GEINet: View-Invariant Gait Recognition Using a Convolutional Neural Network

Kohei Shiraga, Yasushi Makihara, Daigo Muramatsu
The Institute of Scientific and Industrial Research, Osaka University
8-1 Mihogaoka, Ibaraki, Osaka, Japan
{shiraga, makihara, muramatsu}@am.sanken.osaka-u.ac.jp

Tomio Echigo
Osaka Electro-Communication University
18-8 Hatsucho, Neyagawa, Osaka, Japan
echigo@osakac.ac.jp

Yasushi Yagi
Osaka University
1-1 Yamada-oka, Suita, Osaka, Japan
yagi@sanken.osaka-u.ac.jp

Abstract

This paper proposes a method of gait recognition using a convolutional neural network (CNN). Inspired by the great successes of CNNs in image recognition tasks, we feed in the most prevalent image-based gait representation, that is, the gait energy image (GEI), as an input to a CNN designed for gait recognition called GEINet. More specifically, GEINet is composed of two sequential triplets of convolution, pooling, and normalization layers, and two subsequent fully connected layers, which output a set of similarities to individual training subjects. We conducted experiments to demonstrate the effectiveness of the proposed method in terms of cross-view gait recognition in both cooperative and uncooperative settings using the OU-ISIR large population dataset. As a result, we confirmed that the proposed method significantly outperformed state-of-the-art approaches, in particular in verification scenarios.

1. Introduction

Gait has been regarded as a behavioral biometric within the biometrics research community since early-stage studies on gait recognition based on biological motion [16] and image processing [36], and a subsequent large body of recent studies on gait recognition during the last two decades [27]. Because gait biometrics are available even at a distance from a camera, as well as for an uncooperative subject, unlike other biometrics such as a fingerprint, finger and hand veins, iris, and face, the gait biometrics are expected to be applied to surveillance and criminal investigation using closed-circuit televisions (CCTVs) installed in public and private spaces. In fact, gait recognition has started to be used in practical cases in criminal investigation [3].

However, the availability of gait recognition for uncooperative subjects induces problematic covariates, including the view angle, walking speed, clothing, surface, carrying status, shoe, and time elapse. Therefore, for further progress, it is essential that gait recognition is more robust against these covariates.

Robust approaches to gait recognition are designed for either a specific covariate (e.g., view-invariant gait recognition) or a generic covariate. Additionally, such robust approaches are considered mainly at two stages: feature extraction and matching.

Regarding feature extraction, a silhouette-based representation is dominant within the gait recognition community [1, 5, 8, 21, 23, 43]. A typical example is gait energy image (GEI) [8], also known as an averaged silhouette [23], which is regarded as a mixture of static and dynamic features. Additionally, several approaches focus more on dynamic parts (e.g., gait entropy image (GEnI) [1] and masked GEI based on GEnI [2]) to mitigate the effect of appearance change caused by clothing and carrying status. These types of handcrafted gait features do not, however, guarantee optimality in terms of recognition.

Regarding matching, robust approaches against the covariates are categorized mainly into two families: generative and discriminative. Given a matching pair of gait features under different conditions (e.g., different views), generative approaches generate the gait features under the same condition for better matching [7, 17, 20, 28]. Generative approaches do not, however, guarantee optimality in terms of recognition because they essentially optimize the accuracy of the generated gait features and not the discrimination capability itself.

978-1-5090-1869-7/16/$31.00 ©2016 IEEE

Discriminative approaches aim at optimizing the discrimination capability. This family usually unfolds an image-based gait feature into a feature vector, where each dimension corresponds to each pixel, and then applies machine learning techniques, such as linear discriminant analysis (LDA) [2, 8], primal rank support vector machine (SVM) [30], and multi-view discriminant analysis (MvDA) [29]. Unfolded feature vector-based discriminative approaches, however, easily result in overtraining because each pixel is represented by each independent axis in the feature space. Hence, spatial proximity in the image structure is never considered.

From the broader viewpoint of relevant research fields, deep learning [9] has achieved great successes in many areas. An advantage of deep learning is that it simultaneously executes feature extraction and recognition within a unified framework using a large amount of training samples. In particular, a convolutional neural network (CNN), which considers spatial proximity using a convolution operation, significantly improves the accuracy of image recognition as demonstrated through a series of ImageNet Large Scale Visual Recognition Challenges [12]. Moreover, the effectiveness of a CNN has been demonstrated in research fields that are more relevant with gait recognition, such as action recognition [14], video classification [18], and face recognition [42].

There are, however, few gait recognition studies that use a deep learning framework because deep learning requires a large number of training samples and it is difficult to collect a large number of training gait samples. To the best of our knowledge, deep learning-based gait recognition has been performed only by Hossain and Chetty [11] and Wu et al. [44]. While Hossain and Chetty reported that the their proposed deep learning feature outperformed the benchmark, there are still some concerns: (1) they did not employ a CNN that is suitable for image recognition tasks because of the consideration of spatial proximity, but a restricted Boltzmann machine (RBM) that is available for general data but not necessarily suitable for image data, and (2) they used only 10 subjects for training, which is completely insufficient for a deep learning framework. On the other hand, Wu et al. employed CNN for gait recognition, and reported better recognition accuracy than those of benchmarks. One concern of their work is that they represented gait by randomly selected silhouette image set, and used it as an input to the CNN. This representation is not appropriate for gait recognition, because useful dynamic information for gait feature cannot be considered in the representation.

We therefore propose a method of gait recognition using a CNN with appropriate gait representation and also demonstrate its effectiveness in cross-view gait recognition tasks using the largest publicly available gait database. The contributions of this paper are twofold.

1. GEINet: CNN structure for robust gait recognition

While Hossain and Chetty [11] employed an RBM, we employ a CNN, which is more suitable for image classification tasks. More specifically, we design GEINet composed of two sequential triplets of convolution, pooling, and normalization layers, and two subsequent fully connected layers, which outputs a set of similarities to individual training subjects given a GEI as an input.

2. Application to cross-view gait recognition with the largest gait database

We use the OU-ISIR large population dataset [13], which is the largest publicly available gait database and is composed of more than 1000 subjects, to appropriately train our GEINet as well as to evaluate cross-view gait recognition performance in a statistically reliable way. Note that the proposed GEINet itself is potentially applicable not only to view variations but also to general covariate conditions, such as clothing and carrying conditions. Additionally, because we feed training samples with view variations into the same GEINet, the covariate conditions for the matching pair do not need to be provided before matching and are therefore easily applied not only in a cooperative setting where a gallery is enrolled under the same condition but also an uncooperative setting where a gallery is enrolled under different conditions, whereas several discriminative approaches, such as MvDA [29], require covariate conditions. We demonstrate through experiments that the proposed method significantly outperforms state-of-the-art approaches, particularly in verification scenarios.

2. Related work

Gait features

The most prevalent gait feature is GEI [8], which is obtained by simply aggregating the silhouette sequence over one gait period. While a GEI represents a mixture of static and dynamic parts, some approaches focus more on dynamic parts to mitigate clothing and carrying status variation because such variations mainly affect the static parts.

A typical approach is a contour-based representation. For example, Mowbray and Nixon [31] represented a temporal sequence of a close-curve contour using a Fourier descriptor, whereas Wang et al. [43] proposed a chrono-gait image (CGI), which aggregates a binary image sequence of contours over a one-fourth gait period with color encoding based on phase information. Moreover, Lam et al. [21] aggregated a binary image sequence of moving contours over one gait period as a gait flow image (GFI).

Additionally, other approaches attempt to extract dynamic parts from a GEI. For example, Bashir et al. [1] extracted GEnI from a GEI, which highlights dynamic parts based on information entropy theory (i.e., high and low intensities are returned for dynamic and static parts, regardless of the foreground and background). However, a masked

GEI [2] is constructed by first making a masking image using thresholding GEnI to preserve only the dynamic parts and masking the static part of the GEI.

As discussed in section 1, these types of handcrafted gait features do not, however, guarantee optimality from the viewpoint of the entire recognition process.

Generative approaches to robust gait recognition

Generative approaches are designed for a specific covariate, in particular, for view angle and walking speed. Kale et al. [17] assumed that a silhouette is represented by an approximation of billboard of a sagittal plane and projected it into a common canonical view (i.e., side view) for matching, whereas Goffredo et al. [7] used a head top and foot bottom trajectory projected on the piecewise linear sagittal plane using self-calibration. While the aforementioned methods are built on the geometric property, there are several machine learning-based approaches without the geometric property. Such a typical approach is the view transformation model (VTM) based on matrix factorization using singular value decomposition [28]. Several variants of the VTM have been proposed, including a VTM using support vector regression for motion correlated parts [20] and a VTM for an arbitrary view using 3D training gait models [34].

As discussed in section 1, generative approaches do not, however, guarantee optimality in terms of recognition because they aim at synthesizing the feature.

