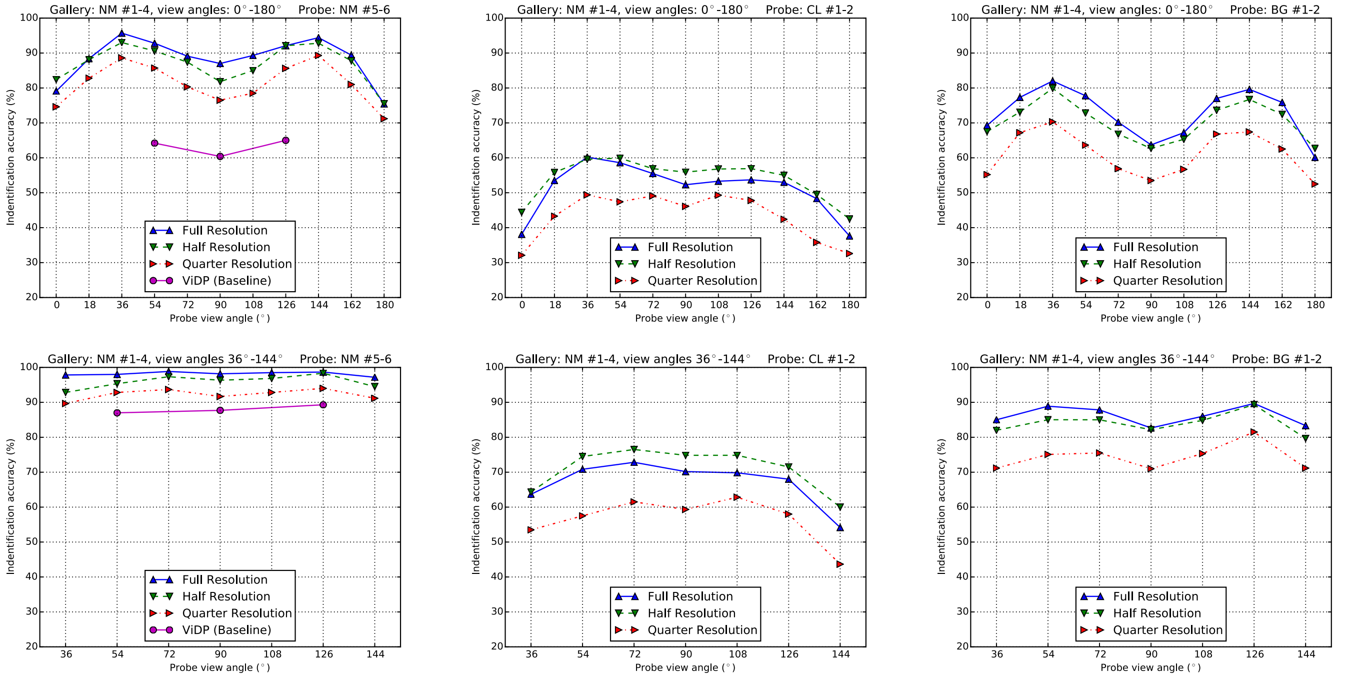


Fig. 15. Impact of input resolutions evaluated on CASIA-B. Given a probe angle, we test all allowed gallery angles and report the average accuracies. ViDP [27] is the previous best performer. NM: Normal walking. CL: With a coat. BG: With a bag.

receptive fields. Note that the GEIs vary drastically in appearances, as shown in Fig. 2. In the network with half resolution, the side length of the receptive field on Layer C3 is doubled, becoming 92 × 92 in the original resolution. That will cover a big part of a GEI in size 80 × 128, allowing the network to apply more complex matching. However, further enlarging the receptive field into size 184×184 has no merits, since in this resolution all differences seem too subtle to recognize.

5.7 Impact of Input Features

As shown in Fig. 16, the proposed mixture of GEIs can further improve the performance, especially under the clothing variation condition. We can take advantage of it to pursue the best recognition rate, as listed in Table 3. The result suggests that CNNs can learn more if we present richer data instead of only a single GEI for each gait sequence.

To evaluate the importance of fine alignment, we randomly shift the silhouettes before averaging them to

