Cross-View Gait Recognition Based on U-Net

Israel Raul Tiñini Alvarez
Centro de Investigación, Desarrollo e
Innovación en Ingeniería Mecatrónica
Universidad Católica Boliviana "San Pablo"
La Paz, Bolivia
ir.tinini@acad.ucb.edu.bo

Guillermo Sahonero-Alvarez
Centro de Investigación, Desarrollo e
Innovación en Ingeniería Mecatrónica
Universidad Católica Boliviana "San Pablo"
La Paz, Bolivia
g.sahonero@acad.ucb.edu.bo

***Abstract*—Gait based recognition systems allow automatic subjects' recognition by using the way of walking. However, the performance of these systems is often degraded by some covariate factors such as walking direction, appearance changes, occlusions, among others. From these, it has been shown that change in appearance is the most influent covariant by drastically affecting the recognition performance. Consequently, inspired by the great successes of GANs in image translation tasks, we propose a method of gait recognition using a conditional generative model to generate view-invariant features. The proposed method is evaluated on one of the largest datasets available under the variations of view, clothing and carrying conditions: CASIA gait database B. Experimental results show that the proposed method outperforms state-of-the-art methods specially in carrying-bag and wearing-coat sequences. The full implementation and trained networks are available at** https://gitlab.com/IsRaTiAl/gait

Keywords—gait recognition, conditional generative model, cross-view recognition.

I. INTRODUCTION

Interest in developing automatic systems for automatic human identification has increased recently due to the growth in the number of crimes and security breaches occurred in past years [1], [2]. As result, the community of researchers and companies of security systems have begun to focus on expanding and improving surveillance systems [3]. Automatic surveillance systems can be classified according to their task: pre-detection, detection/tracking, and classification or identification [4]. Among these, the most attractive for intelligent surveillance is the last one, which would allow not only identify possible risks in the monitored spaces but also, access control, and personal verification [5].

Currently, recognition systems use biometrics traits such as voice, iris, face, and fingerprints, or similar, to automatically identify subjects. These features have become popular due to their unicity, universality, and permanence [1], producing an increase on the number of available datasets. Nevertheless, these biometrics technologies feature two main disadvantages: first, their performance decreases when using low-resolution images, and second, user cooperation is required to achieve high performance [6]–[10]. These limitations are being addressed by proposing and employing new biometrics technologies, such as gait recognition. The signatures found from gait are hard to fake, and the respective required image sequences can be acquired from relatively large distances in uncontrolled scenarios by using simple security cameras since they do not require cooperation from subjects [11].

Gait can be formally defined as a combination of the way of walking and posture [12]. Recently, the gait biometric trait was shown to be feasible to use for human identification despite distance [13]. In fact, it is considered the most suitable biometric for surveillance systems. For these reasons, this technology has attracted significant attention in the research community as it is seen as a valid and even a better alternative for people identification [14].

Gait recognition techniques can be classified into two main categories: model-based and model-free approaches. Model-based methods model the person's body structure; however, this process gets computationally expensive, error-prone, and requires high-resolution images [15]. On the other hand, model-free or appearance-based approaches build signatures from silhouettes extracted from video sequences. The work presented in this paper belongs to the second category. The motivation to use silhouettes for gait recognition is to avoid being affected by clothes' colors and textures, illumination changes, among others [16]. These silhouettes are used as features in the recognition stage. Hence, less computation is required, and its performance does not reduce substantially when using low-resolution images, implying less memory usage [15].

Due to these advantages, current trends in gait recognition methods seem to focus on model-free approaches. Unfortunately, automatic gait recognition remains a challenging task because of the variations that can drastically alter the human appearance, such as viewpoint, clothing, carrying, shoe, surface, and even time conditions [17]. Among them, cross-view is the most difficult to deal with, since we cannot control the walking directions of subjects in real applications [18]. It must be noted that multi-view and cross-view are different. Multi-view settings imply the gallery set to be composed of information from multiple views and conditions. In contrast, cross-view methods suppose that only one view angle is provided as gallery.

In this work, we focus on solving the appearance change by using a conditional generative model, which generates view-invariant representations. These representations are reduced in dimensionality by applying Principal Component Analysis. Then, a discriminant classifier is employed to perform the categorization task. We evaluate the proposed method using CASIA-B benchmarked dataset by measuring the rate of correctly classified examples. Our contributions are as follow:

- We propose to use a combined cross-view gait recognition approach by using a generative model to produce side-view gait energy images and utilize a discriminative model for the classification task.
- We provide an extensive experimental evaluation of the proposed model, which is designed to address appearance changes due to changes in the view.