Discriminative approaches to robust gait recognition

Discriminative approaches are usually designed for general covariates. A simple yet effective strategy is to apply LDA, which follows PCA if necessary for dimension reduction to avoid the curse of dimension, to extract gait features. In fact, this strategy is adopted in many gait features, such as GEI [8], key-pose silhouettes using gait dynamics normalization [24], CGI [43], masked GEI [2], and GFI [21]. Moreover, variants of LDA are employed for cross-view gait recognition, such as uncorrelated discriminant simplex analysis [25] and MvDA [29]. In addition to the family of discriminant analyses, metric learning approaches have been proposed. Martin-Felez and Xiang [30] introduced the primal rank support vector machine (SVM) for robust gait recognition, which is available regardless of the covariate types. As discussed in section 1, the aforementioned methods do not consider the image structure. Hence, they suffer from overtraining.

Gait databases

A key to the success of deep learning-based robust gait recognition partly relies on a sufficient number of training samples. Hence, large-scale gait databases are essential. For this purpose, the gait database should contain a large number of subjects as well as a variety of covariate conditions. Some publicly available gait databases, such as the USF gait database [38], SOTON gait database [40], CASIA gait database [46], OU-ISIR Treadmill dataset [26], and TUM GAID [10], contain several covariate conditions, such as view angle, clothing, and carrying conditions. They do not, however, contain a sufficient number of subjects (i.e., they contain at most 100 orders) for deep learning. The only exception is the OU-ISIR large population dataset [13], which contains over 4000 subjects and has view variations from an oblique view to a side view. We therefore use the OU-ISIR large population dataset to train our proposed GEINet as well as for performance evaluation.

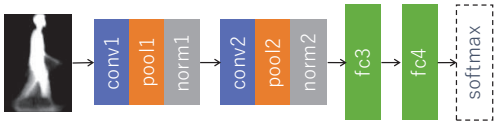

Figure 1: The structure of GEINet.

Table 1: Layer configurations for GEINet. Act. denotes the activation function.

Layer	#Kernels	Size/stride	Act.	Pooling
conv1	18	7×7×1/1	ReLU	
pool1		2×2/2		Max pooling
conv2	45	5×5×18/1	ReLU	
pool2		3×3/2		Max pooling

3. Proposed method

3.1. Input data

An appearance-based approach to gait recognition usually follows three steps: silhouette extraction, feature extraction, and matching. The motivation of silhouette extraction for gait recognition is to avoid being affected by clothes' colors and textures, unlike the person re-identification task. We therefore consider using silhouette-based representation as input data for our CNN structure. Moreover, because it is difficult to directly handle the temporal aspect of a gait silhouette sequence (e.g., intra-subject and/or inter-subject differences of gait stances (phases) at the first frame and also during the gait period), we then adopt a GEI, as input data for our CNN structure. Additionally, because the CNN structure usually requires fixed-size input data, we set the GEI size as 88 × 128 pixels.

3.2. Network structure

Because our CNN is built on GEI, we call it GEINet in this paper. GEINet is structured as an eight-layered CNN network, where the leading six layers are two sequential triplets of convolution, pooling, and normalization layers as shown in Fig. 1. The configurations for each convolution and pooling layer are provided in Table 1. Local response normalization (LRN) [19] is employed for normalization layers norm1 and norm2. Following layer fc3 is a fully connected layer that has 1024 units with dropout [41], where the ReLU [35] activation function is used. Another fully

connected layer, fc4, has N units, where N corresponds to the number of training subject IDs. On learning phase, a set of similarity to individual training subjects are calculated using the softmax function. More specifically, the i-th unit of the last layer ideally returns 1 for the i-th subject's GEI, otherwise, it returns 0.

The designed GEINet has a similar structure, to some extent, as one of the most successful network structures, that is, AlexNet for image classification [19], which has leading convolution layers that optionally follow pooling and normalization layers, and several fully connected layers with a final softmax. However, GEINet is not as deep compared with recent deep neural networks such as AlexNet. This may be partly caused by the fact that AlexNet focuses on the image classification task of fairly different objects (e.g., container ship, cherry, and leopard) under large spatial displacement, whereas GEINet focuses on subtle inter-subject differences within the same action class, that is, gait, using spatially well-aligned silhouettes (i.e., scale normalization and region center alignment).

In fact, DeepFace [42], which is a seminal work using deep learning in the biometric research field, employs a relatively shallow network structure, that is, an eight-layer network composed of single convolution-pooling-convolution filtering (three layers), followed by three locally connected layers and two fully connected layers. Moreover, through preliminary experiments, we have confirmed that additional convolutional layers following two sequential triplets of convolution, pooling, and normalization layers do not improve gait recognition accuracy. As a result, we conclude that the designed GEINet is sufficient for gait recognition, that is, person classification from the same gait class under a spatially well-aligned condition.

3.3. Learning

Because GEINet contains a set of weighting parameters $\boldsymbol{w}$, we need to optimize them from training samples.

Let $\{I_1, \ldots, I_M\}$ be a set of training GEIs and $\{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_M\}$ a set of corresponding indicator vectors of training labels (i.e., subject IDs), where M is the number of training samples. Note that the indicator vector for the m-th training sample is defined as $\boldsymbol{d}_m = [\delta_{y_m,1}, \ldots, \delta_{y_m,N}]^T$, where y_m is the subject ID for the m-th training sample and δ is Kronecker ' s delta.

We then define an N-dimensional vector at layer fc4 as a mapping function from the input GEI I, with the set of weighting parameters $\boldsymbol{w}$ as $\boldsymbol{v}_{fc4}(I; \boldsymbol{w})$. We subsequently compute a softmax of the n-th unit at layer fc4 as the final output as

$$v'(I; \boldsymbol{w})_n = \frac{\exp(\boldsymbol{v}_{fc4}(I; \boldsymbol{w})_n)}{\sum_{j=1}^{N} \exp(\boldsymbol{v}_{fc4}(I; \boldsymbol{w})_j)}, \quad (1)$$

where the subscript n denotes the n-th unit in the N-dimensional vector. Learning is then performed by minimizing the following cross-entropy loss $L(\boldsymbol{w})$ as

$$L(\boldsymbol{w}) = -\sum_{m=1}^{M}\sum_{n=1}^{N} d_{mn} \log v'(I_m; \boldsymbol{w})_n. \quad (2)$$

Finally, the set of weighting parameters $\boldsymbol{w}$ is updated using the stochastic gradient descent (SGD) algorithm [4].

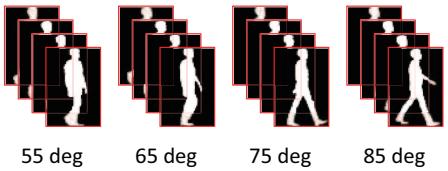

Figure 2: Examples of gait image sequences with four observation views in the OU-ISIR dataset

3.4. Recognition

Given a matching pair of a probe GEI I_p and gallery GEI I_g, we compute a dissimilarity score between them through a trained GEINet. More specifically, we adopt a set of outputs at the fully connected layer fc4 as a feature vector, that is, $\boldsymbol{v}_{fc4}(I; \boldsymbol{w}) \in \mathbb{R}^N$, which is regarded as a set of similarity scores for each training subject. We then simply compute the L2 distance between a probe GEI I_p and gallery GEI I_g as

$$dist(I_p, I_g; \boldsymbol{w}) = \|\boldsymbol{v}_{fc4}(I_p; \boldsymbol{w}) - \boldsymbol{v}_{fc4}(I_g; \boldsymbol{w})\|_2, \quad (3)$$

where $\|\cdot\|_2$ means L2 norm.

4. Experiments

4.1. Overview

We evaluated the effectiveness of the proposed method during a cross-view gait recognition task. More specifically, we considered two settings: recognition with a cooperative gallery subject (i.e., *cooperative setting*) and with an uncooperative gallery subject (i.e., *uncooperative setting*). Views of the gallery gait image sequences were the same among enrolled subjects in the cooperative setting, whereas they may have differed for each subject in the uncooperative setting.

4.2. Setup

We used a subset of the OU-ISIR large population dataset [13] (Publicly available at `http://www.am.sanken.osaka-u.ac.jp/BiometricDB/GaitLP.html`) both for training and testing. The subset was composed of two gait image sequences from 1912 subjects. Four silhouette image subsequences of one gait cycle with different observation views were extracted from each gait image sequence. Figure 2 shows examples of the silhouette images of a subject with four observation views: 55, 65, 75, and 85 deg.

Moreover, we divided the 1912 subjects into two disjoint groups of the same size, that is, 956 training and

956 testing subjects in order to meet protocols of benchmarks [29, 32, 33] described later. We used the gait image sequences from the 956 training subjects for learning the GEINet, which means the number of units on the fully connected layer fc4, that is, the dimensionality N of the gait feature for recognition, equaled 956. Additionally, as stated previously, each subject had eight GEIs: four view-angle variations and two types of sequences (gallery and probe). Hence, a total of 7648 training pairs of GEIs and its corresponding indicator vector $\boldsymbol{d}$ of a training subject ID were provided to train GEINet.

Regarding the hyper-parameter setting, we empirically selected the size of mini-batches for the SGD as 239. Moreover, we set the initial learning rate to 0.02, which was decreased by multiplying it by $\gamma = \sqrt[4]{0.01}$ at every 10,000th iteration. The total number of iterations was 50,000, which was approximately equivalent to 1500 epochs. GEINet was trained and tested using Caffe [15] on a NVIDIA GeForce GTX TITAN X.

4.3. Evaluation criteria

In each setting, we evaluated the recognition accuracy for two tasks: verification and identification. For the verification task, we calculated false acceptance rates (FARs) and false rejection rates (FRRs), and plotted receiver operating characteristics (ROC) curves. We then calculated equal error rates (EERs) as a criterion of the verification capability. Moreover, we plotted cumulative match characteristic (CMC) curves and calculated the rank-1 identification rates as a criterion of the identification capability.

4.4. Results for a cooperative setting

We evaluated the recognition accuracy of the proposed method using two protocols, which were the same as those adopted in previous studies [29, 32, 33]. We compared the accuracy of the proposed method with both those of generative approaches under Protocol 1 and discriminative approaches under Protocol 2.

Protocol 1: Comparison with generative approaches

Protocol 1 was considered in works by Muramatsu et al. [32, 33]. In this protocol, the frequency-domain feature (FDF) [28] is extracted from a gait image sequence as a gait feature and all pairs of cross-view matching are evaluated. For the evaluation, two cross-validations were employed: the subjects in the subset were randomly divided into two disjoint groups of the same size and gait image sequences associated with each group were used for training and testing alternately. To reduce any effect of random grouping, the two cross-validations were repeated five times using different subject groupings. We used the same subject lists as those in works by Muramatsu et al. [32, 33] (The lists are provided by Muramatsu et al. [32, 33] at `http://www.am.sanken.osaka-u.ac.jp/BiometricDB/dataset/GaitLP/Benchmarks.html`). Evaluation using this protocol enabled us to compare the recognition accuracy of the proposed method with that of generative approaches, such as VTM with transformation consistency measures [32] (refer *TCM+*) and quality-dependent VTM [33] (refer *wQVTM*). We also provided a result for a variant of GEINet, that is, GEINet with FDF (refer *w/ FDF*), where the input data type was FDF and weighting parameter training was also performed

Table 2: Comparison of EERs [%] with generative approaches in a cooperative setting. The best and the second best results are indicated by a bold and italic bold font, respectively, which also applies to the tables that follow.

Gallery view	Method	Probe view 55	65	75	85
55	GEINet	(1.3)	**1.4**	**1.7**	**2.5**
	w/ FDF	(1.9)	***2.0***	***2.3***	***2.9***
	TCM+		3.2	4.0	5.7
	wQVTM		3.6	4.8	6.5
65	GEINet	**1.2**	(1.0)	**1.3**	**1.6**
	w/ FDF	***1.7***	(1.4)	***1.7***	***2.2***
	TCM+	3.0		3.4	4.2
	wQVTM	3.5		3.4	5.1
75	GEINet	**1.5**	**1.2**	(1.2)	**1.4**
	w/ FDF	***2.0***	***1.5***	(1.6)	***1.7***
	TCM+	4.0	3.4		3.8
	wQVTM	4.7	3.7		3.8
85	GEINet	**2.4**	**1.6**	**1.2**	(1.1)
	w/ FDF	***2.5***	***1.9***	***1.6***	(1.4)
	TCM+	5.5	4.4	3.7	
	wQVTM	6.5	4.9	3.7	

Table 3: Comparison of rank-1 identification rates [%] with generative approaches in a cooperative setting.

Gallery view	Method	Probe view 55	65	75	85
55	GEINet	(94.7)	**93.2**	**89.1**	***79.9***
	w/ FDF	(92.7)	***91.4***	***87.2***	**80.0**
	TCM+		79.9	70.8	54.5
	wQVTM		78.3	64.0	48.6
65	GEINet	**93.7**	(95.1)	**93.8**	**90.6**
	w/ FDF	***92.3***	(93.9)	***92.2***	***88.6***
	TCM+	81.7		79.5	70.2
	wQVTM	81.5		79.2	67.5
75	GEINet	**90.1**	**94.1**	(95.2)	**93.8**
	w/ FDF	***88.8***	***92.6***	(93.4)	***91.9***
	TCM+	71.9	80.0		79.0
	wQVTM	70.2	80.0		78.2
85	GEINet	**81.4**	**91.2**	**94.6**	(94.7)
	w/ FDF	***80.9***	***88.4***	***92.2***	(93.2)
	TCM+	53.7	73.0	79.4	
	wQVTM	51.1	68.5	79.0	

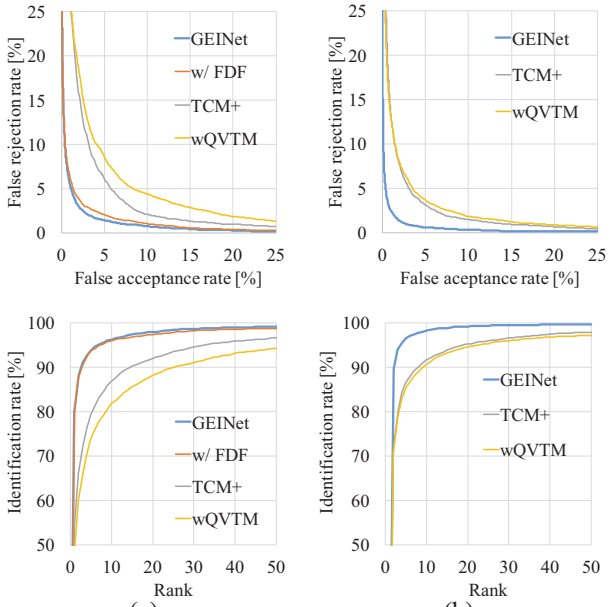

Figure 3: ROC curves (top) and CMC curves (bottom). (a) Comparison with generative approaches under 55-deg gallery vs. 85-deg probe. (b) Comparison in an uncooperative setting.

based on FDFs, but the network structure remained fixed to the original GEINet trained by GEIs except for the number of channels of input data and the kernel size of conv1 layer. Note that this is reference data to demonstrate that the difference in input data type between the proposed method (GEI) and the benchmarks [32, 33] (FDF) does not affect the accuracy and we do not intend to compare GEINet to GEINet with FDF because the structure of GEINet is not optimized for FDF.

Figures 3(a) show the ROC and CMC curves for the matching of the 55-deg gallery vs. 85-deg probe. We also summarize the EERs and rank-1 identification rates of all the cross-view pairs in Tables 2 and 3.

As a result, we can see that the proposed method significantly outperformed the benchmarks (e.g., the EERs of the proposed method are lower than the halves of the EERs of the benchmarks for all the view combinations). Additionally, it is noticeable that this accuracy improvement was not due to an input data type difference because even GEINet with FDF significantly outperformed the benchmarks, which had the same gait feature for the input data as the benchmarks.

Protocol 2: Comparison with discriminative approaches

We used the same protocol as the paper by Mansur et al. [29] (we call it Protocol 2). The subjects in the subset were randomly divided into two disjoint groups of the same size for training and testing, respectively (Mansur et al. [29] provided the subject lists for training and testing on their website: `http://www.am.sanken.osaka-u.ac.jp/~mansur/files/list_train_test.txt`). We evaluated the proposed method under cross-view matching between the 85-deg gallery and each 55-deg, 65-deg, and 75-deg probe, and compared it with benchmarks including discriminative approaches: LDA [37], DATER [45], MvDA [29], GMLDA [39], and CCA [22].

Table 4: Comparison of EERs [%] and rank-1 identification rates [%] in a cooperative setting in Protocol 2 for each probe view. Note that significant figures for the benchmarks are limited because we used values from the original paper by Mansur et al. [29]. Methods marked with (*) require view information for a matching pair.

	EER			Rank-1		
Method	55	65	75	55	65	75
GEINet	**2.7**	**1.8**	**1.0**	***80.4***	***91.5***	94.8
LDA	8	***5***	***4***	56	91	***96***
DATER	30	22	16	10	29	65
GMLDA(*)	12	9	5	68	82	95
MvDA(*)	***7***	***5***	***4***	**88**	**96**	**97**
CCA(*)	21	13	8	52	81	92

We summarize the EERs and rank-1 identification rates in Tables 4. For rank-1 identification rates, MvDA yielded the best accuracy for all the settings, although MvDA cannot be applied when the view information for a matching pair is not provided in advance, as well as the generative approaches such TCM+ [32] and wQVTM [33]. The proposed method yielded the best accuracy among the approaches that do not require view information, except for LDA, which is slightly better for quite close-view settings, that is, 85-deg gallery vs. 75-deg probe.

Moreover, the proposed method significantly outperformed the benchmarks with respect to EERs. More specifically, the EERs of the proposed method were lower than the halves of those of the benchmarks for all the view combinations.

The inconsistency between verification and identification accuracies is often the case with soft biometrics, such as gait, because the identification performances are determined using probe-dependent rank statistics, whereas the verification performances are determined using aggregated score distributions, as discussed in the paper by DeCann and Ross [6].

4.5. Results for an uncooperative setting

We employed the same subject lists for cross-validation as Protocol 1 to evaluate the uncooperative setting. Specifically, we first drew a gallery view of each subject randomly from {55, 65, 75, and 85 deg}. We then drew a probe view of a test subject randomly from the other three views (e.g., if a 55-deg gallery view was drawn, the probe view was randomly drawn from 65, 75, or 85 deg. For each cross-

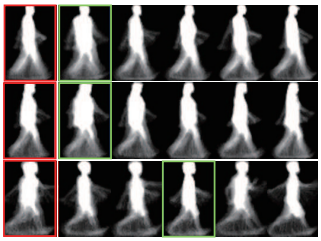

Figure 4: Probe features (left) with the top five gallery features (right). The gallery feature with a green bounding box indicates a true-match subject.

validation, we employed the view selection independently. For comparison with TCM+ and wQVTM, we used distance matrices provided by Muramatsu et al. [32, 33].

As a qualitative evaluation, we show probe features with the top five gallery features in Fig. 4. We can observe that the proposed method can correctly select the gait feature of the true-match subject with a different view as the first rank in the first and second rows, and the third rank in the third row, even though the views of these gait features are different from those of the probe.

As a quantitative evaluation, we show ROC and CMC curves in the uncooperative setting in Fig. 3(b). Additionally, EERs of the proposed method, TCM+, and wQVTM are 1.6%, 4.0%, and 4.3%, respectively, and the corresponding rank-1 identification rates are 89.7%, 71.5%, and 70.6%, respectively. As a result, the proposed method also outperforms the benchmarks in an uncooperative setting.

5. Conclusion

This paper described a method of gait recognition using a CNN. More specifically, we designed GEINet composed of two sequential triplets of convolution, pooling, and normalization layers, and two subsequent fully connected layers, which outputs a set of similarities to individual training subjects given a GEI as an input. As a result of experiments for cross-view gait recognition in both cooperative and uncooperative settings using the OU-ISIR large population dataset, we confirmed that the proposed method significantly outperformed the state-of-the-art approaches, that is, it yielded approximately two times better accuracy in verification scenarios.

A path for future research is to validate the proposed method against a wider view variation in conjunction with the construction of a new database because the view variation in the OU-ISIR large population dataset is relatively limited.

Acknowledgment

This work was partly supported by JSPS Grants-in-Aid for Scientific Research (A) 15H01693, and the JST CREST "Behavior Understanding based on Intention-Gait Model" project.

References

[1] K. Bashir, T. Xiang, and S. Gong. Gait recognition using gait entropy image. In *Proc. of the 3rd Int. Conf. on Imaging for Crime Detection and Prevention*, pages 1–6, Dec. 2009.

[2] K. Bashir, T. Xiang, and S. Gong. Gait recognition without subject cooperation. *Pattern Recognition Letters*, 31(13):2052–2060, Oct. 2010.

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

[4] O. Bousquet and L. Bottou. The tradeoffs of large scale learning. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 161–168. Curran Associates, Inc., 2008.

[5] B. Decann and A. Ross. Gait curves for human recognition, backpack detection, and silhouette correction in a nighttime environment. In *Proc. of the SPIE, Biometric Technology for Human Identification VII*, volume 7667, pages 76670Q–76670Q–13, 2010.

[6] B. DeCann and A. Ross. Relating roc and cmc curves via the biometric menagerie. In *Proc. of the 6th IEEE Int. Conf. on Biometrics: Theory, Applications and Systems*, pages 1–8, Sept 2013.

[7] M. Goffredo, I. Bouchrika, J. Carter, and M. Nixon. Self-calibrating view-invariant gait biometrics. *IEEE Trans, on Systems, Man, and Cybernetics, Part B: Cybernetics*, 40(4):997–1008, 2010.

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

[9] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural Comput.*, 18(7):1527–1554, July 2006.

[10] M. Hofmann, J. Geiger, S. Bachmann, B. Schuller, and G. Rigoll. The tum gait from audio, image and depth (gaid) database: Multimodal recognition of subjects and traits. *J. Vis. Comun. Image Represent.*, 25(1):195–206, Jan. 2014.

[11] E. Hossain and G. Chetty. Multimodal feature learning for gait biometric based human identity recognition. In *Proc. of the 20th Int. Conf. on Neural Information Processing*, pages 721–728, 2013.

[12] Ilsvrc, imagenet large scale visual recognition challenge. `http://www.image-net.org/challenges/LSVRC/`.

[13] H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi. The OU-ISIR gait database comprising the large population dataset and performance evaluation of gait recognition. *IEEE Trans. on Information Forensics and Security*, 7(5):1511–1521, Oct. 2012.

[14] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(1):221–231, Jan 2013.

[15] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proc. of the 22Nd ACM Int. Conf, on Multimedia*,, pages 675–678. ACM, 2014.

[16] G. Johansson. Visual motion perception. *Scientific American*, 232:75–88, Jun. 1976.

[17] A. Kale, A. Roy-Chowdhury, and R. Chellappa. Towards a view invariant gait recognition algorithm. In *Proc. of IEEE Conf. on Advanced Video and Signal Based Surveillance*, pages 143–150, 2003.

[18] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1725–1732, June 2014.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

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

[21] T. H. W. Lam, K. H. Cheung, and J. N. K. Liu. Gait flow image: A silhouette-based gait representation for human identification. *Pattern Recognition*, 44:973–987, April 2011.

[22] N. Liu, J. Lu, and Y.-P. Tan. Joint subspace learning for view-invariant gait recognition. *IEEE Signal Processing Letters*, 18(7):431–434, 2011.

[23] Z. Liu and S. Sarkar. Simplest representation yet for gait recognition: Averaged silhouette. In *Proc. of the 17th Int. Conf. on Pattern Recognition*, volume 1, pages 211–214, Aug. 2004.

[24] Z. Liu and S. Sarkar. Improved gait recognition by gait dynamics normalization. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(6):863–876, 2006.

[25] J. Lu and Y.-P. Tan. Uncorrelated discriminant simplex analysis for view-invariant gait signal computing. *Pattern Recognition Letters*, 31(5):382–393, 2010.

[26] Y. Makihara, H. Mannami, A. Tsuji, M. Hossain, K. Sugiura, A. Mori, and Y. Yagi. The OU-ISIR gait database comprising the treadmill dataset. *IPSJ Trans. on Computer Vision and Applications*, 4:53–62, Apr. 2012.

[27] Y. Makihara, D. S. Matovski, M. S. Nixon, J. N. Carter, and Y. Yagi. *Gait Recognition: Databases, Representations, and Applications*, pages 1–15. John Wiley & Sons, Inc., 1999.

[28] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi. Gait recognition using a view transformation model in the frequency domain. In *Proc. of the 9th European Conf. on Computer Vision*, pages 151–163, May 2006.

[29] A. Mansur, Y. Makihara, D. Muramatsu, and Y. Yagi. Cross-view gait recognition using view-dependent discriminative analysis. In *IEEE Int. Joint Conf. on Biometrics*, pages 1–8, Sept 2014.

[30] R. Martin-Felez and T. Xiang. Uncooperative gait recognition by learning to rank. *Pattern Recognition*, 47(12):3793 – 3806, 2014.

[31] S. Mowbray and M. Nixon. Automatic gait recognition via fourier descriptors of deformable objects. In *Proc. of the 1st IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, pages 566–573, 2003.

[32] D. Muramatsu, Y. Makihara, and Y. Yagi. Cross-view gait recognition by fusion of multiple transformation consistency measures. *IET Biometrics*, 4:62–73(11), June 2015.

[33] D. Muramatsu, Y. Makihara, and Y. Yagi. View transformation model incorporating quality measures for cross-view gait recognition. *IEEE Trans. on Cybernetics*, 2016 (in press).

[34] D. Muramatsu, A. Shiraishi, Y. Makihara, M. Uddin, and Y. Yagi. Gait-based person recognition using arbitrary view transformation model. *IEEE Trans. on Image Processing*, 24(1):140–154, Jan 2015.

[35] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proc. of the 27th International Conference on Machine Learning*, pages 807–814. Omnipress, 2010.

[36] S. Niyogi and E. Adelson. Analyzing and recognizing walking figures in xyt. In *Proc. of the 7th IEEE Conf. on Computer Vision and Pattern Recognition*, pages 469–474, 1994.

[37] N. Otsu. Optimal linear and nonlinear solutions for least-square discriminant feature extraction. In *Proc. of the 6th Int. Conf. on Pattern Recognition*, pages 557–560, 1982.

[38] S. Sarkar, J. Phillips, Z. Liu, I. Vega, P. G. ther, and K. Bowyer. The humanid gait challenge problem: Data sets, performance, and analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(2):162–177, 2005.

[39] A. Sharma, A. Kumar, H. D. III, and D. W. Jacobs. Generalized multiview analysis: A discriminative latent space. In *Proc of the 25th IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2160–2167, 2012.

[40] J. Shutler, M. Grant, M. Nixon, and J. Carter. On a large sequence-based human gait database. In *Proc. of the 4th Int. Conf. on Recent Advances in Soft Computing*, pages 66–71, Nottingham, UK, Dec. 2002.

[41] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[42] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1701–1708, June 2014.

[43] C. Wang, J. Zhang, L. Wang, J. Pu, and X. Yuan. Human identification using temporal information preserving gait template. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(11):2164 –2176, nov. 2012.

[44] Z. Wu, Y. Huang, and L. Wang. Learning representative deep features for image set analysis. *IEEE Trans. on Multimedia*, 17(11):1960–1968, Nov 2015.

[45] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, and H.-J. Zhang. Discriminant analysis with tensor representation. In *Proc. of the IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, pages 526–532, Jun. 2005.

[46] S. Yu, D. Tan, and T. Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *Proc. of the 18th Int. Conf. on Pattern Recognition*, volume 4, pages 441–444, Aug. 2006.