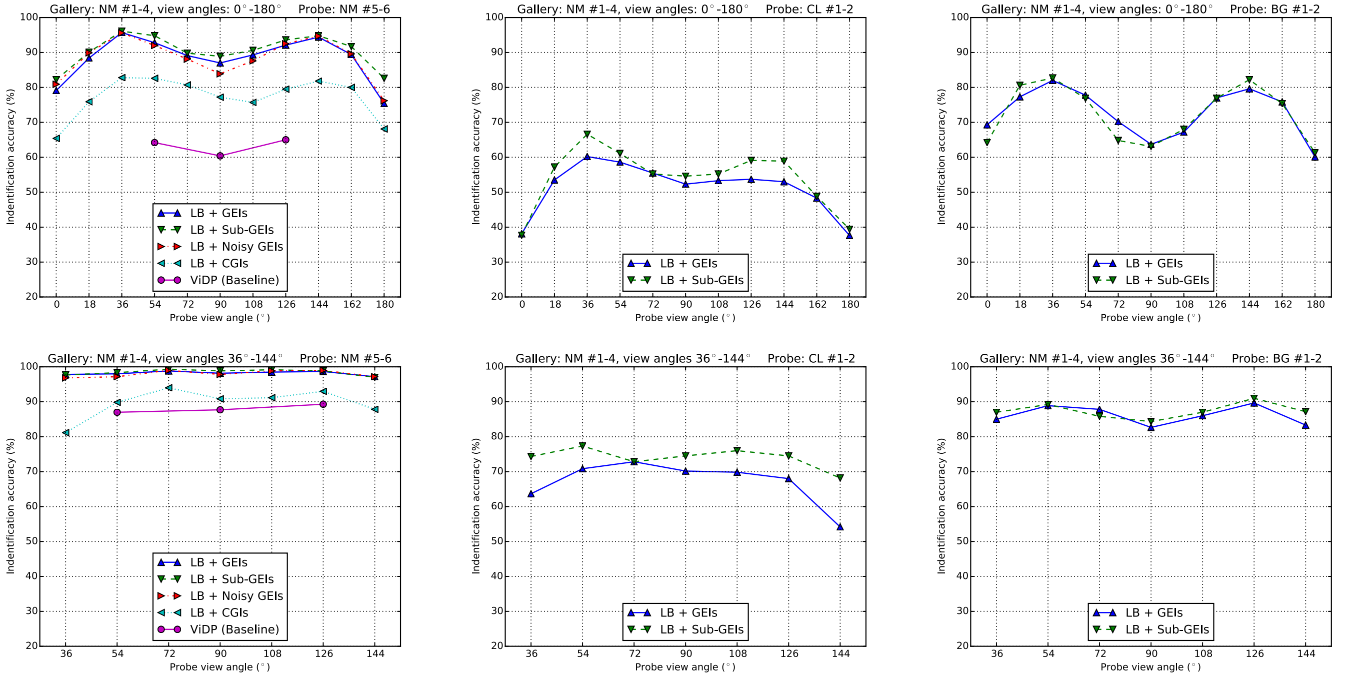


Fig. 16. Impact of input features evaluated on CASIA-B. Given a probe angle, we test all allowed gallery angles and report the average accuracies. ViDP [27] is the previous best performer. NM: Normal walking. CL: With a coat. BG: With a bag.

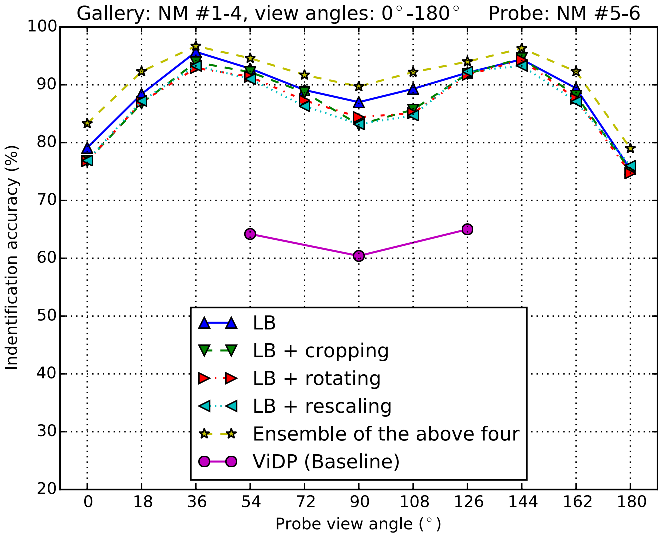


Fig. 17. Impact of data augmentation evaluated on CASIA-B.

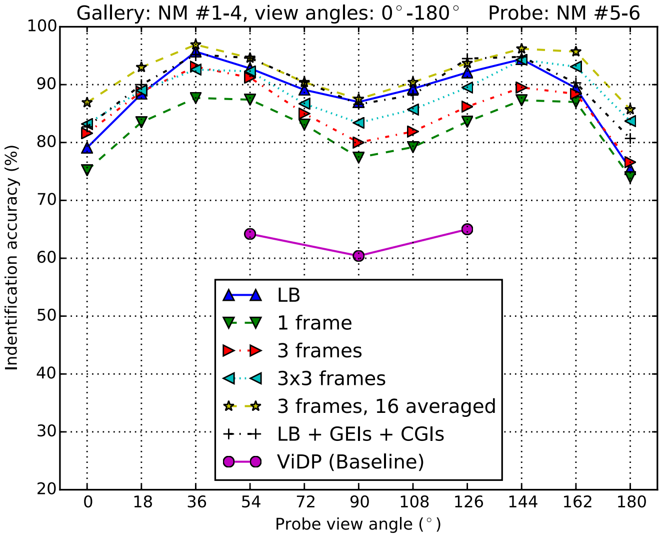


Fig. 18. Impact of temporal information evaluated on CASIA-B.

compute GEIs. They are shifted in horizontal and vertical directions respectively by an offset randomly sampled from a Gaussian distribution with a mean of zero and a standard deviation of 20 (in pixels). Recall that the size of GEIs is 128×88. So this shifting can drastically change the appearances of obtained GEIs. Nevertheless, the results in Fig. 16 (LB + Noisy GEIs) demonstrate that CNNs are pretty robust to this kind of noise.