The remainder of the paper is organized as follows: Section 2 reviews previous works. Section 3 describes the proposed method. Experiments and results are presented in Section 4. Finally, Section 5 shows our conclusions.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

II. Related Works

Cross-view gait recognition approaches can be categorized in two: discriminative and generative models. Discriminative approaches aim at optimizing the discrimination capability of the classifiers. In contrast, generative approaches aim at learning view projections, i.e. they transform gait representations from one view to another.

Discriminative approaches rely on handcrafted characteristics generated from silhouette images, which are fed into machine learning models as features. Typically, traditional classifiers are used in discriminative models such as k-nearest neighbors (KNN) [10], linear discriminant analysis (LDA) [6], canonical discriminant analysis (CDA) [19]. However, other authors [8]–[10] focus on selecting the most important parts from the silhouettes to represent the gait sequences. Nevertheless, currently, there is a trend of employing deep learning techniques to address the issue of appearance and view changes [1]. For example, [20] and [21] employ a 2D image to represent the gait cycle, and a neural network is used as classifier. A similar idea is presented in [22] and [23], where gait representations are used to feed convolutional neural networks (CNN). Although discriminative models work well in specific scenarios, they usually overfit the training data due to the lack of available datasets, so they cannot generalize the gait patterns for changes in view and appearance.

Compared to the first category, generative approaches can be applied to real scenarios since prior knowledge is not entirely needed. For example, in [16], the GEINet is proposed, which is one of the first works to address the cross-view recognition. This work applies modules of the form Conv-Pool-Normalization to transform all the representations to side-view images. Even more sophisticated preprocessing algorithms are employed in [14], where a deep model based on auto-encoders is designed to overcome the issue of view variations. Specifically, Stacked Progressive AutoEncoders (SPAE) of 5 layers were used. The advantage of this model is that it can extract view variant features from any view using only one model; therefore, another algorithm for view estimation is not needed.

In [11], the same idea of SPAE is applied to deal with the problem of view, clothing, and carrying condition variation. Although the results obtained did not outperform the state-of-the-art methods, employing autoencoders to address the problem of covariates is a promising idea that already showed improvements in the field of computer vision. A similar approach is used in [13], which addresses the problem of view variations by employing a GAN model as a regressor to extract invariant gait features. For that, gait data captured with multiple variations is transformed into the side view without prior information. What is interesting in this work is that they use two discriminators, the first one is used to predict whether an image is real, and the second one is used to predict if a generated image has preserved the identity information. Moreover, in [24], the same architecture is used, but this time an improved loss function is applied.

Finally, Zhang et al. [25] aim to address the cross-view problem using a gait representation network and a generative network. They propose a feature learning network based on a VGG16 and Hard Triplet Loss. Due to the variations on view angle, clothing, and carriages, they propose a GAN based on a U-Net, which acts as a regressor to transform gait images captured with any source of variation into a unique view image.

Gait representations are needed due to the large amount of data generated from the recorded gait sequences. One of the most popular gait representations, the gait energy image (GEI) suggested by Han et al. [3], is a spatio-temporal representation of the gait that can be obtained by averaging silhouettes over a gait cycle. It considers an effective balance between computational cost and recognition performance [7]. However, clothing, and carrying variations influence the recognition performance [15]. In fact, Matovski et al. [12] demonstrated that clothing variations drastically affect the recognition performance when using GEI representations.

III. Proposed Method

In this paper, we focus on solving the problem of view-variations caused by the change in walking directions of the subjects. In fact, Yu et al. [11], [13], [14], [24] have demonstrated that the recognition performance is significantly affected when there is a large view variation between the gallery and the probe sets.

To reduce the effect of view-variations, a GAN is employed to generate invariant GEI representations, i.e. gait images at arbitrary views with carrying objects and wearing coats are converted to images at the side view in normal conditions. We used side view images because they allow us to capture more robust information [21]. However, preserving identification information in the generated images is the most challenging part of this task.

Our framework contains four main modules, which are shown in Fig. 1. We first introduce the gait energy image, which is a signature used to represent the gait cycles. Subsequently, the generator and discriminators models are detailed. After that, we describe the feature extraction and selection techniques utilized. Finally, the classification algorithm which uses the selected features from the generated images to predict the identity of one gait sample is described.

