On Input/Output Architectures for Convolutional Neural Network-Based Cross-View Gait Recognition

Noriko Takemura, Yasushi Makihara, Daigo Muramatsu, *Member, IEEE*, Tomio Echigo, *Member, IEEE*, and Yasushi Yagi, *Member, IEEE*

***Abstract*— In this paper, we discuss input/output architectures for convolutional neural network (CNN)-based cross-view gait recognition. For this purpose, we consider two aspects: verification versus identification and the tradeoff between spatial displacements caused by subject difference and view difference. More specifically, we use the Siamese network with a pair of inputs and contrastive loss for verification and a triplet network with a triplet of inputs and triplet ranking loss for identification. The aforementioned CNN architectures are insensitive to spatial displacement, because the difference between a matching pair is calculated at the last layer after passing through the convolution and max pooling layers; hence, they are expected to work relatively well under large view differences. By contrast, because it is better to use the spatial displacement to its best advantage because of the subject difference under small view differences, we also use CNN architectures where the difference between a matching pair is calculated at the input level to make them more sensitive to spatial displacement. We conducted experiments for cross-view gait recognition and confirmed that the proposed architectures outperformed the state-of-the-art benchmarks in accordance with their suitable situations of verification/identification tasks and view differences.**

***Index Terms*— Convolutional neural network, cross-view, gait recognition.**

I. Introduction

BIOMETRIC-based person authentication methods have been extensively studied for various applications, such as access control, surveillance, and forensics. As biometric traits, the face, voice, fingerprints, hand veins, iris, handwriting, and gait are available for such applications. Of these, the gait is one of the most practical traits for video-based surveillance and forensics because it can be obtained from closed-circuit television footage captured at a distance, as well as from an uncooperative subject. In fact, gait recognition has been already used in practical cases in criminal investigations [1]–[3].

However, gait recognition is susceptible to intra-subject variations, such as view, clothing, walking speed, shoes, and carrying status; hence, generative or discriminative approaches to robust gait recognition against these variations have been intensively studied in recent years. Generative approaches usually transform gait features from different conditions into those under the same condition for better matching, for example, in case of cross-view gait recognition, the transformation into a common canonical view [4], [5] or the transformation from a gallery (enrollment) condition into a probe (query) condition [6], [7]. Generative approaches are, however, not necessarily optimal from a classification viewpoint because they essentially aim not at classification but the synthesis of gait features among different conditions. Discriminative approaches, by contrast, aim at directly optimizing the discrimination capability under various conditions by learning discriminant subspaces [8]–[11] or metrics [12]. Furthermore, taking into account the great success of deep learning frameworks in many research areas of computer vision, pattern recognition, and biometrics, discriminative approaches to gait recognition also gear to a convolutional neural network (CNN) frameworks [13]–[17] and have achieved state-of-the-art performance in cross-view gait recognition.

An advantage of the CNN-based approach is that we can flexibly design network architectures for better performance by changing inputs, outputs, and loss functions, for example, a single gait feature as an input, subject labels as outputs, and cross-entropy loss in [15]; and a pair of gait features for matching as an input, genuine or imposter probabilities as outputs, and a contractive loss in [14], also known as the Siamese network [18]. The network architectures, in particular the input, output, and loss functions, should be carefully designed by considering the properties of recognition tasks, that is, verification (one-to-one matching) or identification (one-to-many matching) because the preferable properties for verification and identification are substantially different, as discussed in [19].

Manuscript received January 31, 2017; revised July 30, 2017; accepted September 21, 2017. Date of publication October 9, 2017; date of current version September 4, 2019. This work was supported in part by JSPS Grants-in-Aid for Scientific Research (A) under Grant JP15H01693 and in part by the JST CREST Behavior Understanding Based on Intention-Gait Model Project. This paper was recommended by Associate Editor F. Porikli. *(Corresponding author: Noriko Takemura.)*

N. Takemura, Y. Makihara, D. Muramatsu, and Y. Yagi are with the Mitsubishi Electric Collaborative Research Division for Wide-Area Security Technology, Institute of the Scientific and Industrial Research, Osaka University, Osaka 567-0047, Japan (e-mail: takemura@am.sanken.osaka-u.ac.jp; makihara@am.sanken.osaka-u.ac.jp; muramatsu@am.sanken.osaka-u.ac.jp; yagi@am.sanken.osaka-u.ac.jp).

T. Echigo is with the Department of Engineering Informatics, Osaka Electro-Communication University, Osaka 572-8530, Japan (e-mail: echigo@osakac.ac.jp).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2017.2760835

Another advantage of the CNN-based approach is the robustness (insensitiveness) to spatial displacement of the recognition targets. Specifically, for the cross-view gait recognition case, the convolution and max pooling layers mitigate the spatial displacement of body parts by view differences, that is, intra-subject variations by view differences, to some extent. However, this means that the meaningful spatial displacement derived from inter-subject variations may be excluded at the same time. In particular, this trade-off between intra-subject variations by view differences and inter-subject variations is a serious problem for gait recognition because the inter-subject difference within the same action, that is, gait, is subtle (e.g., spatial displacement by differences in strides, arm swings, body shapes, and postures).

To summarize, we have noticed that existing CNN-based cross-view gait recognition fails to address the following two aspects: (1) suitable network architectures depending on verification or identification tasks, and (2) suitable network architectures by taking into consideration the trade-off between intra-subject spatial displacement caused by view differences and inter-subject spatial displacement.

To clarify the first issue, let us elaborate on the verification and identification tasks as follows: For the verification task, a pair of gait features, typically, one is a probe (query) and the other is a gallery (enrolment), are matched to judge whether they are originated from the same subject or different subjects. More specifically, if a similarity score between a matching pair is beyond a certain acceptance threshold, they are judged to be the same subject and vice versa. The verification task is associated with several applications: matching a perpetrator and suspect with a gait verification system for a criminal investigation [3] and detecting a specific person, such as a wanted criminal and terrorist at border control. The key to the success of the verification task is therefore to obtain higher similarity scores for the same subject pairs than those of different subject pairs. For the identification task, a probe (query) is matched with all the galleries (enrolments) to locate the same subject as the probe. Typically, we compute similarity scores between the probe and all the galleries, respectively, and select the gallery with the highest similarity score as the same subject. This task is associated with several applications: retrieving a person captured by a camera from videos from other cameras, that is, person reidentification (ID), and ID-less access control. The key to the success of the identification task is that a similarity score between a probe and the same subject in the galleries is relatively higher than those between a probe and different subjects. Note that it is not necessary that a similarity score between a probe (let it be probe A) and the same subject in the galleries is higher than that between another probe (let it be probe B) and different subjects in the galleries because identifications for probes A and B are independent of each other. It is therefore satisfactory even if the similarity score between probe A and the same subject is low because of covariate conditions (e.g., view difference), provided the similarity scores between probe A and different subjects are lower than it; that is, the important point is the relative similarity scores between a probe and all the galleries. To summarize, the absolute similarity scores are important for the verification task, and the relative similarity scores between a probe and the galleries are important for the identification task; hence, it is preferable to design effective network architectures of a CNN by taking into account the required properties of the verification and identification tasks.

We subsequently elaborate on the second issue: the trade-off between intra-subject spatial displacement caused by view differences and inter-subject spatial displacement. For the case in which the intra-subject spatial displacement is larger than the inter-subject spatial displacement because of larger view differences, it would be more advantageous to enhance the robustness (insensitiveness) to the spatial displacement for better performance. By contrast, in a case where the inter-subject spatial displacement is larger than the intra-subject spatial displacement because of a smaller view difference or no view difference, it would be better to mitigate the insensitiveness, that is, enhance the sensitiveness to the spatial displacement for better performance.