Besides, we also train a network on top of the chrono-gait images [31]. As shown in Fig. 16, this kind of features on its own performs much worse than GEIs.

5.8 Impact of Data Augmentation

We apply no data augmentation in most of our experiments. The exceptions are three LB networks, which are respectively trained with randomly cropping a 110×110 sub-window, rotating between -8 and 8 degrees, and rescaling between 90 and 110 percent. The related results are shown in Fig. 17. These networks are inferior in terms of performance compared to the standard LB network trained without any data augmentation. This is reasonable since samples transformed drastically seldom appear either in the training or testing dataset, which is true for most of the existing gait recognition datasets. This is very different from those for generic object recognition, e.g., the ImageNet dataset [44], and those for face recognition, e.g., the Labeled Faces in the Wild dataset (LFW) [51]. Nevertheless, combining the above mentioned four networks by averaging their predicted similarity scores, we can achieve apparently better recognition rates, as shown in Fig. 17. This result suggests that the three networks trained specifically for hard cases can supplement the standard one when it fails due to obvious transformations.

5.9 Impact of Temporal Information

Results showing the impact of temporal information are given in Fig. 18. As shown in Fig. 8, the first method is to train a two-stream network on top of GEIs and CGIs (LB + GEIs + CGIs), which performs slightly better than the baseline (LB). This result is expected since the two-stream network takes advantage of the two kinds of gait features, which respectively capture different aspects of information from gait sequences. The second method is to train a network with 3D convolution layers, as shown in Fig. 9. When we increase the number of 3D convolution layers from zero to two, there are significant improvements in terms of recognition rates. Furthermore, when the feature map averaging component is included, the network (3 frames, 16 averaged) outperforms LB with a clear margin. Otherwise, the temporal window of a training sample would be so narrow (nine frames in this paper) that it can hardly carry enough information of a gait sequence. For comparison, recall that a GEI is computed by averaging all frames of a gait sequence, so its temporal window is as wide as the whole sequence.

For clarity, we list the numerical results in Table 3. Generally speaking, the three networks considering temporal information perform better than those five only trained on top of GEIs. And notably, we achieve the best mean recognition rate using the ensemble of these networks. Specifically, we combine the five models based on GEIs to obtain an average recognition rate of 93.0 percent, and combine the three models dealing with temporal information to obtain 93.5 percent. When we combine the above two, the performance improves further. These results suggest the importance of capturing temporal information. More sophisticated models such as the recurrent neural networks [46] can hopefully further improve the performance.

5.10 Comparison on CASIA-B

The comparison of our method with those in the literature is presented in Tables 1, 2, 3 and 4. These methods are listed because they are the most recent and best performers. Their scores are directly taken from the original papers, and the comparison is only conducted between the results obtained with the same division of training and testing data. Generally speaking, our method performs better than these methods with significant margins. For example in Table 3, our average accuracy for the probe angle 126 degree is at least 92.1 percent, compared with 65.0 percent, when we use 74 subjects for training and consider all of the 11 view angles in the gallery. Particularly, by averaging the similarity scores of

TABLE 1
Comparison of Our Method with Previous Ones on CASIA-B by Average Accuracies (%), Excluding Identical-View Cases

Gallery NM #1-4	0°-180°				36°-144°		
Probe NM #5-6	0°	54°	90°	126°	54°	90°	126°
SVR [34]	–	28	29	34	35	44	45
TSVD [33]	–	39	33	42	49	50	54
CMCC [12]	46.3	52.4	48.3	56.9	-	-	-
ViDP [27]	–	59.1	50.2	57.5	83.5	76.7	80.7
Ours	**54.8**	**77.8**	**64.9**	**76.1**	**90.8**	**85.8**	**90.4**

Models are trained with GEIs of the first 24 subjects.

multiple networks, we achieve the best mean recognition rate, i.e., 94.1 percent. In addition, we highlight the four points as follows.

First, our method achieves promising performance even only with the sequences from 24 subjects for training, as shown in Tables 1 and 2. There are six gait sequences per subject, each of which is recorded from 11 view angles. Namely, in this case, the number of GEIs for training is only 1,584. These data are much less than those used in DeepFace, proposed by Taigman et al. [24]. They used 4.4 million labeled faces from 4,030 people for pre-training. Our method can work on such small data for two reasons.

- We feed our networks with pairs of GEIs, so the number of combinations for training can be above a million.
- Networks LB and MT have no fully-connected layers besides the final two-way linear classifiers, which greatly reduces the number of trainable parameters.

Also note that the network pre-trained by identification in Table 3 is our implementation following their pipeline. Namely, we first train features in terms of human classification (identification), and then use these features to evaluate the similarities between pairs of probe and gallery GEIs.

TABLE 2
Comparison with Kusakunniran et al.'s Method [12] and Yu et al.'s Baseline [6] under Normal Walking Conditions on CASIA-B by Accuracies (%)

Probe view	Gallery view	Ours	CMCC [12]	NN [6]
0°	18°	**95.0**	85	23.8
54°	36°	**98.5**	97	29.8
54°	72°	**98.5**	95	21.8
90°	72°	**99.5**	96	81.5
90°	108°	**99.5**	95	87.9
126°	108°	**99.0**	98	37.1
126°	144°	97.0	**98**	43.1
0°	36°	**73.5**	47	4.4
54°	18°	**91.5**	65	8.9
54°	90°	**93.0**	63	17.7
126°	90°	**92.0**	78	15.3
126°	162°	**83.0**	75	2.4
54°	0°	**47.5**	24	4.0
54°	108°	**89.5**	53	16.9
90°	36°	**67.5**	41	6.9
90°	144°	**66.0**	41	1.6
126°	72°	**90.5**	60	21.0
126°	180°	**43.0**	22	3.6

Models are trained with GEIs of the first 24 subjects.

Second, considering the large number of parameters in our deep networks, it is natural to use as more training data as possible. There is a clear margin between models learned with different number of training samples. As listed in Tables 1 and 3, the improvement gained by training data from 50 more subjects with gallery view angles in 0-180 degree is around 20 percent, and the one with gallery view angles in 36-144 degree is around 10 percent. This improvement has enlarged the gap between the previous best method ViDP [27] and ours, from about 20 percent to about 30 percent.

Third, our method performs pretty well even when the cross-view angle is large, e.g., 36 or 54 degree, as shown in Table 2. Here, CMCC [12] is listed for comparison instead of ViDP [27], only because no such numerical results are reported in the ViDP paper. But, their performances are quite comparable, as shown in Table 1. When the cross-view angle is 18 degree, CMCC performs nearly the same with our method. In contrast, for larger cross-view angles, CMCC's performance degrades drastically. These results demonstrate that deep non-linear models have advantages in capturing invariant features, even when the appearances of GEIs have changed drastically due to viewpoint variations.

Finally, the results on the CL subset shown in Table 4 seem less promising than those on the BG subset. According to our empirical study, one possible reason is the lack of training data. Due to larger appearance variations, the CL subset is harder than the NM and BG subsets. Networks may easily over-fit if the training set is not big enough. Recall the large improvements brought by only increasing the training data, as shown in Tables 1 and 3. As for the CL subset, with 34 subjects for training, the average scores of our method are respectively 20.7, 33.8, 34.8 and 33.2 percent, given that the probe angles are 0, 54, 90 and 126 degree. If we use 74 subjects for training, the average scores will be raised up to 37.7 percent for 0 degree and around 60 percent for the rest, as listed in Table 3. Note that we do not thoroughly compare with any baselines on this subset only because there are no comparable results reported in the current literature. The most related work is Hu's RLTDA [37] as given in Table 4, where only several cases are reported, with cross-view angle being 18 degree. Nevertheless, with 74 subjects for training, we have raised the identification scores up to a pretty acceptable level when excluding the four outer view angles. For the CL subset, they are above 70 percent (only except the case when the probe view angle is 144 degree), and for the BG subset, they are all above 80 percent, as shown in Fig. 16.

5.11 Comparison on OU-ISIR

We apply five-fold cross-validation on this dataset. All the subjects are randomly divided into five sets. In each run, keep one set for testing, and train a network with the remaining sets. Finally, the average identification accuracies and their deviations are reported. At this point, there are no previous works reporting cross-view recognition scores on this dataset. So in Table 5, we report the identical-view recognition scores, as well as cross-view scores, in order to compare with those reported by Iwama et al. [28], and to tell the impact

TABLE 3
Comparison of Our Method with Previous Ones on CASIA-B by Average Accuracies (%), Excluding Identical-View Cases