A. Representation

GEI is a spatio-temporal representation of gait cycles, which is produced by averaging the silhouettes extracted over a complete gait cycle. It is currently one of the most used signatures to represent gait for its robustness to noise and efficient computation [3] – Fig. 2 shows a set of sample silhouettes images from one person, and the corresponding GEI rightwards. However, it is vulnerable to appearance changes of the human silhouette due to variations in clothing or view-angle [15] – Fig. 3 exemplifies changes on representation in according to different circumstances.

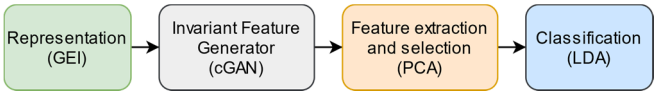

Fig. 1. The proposed framework for gait recognition

Fig. 2. Silhouettes extracted from a gait cycle, where the rightmost correspond to the Gait energy image.

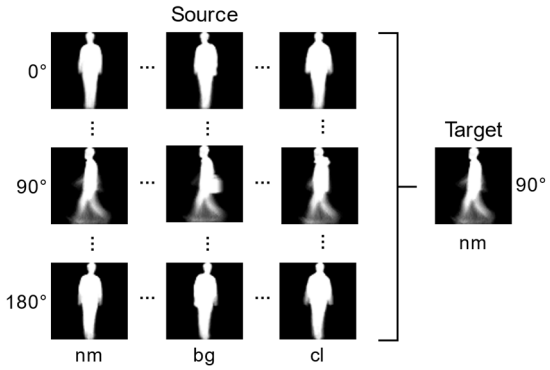

Fig. 3. Source and target images disposition.

B. Invariant feature generator

Inspired by the pixel-level domain transfer in [13], and the architecture presented in [26], in this work, we propose a conditional generative model to transform gait representations from any view and appearance condition to representations at side view in normal conditions by using an architecture based to a U-Net.

1) Input data.

To train the GAN, first, it is necessary to arrange the data. Therefore, we set the GEIs at all the viewpoints in normal walking, carrying a bag, and wearing coat sequences as the source information and the GEIs of normal walking at 90° (side view) as the target data. After this step, 40M pairs of source-target representations were collected to train the GAN.

2) Generative and discriminator models

GANs are models that learn to map from a vector z to an image y, i.e. $G: z \rightarrow y$. In contrast, conditional GANs (cGANs) learn to map from an image x and a vector z to an image y, $G: \{x, z\} \rightarrow y$. These GANs are implemented by a system of two neural networks competing against each other, a generative and a discriminative network. The generative model G is trained to produce outputs that cannot be distinguished from "real" images by a discriminator D, which is trained to do as well as possible at detecting the generator's "fakes" outputs [24].

We adapted our generator and discriminator architectures from those presented in [27]. Both generator and discriminator use modules of the form convolution-BatchNorm-ReLu. However, unlike from many previous solutions which use an encoder-decoder network as generators, we added skip connections between each layer i and layer $n - i$, where n is the total number of layers. These connections basically concatenate the activations from the encoder to the decoder, which duplicates the number of channels in the decoder. This action helped us to relieve the bottleneck problem and increased the recognition accuracy. The structure of the generator is shown in Fig. 4.

As it has been stated before, the discriminator uses similar modules as the generator. However, in contrast to [13], [24], [25], which use a single output unit to predict whether or not the input is real or not, we decided to use a slightly different architecture called PatchGAN [27], see Fig. 5. What is useful about this approach is that it penalizes only certain parts of the images, i.e. it predicts if a certain each patch in an image is real or fake. This helps us to focus on certain regions of the GEI corresponding to the most robust parts against appearance variations, such as the head and feed, which contain the most important features [19], [28].

3) Objective

The objective of our conditional GAN can be expressed as:

$$\mathcal{L}_{cGAN}(G, D) = E_{x,y}[\log D(x, y)] + E_{x,z}\left[\log\left(1 - D\left(x, G(x, z)\right)\right)\right] \quad (1)$$

where the generator G tries to minimize this function. In contrast, the discriminator D tries to maximize it. Furthermore, to generate an output similar to the ground truth, previous works have found beneficial to mix the previous loss with more traditional loss functions. Therefore, we decided to use L1 distance instead of L2 since the latter one induces blur in the output as it is demonstrated in [27] establishing:

$$\mathcal{L}_{L1}(G) = E_{x,z}[\|y - G(x, z)\|_1] \quad (2)$$

The final objective can be defined as:

$$G^* = arg \min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L1}(G) \quad (3)$$

where λ is regularizing hyperparameter. For example, when using $\lambda = 0$ the cGAN generates high defined outputs, but the classification accuracy decreases.