As mentioned previously, effective network architectures of CNNs for cross-view gait recognition are highly dependent on tasks and view differences; hence, we propose suitable approaches to CNN-based cross-view gait recognition for each combination of verification/identification task and small/large view differences. The contributions of this work are four-fold, as follows:

1) **Effective network architectures depending on tasks by contrastive loss/triplet ranking loss.**
   We show that effective input/output architectures of CNN are dependent on verification/ identification task through our experiments on cross-view gait recognition. In particular, a CNN with contrastive loss function (i.e., named *2in* in this paper) is effective for verification task and a CNN with triplet ranking loss (i.e., named *3in* in this paper) is effective for identification task. Although this is not a technical novelty-oriented contribution, we believe this appropriate use of the contrastive loss and the triplet ranking loss depending on tasks, is insightful under situations where most of studies on gait recognition overlook this point.
2) **Different representation of the effective network architectures of 1) by high/low-level difference structure.**
   We constructed new CNNs, *diff* and *2diff*, with low-level difference structures, that is, at the input level, before obtaining the insensitiveness based on *2in* for verification and *3in* for identification respectively. *diff* and *2diff* are effective for the case in which the inter-subject spatial displacement is larger than the intra-subject spatial displacement bacause of a smaller or no view difference, while *2in* and *3in* with high-level difference structures are effective in the case in which the intra-subject spatial displacement is larger than the inter-subject spatial displacement because of large view differences.
3) **High accuracy by Score-level fusion of CNNs with view difference-dependent structures.** Because the degree of view difference is not a binary variable, unlike the task types (i.e., verification or identification),

the aforementioned high-level and low-level difference structures are complementary to each other. We therefore propose a score-level fusion of those view difference-dependent structures to achieve state-of-the-art performance.

4) **Statistically reliable performance evaluation of cross-view gait recognition.** The performance evaluation of previous cross-view gait recognition lacks one of two aspects, a large population (e.g., only hundred-order subjects in the Chinese Academy of Sciences (CASIA) Gait Database [20]) or large view difference (e.g., only a 30° view difference in the Osaka University Institute of Scientific and Industrial Research (OU-ISIR) Gait Database, the Large Population (LP) dataset [21]), whereas this work validates the proposed network architectures with a large-scale as well as a large view-variation gait database that includes more than 10,000 subjects with 14 view variations. This enables us to provide substantially more statistically reliable performance evaluation results than previous work [14], [15].

The outline of the paper is as follows: In Section II, we introduce an existing gait recognition method using the CNN framework, and in Section III, we address the proposed methods for various situations. We describe the performance evaluation for gait recognition in Section IV, and present our conclusions and discuss future work in Section VI.

II. Related Works

Approaches to CNN-based gait recognition are mainly divided into two families in terms of network architectures: (1) a single CNN with a single input (gait feature), and (2) two parallel CNNs with a pair of inputs.

For the CNN with one input, the following methods have been proposed. Shiraga *et al.* [15] designed GEINet, which is a CNN with a single gait feature, that is, the gait energy image (GEI) [8], also known as the averaged silhouette [22] for cross-view gait recognition. Their experimental results demonstrated that GEINet outperformed existing generative and discriminative methods without CNN. Wolf *et al.* [17] proposed a CNN with three-dimensional convolutions using spatio-temporal cuboids of input images with three channels: gray-scale image and optical flow in the horizontal and vertical directions. The approach was evaluated on three datasets that included variations in clothing, walking speeds, and the view angle. Castro *et al.* [23] proposed a CNN using spatio-temporal cuboids of optical flow as the input and showed the robustness of the gait features extracted through the CNN against covariate factors, such as clothing and carrying conditions. A input features of [17] and [23] is significantly differ from a feature of silhouette-based inputs. Since the feature of their inputs includes texture information, performance of gait recognition significantly changes depending on whether subjects' clothes of testing and training data are same or not.

For the CNN with a pair of inputs, the following methods have been proposed. Wu *et al.* [14] provided an extensive empirical evaluation in terms of various scenarios, namely cross-view and cross-walking-condition, with different preprocessing approaches and network architectures, using silhouette-based gait features, that is, GEIs, chrono gait images (CGIs) [24], and silhouette sequences. Their experimental results for their proposed method combining eight variations of CNNs outperformed other approaches, including discriminative approaches with CNN [13]. Zhang *et al.* [16] designed two parallel CNNs with two input GEIs known as the Siamese network [18] with shared parameters.

For the CNN with a single input, the parameters of the CNN are trained so that a soft-max score of the last fully connected layer's node of a true-match subject is high, whereas that those of false-match subjects are low. Specifically, the loss function is defined as the cross-entropy between the soft-max scores and ground truth scores (i.e., one and zero for true-match and false-match subjects, respectively). As the soft-max scores are normalized so that the summation is one, the relative scores among the subjects and not the absolute scores are considered. Because an identification task depends on the relative scores among subjects and not the absolute scores, the CNN with a single input and soft-max loss function is relatively suitable for the identification task.

Additionally, as a more suitable method for the identification task, a triplet of a probe (query), true-match enrolment, and false-match enrolment is considered as an input architecture, and a triplet loss function, which penalizes the case in which the similarity between the probe and the true-match enrolment is less than that between the probe and the false-match enrolment, is also considered [25], [26]. This type of triplet loss has been considered not only in CNN-based approaches but also a variety of metric learning-based approaches (e.g., Rank-SVM for gait recognition [12] and probabilistic relative distance comparison (PRDC) for person reidentification [27]). By contrast, for the CNN with a pair of inputs, the parameters are learned so as to discriminate between whether the pair of inputs are originated from the same subject or not based on the absolute output scores. Specifically, the loss function is defined as a contractive loss, where the small similarity for the genuine pairs and the large similarity for the imposter pairs are penalized. The verification process is performed for each pair of inputs independently and the absolute scores affect performance. Therefore, the CNN with a pair of inputs is effective for the verification task.

Although the CNN with a single input is appropriate for the identification task and the CNN with a pair of inputs is appropriate for the verification task, most previous studies on CNN-based cross-view gait recognition evaluate only rank-1 identification rates, that is, a criterion for the identification task, with the exception of Shiraga *et al.* [15], who evaluated their method both with rank-1 identification rates and criteria, for example, equal error rates (EERs) of false acceptance rates (FARs) and false rejection rates (FRRs). Consequently, no studies have extensively discussed the aforementioned network architecture depending on the gait recognition task.

Furthermore, a trade-off between the intra-subject spatial displacement caused by view differences and inter-subject spatial displacement is also an important aspect because the insensitiveness to spatial displacement increases as the input data passes increasing numbers of convolution and

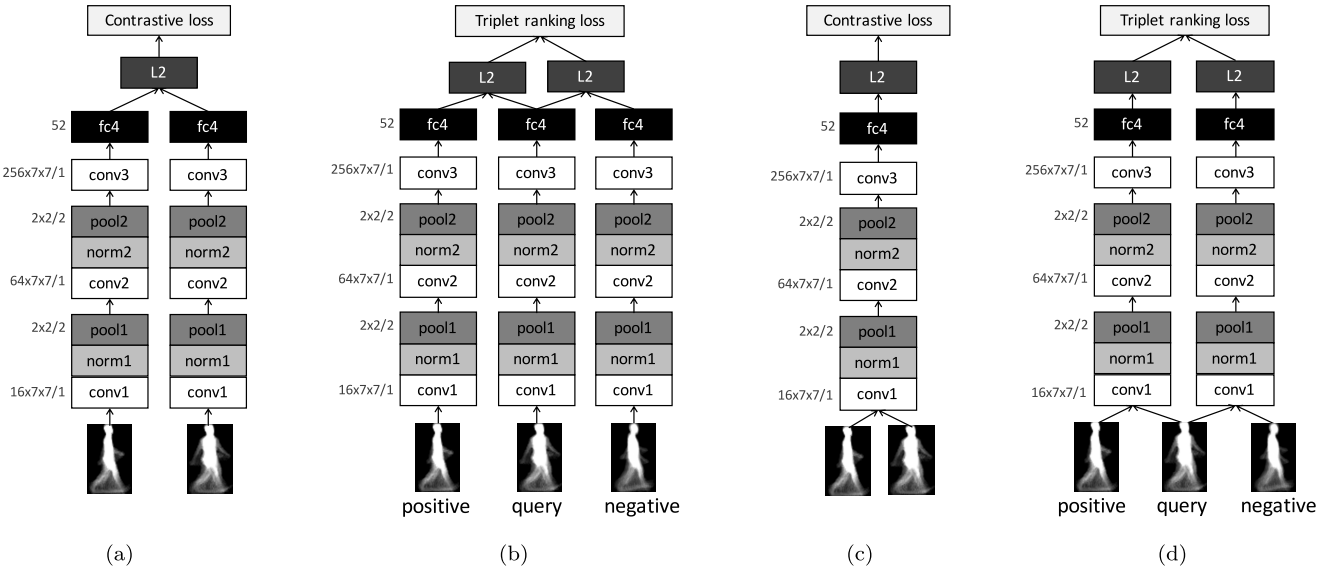

Fig. 1. Network architectures of CNN-based cross-view gait recognition, where conv, norm, pool, fc, and L2 denote convolution layer, normalization layer, pooling layer, fully connected layer, and L2 norm/distance of one/two output from the previous layer, respectively. Numbers written on the left-hand side of the conv, pool, and fc boxes indicate [(#kernel)×(kernel size: width×height)/(stride)], [(#kernel size: width×height)/(stride)], and [(#output node)], respectively. (a) *2in.* (b) *3in.* (c) *diff.* (d) *2diff.*

max-pooling layers in the CNN-based approaches. For example, Wu *et al.* [14] considered three CNNs, where operations (e.g., a difference or fusion) between a matching pair are performed at three different layers, that is, the first, middle, and last layers, but they did not discuss the trade-off between intra-subject spatial displacement caused by view differences and inter-subject spatial displacement from the viewpoints of the layers of operations between the matching pairs. As mentioned above, the previous studies did not address cross-view gait recognition in terms of input/output architecture, that is, the contrastive loss/triplet ranking loss for verification/identification and the high/low-level difference structure for large/small view difference.

Thus, we investigate the effect of input/output architectures on recognition performance for the aforementioned two aspects: identification/verification tasks and the insensitiveness to spatial displacement, which are absent in existing CNN-based gait recognition work.

CNN behavior would change depending on the kind of input image, e.g., silhouettes, gray/color images or optical flow, and depending on the kind of other network architecture of our framework about network architecture in terms of suitability for verification/identification task and small/large view differences, e.g., depth, filter size or branching/coupling. Since our framework can be applied to any exitsting CNN while maintaining its original form in terms of input images and other network architecture of our framework, the CNN can be reconstructed into effective architecture for verification/identification task and small/large view differences while utilizing characteristics of the original CNN. Note that input images should be phase-independent features such as GEI, or gait phases of two inputs should be arranged before feeding in case of phase-dependent feature such as one frame gait image for applying our framework.

III. CNN-Based Cross-View Gait Recognition

In this section, we describe our proposed methods, which are effective depending on the situation of gait recognition. For our methods, GEI is used as the gait feature. GEI is the most prevalent gait feature and is obtained by simply aggregating the silhouette sequence over one gait period. Thus, a GEI represents a mixture of static and dynamic parts.

We first introduce effective methods for each gait recognition task and then introduce effective methods depending on a difference in appearance caused by the effects of view angle difference.

A. General Settings

Our CNNs have three convolution layers, two normalization layers, two pooling layers and one fully connected layer as shown in Fig. 1. In our proposed methods, local response normalization (LRN) [28] is used for the normalization layers norm1 and norm2. A max pooling strategy is used for the pooling layers pool1 and pool2. The ReLU [29] activation function is used for the convolution layers conv1, conv2, and conv3, and the fully connected layer fc. The L2 norm/distance of the one/two output from the fc4 layer is calculated in the last phase of the CNNs. In the training phase, the weight parameters of the CNN are initialized based on Xavier's algorithm presented in [30] and the bias parameters are initialized to zero, and they are updated using the stochastic gradient descent (SGD) algorithm [31]. The momentums for the weights and bias parameters are 0.9, and we use a learning rate that is initialized to 0.01 and divided by 10 four times during the training phase. For the fc4 layers with 52 units, a dropout technique (with probability 0.5) [32] is used. In the testing phase, the L2 distance calculated by a pair of input GEIs is regarded as the dissimilarity between them, and we discriminate between them based on the L2 distance.

B. Methods Depending on the Gait Recognition Task

As mentioned in Section I, there are two tasks for gait recognition: identification and verification. We propose methods for each task and describe their details as follows:

1) Method for the Verification Task: 2in is proposed as the method for the verification task. This method is based on two parallel CNNs known as the Siamese network [16], [18], [33] with shared parameters. In the network, the output of the fc4 layer is regarded as a feature vector of each input GEI. A contrastive loss for each pair is defined as $\boldsymbol{d}^2$ (the squared L2 distance of the feature vectors of the two input GEIs) if they are originated from the same subject, or as the so-called hinge loss if they are different subjects. The contrastive loss function is shown as follows:

$$L_{\text{cont}} = \frac{1}{2M}\sum_{m=1}^{M}(\delta_{y_{1m}y_{2m}} d_m^2 + (1-\delta_{y_{1m}y_{2m}})\max(\text{margin} - d_m^2, 0)), \quad (1)$$

where M denotes the number of input pairs of GEIs for training, δ denotes the Kronecker delta, and y_{1m} and y_{2m} denote the subject IDs of the m-th input GEIs. The value of the margin is empirically determined as three. In the testing phase, we discriminate between whether two inputs are originated from the same subjects or not based on $\boldsymbol{d}^2$. As for Eq. (1), the parameters of *2in* are trained so that the dissimilarity, that is, the values of the contrastive loss function, are smaller for the same subject pair, while those for different subject pairs. Thus, *2in* with contrastive loss is suitable for verification.

2) Method for the Identification Task: 3in is proposed as the method for the identification task. This method is based on three parallel CNNs as the triplet network [25], [26] with shared parameters. In the network, triplet GEIs are fed as input: *query*; *positive*, which is the same subject as *query*; and *negative*, which is different subject from *reference*. The output of the fc4 layer is regarded as a feature vector, as with the case of *2in*. A triplet ranking loss for each triplet is defined as the difference between $\boldsymbol{d}_{\text{pos}}^2$ (the squared L2 distance between feature vectors of *positive* and *query*) and $\boldsymbol{d}_{\text{neg}}^2$ (the squared L2 distance between feature vectors of *negative* and *query*). The triplet ranking loss function is shown as follows:

$$L_{\text{trip}} = \frac{1}{2M}\sum_{m=1}^{M}\max(\text{margin} - d_{\text{neg},m}^2 + d_{\text{pos},m}^2, 0). \quad (2)$$

As Eq. (2), the parameters of *3in* are trained so that the dissimilarity between a query and the same subject, that is, the distance between *query* and *positive*, is relatively lower than that between a query and different subjects, that is, the distance between *query* and *negative*. Thus, *3in* with triplet ranking loss is suitable for identification.

C. Methods Depending on a Difference in Appearance Caused by View Angle Difference

Differences in appearance of input GEIs are caused by the differences of subjects and effects of covariates, for example, view angles. For gait recognition, we discriminate between whether pairs of GEIs are originated from the same subject or not based on the inter-subject spatial displacement while suppressing the influence of the intra-subject spatial displacement caused by view differences.

2in and *3in* with high-level difference structures, that is, considering the difference between a matching pair at the higher level of network architectures after obtaining insensitiveness to spatial displacement, can be robust (insensitive) methods for spatial displacement.

For a case in which the intra-subject spatial displacement is larger than the inter-subject spatial displacement because of larger view differences, a CNN with high-level difference structures (i.e., *2in* and *3in*) would be better. By contrast, in a case where the inter-subject spatial displacement is larger than the intra-subject spatial displacement because of a smaller view difference or no view difference, a CNN with low-level difference structures, that is, taking the difference between a matching pair at the input level before obtaining the insensitiveness, would be better.

Thus, we propose two CNNs for which input GEIs are matched in the form of the images in the first phases of the CNNs: *diff* (see Fig. 1c) and *2diff* (see Fig. 1d) corresponding to *2in* and *3in*, respectively. In *diff*, a subtracted image of pair between input GEIs is fed as input into the CNN, and the squared L2 norm of output of the fc4 layer is treated as $\boldsymbol{d}^2$ in Eq. (1) to calculate the contrastive loss. In *2diff*, on the other hand, two images which are obtained by subtracting *positive* or *negative* from *query* are fed as inputs into each of the two paralle CNNs, and each squared L2 norm is treated as $\boldsymbol{d}_{\text{pos}}^2$ or $\boldsymbol{d}_{\text{neg}}^2$ in Eq. (2) to calculate the triplet ranking loss. *diff* and *2diff* are more directly affected by appearance differences than *2in* and *3in*, both inter and intra-subject, because of the spatial displacement of the corresponding body parts caused by the effects of covariates.

In order to achieve the state-of-the-art performance, we also propose a score-level fusion of those view difference-dependent structures, i.e., fusion of *2in* and *diff* for verification and fusion of *3in* and *2diff* for identification, by averaging L2 distances which are outputs of the fc4 layer of each trained CNN.

D. Sampling Problem

There are some problems regarding sampling training data for the case of using the triplet ranking loss function (Eq. (2)). When the triplet input GEIs include an easy negative sample, that is, a negative sample that is discriminated from the positive sample easily because of a large difference between $\boldsymbol{d}_{\text{pos}}^2$ and $\boldsymbol{d}_{\text{neg}}^2$ because the difference is larger than the margin, the parameters of the CNNs are not updated, that is, the triplet ranking loss is equal to zero. Thus, the parameters can rarely be updated, in particular, for the case of using the gait database that includes many easy negative samples. For such a case, some measures should be taken to mitigate the problem, for example, considering fine tuning using negative samples whose loss values calculated by inputting trained CNN are larger than zero.

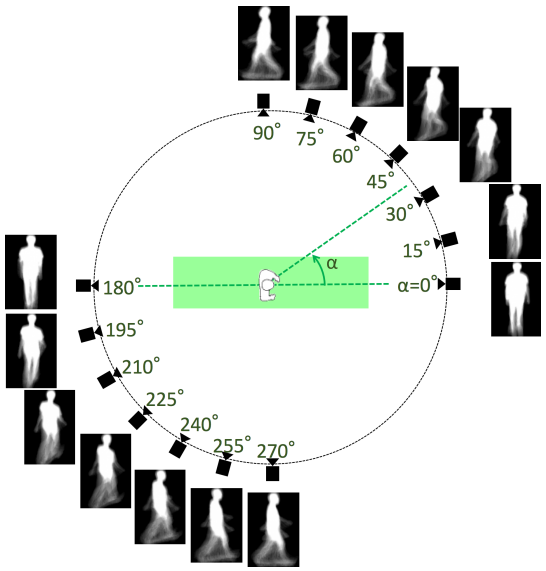

Fig. 2. Examples of GEIs in the OU-ISIR MVLP gait database.

IV. Performance Evaluation

We assumed that the intra-subject spatial displacement was caused by the view angle difference and we evaluated the proposed methods for cross-view gait recognition.

We first evaluated the performance of the proposed methods, that is, *2in*, *3in*, *diff*, *2diff*, and their combinations, for both verification and identification. Second, we compared our method with existing state-of-the-art methods based on various approaches: a generative approach and discriminative approach with or without a CNN framework.

A. Comparison of the Proposed CNNs Depending on the Situation

To evaluate cross-view gait recognition in a statistically reliable way, we built a large gait database with a wide view angle. Using the database, we evaluated the proposed methods in terms of gait recognition tasks and the number of effects of view angle variation.

1) Gait Database: We built the world's largest wide view variation gait database, the OU-ISIR Multi-View Large Population (MVLP) gait dataset, which includes 10,307 subjects (5,114 males and 5,193 females) from 14 view angles: 0°, 15°, 30°, 45°, 60°, 75°, 90°, 180°, 195°, 210°, 215°, 240°, 255°, and 270°, as shown in Fig. 2. The database contains two sequences for each subject and view angle. The gait data were collected in conjunction with a long-running exhibition at a science museum (Miraikan), and informed consent was obtained electronically using the gait collecting framework introduced in [34]. We extracted GEIs as the gait feature for performance evaluation using the method in [21].

2) Evaluation Method: We divided the 10,307 subjects into two disjoint groups of approximately equal size, that is, 5,153 training and 5,154 testing subjects. Additionally, as mentioned previously, each subject had 28 GEIs (14 view angles × 2 sequences).

To approximately align the walking direction in the GEIs, GEIs with more than a 180° view angle were flipped right-to-left based on a perspective projection assumption [35], and the parameters of the CNN were trained for data that included all view angles simultaneously. Note that an approximate walking direction estimation (leftward vs. rightward) is relatively easily performed compared with the detailed pairwise view estimation (30° vs. 45°) required.

For each setting, we evaluated the recognition accuracy for two tasks: verification, that is, one-to-one matching, and identification, that is, one-to-many matching. For the verification task, we calculated false acceptance rates (FARs) and false rejection rates (FRRs). We then calculated the equal error rates (EERs) of the FAR and FRR as a trade-off criterion of the verification capability. Moreover, we calculated the rank-1 identification rate as a criterion of the identification capability.

3) Experimental Results: We evaluated the recognition accuracy of our proposed methods for all pairs of the four typical view angles: 0°, 30°, 60°, and 90°. The rank-1 identification rates and EERs are shown in Table I.

a) 2in, diff versus 3in,2diff: According to Table I, in terms of rank-1 identification rates, *3in* and *2diff* were better than *2in* and *diff*, respectively. In terms of EERs, by contrast, *2in* and *diff* were better. Thus, *2in* and *diff* were effective for the verification task, and *3in* and *2diff* were effective for the identification task.

b) 2in, 3in versus diff, 2diff: In *2in* and *3in* with high-level difference structures, pair or triplet GEIs were matched in the form of 52-dimensional feature vectors in the last phases of the CNNs. By contrast, in *diff* and *2diff* with low-level difference structures, the GEIs were matched in the form of images in the first phases of the CNNs. Thus, *diff* and *2diff* were more directly affected by appearance differences than *2in* and *3in*. As a result, as shown in Table I, in the case of the small (or no) view angle difference, the accuracy of *diff* and *2diff* was higher than that of *2in* and *3in* because leveraging the inter-subject difference was more effective than mitigating the intra-subject difference caused by view angle differences, and *2in* and *3in* were better than *diff* and *2diff* because mitigating the intra-subject difference caused by view angle difference was more effective than leveraging the inter-subject difference. Note that the rank-1 identification rates of *2in* and *diff*, and the EERs of *3in* and *2diff* do not necessarily correspond with the tendency mentioned above because *2in* and *diff* were trained for verification and *3in* and *2diff* were trained for identification.

Considering the relation between *2in* and *diff* in more detail, we compared the distribution of the normalized L2 distance, which was calculated in the last phase of the CNN in *2in* and *diff*, shown in Fig. 3. According to Fig. 3, the overlapping area (i.e., confusing areas of positive and negative samples)in *diff* expanded drastically as the angular difference increased, whereas the overlapping area in *2in* expanded slightly. This also shows that *2in* was robust and *diff* was sensitive to spatial displacement.

c) Combination (2in+diff, 3in+2diff): As mentioned previously, *2in* and *diff* were effective for verification and *3in* and *2diff* were effective for identification, and the properties

TABLE I

RECOGNITION ACCURACY FOR ALL PAIRS OF FOUR TYPICAL VIEW ANGLES. THE BEST AND THE SECOND BEST RESULTS ARE INDICATED BY A BOLD FONT WITH UNDERLINING AND BOLD FONT, RESPECTIVELY, FOR EACH VIEW CONDITION, WHICH ALSO APPLIES TO THE TABLES THAT FOLLOW. NOTE THAT ACTUAL VALUES, NOT ROUNDED VALUES, WERE USED FOR RANKING. (a) RANK-1 IDENTIFICATION RATES (%). (b) EER (%)

(a)

(a-1) *2in*

Gallery \ Probe	0	30	60	90	mean
0	58.2	32.9	16.5	17.4	32.1
30	27.4	81.0	39.0	35.6	45.8
60	12.7	39.4	78.9	46.2	44.3
90	12.4	34.7	42.4	84.1	43.4
mean	27.7	47.0	44.2	45.8	41.2

(a-2) *3in*

Gallery \ Probe	0	30	60	90	mean
0	77.6	**45.4**	**18.1**	17.3	**39.6**
30	**37.8**	89.4	**49.4**	36.7	**53.3**
60	14.3	**51.4**	87.3	**52.2**	**51.3**
90	**14.6**	36.0	**50.4**	88.7	**47.4**
mean	**36.1**	**55.5**	**51.3**	**48.7**	**47.9**

(a-3) *diff*

Gallery \ Probe	0	30	60	90	mean
0	63.2	23.1	5.2	5.7	24.3
30	21.7	83.1	32.4	19.1	39.1
60	5.1	34.2	70.8	42.8	38.2
90	4.8	17.9	38.1	77.0	34.5
mean	23.7	39.6	36.6	36.2	34.0

(a-4) *2diff*

Gallery \ Probe	0	30	60	90	mean
0	**83.6**	34.4	10.3	8.6	34.2
30	30.6	**92.5**	43.8	26.6	48.4
60	8.5	42.0	**89.6**	49.7	47.4
90	6.9	24.9	44.2	**90.8**	41.7
mean	32.4	48.5	47.0	43.9	42.9

(a-5) *2in+diff*

Gallery \ Probe	0	30	60	90	mean
0	66.1	36.4	17.4	**18.0**	34.5
30	30.8	84.6	42.5	**37.6**	48.9
60	**13.0**	42.6	82.0	50.1	46.9
90	13.3	**36.5**	47.0	87.1	46.0
mean	30.8	50.0	47.2	48.2	44.1

(a-6) *3in+2diff*

Gallery \ Probe	0	30	60	90	mean
0	**83.9**	**51.6**	**20.4**	**18.9**	**43.7**
30	**44.8**	**92.1**	**55.9**	**42.2**	**58.7**
60	**16.2**	**57.1**	**90.3**	**60.8**	**56.1**
90	**15.6**	**41.3**	**59.6**	**91.9**	**52.1**
mean	**40.1**	**60.5**	**56.6**	**53.5**	**52.7**

(b)

(b-1) *2in*

Gallery \ Probe	0	30	60	90	mean
0	1.9	2.7	4.7	**4.1**	3.4
30	3.0	1.1	2.2	**1.9**	2.1
60	5.2	2.1	1.2	**1.9**	2.6
90	**4.7**	**2.1**	2.0	0.9	**2.4**
mean	3.7	2.0	2.5	**2.2**	2.6

(b-2) *3in*

Gallery \ Probe	0	30	60	90	mean
0	1.7	**2.4**	4.8	4.5	3.3
30	2.8	1.0	2.2	2.5	2.1
60	5.1	2.3	1.3	2.1	2.7
90	5.0	2.6	2.2	1.1	2.7
mean	3.6	2.1	2.6	2.5	2.7

(b-3) *diff*

Gallery \ Probe	0	30	60	90	mean
0	1.7	3.7	7.8	7.0	5.0
30	3.8	**1.0**	2.9	3.4	2.8
60	8.0	2.9	**1.0**	2.3	3.6
90	7.4	3.5	2.3	**0.6**	3.5
mean	5.2	2.8	3.5	3.3	3.7

(b-4) *2diff*

Gallery \ Probe	0	30	60	90	mean
0	2.5	4.6	8.5	8.5	6.0
30	4.6	1.4	3.7	4.7	3.6
60	8.9	3.8	1.7	3.7	4.5
90	8.6	4.4	3.7	1.4	4.5
mean	6.1	3.6	4.4	4.6	4.7

(b-5) *2in+diff*

Gallery \ Probe	0	30	60	90	mean
0	**1.4**	2.5	**4.6**	**4.0**	**3.1**
30	**2.6**	**0.9**	**1.9**	**1.9**	**1.8**
60	**5.0**	**1.9**	**1.0**	**1.7**	**2.4**
90	**4.4**	**2.0**	**1.7**	**0.6**	**2.2**
mean	**3.4**	**1.8**	**2.3**	**2.1**	**2.4**

(b-6) *3in+2diff*

Gallery \ Probe	0	30	60	90	mean
0	**1.4**	**2.2**	**4.7**	4.4	**3.2**
30	**2.6**	1.0	**2.0**	2.4	**2.0**
60	**5.1**	**2.1**	1.1	2.0	**2.6**
90	4.9	2.4	**2.0**	0.9	2.5
mean	**3.5**	**1.9**	**2.4**	2.4	**2.6**

of *2in* and *3in* with high-level difference structures were substantially different from those of *diff* and *2diff* with low-level difference structures. Therefore, combining these two complementary CNNs improved the accuracy of gait recognition for each task. Note that two CNNs were combined by averaging the L2 distances, which were outputs of the last layer of each trained CNN.

As shown in Table I, the accuracy of the combined methods, that is, *2in+diff* and *3in+2diff*, was better than that of the methods that used one CNN. Furthermore, the EERs of *2in+diff* were better than *3in+2diff*. By contrast, the rank-1 identification rates of *3in+2diff* were better, that is, *2in+diff* and *3in+2diff* were effective for verification and identification, respectively.

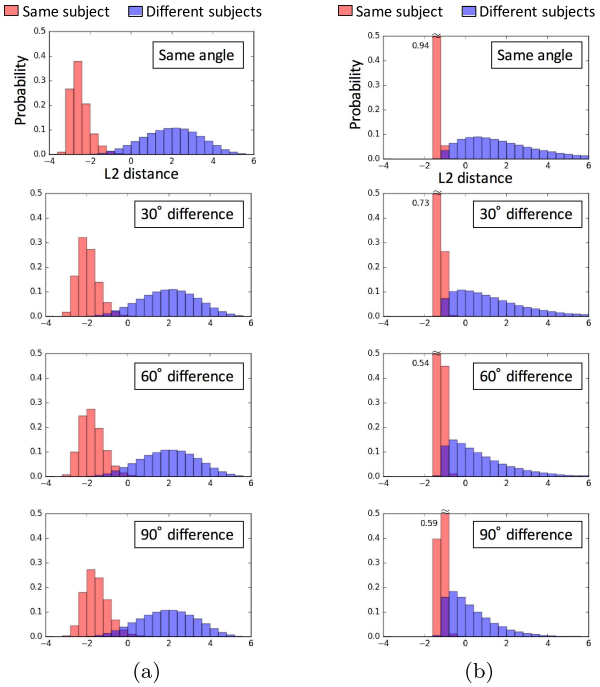

Fig. 3. L2 distance distribution in *2in* and *diff*. (a) *2in*. (b) *diff*.

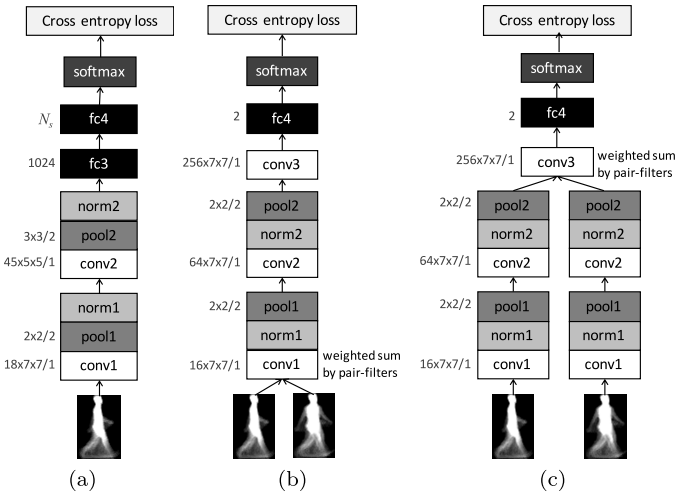

Fig. 4. Network architectures of CNN-based benchmarks. N_S in (a) indicates the number of training subjects. (a) *GEINet*. (b) *LB*. (c) *MT*.

B. Comparison With Benchmarks

We evaluated the performance of existing methods and compared it with the results of our methods using two databases: the OU-ISIR MVLP dataset and OU-ISIR LP dataset [21]. Although OU-ISIR MVLP should be used for a statistically reliable evaluation because the database is the world's largest wide view variation gait database, the existing methods, which are difficult to implement correctly, cannot be evaluated. Therefore, we compared the results of the existing methods using OU-ISIR LP, the second largest (approximately 4,000 subjects) database, reported in the published paper [21].

1) Benchmarks: In this section, we describe the four existing methods used for the evaluation experiments. Each of them is a state-of-the-art method for the generative approach and discriminative approach with or without CNN, respectively. In terms of the discriminative approaches with CNN, we compared our results with two methods: a CNN with one input GEI and CNNs with two inputs. We also compared the method of direct matching (*DM*) without any training, that is, the Euclidean distance between GEIs, as a baseline.

- Direct matching (*DM*) [21]:
  This method is based on direct matching, that is, the L2 distance in the original feature space. We regarded a GEI as a feature vector whose dimension was equal to the number of its pixels, and then computed the L2 distance of two GEIs as the dissimilarity.
- View transformation model (*VTM*) [6]:
  As a general approach to cross-view gait recognition, the VTM family of methods [7], [36]–[39] have been widely studied and a singular value decomposition-based VTM [6] is the most basic approach. Hence, we exploited it as a baseline for the generative approach. This method obtains the VTM using the training data of multiple subjects from multiple view angles. In the recognition phase, the VTM transforms gallery features into the same view angle as those of an input feature, and the features are matched under the same view.
- Linear discriminant analysis (*LDA*) [40]:
  This method is based on LDA, which is the most conventional approach and hence widely exploited in many fields (e.g., Fisher face [41]). Therefore, we adopted this method as the baseline for the discriminative approach. Specifically, we first applied principal component analysis (PCA) to an unfolded GEI feature vector to reduce its dimensions and subsequently applied LDA to obtain the discriminant features.
- *GEINet* [15](See Fig.4(a)):
  This method is based on one of the simplest CNNs, where one input GEI is fed into the network and the number of nodes in the final layer (fc4) is equal to the number of training subjects. The soft-max value, calculated from the output of the final layer, is regarded as a type of probability that the input matches for a corresponding subject.
- Wu's method proposed in TPAMI, 2016 [14] (*Wu*):
  In this method, pairs of images are discriminated between based on the combined similarity of the following eight CNNs, which is the average of the similarity calculated for each CNN:
  - local @ bottom with three convolution layers (*LB*)
  - *LB* with group sparsity
  - Siamese network with three convolution layers
  - mid-level @ top with three convolution layers (*MT*)
  - *MT* with group sparsity
  - two-stream network with GEI and CGI
  - network with one-frame input (averaging 16 samples)
  - network with three-frame input (averaging 16 samples)

  Because it was difficult for us to implement all CNNs correctly, we implemented *LB* and *MT* for performance

TABLE II

RECOGNITION ACCURACY COMPARING OUR METHOD WITH THE BENCHMARKS USING OU-ISIR MVLP. (a) RANK-1 IDENTIFICATION RATES (%). (b) EERS (%)

(a)

	Angular difference				
	0	30	60	90	mean
DM	77.4	2.4	0.2	0.0	20.3
LDA	81.6	10.1	0.8	0.1	24.4
VTM	77.4	2.7	0.6	0.2	20.5
GEINet	85.7	40.3	13.8	5.4	40.7
LB (*Wu*)	**89.9**	42.2	15.2	4.5	42.6
MT (*Wu*)	89.3	**49.0**	20.9	8.2	46.9
2in	75.5	37.9	24.9	14.9	41.2
3in	85.7	47.8	**26.3**	**15.9**	**47.9**
diff	73.6	32.1	11.8	5.2	34.0
2diff	89.1	40.8	17.6	7.8	42.9
2in+diff	80.0	41.5	26.1	15.6	44.1
3in+2diff	**89.5**	**55.0**	**30.0**	**17.3**	**52.7**

(b)

	Angular difference				
	0	30	60	90	mean
DM	6.5	25.2	41.4	46.2	27.2
LDA	6.2	22.7	35.7	40.1	24.0
VTM	6.5	26.8	34.2	38.5	25.0
GEINet	2.4	5.9	12.7	17.2	8.1
LB (*Wu*)	1.0	3.3	6.7	9.3	4.3
MT (*Wu*)	**0.9**	2.5	5.2	7.0	3.3
2in	1.3	2.4	**3.5**	**4.4**	2.6
3in	1.3	2.3	3.7	4.7	2.7
diff	1.1	3.0	5.7	7.2	3.7
2diff	1.8	4.0	6.6	8.5	4.7
2in+diff	**1.0**	**2.0**	**3.4**	**4.2**	**2.4**
3in+2diff	1.1	**2.2**	3.6	4.6	**2.6**

TABLE III

RECOGNITION ACCURACY COMPARING OUR METHODS WITH THE BENCHMARKS USING OU-ISIR LP. (a) RANK-1 IDENTIFICATION RATES (%). (b) EERS (%)

(a)

	Angular difference				
	0	10	20	30	mean
DM	91.5	49.5	11.2	2.8	44.6
LDA	97.8	97.1	93.4	82.9	94.6
VTM	91.5	64.0	37.2	20.5	58.7
GEINet	96.5	95.8	92.5	84.9	93.8
Wu [14]	98.9	95.5	92.4	85.3	94.3
2in	97.9	97.6	95.6	92.0	96.5
3in	98.5	98.2	96.4	92.3	97.1
diff	98.7	98.5	97.2	94.7	97.7
2diff	99.1	99.0	98.0	95.1	98.3
2in+diff	**99.3**	**99.2**	**98.6**	**96.9**	**98.8**
3in+2diff	**99.2**	**99.2**	**98.6**	**97.0**	**98.8**

(b)

	Angular difference				
	0	10	20	30	mean
DM	4.3	8.4	20.2	31.3	13.2
LDA	2.1	2.5	3.7	5.7	3.1
VTM	4.3	10.5	14.8	18.9	11.1
GEINet	1.9	2.1	3.0	4.9	2.6
Wu [14]	-	-	-	-	-
2in	0.3	0.3	0.5	0.7	0.4
3in	0.7	0.8	1.0	1.4	0.9
diff	**0.3**	**0.3**	**0.4**	**0.7**	**0.4**
2diff	1.8	2.0	2.7	3.9	2.4
2in+diff	**0.2**	**0.2**	**0.2**	**0.4**	**0.2**
3in+2diff	1.0	1.1	1.4	1.9	1.3

evaluation for the case of using the OU-ISIR MVLP database. Note that for the case of using the OU-ISIR LP database, the results of *Wu* were the same as those provided in [14]. We describe these two methods as follows:

- *LB* (See Fig.4(b)):
  This method is based on two parallel CNNs with shared parameters, where two nonlinear projections are applied before computing the differences between pairs of images on the conv1 layer. The pixel-wise summation of weighted entries using a pair-filter is calculated as the output of the conv1 layer. The softmax loss function is used for training, where the softmax value indicates the probability of the event that they are the same subject.
- *MT* (See Fig.4(c)):
  This method is based on a CNN whose structure is similar to that of *LB* except for a layer in which we compute the differences between pairs of images. A linear projection is applied before computing the differences between pairs of GEIs on the conv3 layer.

2) Experimental Results Using OU-ISIR MVLP: We summarize the accuracy of benchmarks and our proposed methods in Table II by averaging the results for each angular difference for all possible pairs of four typical view angles, for example, the result of the 60° difference was calculated by averaging the results of four pairs, (probe view, gallery view) = (0°, 60°), (30°, 90°), (60°, 0°), and (90°, 30°). Note that for the generative approaches, that is, *VTM*, we assumed that the probe and gallery view angles were known and the VTM from a source view to a target view were trained using the gait features from a possible view angle. The results of our methods were the same as those of Table I.

According to Table II, the discriminative approach with CNN outperformed the other approaches for both verification and identification. Furthermore, the EERs of *2in+diff* were the best for verification and the rank-1 rates of *3in+2diff* were the best for identification. The results of *LB* and *MT* are comparable to the results of the score-level fusion in case with no view difference, i.e., the case where intra-subject spatial displacement is sufficiently small compared to inter-subject spatial displacement. In contrast, the results of *LB* and *MT* are worse than those of the score-level fusion in the case with larger view difference.

3) Experimental Results Using OU-ISIR LP: We used the same protocol as Wu *et al.* [14] for performance evaluation compared with *Wu*. We divided 3,844 subjects into five disjoint groups of approximately equal size and conducted five-fold cross validation using the same subject ID list as Wu *et al.* Additionally, because OU-ISIR LP included many easy samples for discrimination because of its small view angle variation, we used fine tuning for *3in* and *2diff* to manage the sampling problem mentioned in Section III-D.

The results of the benchmarks and our methods are shown in Table III. Comparing our proposed methods, the same tendency as the results using OU-ISIR MVLP, that is, *2in*, *diff*, and their combination, were effective for verification, and *3in*, *2diff*, and their combination were effective for identification. Furthermore, *2in* and *3in* were better than *diff* and *2diff* for cases with large angular differences, and the opposite

TABLE IV
EER COMPARING *diff* WITH *GoogLeNet*

	Angular difference				
	0	30	60	90	mean
diff	1.1	3.0	5.7	7.2	3.7
GoogLeNet	1.3	3.6	6.7	8.7	4.4

TABLE V
EER COMPARING *diff* WITH *LB*

(Matching method) - (Loss function)	Angular difference				
	0	30	60	90	mean
Sub - Contrastive (*diff*)	1.1	3.0	5.7	7.2	3.7
Sub - Softmax	1.1	3.3	6.8	9.3	4.3
wSum - Contrastive	1.0	3.1	5.9	7.7	3.8
wSum - Softmax (*LB*)	1.0	3.3	6.7	9.3	4.3

result was obtained for cases with small angular differences. Moreover, our combined methods outperformed the other existing approaches.

V. DISCUSSION

A. Comparison With Deeper CNN

To investigate an effect of CNN depths, we have compared results of a state-of-the-art deep network, *GoogLeNet* [42] which is the winner of ILSVRC (ImageNet Large Scale Visual Recognition Challenge) 2014, with those of the simplest one of the proposed CNNs, *diff*. The depth of *GoogLeNet* is 22 (21 convolutional layers and 1 fully connected layer) while the depth of *diff* is 4 (3 convolutional layers and 1 fully connected layer). A difference image of two GEIs same as inputs of *diff* is fed to *GoogLeNet*, and loss functions of *GoogLeNet* are changed to contrastive loss function same as *diff*.

Since CNN with contrastive loss function is effective for verification, EER is used for comparing their performance. EERs of *GoogLeNet* and *diff* are shown in Table IV. As the results, EER of *GoogLeNet* is worse than that of *diff*. Although a deeper network has advantage of having higher expression capability of gait feature, it means that this also has disadvantage of being easier to overfit. In a case that variation of input image is large such as object recognition, the effect of advantage can be larger than that of disadvantage. On the other hand, in a case that variation of input image is small such as gait recognition, the effect of advantage can be smaller. This can be why result of *GoogLeNet* become worse. According to [14] which also mentioned about impact of CNN depths, it was confirmed that performance of CNN with three convolutional layers is the best among CNNs with two, three or five convolutional layers for silhouette-based cross-view gait recognition. From the above results, CNN with three convolution layers would be effective for silhouette-based cross-view gait recognition.

B. Comparison diff With LB

LB is very similar to *diff*. *diff* differ from *LB* in that its loss is contrastive loss (*Contrastive*) instead of cross entropy loss based on soft-max score (*Softmax*), and in that it takes difference between two input images directly (*Sub*) for matching them instead of pixelwise weighten summation using pair filter (*wSum*). We conducted expreiments of using all possible combination of these loss function (*Contrastive*/*Softmax*) and these matching methods (*Sub*/*wSum*). Table V shows EERs for each of them.

Since *wSum* has higher expression capability of gait feature covering that of *Sub*, i.e., *wSum* can expression any feature expressed by *Sub*, EERs of *wSum* are slightly better than those of *Sub* in the case with no angular difference, i.e., the case where intra-subject spatial displacement is sufficiently small compared to inter-subject spatial displacement. On the other hand, in the cases with large angular difference, i.e., the case where intra-subject spatial displacement is larger than inter-subject spatial displacement due to large view differences, EERs of *wSum* become worse than those of *Sub* because of overfitting.

Both of *diff* and *LB* deal with two-class classification whether two inputs are originated from the same subject or different subjects. Comparing *Softmax* with *Contrastive* in detail, *Softmax* handles two classes with same weight while *Contrastive* ignoring the case of different subjects with large difference between their features, i.e., focusing around a boundary of two classes. Since variation of differences between their features in the case of different subjects is much larger than that of same subject as a matter of fact, the results of *Contrastive* are better than those of *Softmax* in Table V. Thus, *Contrastive* would be more effective for gait verification.

VI. CONCLUSION

This paper described a method of cross-view gait recognition using CNN. More specifically, we designed a Siamese network for verification and a triplet network for identification as CNNs with high-level difference structures. We also designed CNNs with low-level difference structures, that is, considering the difference at the input level, before obtaining the insensitiveness corresponding to them. We then proposed a score-level fusion of these structures, which are complementary to each other. As a result of the experiments for cross-view gait recognition using the OU-ISIR MVLP dataset and the OU-ISIR LP dataset, we confirmed that the hypothesis regarding network architecture was correct, that is, (1) the Siamese network was suitable for verification, (2) the triplet network was suitable for identification, (3) a CNN with a high-level difference structure was effective for the case in which the intra-subject spatial displacement was larger than the inter-subject spatial displacement because of large view differences, (4) a CNN with a low-level difference structure was effective for the case in which the inter-subject spatial displacement was larger than the intra-subject spatial displacement because of a smaller or no view difference. We also confirmed that the proposed score-level fusion method significantly outperformed the state-of-the-art approaches for both verification and identification.

In this paper, we evaluated cross-view gait recognition by considering the spatial displacement caused by view angle difference. However, spatial displacement is caused by not

only view difference but also walking speed difference, carrying status difference, clothing difference, and other factors. Thus, evaluating our proposed method for gait recognition with spatial displacement caused by other covariates will be considered in future work.

ACKNOWLEDGMENT

This work was performed at the Mitsubishi Electric Collaborative Research Division for Wide-area Security Technology, the Institute of the Scientific and Industrial Research, Osaka University.

REFERENCES

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

[2] N. Lynnerup and P. Larsen, "Gait as evidence," *IET Biometrics*, vol. 3, no. 2, pp. 47–54, 6 2014.

[3] H. Iwama, D. Muramatsu, Y. Makihara, and Y. Yagi, "Gait verification system for criminal investigation," *IPSJ Trans. Comput. Vis. Appl.*, vol. 5, pp. 163–175, Oct. 2013.

[4] A. Kale, A. K. R. Chowdhury, and R. Chellappa, "Towards a view invariant gait recognition algorithm," in *Proc. IEEE Conf. Adv. Video Signal Based Surveill.*, Jul. 2003, pp. 143–150.

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

[6] 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.*, May 2006, pp. 151–163.

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

[8] 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.

[9] K. Bashir, T. Xiang, and S. Gong, "Gait recognition without subject cooperation," *Pattern Recognit. Lett.*, vol. 31, no. 13, pp. 2052–2060, Oct. 2010.

[10] A. Mansur, Y. Makihara, D. Muramatsu, and Y. Yagi, "Cross-view gait recognition using view-dependent discriminative analysis," in *Proc. 2nd IEEE Int. Joint Conf. Biometrics (IJCB)*, Sep. 2014, pp. 1–8.

[11] Y. Guan, C. T. Li, and F. Roli, "On reducing the effect of covariate factors in gait recognition: A classifier ensemble method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 7, pp. 1521–1528, Jul. 2015.

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

[13] Z. Wu, Y. Huang, and L. Wang, "Learning representative deep features for image set analysis," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1960–1968, Nov. 2015.

[14] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan, "A comprehensive study on cross-view gait based human identification with deep CNNs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 209–226, Feb. 2017.

[15] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "GEINet: View-invariant gait recognition using a convolutional neural network," in *Proc. 9th IAPR Int. Conf. Biometrics*, Jun. 2016, pp. 1–8.

[16] C. Zhang, W. Liu, H. Ma, and H. Fu, "Siamese neural network based gait recognition for human identification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 2832–2836.

[17] T. Wolf, M. Babaee, and G. Rigoll, "Multi-view gait recognition using 3D convolutional neural networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 4165–4169.

[18] S. Chopra, R. Hadsell, and Y. Lecun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 539–546.

[19] B. DeCann and A. Ross, "Relating ROC and CMC curves via the biometric menagerie," in *Proc. 6th IEEE Int. Conf. Biometrics, Theory, Appl. Syst.*, Sep. 2013, pp. 1–8.

[20] 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 Recognit.*, vol. 4. Aug. 2006, pp. 441–444.

[21] 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.

[22] Z. Liu and S. Sarkar, "Simplest representation yet for gait recognition: Averaged silhouette," in *Proc. 17th Int. Conf. Pattern Recognit.*, vol. 1. Aug. 2004, pp. 211–214.

[23] F. M. Castro, M. J. Marín-Jiménez, N. Guil, and N. P. de la Blanca, "Automatic learning of gait signatures for people identification," in *Proc. 14th Int. Work-Conf. Artif. Neural Netw.*, Jun. 2017, pp. 257–270.

[24] C. Wang, J. Zhang, L. Wang, J. Pu, and X. Yuan, "Human identification using temporal information preserving gait template," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2164–2176, Nov. 2012.

[25] J. Wang *et al.*, "Learning fine-grained image similarity with deep ranking," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1386–1393.

[26] X. Wang and A. Gupta, "Unsupervised learning of visual representations using videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2015, pp. 2794–2802.

[27] W. S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *Proc. CVPR*, Jun. 2011, pp. 649–656.

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

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

[30] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. Artif. Intell. Stat. (AISTATS)*, 2010, pp. 249–256.

[31] O. Bousquet and L. Bottou, "The tradeoffs of large scale learning," in *Advances in Neural Information Processing Systems*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. Red Hook, NY, USA: Curran Associates, 2008, pp. 161–168.

[32] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[33] N. Takemura, K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "View-invariant gait recognition using convolutional neural network," (in Japanese) *IEICE Trans. Fundam.*, vol. J99-A, no. 12, pp. 440–450, Dec. 2016.

[34] Y. Makihara *et al.*, "Gait collector: An automatic gait data collection system in conjunction with an experience-based long-run exhibition," in *Proc. 9th IAPR Int. Conf. Biometrics*, Jun. 2016, pp. 1–8.

[35] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi, "Which reference view is effective for gait identification using a view transformation model?" in *Proc. IEEE Comput. Soc. Workshop Biometrics*, New York, NY, USA, Jun. 2006, p. 45.

[36] 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. 2nd IEEE Int. Workshop Tracking Hum. Eval. Their Motion Image Sequences*, Oct. 2009, pp. 1058–1064.

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

[38] D. Muramatsu, Y. Makihara, and Y. Yagi, "Cross-view gait recognition by fusion of multiple transformation consistency measures," *IET Biometrics*, vol. 4, pp. 62–73, Jun. 2015.

[39] D. Muramatsu, Y. Makihara, and Y. Yagi, "View transformation model incorporating quality measures for cross-view gait recognition," *IEEE Trans. Cybern.*, vol. 46, no. 7, pp. 1602–1615, Jul. 2016.

[40] N. Otsu, "Optimal linear and nonlinear solutions for least-square discriminant feature extraction," in *Proc. 6th Int. Conf. Pattern Recognit.*, 1982, pp. 557–560.

[41] P. N. Belhumeur, J. P. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[42] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

Noriko Takemura received the B.S., M.S., and Ph.D. degrees in engineering from Osaka University, Osaka, Japan, in 2006, 2007, and 2010, respectively. She is currently an Associate Professor with the Institute of Datability Science, Osaka University. Her current research interests include gait recognition, ambient intelligence, and emotion estimation. She is a member of the SICE, RSJ, and VRSJ.

Yasushi Makihara received the B.S., M.S., and Ph.D. degrees in engineering from Osaka University, Osaka, Japan, in 2001, 2002, and 2005, respectively. He is currently an Associate Professor with the Institute of Scientific and Industrial Research, Osaka University. His current research interests include gait recognition, morphing, and temporal super resolution. He is a member of the IPSJ, RJS, and JSME.

Daigo Muramatsu (M'05) received the B.S., M.E., and Ph.D. degrees in electrical, electronics, and computer engineering from Waseda University, Tokyo, Japan, in 1997, 1999, and 2006, respectively. He is currently an Associate Professor with the Institute of Scientific and Industrial Research, Osaka University. His current research interests include gait recognition, signature verification, and biometric fusion. He is a member of the IEICE, IPSJ, and IIEEJ.

Tomio Echigo (M'09) received the B.S. and M.S. degrees in electrical engineering from the University of Osaka Prefecture, Japan, in 1980 and 1982, respectively, and the Ph.D. degree in engineering science from Osaka University, Japan, in 2003. He joined IBM Japan, Ltd., in 1982, where he was an Advisory Researcher with the Tokyo Research Laboratory, IBM Research, Kanagawa, Japan. From 2003 to 2006, he was a Visiting Professor with Osaka University. Since 2006, he has been with Osaka Electro-Communication University, where he is currently a Professor with the Department of Engineering Informatics. His research interests include image and video processing, medical imaging, multimedia, and robot vision. He is a member of IPSJ, IEICE, RSJ, and ITE.

Yasushi Yagi (M'91) received the Ph.D. degree from Osaka University in 1991. In 1985, he joined the Product Development Laboratory, Mitsubishi Electric Corporation, where he was involved in robotics and inspections. He became a Research Associate in 1990, a Lecturer in 1993, an Associate Professor in 1996, and a Professor in 2003 with Osaka University. He was the Director of the Institute of Scientific and Industrial Research, Osaka University, from 2012 to 2015. He has been the Executive Vice President of Osaka University since 2015. His research interests are computer vision, pattern recognition, biometrics, human sensing, medical engineering, and robotics. He is a fellow of IPSJ and a member of IEICE and RSJ. He received the ACM VRST2003 Honorable Mention Award, the IEEE ROBIO2006 Finalist of the T.J. Tan Best Paper in Robotics, the IEEE ICRA2008 Finalist for the Best Vision Paper, the PSIVT2010 Best Paper Award, the MIRU2008 Nagao Award, the IEEE ICCP2013 Honorable Mention Award, the MVA2013 Best Poster Award, the IWBF2014 IAPR Best Paper Award, and the *IPSJ Transactions on Computer Vision and Applications* Outstanding Paper Award in 2011 and 2013. He has served as the Chair for international conferences, including ROBIO2006 (PC), ACCV (2007PC, 2009GC), PSVIT2009 (FC), and ACPR (2011PC, 2013GC). He has also served as the Editor for the IEEE ICRA Conference Editorial Board from 2008 to 2011. He is a member of the Editorial Board of the *International Journal of Computer Vision*, an Editor-in-Chief of the *IPSJ Transactions on Computer Vision and Applications*, and the Vice-President of the Asian Federation of Computer Vision Societies.