Gallery NM #1-4	0°-180°												36°-144°		
Probe NM #5-6	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean	54°	90°	126°
CCA [36]	–	–	–	–	–	–	–	–	–	–	–	–	66	66	67
ViDP [27]	–	–	–	64.2	–	60.4	–	65.0	–	–	–	–	87.0	87.7	89.3
Network pre-trained by identification	60.1	66.4	74.7	72.7	65.7	64.8	69.5	71.4	72.8	66.8	58.5	67.6	82.5	83.0	85.5
LB network with 3 convolution layers	79.1	88.4	95.7	92.8	89.1	87.0	89.3	92.1	94.4	89.4	75.4	88.4	98.0	**98.2**	98.7
LB network with group sparsity	82.6	90.3	**96.1**	94.3	90.1	**87.4**	89.9	94.0	94.7	91.3	78.5	89.9	98.0	98.0	99.2
Siamese network with 3 convolution layers	85.3	**93.1**	94.6	**94.8**	89.7	87.0	88.3	93.8	94.6	**93.3**	82.9	90.7	**98.5**	97.8	99.0
MT network with 3 convolution layers	**87.7**	92.0	95.3	94.2	89.9	86.5	**90.2**	**95.0**	96.5	92.9	82.9	**91.2**	97.5	96.7	99.3
MT network with group sparsity	85.8	91.6	95.4	93.6	**90.2**	84.2	88.5	93.8	**96.9**	91.4	81.3	90.2	98.3	96.8	**99.7**
Two-stream network with GEI and CGI	82.7	90.0	95.2	94.6	**90.5**	86.6	88.2	**94.5**	94.8	90.3	80.7	89.8	98.0	**97.8**	**99.7**
Network with 1-frame input (averaging 16 samples)	86.9	92.5	95.7	**95.5**	88.9	86.9	88.5	92.5	96.2	94.5	**86.7**	91.3	**98.5**	96.7	98.3
Network with 3-frame input (averaging 16 samples)	**87.1**	**93.2**	**97.0**	94.6	90.2	**88.3**	**91.1**	93.8	**96.5**	**96.0**	85.7	**92.1**	97.3	97.7	98.8
Ensemble of LB networks (different data augmentation)	83.3	92.3	96.7	94.6	91.7	89.7	92.2	94.0	96.3	92.3	79.0	91.1	98.0	99.0	99.7
Ensemble of networks with GEI	87.7	93.3	97.3	95.6	93.4	90.5	92.9	96.2	97.5	93.8	85.1	93.0	**98.7**	98.8	100.0
Ensemble of networks with temporal information	88.1	93.3	98.0	96.1	93.6	90.8	92.0	96.0	98.1	95.6	**86.4**	93.5	98.5	**99.2**	100.0
Ensemble of the above two	**88.7**	**95.1**	**98.2**	**96.4**	**94.1**	**91.5**	**93.9**	**97.5**	**98.4**	**95.8**	85.6	**94.1**	98.5	99.0	100.0
Probe CL #1-2	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean	54°	90°	126°
LB network with sub-GEIs	37.7	57.2	66.6	61.1	55.2	54.6	55.2	59.1	58.9	48.8	39.4	53.98	77.3	74.5	74.5
Probe BG #1-2	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean	54°	90°	126°
LB network with sub-GEIs	64.2	80.6	82.7	76.9	64.8	63.1	68.0	76.9	82.2	75.4	61.3	72.4	89.2	84.3	91.0

Models are trained with gait sequences of the first 74 subjects.

of viewpoint variations. The results show that our method can generalize well for such a large-scale dataset. Notably, there are hundreds of subjects whose GEIs of view angle 85 degree are missing. We did not balance the training samples across different view angles. As a result, there are relatively less GEIs of view angle 85 degree in the training dataset, which should be the reason why the corresponding cross-view recognition scores drop more than the other cases.

5.12 Comparison on USF

According to the official setting, the considered conditions should include one for gallery and 12 for probe (from ProbeA to ProbeL). Our method requires an auxiliary set of data to train networks for gait verification, i.e., predicting the similarity between a pair of gait features. To this end, we follow the strategy proposed by Martín-Félez and Xiang [52]. It amounts to randomly splitting the subjects into two groups, each of which has almost the same number of subjects within the gallery and probe sets respectively. For example, there are 61 subjects in both groups for the gallery set, while there are respectively 16 and 17 subjects for the ProbeK set. We repeat the above described splitting for five times, train five models separately and report their average performances.

TABLE 4
Comparison with Hu's Method [37] and Yu et al.'s Baseline [6] under Different Walking Conditions on CASIA-B by Accuracies (%)

Probe	Gallery	Ours			RLTDA [37]		NN [6]	
		BG	CL	CL74	BG	CL	BG	CL
54°	36°	**92.7**	49.7	84.0	80.8	**69.4**	22.5	16.4
54°	72°	**90.4**	**62.0**	91.0	71.5	57.8	9.0	11.2
90°	72°	**93.3**	**78.3**	98.0	75.3	63.2	31.8	23.5
90°	108°	**88.9**	**75.6**	94.0	76.5	72.1	42.5	25.9
126°	108°	**93.3**	58.1	86.0	66.5	**64.6**	24.6	16.7
126°	144°	**86.0**	51.4	78.0	72.3	**64.2**	33.1	19.2

CL: With a coat. BG: With a bag. Supervised models are trained with GEIs of the first 34 subjects, except for CL74, which is trained with the first 74 subjects.

Among the 12 probe sets, the ProbeA is only different from the gallery set by viewpoint. On this probe set, an LB network can achieve an identification accuracy of 96.7 ± 0.5%, which is obviously better than the previous best performer's 93 percent [37].

6 Discussion

6.1 Datasets for Uncooperative Gait Recognition

It is probably the time to move to gait recognition in natural surveillance videos, considering the promising performances achieved by deep CNNs. A subject may halt, or turn around, so his/her gait sequence is not consecutive. There may be multiple subjects at the same time, and moving objects in the background, so it is harder to extract silhouettes. The cameras may be above the subjects, so more viewpoints should be considered. In the gait community, the above aspects are seldom studied, although they are the very problems to face with if we want to apply gait recognition in practice. Especially, there are no such datasets for gait recognition. In the literature, there are some datasets which are originally designed for cross-camera human tracking or re-identification and can somehow act for this purpose. For example, Bialkowski et al. [53] collected such a dataset with multi-camera surveillance networks. However, there are only 150 sequences in this dataset, compared with 13,640 in CASIA-B. It would be very hard to train cross-view gait recognition models on so small a dataset due to severe over-fitting. Furthermore, there are no gait recognition results reported on this dataset in the literature.