C. Feature extraction and selection

Once the generator has been trained, and it is able to produce view-invariant GEI representations, a feature extraction stage is required since GEI images are considered high-dimensional data. This step allows us not only select those characteristics that contain more information with minimum redundancy but also improve the performance of our classifier and speed up the computation [29]. For this stage, we chose PCA, which is a common preprocessing technique used to perform dimensionality reduction [30], [31].

Having extracted the most robust components with PCA, the selection of the number of components to feed the classifier was determined according to the number of samples. To determine the number of components to be used, we considered [3], where it is suggested to retain 2c components where c corresponds to the number of classes.

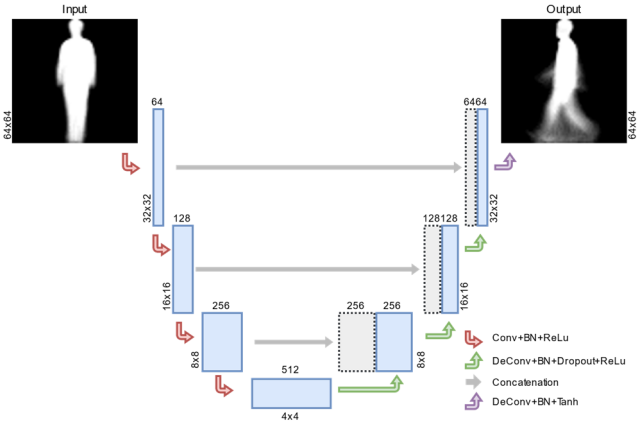

Fig. 4. View-invariant generator architecture.

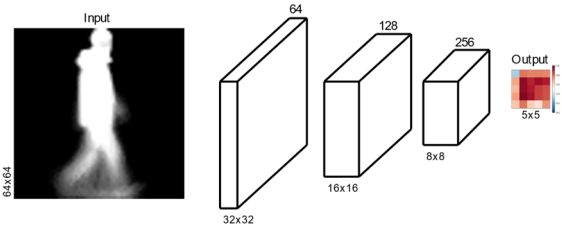

Fig. 5. View-invariant discriminator architecture.

D. Classification

The classification is performed to decide whether a subject belongs to a class in the database or not. Based on the bibliographic revision we have made, we noticed that simple classifiers are preferred due to the limited number of samples for training in CASIA-B [11]. However, unlike the majority of the reviewed works which use KNN for classification stage [4], we used LDA, that despite being a simple technique, is an effective strategy [16].

The discriminative approach has been adopted in several works, which showed its effectiveness and robustness. In fact, Rida et al. [8], [9] implemented a discriminant model as the classifier combined with PCA obtaining a good performance in side-view configuration. Moreover, the PCA - LDA strategy has shown to achieve the best data representation and the best class separability simultaneously [3].

The performance of our method has been measured by the correct classification rate (CCR) that corresponds to the ratio of the number of correctly classified samples over the total number of samples.

IV. Experiments and Results

In this section, we introduce the dataset used the experiments. Next, the experiment settings are described. Following, the architectures of the generator and discriminator are detailed. Finally, we compare the performance of the proposed method with state-of-the-art methods.

A. Dataset description

We used CASIA to evaluate our model. CASIA-B [32] is one of the largest datasets available for benchmarking gait recognition techniques, which has been collected by The Institute of Automation Chinese Academy of Sciences. It is an indoor gait dataset and comprises 124 subjects captured from 11 views. The view range goes from 0° to 180° with intervals of 18° between two nearest views. For each subject, there are 10 walking sequences consisting of 6 normal walkings ("nm"), 2 carrying-bag sequences ("bg"), and 2 wearing-coat sequences ("cl"). Fig. 6. shows GEIs from one subject in all the conditions at 11 views. Since our experiments mainly focus on view, clothing, and carrying condition variations in gait recognition, all the GEI images were resized to 64x64.

B. Experiments designs

To fairly compare the proposed method with the state-of-the-art approaches, we adopted the experimental setup proposed in [11], [13], [14]. Therefore, we divided the dataset into two subsets. The training set was composed of the six normal, the two carrying-bag, and the two wearing-coat sequences of the first 62 subjects. In the test stage, the remaining 62 subjects were employed. Consequently, the first four normal sequences denoted as ("nm1") were used as gallery set whereas the two left sequences ("nm2") along with the "bg" and "cl" sequences were used as prove set to test the variations on view, carrying, and clothing conditions as it is shown in Table I.

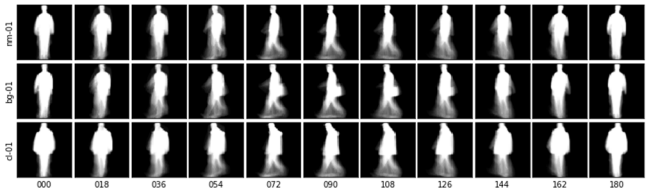

Fig. 6. Walking sequences at all the conditions at the 11 views from CASIA-B dataset.

TABLE I. Experimental Design.

Train set Subjects: 1-62		nm01-nm06, bg01, bg02, cl01, cl02
Test set Subjects: 63-124	Gallery set	nm01-nm04
	Prove "nm"	nm05, nm06
	Prove "bg"	bg01, bg02
	Prove "cl"	cl01, cl02

C. Model Parameters

As we stated before, modules of convolution-BatchNorm-ReLU were used to build both the generator and the discriminator. To denote these modules, `Ck` will represent a module composed of Convolution-BatchNorm-ReLU with `k` filters. Whereas `CDk` states for a module of Convolution-BatchNorm-Dropout-ReLU with `k` layers and a dropout of 0.5.

Since the architectures used here were adapted from those presented in [27], filters of 4x4 with a stride of 2 were used in all convolutions and deconvolutions layers. Furthermore, we implemented Leaky ReLUs in the encoder and discriminator layers with a slope of 0.2, whilst in the decoder network, simple ReLUs were used.

1) Generator architecture

The generator can be divided into two parts: an encoder and a decoder. In the encoder, the convolutions downsampled the feature maps by a factor of 2. In contrast, the decoder upsampled them by the same factor. These networks keep the next structure:

Encoder: C64-C128-C256-C512.

Decoder: C512- C256- C128.

Since we are using a U-Net architecture, skip connections concatenate activations from layer *i* to layer *n-i*, so the number of channels in the decoder duplicated. Different from the rest, the BatchNorm is not used for the first layer in the encoder. Additionally, after the last layer in the decoder, a convolution followed by a Tanh function is applied to match the number of the output's channel.

2) Discriminador architecture

The discriminator, which is trained to determinate if an area of the generated output is "real" or not, follows the next structure:

Discriminator: C64-C128-C256.

As an exception to the above notation, after the last layer, a convolution is applied to map to a 1-dimensional output, followed by a Sigmoid activation function. Besides, like the encoder, BatchNorm is not applied to the first layer. Additionally, the filters of the last layer were of the size 4x4 with a stride of 1.

D. Training detaills

To train both the generator and the discriminator, Adam optimized was employed, with a learning rate of 0.0002 and momentum parameters of $\beta 1 = 0.5$, $\beta 2 = 0.999$. We noticed that good performance was achieved after 20 epochs of training when using $\lambda = 100$.

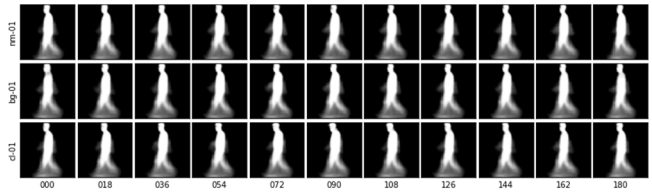

Fig. 7. Generated walking sequences at all the conditions from CASIA-B dataset.

TABLE II. RECOGNITION RATES WHEN THE PROBE DATA IS NORMAL WALKING SEQUENCES.

		Prove set view (nm05, nm06)										
		0	18	36	54	72	90	108	126	144	162	180
Gallery set view	0	96.77	64.52	50.00	42.74	29.03	22.58	23.39	27.42	37.90	46.77	73.39
	18	76.61	98.39	91.94	70.97	48.39	37.90	41.94	46.77	51.61	54.03	51.61
	36	49.19	90.32	97.58	92.74	72.58	58.87	59.68	66.94	67.74	54.03	34.68
	54	37.10	68.55	91.94	95.97	92.74	85.48	84.68	79.84	72.58	46.77	26.61
	72	24.19	38.71	71.77	91.13	98.39	96.77	95.16	85.48	64.52	37.90	20.97
	90	21.77	29.84	53.23	79.03	96.77	98.39	98.39	83.06	53.23	33.87	19.35
	108	24.19	40.32	62.90	83.06	94.35	95.97	98.39	93.55	75.81	48.39	22.58
	126	19.35	51.61	59.68	75.81	82.26	80.65	91.94	96.77	95.97	54.84	26.61
	144	34.68	54.84	64.52	67.74	67.74	58.87	76.61	96.77	99.19	79.84	37.90
	162	52.42	64.52	60.48	48.39	35.48	38.71	40.32	59.68	86.29	99.19	73.39
	180	69.35	44.35	30.65	20.16	18.55	17.74	16.13	23.39	33.06	73.39	99.19

TABLE III. RECOGNITION RATES WHEN THE PROBE DATA IS CARRYING A BAG DATA.

		Prove set view (bg01, bg02)										
		0	18	36	54	72	90	108	126	144	162	180
Gallery set view	0	83.06	53.23	35.48	24.19	17.74	20.97	16.94	18.55	24.19	33.06	50.81
	18	57.26	87.90	74.19	47.58	38.71	30.65	32.26	37.90	45.16	43.55	35.48
	36	37.90	78.23	83.06	70.16	54.84	39.52	36.29	42.74	51.61	41.94	25.00
	54	29.03	54.84	79.84	84.68	79.03	58.87	56.45	54.84	56.45	37.90	24.19
	72	20.97	33.06	54.03	74.19	93.55	81.45	79.03	69.35	44.35	26.61	20.97
	90	17.74	33.06	49.19	59.68	82.26	83.06	79.84	62.90	35.48	25.00	20.97
	108	14.52	32.26	48.39	59.68	79.03	78.23	86.29	79.03	67.74	29.84	14.52
	126	23.39	30.65	43.55	54.84	62.90	58.06	74.19	81.45	76.61	40.32	22.58
	144	29.03	39.52	45.97	41.13	48.39	45.97	53.23	75.00	84.68	54.84	34.68
	162	37.10	39.52	35.48	28.23	26.61	29.84	26.61	38.71	57.26	81.45	50.00
	180	51.61	28.23	25.00	20.16	12.10	9.68	4.84	12.90	26.61	40.32	87.90

TABLE IV. RECOGNITION RATES WHEN THE PROBE DATA IS COAT WEARING DATA.

		Prove set view (cl01, cl02)										
		0	18	36	54	72	90	108	126	144	162	180
Gallery set view	0	41.13	25.81	18.55	16.13	12.10	11.29	8.87	15.32	18.55	20.97	25.00
	18	25.00	45.16	46.77	27.42	20.16	19.35	16.94	18.55	25.81	21.77	17.74
	36	23.39	41.13	58.06	51.61	28.23	20.16	19.35	31.45	29.03	27.42	16.94
	54	17.74	29.03	48.39	58.06	46.77	37.90	42.74	36.29	31.45	16.94	11.29
	72	21.77	26.61	35.48	46.77	63.71	50.81	47.58	47.58	32.26	16.13	8.06
	90	18.55	25.00	30.65	41.13	59.68	53.23	50.00	46.77	31.45	20.16	9.68
	108	17.74	28.23	34.68	45.16	58.87	49.19	60.48	58.06	39.52	25.00	7.26
	126	9.68	24.19	29.84	33.06	45.97	40.32	45.16	60.48	53.23	29.84	12.10
	144	13.71	23.39	28.23	36.29	29.84	24.19	37.10	51.61	55.65	39.52	15.32
	162	15.32	19.35	29.03	20.16	18.55	11.29	16.13	27.42	35.48	41.94	24.19
	180	14.52	10.48	8.87	8.06	2.42	6.45	6.45	12.10	15.32	26.61	40.32

All networks were trained from scratch, so the weights were initialized from a Gaussian distribution with a mean 0 and a standard deviation of 0.02. However, since the number of sequences per subject is limited, data augmentation was necessary. For this purpose, the random jitter technique was applied before the training stage. All the GEI representations have been resized from 64x64 to 67x67 and then randomly cropped back to their original size.

The different configurations of our model were implemented in Python using Keras and TensorFlow framework. The training and test stages were performed using the Google Colab tool. The virtual machine used in the experiments had a 2.3 GHz dual-core processor, 25 GB of RAM, and an NVIDIA Tesla K80 graphic card with 12 GB of memory, and 2496 CUDA cores.

E. Results and analysis

In order to demonstrate the performance of our method, first, we present the generated GEIs in Fig. 7. As can be seen, the generator effectively transforms representations from any view and condition to GEIs in normal conditions at side view. Indeed, the generator is capable to keep the subject's identity information (posture) as seen in the last row of both images.

To evaluate the performance of our model, three covariates have been considered: view, clothing, and carrying variations. The performance of our model under these conditions is shown in Tables II-IV. For each table, each row corresponds to a view angle of the gallery set, whereas each column corresponds to the view angle of the probe set. Since CASIA-B contains walking sequences of 11 views, there are 121 pairs of combinations.

Table II shows the CCR of our model when "nm1" sequences at a specific view are put into the gallery set, and the "nm2" at another view are put into the prove set. For the results in Table III, "nm1" sequences are put into the gallery whereas the probe set contains images of people carrying bags "bg". Similar configuration is used for Table IV, yet in this case sequences wearing a coat "cl" are put into the probe set.

It can be seen from Tables II-IV that the performance of our model achieves outstanding results when the gallery and prove set correspond to a similar view angle. Therefore, we can state that the framework proposed is quite effective when there is not a large view difference.

F. *Comparison with the state-of-the-art*

To fairly illustrate the performance of the proposed method, we also compare our method against the reported by other state-of-the-art methods when the gallery images are different from those in the prove set, and when they are the same. we compare our method with GEI+PCA [3], SPAE [14], GaitGANv1 [13], and GaitGANv2 [24]. We have chosen these works since we use a similar experiment configuration in terms of gallery and probe set.

First, we compared the recognition rates without view changes. To compute this, we averaged the CCRs on the main diagonal of Tables II-IV, which correspond to the recognition rates without view changes. The corresponding average rates of GEI+PCA, SPAE, GaitGANv1, and GaitGANv2 were obtained from their respective works. The comparison is shown in Fig. 8.

We carried out a similar experiment for the results shown in Fig. 9. However, this time, to compute the CCR, we averaged all the results of Tables II-IV excepting those of the main diagonal. As shown in the figures above, the CCR of our proposed method is slightly lower in the normal condition but considerably higher in the carrying-bag and waring-coat conditions. In real life, people wear different clothes depending on days (cool or warm days) and the season (winter or summer); therefore, clothing variants are always present, being this an important issue [12]. The proposed representation performs very well in view, carrying, and clothing variation problem since it outperforms other approaches.

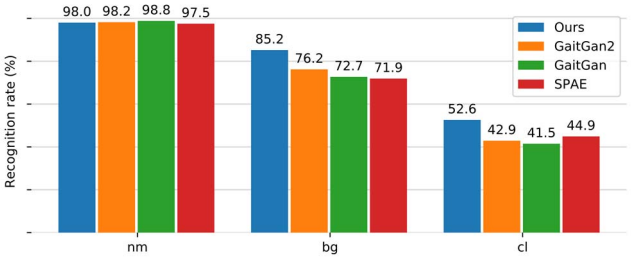

Fig. 8. Comparison of CCR with some state-of-the-art methods without view variations at three different conditions.

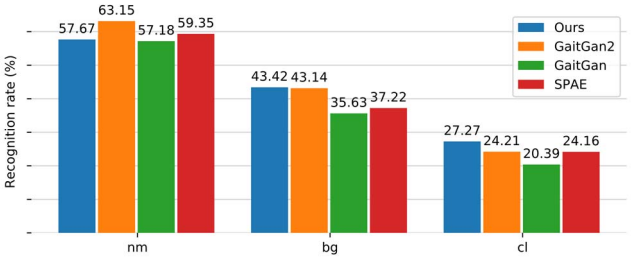

Fig. 9. Comparison of CCR with some state-of-the-art methods without taking account the main diagonal at three different conditions.

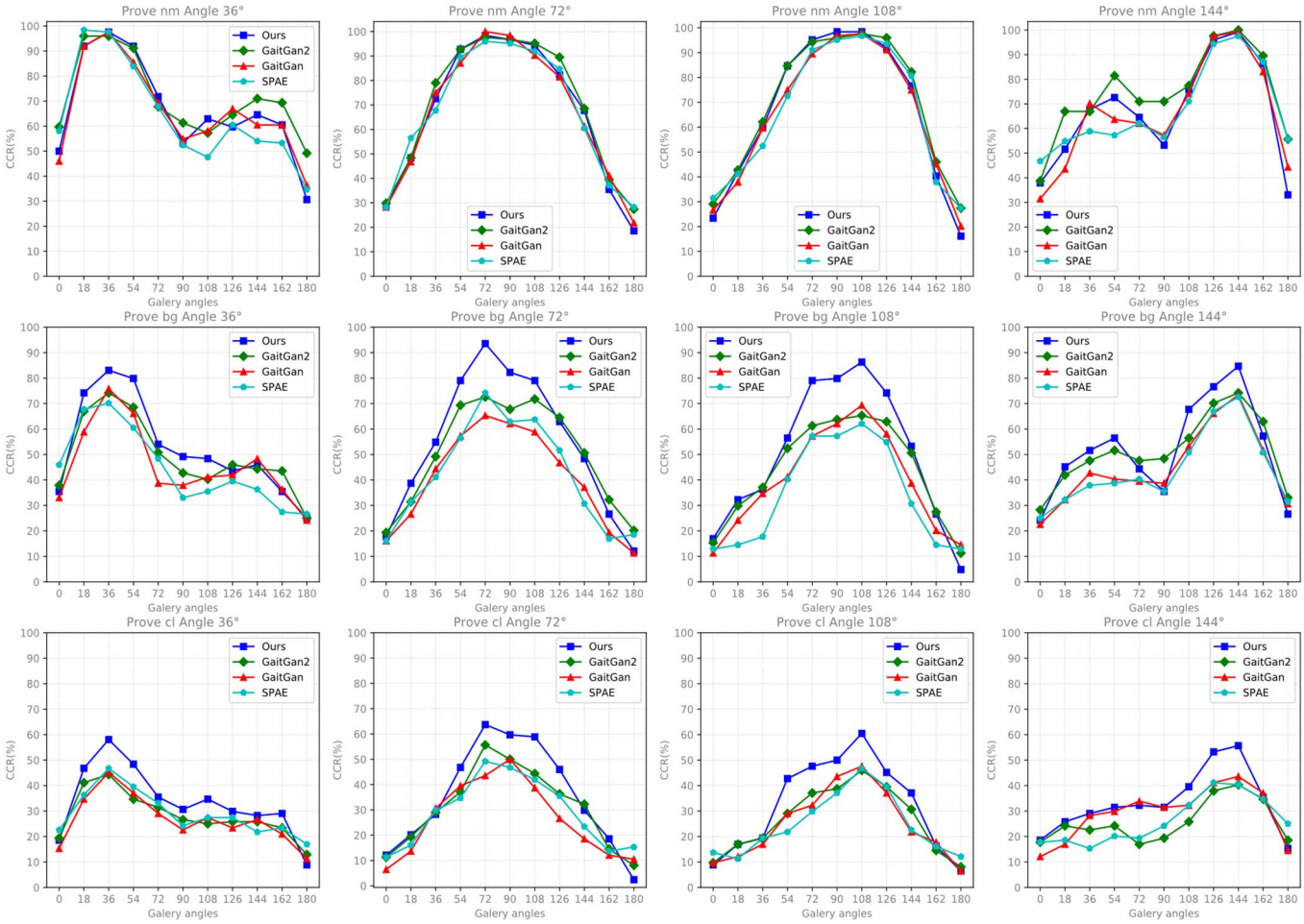

Fig. 10. Comparison of CCR with SPAE, GaitGANv1, and GaitGANv2 at different probe angles. Rows represent different walking conditions and columns represents a probe angle.

Finally, we compared the techniques at different viewpoints. To show the performance of our method, we plot the CCR when the prove set contented regenerated images from 36°, 72°, 108°, and 144°. The techniques are compared in Fig. 10 based on their CCR. The first row of the figure corresponds to the recognition rates at different probe angles in normal walking sequences. The second row and third show the accuracy in the carrying-bag and wearing-coat sequences. