Research Article

Wanjiang Xu*

Deep Large Margin Nearest Neighbor for Gait Recognition

https://doi.org/10.1515/jisys-2020-0077
Received Aug 17, 2020; accepted Feb 08, 2021

Abstract: Gait recognition in video surveillance is still challenging because the employed gait features are usually affected by many variations. To overcome this difficulty, this paper presents a novel Deep Large Margin Nearest Neighbor (DLMNN) method for gait recognition. The proposed DLMNN trains a convolutional neural network to project gait feature onto a metric subspace, under which intra-class gait samples are pulled together as small as possible while inter-class samples are pushed apart by a large margin. We provide an extensive evaluation in terms of various scenarios, namely, normal, carrying, clothing, and cross-view condition on two widely used gait datasets. Experimental results demonstrate that the proposed DLMNN achieves competitive gait recognition performances and promising computational efficiency.

Keywords: gait recognition; subspace learning; distance metric learning

1 Introduction

Gait recognition, aiming to identify humans at a distance by inspecting their walking manners, has recently received increasing attentions [17]. Compared with other biometrics (e.g., facial, iris, fingerprint), human gait has some important advantages: 1) it can work well at a distance when other biometrics are obscured or the resolution is insufficient; 2) it is difficult to imitate or camouflage because it is people’s long standing habit; 3) it is non-intrusive as it does not require the cooperation of the subject. These properties make gait be suitable for security, surveillance applications perfectly [4].

There has already been a lot of works on gait recognition. One of the famous methods is Gait Energy Image (GEI) [7]. GEI is formed by averaging properly aligned human silhouettes of a gait period. Figure 1 shows example GEIs of two subjects. Unfortunately, there are some covariate factors (such as clothing, carrying, viewpoint and so on) affecting the appearance of GEI drastically. As seen in Figure 1, GEIs vary greatly in different conditions even if they belong to the same person. As a result, there will be a drastic negative impact on gait recognition [6].

To improve the accuracy of successful matching gait features, a distance metric learning method such as large margin nearest neighbor (LMNN) [24] can be applied to reduce the intra-subject variation and increase the inter-subject variation. A linear mapping function is often used to transform feature space into a distance metric space, in which gait similarity is measured for recognition. However, when gait features are highly nonlinear distributed, linear methods are difficult to extract effectively gait features.

In recent years, deep learning (DL) [5, 9, 20, 25] has achieved excellent success in various computer vision and pattern recognition tasks. In fact, deep neural network is a highly non-linear model which could extract rich and discriminant features [25]. Benefit from DL, in this paper, we employ deep convolutional neural networks instead of linear transformation of LMNN to learn the metric space, which is termed as Deep Large Margin Nearest Neighbor (DLMNN). As shown in Figure 2, DLMNN learns a deep discriminant distance metric space, under which the similarities of gait samples can be measured properly for classification.
Deep Large Margin Nearest Neighbor for Gait Recognition

Figure 1: Example GEIs of two persons in CASIA-B gait dataset [30]. The leftmost column is the GEIs under viewing angle 90 in normal condition, while the rest are GEIs with covariates such as clothing, carrying and view.

Figure 2: Schematic illustration of our proposed DLMNN. Deep neural network transforms samples from input space into feature space so that positive samples lie within a small radius and negative samples lie outside with a margin.

The contributions of this paper are as follows. (1) We propose a new deep learning based distance metric learning method, called Deep Large Margin Nearest Neighbor, which is the improvement of the famous LMNN. (2) An elaborate learning framework and training algorithm are provided for DLMNN. (3) DLMNN is applied for gait recognition and achieves competitive performance on a set of evaluation experiments.

The rest of the paper is organized as follows. Section 2 discusses related works. Section 3 reviews a distance metric learning approach Large Margin Nearest Neighbor who motivates our work. Section 4 describes the framework of the proposed method and its training process. Section 5 presents experimental results on two benchmark datasets. Section 6 gives the conclusion.

2 Related Works

Many gait recognition techniques have been developed in recent years, which can be generally classified into two typical categories: model-based methods [1, 23, 28] and appearance-based methods [7, 12, 15, 22]. The model-based methods generally characterize kinematics of human joints to measure physical gait parameters such as trajectories, limb lengths, and angular speeds. However, human body is a highly flexible structure, and it is difficult to precisely restore body structures from images or videos in many scenarios. Without explicitly considering the underlying structure appearance-based methods extract gait features directly from videos. Generally, appearance-based methods first detect and crop human silhouettes from all frames in one video, then convert a sequence of frames into one gait template image for similarity measurement. Several gait
templates have been proposed over the last decades, such as GEI [7], GEnI [12], GFI [15] and CGI [22]. These template images reserve rich motion and shape information of human walking. Han and Bhanu [7] proposed gait energy image (GEI) as the feature representation by averaging silhouettes over one gait cycle. Bashir et al. [12] proposed gait entropy image (GEnI) encoding the randomness of pixel values in the silhouette images over a complete gait cycle. Lam et al. [15] proposed gait flow image (GFI) using an optical flow field to emphasize timing information in a gait cycle. Wang et al. [22] proposed Chrono-Gait image (CGI) encoding the temporal information via color mapping. Recently, Iwama et al. [11] illustrated that GEI was the most effective gait template by comprehensive gait recognition experiments on their proposed gait dataset consisting of more than 3,000 subjects. However, they also found that GEI performs well when there are no covariates, while it is error-prone when covariates exist.

Many researchers have studied various feature extractors to learn discriminant gait feature to cope with different covariates. Guan et al. [6] proposed a classifier ensemble method based on random subspace method and majority voting for clothing-invariant gait recognition. Huang and Boulgouris [10] developed shifted energy image and gait structural feature extraction algorithm to address carrying factor. Ben et al. [3] proposed a Coupled Patch Alignment (CPA) algorithm for cross-view gait recognition. These works have satisfactory performance against one specific covariate. However, their recognition precisions would drop drastically when other covariates exist. These methods are traditional machine learning methods which are mostly based on linear transformation. As a consequence, they may not work well in much complicated multi-covariate cases.

Recently, deep learning has made rapid progress in the past few years in many areas. Particularly, the deep convolutional neural networks (CNN) were used to tackle with complicated computer vision tasks [5, 20, 30], updating the record scores one after another. As for gait recognition work, Shiraga et al. [21] proposed GEINet based on CNN and GEI. CNN can learn rich feature in a discriminative manner due to its deep and highly non-linear model. However, they employ traditional softmax loss function which is more suitable for image classification rather than for similarity measurement. Wu et al. [25] adopted CNN to measure similarity of any two GEIs and achieved best performance in their cross-view gait recognition experiments. However, the input of their network is a pair of GEIs, one gallery and one probe. That means in testing phase it incurs much high computational cost for measuring all pairs of GEIs. Yu et al. [29] proposed GaitGAN to transform gait data from any viewing, clothing, and carrying conditions to the side view with normal condition. They adopted Generative Adversarial Networks (GAN) as a regressor to generate invariant gait images. However, the generated gait images contain lots of noise information which may decrease recognition precision. Zhang et al. [31] developed a Siamese neural network framework with contrastive loss function for gait recognition. Their method is based on distance metric learning which can learn effective features automatically, leading to good recognition performance. Our proposal in this paper also adopts distance metric learning based on CNN, and we find that the proposed method can extract robust and discriminative gait features.

3 Large Margin Nearest Neighbor

In this section, we briefly introduce distance metric learning (DML) and the learning framework of Large Margin Nearest Neighbor (LMNN) classifier.

3.1 Distance Metric Learning

Distance Metric Learning [26] aims to learn a distance metric for the input space of data from a given collection of pair of similar/dissimilar samples that preserves the distance relation among the training data. Let \( \mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \) be the training set, where \( \mathbf{x}_i \in \mathbb{R}^d \) is the \( i \)th training sample and \( n \) is the total number of training samples. A typical distance metric learning aims to seek a square matrix \( \mathbf{M} \in \mathbb{R}^{d \times d} \) from the training set \( \mathbf{X} \), under which the distance between two samples \( \mathbf{x}_i \) and \( \mathbf{x}_j \) can be measured as:

\[
d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j)}
\] (1)
The matrix $M$ is a positive semi-definite matrix. It can be factorized as $M = W^T W$, where $W \in R^{p \times d}$ and $p < d$. Therefore, $d_M(x_i, x_j)$ can be denoted as follows:

$$d_M(x_i, x_j) = \sqrt{(x_i - x_j)^T M(x_i - x_j)} = \sqrt{(x_i - x_j)^T W^T W (x_i - x_j)} = \|W(x_i - x_j)\|_2$$

(2)

Learning such a distance metric is equivalent to finding a projection matrix $W$. The matrix can map input space to the metric space, in which the Euclidean metric is applied for measurement.

### 3.2 Large Margin Nearest Neighbor

Large Margin Nearest Neighbor (LMNN) [24] is one of the most famous DML based methods, which learns a matrix $W$ that minimizes the distance between each training sample and its $K$ nearest similarly labeled neighbors, while maximizes the distance between all differently labeled samples. The objective of LMNN is shown as follows, that consists of two terms, one which acts to pull same-class neighbors closer together, and another which acts to push different-class samples further apart.

$$L_{LMNN} = \sum_{i,j \rightarrow i} \sum_{l} (1 - y_{il}) \left[ \tau + d_M(x_i, x_l) - d_M(x_i, x_j) \right] + \gamma \sum_{i,j \rightarrow i} d_M(x_i, x_j)$$

(3)

where $y_{ij}$ is indicator variable $y_{ij} = 1$ if and only if $x_i$ and $x_j$ have the same label, and $y_{ij} = 0$ otherwise; $j \rightarrow i$ denotes that $x_i$ is similarly labeled neighbor of $x_j$; $\lfloor \cdot \rfloor_+ = \max(\cdot, 0)$ denotes the standard hinge loss; $\tau$ is the predefined margin; $\gamma$ is a balance parameter.

There are two kinds of distances in LMNN: one for same-class pairs (input sample and its similarly labeled samples), and another for different-class pairs (input sample and its differently labeled samples). The first term in Eq. (3) is the inter-class loss which penalizes small distances between differently labeled samples. In the metric space, the distances between objective sample and differently labeled samples should be larger than the distances between objective sample and similarly labeled sample with a large margin. The second term is the intra-class loss which penalizes large distances between each input sample and its similarly labeled neighbors. In the metric space, these distances should be as small as possible. The balance parameter $\gamma$ balances the two goals. Finally, the overall objective of Eq. (3) maximizes the margin by pulling same-class pairs of samples together and pushing different-class pairs further apart.

### 4 Proposed Approach

#### 4.1 Deep Distance Metric Learning

As discussed in section 3, the conventional distance metric learning method (such as LMNN) only seeks for an optimal linear projection matrix to project original input space into the metric space. In this work, we apply a deep convolutional neural network (CNN), instead of linear matrix as the projection function $f(\cdot)$.

Given a pair of samples $x_i$ and $x_j$, they can be represented as $f(x_i)$ and $f(x_j)$ when they are passed through a deep convolutional neural network. Their distance can be measured by computing the squared Euclidean distance between $f(x_i)$ and $f(x_j)$, which is defined as follows:

$$d_f^2(x_i, x_j) = \|f(x_i) - f(x_j)\|_2^2$$

(4)

Based on Eq. (4), different objective (loss) functions can be provided to obtain deep non-linear mapping function $f(\cdot)$. With function $f(\cdot)$, each sample is projected onto the metric space. Because of the great success of LMNN in pattern recognition area, a similar loss function (DLMNN loss) is applied to minimize the distance between same-class samples and maximize the distance between different-class samples simultaneously.
4.2 DLMNN framework

As described in section 3, there are two kinds of distances in LMNN: the distance between two same-class samples and the distance between two different-class samples. In this work, to obtain the two distances in a deep CNN based model, we use three CNNs to compute the representations of two similarly labeled samples and one differently labeled sample, respectively.

The framework is shown in Figure 3. Triplet samples (GEIs) are as input of the proposed method. Three GEIs forms the i-th triplet, denoted by a triplet \( < x_i^0, x_i^+, x_i^- > \), where \( x_i^0 \) and \( x_i^+ \) are from the same person, while \( x_i^- \) is from a different person. The three GEIs are passed to three CNNs which share the same parameters, i.e., weights and bias. Through the three CNNs, we map the three GEIs from input space into feature space, where \( < x_i^0, x_i^+, x_i^- > \) is represented as \( < f(x_i^0), f(x_i^+), f(x_i^-) > \).

![Figure 3: The training framework of the proposed DLMNN method for gait recognition.](image)

Similar to LMNN, the learned space in our method will have the property that the distance between same-class samples \( f(x_i^0) \) and \( f(x_i^+), \) denoted as \( d_i^2(x_i^0, x_i^+) \), is small enough, meanwhile the distance between different-class samples \( f(x_i^0) \) and \( f(x_i^-), \) denoted as \( d_i^2(x_i^0, x_i^-) \), is larger than \( d_i^2(x_i^0, x_i^+) \) with a predefined margin. As a consequent, our DLMNN loss function is defined as follows:

\[
L_i = \frac{1}{2} \left[ \tau + d_i^2(x_i^- - x_i^+) - d_i^2(x_i^0 - x_i^-) \right]^+ + \frac{\gamma}{2} d_i^2(x_i^0 - x_i^-)
\]

where \([\cdot]^+\) is the function \( \max(\cdot, 0) \), \( \tau \) is the predefined margin, and \( \gamma \) is a balance factor to balance the two terms. The loss function aims to pull the samples of same person closer, and meanwhile put the samples of different person father from each other in the learned space.

4.3 The Training Algorithm

We use stochastic gradient decent algorithm to train the proposed CNN architecture model with the DLMNN loss function. Three CNNs are used to extract gait feature. The derivative of \( f(x_i^0) \) can be computed as follows:

\[
\frac{\partial L_i}{\partial f(x_i^0)} = \frac{1}{2} \frac{\partial}{\partial f(x_i^0)} \left[ \tau + ||f(x_i^-) - f(x_i^0)||^2 - ||f(x_i^+) - f(x_i^0)||^2 \right]^+ + \gamma (f(x_i^-) - f(x_i^0))
\]

\[
= \begin{cases} 
\gamma (f(x_i^-) + f(x_i^0)) - (1 + \gamma) \cdot f(x_i^0) & \text{if } \tau + ||f(x_i^-) - f(x_i^0)||^2 - ||f(x_i^+) - f(x_i^0)||^2 > 0 \\
\gamma f(x_i^-) - \gamma f(x_i^0) & \text{elsewise}
\end{cases}
\]
The derivative of $f(x_i)$ can be computed as follows:

$$
\frac{\partial L_i}{\partial f(x_i)} = \frac{1}{2} \frac{\partial}{\partial f(x_i)} \left[ \tau + ||f(x_i) - f(x_i')||^2 - ||f(x_i') - f(x_i'')||^2 \right] - \gamma \cdot (f(x_i') - f(x_i'')) \\
= \begin{cases} 
(1 + \gamma) \cdot (f(x_i') - f(x_i'')) & \text{if } \tau + ||f(x_i') - f(x_i'')||^2 - ||f(x_i') - f(x_i'')||^2 > 0 \\
\gamma f(x_i') - \gamma f(x_i'') & \text{otherwise}
\end{cases}
$$

And the derivative of $f(x_i')$ can be computed as follows:

$$
\frac{\partial L_i}{\partial f(x_i')} = \frac{1}{2} \frac{\partial}{\partial f(x_i')} \left[ \tau + ||f(x_i') - f(x_i)||^2 - ||f(x_i') - f(x_i)||^2 \right], \\
= \begin{cases} 
f(x_i') - f(x_i') & \text{if } \tau + ||f(x_i') - f(x_i)||^2 - ||f(x_i') - f(x_i)||^2 > 0 \\
0 & \text{otherwise}
\end{cases}
$$

Because the three CNNs share the same weights, the derivatives of the weights $w$ can be computed as follows:

$$
\frac{\partial L_i}{\partial w} = \frac{\partial L_i}{\partial f(x_i)} \frac{\partial f(x_i)}{\partial w} + \frac{\partial L_i}{\partial f(x_i')} \frac{\partial f(x_i')}{\partial w} + \frac{\partial L_i}{\partial f(x_i'')} \frac{\partial f(x_i'')}{\partial w}
$$

From above derivations, it is clear that the gradient on each input triplet can be easily computed given the values of $f(x_i')$, $f(x_i)$, $f(x_i'')$ and $\frac{\partial f(x_i)}{\partial w}$, $\frac{\partial f(x_i')}{\partial w}$, $\frac{\partial f(x_i'')}{\partial w}$. They can be obtained by running standard forward and backward propagations for each image in the triplet examples. For each iteration, we exploit mini-batch stochastic gradient descent algorithm, which needs to go through all triplets in each batch to accumulate the gradients. Algorithm 1 shows the main process of the training algorithm.

**Algorithm 1:** DLMNN training algorithm

1: Initialize the network parameters $w$, $t = 0$.
2: while $t <$ Maximum iterative number $T$ do
3:   select K sample triplets in training set $X$ to form a training subset $D$.
4: end while
5: for all training triplet samples $< x_i', x_i'', x_i'' >$ in subset $D$ do
6:   Calculate $f(x_i')$, $f(x_i)$, $f(x_i'')$ by forward propagation.
7: end for
9: Calculate $\frac{\partial L_i}{\partial w}$ according to Eq. (9).
10: Update the parameters $w^{(t)} = w^{(t-1)} - \lambda \frac{\partial L_i}{\partial w}$
11: $t = t + 1$.
12: return $w$

**4.4 Discussions**

To further clarify the effect of our method, this section will discuss in detail the differences between our method and previous closely related methods. For better illustration, we present the 2D distribution of feature learned by these methods on MNIST [16] dataset as shown in Figure 4.

Difference form Discriminative Deep Metric Learning [9] and Contrastive loss [31]. Contrastive loss is formulated as $L_{cont} = d_f(x_i' - x_i') + [\tau - d_f(x_i - x_i')]_+$, and DDML loss is formulated as $L_{DDML} = [1 - l_{ij}(\tau - d_f(x_i - x_i))]_+$, where the value of $l_{ij}$ is 1 or -1. Each pair of samples is independently penalized in their networks. Conversely, positive pair and negative pair are simultaneously penalized in our method.
Large margin between the distance of positive pair and that of negative pair is kept for a better classification. As shown in Figure 4, compared to DDML and Contrastive loss, our DLMNN could learn a more discriminative subspace with large between-class scatter.

Difference from Triplet loss [19]. Triplet loss is formulated as $L_{tri} = [\tau + d_f(x^+_i - x^-_i) - d_f(x^+_i - x^-_i)]^+$. It is a part of DLMNN loss compared with formula (5). Our DLMNN not only maintains the large margin between the distance of negative pair and that of positive pair, but also shrinks the distance of positive pair continuously. As shown in Figure 4, DLMNN delivers smaller within-class scatter than triplet loss. And it is very beneficial for discriminative feature learning.

5 Experiments

5.1 Parameter setting

5.1.1 Datasets

Extensive experiments have been conducted on the two largest benchmark gait datasets: CASIA-B [30] and OU-ISIR-LP [11]. CASIA-B dataset [30] is one of the most widely used gait dataset to evaluate gait recognition across different viewing angles. This database contains 128 subjects from 11 views ($0^\circ$, $18^\circ$, $\ldots$, $180^\circ$). There are six normal, two carrying, and two wearing gait sequences for each subject under each view. Figure 5 shows the examples at 11 different views from a subject of normal walking.

The second dataset is OU-ISIR-LP gait dataset [11]. OU-ISIR-LP is the largest gait dataset which was created by Institute of Scientific and Industrial Research, Osaka University. In OU-ISIR-LP, there are 4,007 subjects (2,135 males and 1,872 females) with ages ranging from 1 to 94 years old. Gait data was captured using a single camera placed at a 5-meter distance from the course. For each subject, there are two sequences available, one in the gallery and the other as a probe sample. Example images of the subjects are shown in Figure 6.
5.1.2 Gait Feature Representation

In this work, we use Gait Energy Image (GEI) [7] to represent gait. As shown in Figure 7, firstly, extract human silhouettes from a raw sequence using image segmentation algorithm [8]. Then, align and scale each human silhouette to standard size. Finally, average the silhouettes along temporal dimension to get a GEI.

Specifically, let \( I(x, y, t) \) represent a normalized and aligned walking binary silhouette sequence. The grey-level GEI \( G(x, y) \) is defined as follows.

\[
G(x, y) = \frac{1}{T} \sum_{t=1}^{N} I(x, y, t)
\]

where \( N \) is the number of frames in complete cycles of the sequence, \( t \) is the frame number of the sequence, \( x \) and \( y \) are values in the 2D image coordinate. GEI contains rich information of human gait including human shape, motion frequency, temporal and spatial changes of human body.
5.1.3 Classifier

To perform recognition, we have gait templates of subjects as our gallery gait $x^g_l$ ($l = 1, 2, \ldots, n$). Any probe gait $y^p$ can now be recognized as the same subject in the gallery. The projection function $f(\cdot)$ uses CNN for feature extraction. The identity is estimated by the nearest neighbor classifier, which can be written as

$$\arg\min_{l=1,2,\ldots,n} ||f(y^p) - f(x^g_l)||$$

(11)

where $n$ is the amount of gallery samples.

5.1.4 Network Parameters

The CNN architecture of DLMNN in this work is shown in Figure 8. Each convolutional kernel size is $3 \times 3$. Each convolutional layer is followed by a rectified linear unit (ReLU) except the last one (Conv52). The first four pooling layers use max operator. To generate a compact and discriminative feature representation, we use average pooling for the last pooling layer (pool5). The feature dimensionality of pool5 is thus equal to the number of channels of Conv52 which is 320. The last layer is fully connected layer FC6, which is used for gait feature representation. The extracted features are further L2-normalized into unit length before metric learning stage. By the CNN, the dimensions of gait feature are reduced from $128 \times 88$ to 128.

Figure 7: Pipeline of generating GEI.

Figure 8: Network backbone for gait recognition.
The weights are initialized using Gaussian distribution with a mean of zero and a standard deviation of 0.001. The bias terms are set to 0. 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 ten at 50,000th iteration and 200,000th iteration, respectively. The total number of iterations was 500,000. We use the standard batch size 128 for the training phase. Each element in the batch is a triplet containing two same-class samples and one different-class sample. We select one person with two of his (her) GEIs randomly and select one GEI from the remaining persons randomly to form a triplet. Our DLMNN network was trained and tested using Caffe on a Nvidia GTX 960 GPU.

5.2 Experimental Design

Firstly, we experiment on the CASIA-B gait database to evaluate the performance of the proposed method. We put the six normal, two clothing coats and two carrying bags sequences of the first 74 subjects into training set and the remaining 50 subjects into testing set. In test set, the first 4 normal walking sequences of each subject are put into gallery set and the other into probe set. Table 1 lists the experimental design. In the following experiments, we evaluate the proposed method on no-covariate, clothing-covariate, carrying-covariate, and view-covariate gait recognition, respectively.

Table 1: Experimental design on CASIA-B dataset.

| Training | Test |   |
|----------|------|---|
| Gallery set | Probe set |   |
| ID: 001-074 | ID: 075-124 | ID: 075-124 |
| Seqs: nm01-06 | Seqs: nm01-04 |   |
| bg01-02, cl01-02 | bg01-02, cl01-02 |   |

The second gait database which we employ to evaluate the proposed method is OU-ISIR-LP. There are two sequences for each subject in the dataset: gallery and probe. The experimental design on OU-ISIR-LP database is shown in Table 2. In the experiment, gallery set is used for training. Because there is only view variation (viewing angle is range from 55° to 85°) considered in this dataset, we evaluate our method on no-variation and view-variation gait recognition respectively in the following experiments.

Table 2: Experimental design on OU-ISIR-LP dataset.

| Training | Test |
|----------|------|
| Gallery set | Probe set |
| gallery sequences | gallery sequences |
|                     | probe sequences |

5.3 Experiments on no-variation gait recognition

For no-variation gait recognition on CASIA-B dataset, we put the first 4 normal sequences at a specific view into the gallery set, and the rest 2 normal condition sequences into probe set. Table 3 shows the recognition results of different methods at each view in normal condition. Three typical feature extraction methods PCA [13], LDA [2] and one DML based method LMNN [24] are used for comparison. There are 11 views in the dataset so that 11 recognition rates are achieved by each methods. From Table 3, we can see that all methods achieve
Table 3: The recognition rates (%) of different methods in no-variation condition evaluated on CASIA-B.

| Methods | Probe viewing angle |
|---------|---------------------|
|         | 0° | 18° | 36° | 54° | 72° | 90° | 108° | 126° | 144° | 162° | 180° |
| PCA     | 100 | 99  | 97  | 96  | 96  | 94  | 96   | 96   | 98   | 98   | 99   |
| LDA     | 100 | 100 | 98  | 99  | 99  | 99  | 97   | 79   | 98   | 99   |      |
| LMNN    | 97  | 98  | 96  | 97  | 97  | 98  | 98   | 97   | 97   | 98   |      |
| DLMNN   | 100 | 100 | 99  | 99  | 100 | 100 | 100  | 99   | 99   | 100  | 100  |

pretty performances. This illustrates gait is a good biometric feature for person identification in computer vision when there are no intra-subject variations.

The experimental results on OU-ISIR-LP dataset are shown in Table 4. SiaNet [31] is deep learning based metric learning method using Siamese net and contrastive loss. The sores of SiaNet are directly taken from the original paper, and the comparison is only conducted between the results obtained with the same division of the training and testing data. Generally speaking, our method performs better than other methods.

Table 4: The recognition rates (%) of different methods in no-variation condition evaluated on OU-ISIR-LP.

| Methods | Probe viewing angle |
|---------|---------------------|
|         | 55° | 65° | 75° | 85° |
| PCA     | 84.7 | 86.63 | 86.91 | 85.72 |
| LDA     | 77.28 | 77.95 | 73.77 | 57.74 |
| SiaNet  | 90.12 | 91.14 | 91.18 | 90.43 |
| DLMNN   | 92.55 | 94.3 | 95.81 | 94.13 |

Figure 9: The recognition rates of all methods in clothing condition.
5.4 Experiments on clothing-covariate gait recognition

We carry out clothing-covariate gait recognition experiments on CASIA-B dataset. The methods for comparison are PCA [13], LDA [2], SRC [27], SRC-V [27], and LMNN [24]. SRC is a sparse representation based classifier and SRC-V is a SRC method with external variation dictionary. From Figure 9, we can see that the remarkable improvements of recognition rates have been achieved by the proposed method in all probe viewing angles.

5.5 Experiments on carrying-covariate gait recognition

The results in Figure 10 evaluate carrying covariate. The adopted database is CASIA-B. The two carrying condition gait sequences at each view are put into probe set. As shown in Figure 10, SRC-V [27] and our method perform better than other methods. And DLMNN performs best generally. LMNN and our DLMNN are both metric learning based method. They have similar objective function. The recognition rates of the two methods are quite different. Compared to LMNN, the proposed method is based on deep learning, which learns a more discriminant metric space. As a result, DLMNN makes a great improvement.

![Performance in carrying bag condition](image)

*Figure 10:* The recognition rates of all methods in carrying condition.

5.6 Experiments on view-variation gait recognition

We evaluate the proposed method in cross-view gait recognition task since viewing angle change is the most common factor impacting gait recognition performance. There are 11 different views in CASIA-B database. Therefore, there are $11 \times 10$ cross-view gait recognition rates totally. We select one view as probe view when the rest views as gallery views. The methods for comparison are PCA [13], VTM [14] and LMNN [24]. VTM method is a state-of-the-art method for cross-view gait recognition. VTM [14] uses view transform model transforming gait feature from one view to another view, to recognize gait across different views. The experimental results are shown in Figure 11. Generally, the two distance metric learning based methods, DLMNN and LMNN, perform better than PCA at all probe angle and gallery angle pairs. Distance metric learning aims to learn a metric space in which same-class samples are clustered and different-class samples are separated. Therefore, DML based
methods are suitable for gait recognition or classification task. Compared to LMNN, our proposed DLMNN method provides a significant improvement in the cross-view recognition results.

Figure 11: Comparison with PCA, VTM and LMNN at different probe viewing angles.

Figure 12: Comparison of the cross-view matching approaches on different groups. Group A D stand for (65,75), (75,65), (75,85), and (85,75).
We also evaluate the proposed method on OU-ISIR-LP dataset. There are 4 viewing angles in the dataset, producing $4 \times 3$ cross-view recognition results totally. We select 4 pairs of cross-view tests for comparison with VTM [14] and SiaNet [31]. As shown in Figure 12, the performance of the proposed method is best. It demonstrates that the proposed method is robust to view-change variations in this four testing groups. Compared to traditional method VTM [14], SiaNet [31] and DLMNN improve the recognition rate obviously because they can automatically learn commendable features with the non-linear projections of deep CNN. Our proposed DLMNN outperforms the state-of-the-art method SiaNet. The large margin constraint used in deep metric learning brings a more discriminant subspace.

Moreover, cumulative match score (CMS) curves are used to further demonstrate the performance of cross-view gait recognition as seen in Figure 13. It is noted that horizontal axis is rank (top n matches) and the vertical axis is the recognition rate. In this experiment, gallery view is $55^\circ$, and probe view is $65^\circ$, $75^\circ$, $85^\circ$ respectively. It can be seen that our proposed method is a more effective strategy to improve the recognition performance for cross-view gait data.

5.7 Comparison with the state-of-the-art

For better illustration, we further compare the proposed method with some CNN-based state-of-the-art methods including LBNet [25], PoseGait [31], GaitGAN [29]. The experimental results are listed in Table 5.

| Methods   | Walking Scenes |
|-----------|----------------|
|           | NM  | BG  | CL  | Cross-view |
| LBNet     | 99.13 | 72.40 | 53.98 | 88.40 |
| PoseGait  | 96.62 | 44.50 | 35.95 | 66.54 |
| GaitGAN   | 98.75 | 72.73 | 41.50 | 62.90 |
| DLMNN     | 99.63 | 82.92 | 54.63 | 80.67 |

From the results we can find that the proposed method outperforms others in NM, BG, CL sets. It only second to LBNet on cross-view gait recognition. LBNet directly measure the similarity of any two GEIs. It seems particularly effective against large view change. In contrast, our method can work well in different scenarios.
This is because our method learns a feature metric subspace in which intra-variances is reduced effectively. The comparison results verify that the proposed method is more dependable.

### 5.8 Runtime Speed

System efficiency is an essential metric for many vision systems including gait recognition. We calculate the efficiency of five CNN-based methods for recognizing one sample on Intel i7-4720HQ CPU and Geforce GTX960M GPU. As shown in Table 6, GEINet [23], SiaNet [21] and ours are more efficient than other two. In PoseGait, most of computational cost is from 2D pose estimation and 3D transformation. LBNet, with the highest of computational costs, has to compute similarities of all pairs of probe and gallery using CNN, while GEINet, SiaNet and our method only carry out forward-network once.

| Methods     | PoseGait [18] | LBNet [25] | GEINet [21] | SiaNet [31] | DLMNN     |
|-------------|---------------|------------|-------------|-------------|-----------|
| Run time (s)| 0.307         | 1.896      | 0.035       | 0.041       | 0.0437    |

### 6 Conclusion and future work

In this paper, we propose a Deep Large Margin Nearest Neighbor (DLMNN) method to extract robust and discriminant features for gait recognition. After analyzing the related gait recognition techniques, we notice that the CNN-based methods make great strides in robust gait recognition. However, the existing CNN-based methods pay more attention to network architecture design rather than discriminant feature learning. Instead, the proposed DLMNN aims to pull the samples of the same person closer, meanwhile, to push the samples belonging to different subjects further from each other in the learned deep feature space. The feature space is learned in a triplet networks with a novel loss function which is named DLMNN loss. We discuss in detail the effect of the DLMNN loss in this work and demonstrate that it delivers smaller within-class scatter and larger between-class scatter, which is beneficial to discriminative feature learning. Comprehensive performance evaluations under various covariation conditions on two benchmark databases are provided. And experimental results demonstrated the outstanding performance of the proposed DLMNN method.

Future research will consider refining the feature learning. For instance, we may apply attention mechanism into the proposed DLMNN, by which we can select attention regions from GEIs and then learn metric subspace for each region. Furthermore, we will continue to seek better deep DML-based Loss function for the task of gait recognition.

**Acknowledgement:** This work is jointly supported by National Natural Science Foundation of China (61906163, 11871417) and Natural Science Foundation of the Jiangsu Higher Education Institutions of China (19KJB520018).

### References

[1] Ariyanto, G., Nixon, M.S.: Model-based 3d gait biometrics. In: International Joint Conference on Biometrics (2011)
[2] Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. fisherfaces: recognition using class specific linear projection. Publication 19(7), 711–720 (2002)
[3] Ben, X., Gong, C., Zhang, P., Jia, X., Wu, Q., Meng, W.: Coupled patch alignment for matching cross-view gaitis. IEEE Transactions on Image Processing PP(6), 1–1 (2019)
Deep Large Margin Nearest Neighbor for Gait Recognition

Bouchrika, I., Goffredo, M., Carter, J., Nixon, M.: On using gait in forensic biometrics. Journal of Forensic Sciences 56(4), 882–889 (2011)

Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N.: Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In: Computer Vision and Pattern Recognition, pp. 1335–1344 (2016)

Guan, Y., Li, C.T., Roli, F.: On reducing the effect of covariate factors in gait recognition: A classifier ensemble method. IEEE Transactions on Pattern Analysis and Machine Intelligence (2015)

Han, J., Bhanu, B.: Individual recognition using gait energy image. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(2), 316–322 (2005)

He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: 2017 IEEE International Conference on Computer Vision (ICCV) (2017)

Hu, J., Lu, J., Tan, Y.P.: Discriminative deep metric learning for face verification in the wild. In: Computer Vision and Pattern Recognition, pp. 1875–1882 (2014)

Huang, X., Boulgouris, N.V.: Gait recognition with shifted energy image and structural feature extraction. IEEE Trans Image Process 21(4), 2256–2268 (2012)

Iwama, H., Okumura, M., Makihara, Y., Yagi, Y.: The ou-isir gait database comprising the large population dataset and performance evaluation of gait recognition. IEEE Transactions on Information Forensics and Security 7(5), 1511–1521 (2012)

Khalid, B., Xiang, T., Gong, S.: Gait recognition using gait entropy image (2009)

Kshirsagar, V.P., Baviskar, M.R., Gaikwad, M.E.: Face recognition using eigenfaces. In: International Conference on Computer Research and Development, pp. 586–591 (2011)

Kusakunniran, W., Wu, Q., Li, H., Zhang, J.: Multiple views gait recognition using view transformation model based on optimized gait energy image. In: IEEE International Conference on Computer Vision Workshops, pp. 1058–1064 (2010)

Lam, T.H.W., Cheung, K.H., Liu, J.N.K.: Gait flow image: A silhouette-based gait representation for human identification. Pattern Recognition 44(4), 973–987 (2011)

LeCun, Y., Cortes, C., Burges, C.: The mnist database of handwritten digits (1998)

Lee, C.S., Elgammal, A.: Gait style and gait content: bilinear models for gait recognition using gait re-sampling. In: IEEE International Conference on Automatic Face and Gesture Recognition, pp. 147–152 (2004)

Liao, R., Yu, S., An, W., Huang, Y.: A model-based gait recognition method with body pose and human prior knowledge. Pattern Recognition 98, 107069 (2019)

Martínez-Díaz, Y., Méndez-Vázquez, H., Nicolás-Díaz, M., García, L.S.L., Gonzalez-Mendoza, M.: Shufflefacenet: A lightweight face architecture for efficient and highly-accurate face recognition. In: The IEEE International Conference on Computer Vision (ICCV) Workshops 2019 (2019)

Shiqi, Y., Haifeng, C., Qing, W., Linlin, S., Yongzhen, H.: Invariant feature extraction for gait recognition using only one uniform model. Neurocomputing 239(C), 81–93 (2017)

Shiraga, K., Makihara, Y., Muramatsu, D., Echigo, T., Yagi, Y.: Geinet: View-invariant gait recognition using a convolutional neural network. In: International Conference on Biometrics, pp. 1–8 (2016)

Wang, C., Zhang, J., Pu, J., Yuan, X., Wang, L.: Chrono-gait image: A novel temporal template for gait recognition. In: European Conference on Computer Vision (2010)

Wang, L., Ning, H., Tan, T., Hu, W.: Fusion of static and dynamic body biometrics for gait recognition. In: Proceedings Ninth IEEE International Conference on Computer Vision (2008)

Weinberger, K.Q.: Distance metric learning for large margin nearest neighbor classification. Jmlr 10 (2009)

Wu, Z., Huang, Y., Wang, L., Wang, X., Tan, T.: A comprehensive study on cross-view gait based human identification with deep cnns. IEEE Transactions on Pattern Analysis and Machine Intelligence 39(2), 209–226 (2016)

Xing, E.P., Ng, A.Y., Jordan, M.I., Russell, S.J.: Distance metric learning with application to clustering with side-information. In: International Conference on Neural Information Processing Systems (2002)

Xu, W., Luo, C., Ji, A., Zhu, C.: Robust gait recognition based on collaborative representation with external variant dictionary. In: Chinese Conference on Biometric Recognition, pp. 409–415 (2015)

Yam, C.Y., Nixon, M.S., Carter, J.N.: Automated person recognition by walking and running via model-based approaches. Pattern Recognition 37(5), 1057–1072 (2004)

Yu, S., Chen, H., Reyes, E.B.G., Poh, N.: Gaitgan: Invariant gait feature extraction using generative adversarial networks pp. 532–539 (2017)

Yu, S., Tan, D., Tan, T.: A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In: International Conference on Pattern Recognition, pp. 441–444 (2006)

Zhang, C., Liu, W., Ma, H., Fu, H.: Siamese neural network based gait recognition for human identification. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2832–2836 (2016)