TABLE 5
Cross-View Gait Recognition Results (%) Obtained on OU-ISIR with Our Method and Comparison with the Baseline Reported by the Dataset Authors [28]

Probe angle	Gallery angle				Mean	Identical angle	
	55°	65°	75°	85°		Ours	NN [28]
55°	–	98.3 ± 0.1	96.0 ± 0.1	80.5 ± 0.4	91.6 ± 0.2	98.8 ± 0.1	84.7
65°	96.3 ± 0.2	–	97.3 ± 0.0	83.3 ± 0.3	92.3 ± 0.1	98.9 ± 0.2	86.6
75°	94.2 ± 0.2	97.8 ± 0.2	–	85.1 ± 0.2	92.4 ± 0.1	98.9 ± 0.0	86.9
85°	90.0 ± 0.5	96.0 ± 0.3	98.4 ± 0.1	–	94.8 ± 0.3	98.9 ± 0.1	85.7

After all, it is not established for the gait recognition purpose. So, it is not suitable to compare with previous methods on this dataset. Besides, considering the above mentioned factors, to re-identify a person in unscripted surveillance videos only relying on gait recognition, there still seems a long way to go. Probably, such a dataset with enough number of training data can push us forward to this goal.

6.2 Less Heuristic Preprocessing

There are many methods in the literature which can be used to improve our preprocessing. For example, pedestrian detection methods [54] can locate a subject from complex backgrounds, pixel-wise labeling methods [55], [56] can extract silhouettes from raw images, and pose estimation methods [57] can provide auxiliary information or help refining the silhouettes. Without these comprehensive methods, it would be intractable to deal with the above discussed kind of datasets for uncooperative gait recognition. But in this paper, preprocessing is not our main concern, so we keep it as our future work. In a preliminary experiment, we use a Fast R-CNN [58] model pre-trained with the ImageNet dataset to align silhouettes in the CASIA-B dataset [6]. However, the obtained GEIs are of apparently lower quality. The cause behind this is that the two tasks have very different objectives. In the object detection task, a bounding box with an overlap rate of no less than 0.5 will be treated as a positive sample. However, in gait recognition, we would anticipate more accurate localization. As a result, the existing pedestrian detection models would require more sophisticated adjustment before they can be used for the purpose of gait recognition.

7 Conclusion

This paper has studied a CNN-based gait recognition method, with an extensive empirical evaluation in terms of different recognition tasks, preprocessing approaches and network architectures. With this method, we have updated the best recognition rates on three challenging datasets, showing its robustness to viewpoint and walking condition variations, and its generalization ability to large datasets and complex backgrounds.

Acknowledgments

This work is jointly supported by National Basic Research Program of China (2012CB316300), National Natural Science Foundation of China (61525306, 61420106015). Y. Huang and T. Tan are the corresponding authors of this paper.

References

[1] P. Larsen, E. Simonsen, and N. Lynnerup, "Gait analysis in forensic medicine," *J. Forensic Sci.*, vol. 53, pp. 1149–1153, 2008.

[2] I. Bouchrika, M. Goffredo, J. Carter, and M. S. Nixson, "On using gait in forensic biometrics," *J. Forensic Sci.*, vol. 56, no. 4, pp. 882–889, 2011.

[3] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Comput. Vis. Image Understanding*, vol. 104, no. 2-3, pp. 249–257, 2006.

[4] A. Farhadi and M. K. Tabrizi, "Learning to recognize activities from the wrong view point," in *Proc. 10th Eur. Conf. Comput. Vis.*, 2008, pp. 154–166.

[5] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, Feb. 2006.

[6] 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. 18th Int. Conf. Pattern Recog.*, 2006, pp. 441–444.

[7] G. Zhao, G. Liu, H. Li, and M. Pietikainen, "3D gait recognition using multiple cameras," in *Proc. Int. Conf. Automat. Face Gesture Recog.*, 2006, pp. 529–534.

[8] G. Ariyanto and M. Nixon, "Model-based 3D gait biometrics," in *Proc. Int. Joint Conf. Biometrics*, 2011, pp. 1–7.

[9] M. Goffredo, I. Bouchrika, J. Carter, and M. Nixon, "Self-calibrating view-invariant gait biometrics," *IEEE Trans. Syst., Man, Cybern., Part B*, vol. 40, no. 4, pp. 997–1008, Aug. 2010.

[10] W. Kusakunniran, Q. Wu, J. Zhang, Y. Ma, and H. Li, "A new view-invariant feature for cross-view gait recognition," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 10, pp. 1642–1653, Oct. 2013.

[11] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi, "Gait recognition using a view transformation model in the frequency domain," in *Proc. 9th Eur. Conf. Comput. Vis.*, 2006, pp. 151–163.

[12] W. Kusakunniran, Q. Wu, J. Zhang, H. Li, and L. Wang, "Recognizing gaits across views through correlated motion co-clustering," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 696–709, Feb. 2014.

[13] Y. LeCun, K. Kavukvuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proc. Int. Symp. Circuits Syst.*, 2010, pp. 253–256.

[14] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. 25*, 2012, pp. 1106–1114.

[15] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "OverFeat: Integrated recognition, localization and detection using convolutional networks," arXiv:1312.6229, 2013.

[16] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, Aug. 2013.

[17] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F. Li, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 1725–1732.

[18] H. Yang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3551–3558.

[19] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *arXiv:1406.2199*, 2014.

[20] Z. He, T. Tan, Z. Sun, and X. Qiu, "Towards accurate and fast iris segmentation for iris biometrics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 9, pp. 1670–1684, Sep. 2008.

[21] F. Liu, D. Zhang, and L. Shen, "Study on novel curvature features for 3d fingerprint recognition," *Neurocomputing*, vol. 168, no. 1, pp. 599–608, 2015.

[22] X. Shi, Z. Guo, and Z. Lai, "Face recognition by sparse discriminant analysis via joint l2,1-norm minimization, pattern recognition," *Pattern Recog.*, vol. 47, no. 7, pp. 2447–2453, 2014.

[23] Z. Lai, Y. Xu, Z. Jin, and D. Zhang, "Human gait recognition via sparse discriminant projection learning," *IEEE Trans. Circits Syste. Video Technol.*, vol. 24, no. 10, pp. 1651–1662, Oct. 2014.

[24] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 1701–1708.

[25] R. H. S. Chopra and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2005, pp. 539–546.

[26] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 1891–1898.

[27] M. Hu, Y. Wang, Z. Zhang, J. Little, and D. Huang, "View-invariant discriminative projection for multi-view gait-based human identification," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 12, pp. 2034–2045, Dec. 2013.

[28] 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. Inf. Forensics Security*, vol. 7, no. 5, pp. 1511–1521, Oct. 2012.

[29] 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. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 162–177, Feb. 2005.

[30] R. Bodor, A. Drenner, D. Fehr, O. Masoud, and N. Papanikolopoulos, "View-independent human motion classification using image-based reconstruction," *Image Vis. Comput.*, vol. 27, no. 8, pp. 1194–1206, Jul. 2009.

[31] C. Wang, J. Zhang, J. Pu, X. Yuan, and L. Wang, "Chrono-gait image: A novel temporal template for gait recognition," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 257–270.

[32] T. Lam, K. Cheung, and J. Liu, "Gait flow image: A silhouette-based gait representation for human identification," *Pattern Recog.*, vol. 44, no. 4, pp. 973–987, Apr. 2010.

[33] W. Kusakunniran, Q. Wu, H. Li, and J. Zhang, "Multiple views gait recognition using view transformation model based on optimized gait energy image," in *Proc. Workshop Tracking Humans Eval. Their Motion Image Sequences*, 2009, pp. 1058–1064.

[34] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li, "Support vector regression for multi-view gait recognition based on local motion feature selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 974–981.

[35] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li, "Gait recognition under various viewing angles based on correlated motion regression," *IEEE Trans. Circits Syst. Video Technol.*, vol. 22, no. 6, pp. 966–980, Jun. 2012.

[36] K. Bashir, T. Xiang, and S. Gong, "Cross-view gait recognition using correlation strength," in *Proc. Brit. Mach. Vis. Conf.*, 2010, pp. 1–11.

[37] H. Hu, "Enhanced gabor feature based classification using a regularized locally tensor discriminant model for multiview gait recognition," *IEEE Trans. Circits Syst. Video Technol.*, vol. 23, no. 7, pp. 1274–1286, Jul. 2013.

[38] Y. LeCun, B. Boser, J. Denker, and D. Henderson, "Handwritten digit recognition with a back-propagation network," in *Proc. Adv. Neural Inf. Proces. Syst.*, 1989, pp. 396–404.

[39] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[40] G. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv:1207.0580*, 2012.

[41] Y. Sun, X. Wang, and X. Tang, "Hybrid deep learning for face verification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1489–1496.

[42] A. Bissacco and S. Soatto, "Hybrid dynamical models of human motion for the recognition of human gaits," *Int. J. Comput. Vis.*, vol. 85, no. 1, pp. 101–114, May 2009.

[43] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[44] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual rcognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[45] (2014). The first international workshop on action recognition with a large number of classes [Online]. Available: http://crcv.ucf.edu/ICCV13-Action-Workshop

[46] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Syst., Man, Cybern., Part B*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[47] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[48] A. Grushin, D. Monner, J. Reggia, and A. Mishra, "Robust human action recognition via long short-term memory," in *Proc. Int. Joint Conf. Neural Netw.*, 2013, pp. 1–8.

[49] V. Veeriah, N. Zhuang, and G. Qi, "Differential recurrent neural networks for action recognition," *arXiv:1504.06678*, 2015.

[50] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2006, pp. 2169–2178.

[51] T. B. G. Huang, M. Ramesh and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Proc. ECCV Workshop Faces Real-life Images: Detection, Alignment, and Recognition*, 2008, pp. 1–11.

[52] R. Martín-Félez and T. Xiang, "Uncooperative gait recognition by learning to rank," *Pattern Recog.*, vol. 47, no. 12, pp. 3793–3806, Dec. 2014.

[53] A. Bialkowski, S. Denman, S. Sridharan, and C. Fookes, "A database for person re-identification in multi-camera surveillance networks," in *Proc. Digit. Image Comput. Tech. Appl.*, 2012, pp. 1–8.

[54] B. Yang, J. Yan, Z. Lei, and S. Li, "Convolutional channel features: Tailoring CNN to diverse tasks," *arXiv:1504.07339*, 2015.

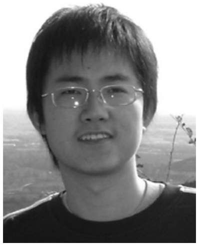
[55] Z. Wu, Y. Huang, Y. Yu, L. Wang, and T. Tan, "Early hierarchical contexts learned by convolutional networks for image segmentation," in *Proc. 22nd Int. Conf. Pattern Recog.*, 2014, pp. 1538–1543.

[56] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *arXiv:1411.4038*, 2014.

[57] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 1653–1660

[58] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

Zifeng Wu received the BSc and MSc degrees from the School of Mechanical Engineering and Automation at Beihang University, Beijing, China, in 2006 and 2009, respectively, the PhD degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2015. He is now a researcher at the University of Adelaide, Adelaide, South Australia, Australia. His research interests include computer vision and deep learning.



Yongzhen Huang received the BE degree from the Huazhong University of Science and Technology in 2006 and the PhD 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. He is a member of the IEEE.

Liang Wang received both the BEng and MEng degrees from Anhui University in 1997 and 2000, respectively, and the PhD degree from the Institute of Automation, Chinese Academy of Sciences (CASIA) in 2004. From 2004 to 2010, he has been a research assistant at Imperial College London, United Kingdom and Monash University, Australia, a research fellow at the University of Melbourne, Australia, and a lecturer at the University of Bath, United Kingdom, respectively. Currently, he is a full professor of Hundred Talents Program at the National Lab of Pattern Recognition, CASIA. His major research interests include machine learning, pattern recognition, and computer vision. He has widely published at highly-ranked international journals such as *IEEE Transactions on Pattern Analysis and Machine Intelligence* and *IEEE Transactions on Image Processing*, and leading international conferences such as CVPR, ICCV, and ICDM. He is an associate editor of *IEEE Transactions on SMC-B*, *International Journal of Image and Graphics*, *Signal Processing*, *Neurocomputing,* and *International Journal of Cognitive Biometrics*. He is currently a senior member of the IEEE and a fellow of the IAPR.

Xiaogang Wang received the BS degree in electronic engineering and information science from the Special Class for Gifted Young, University of Science and Technology of China in 2001, the MPhil degree in information engineering from the Chinese University of Hong Kong in 2004, and the PhD degree in computer science from the Massachusetts Institute of Technology. He is currently an associate professor in the Department of Electronic Engineering at the Chinese University of Hong Kong. He was the area chairs of ICCV 2011 and 2015, ECCV 2014 and 2016, and ACCV 2014 and 2016. He is an associate editor of the *Image and Visual Computing Journal*. His research interests include computer vision and machine learning. He is a member of the IEEE.

Tieniu Tan received the BSc degree in electronic engineering from Xi'an Jiaotong University, China, in 1984, and the MSc and PhD degrees in electronic engineering from Imperial College London, United Kingdom, in 1986 and 1989, respectively. In October 1989, he joined the Computational Vision Group at the Department of Computer Science, The University of Reading, Reading, United Kingdom, where he was a research fellow, senior research fellow, and lecturer. In January 1998, he returned to China to join the National Laboratory of Pattern Recognition (NLPR), Institute of Automation of the Chinese Academy of Sciences (CAS), Beijing, China, where he is currently a professor and the director of the Center for Research on Intelligent Perception and Computing (CRIPAC), and was a former director from 1998 to 2013, of the NLPR and director general of the Institute from 2000 to 2007. He is currently also a vice president of the Chinese Academy of Sciences. He has served as a chair or program committee member for many major national and international conferences. He is or has served as an associate editor or member of editorial boards of many leading international journals including *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Automation Science and Engineering*, *IEEE Transactions on Information Forensics and Security*, *IEEE Transactions on Circuits and Systems for Video Technology*, *Pattern Recognition*, *Pattern Recognition Letters*, *Image and Vision Computing*, etc. He is an editor-in-chief of the *International Journal of Automation and Computing*. He was a founding chair of the IAPR Technical Committee on Biometrics, the IAPR-IEEE International Conference on Biometrics, the IEEE International Workshop on Visual Surveillance and Asian Conference on Pattern Recognition (ACPR). He is currently the immediate past president of the IEEE Biometrics Council and deputy president of Chinese Artificial Intelligence Association. His current research interests include biometrics, image and video understanding, and information content security. He is a fellow of CAS, TWAS (The World Academy of Sciences for the advancement of science in developing countries), IEEE, and IAPR (the International Association of Pattern Recognition), and an international fellow of the UK Royal Academy of Engineering.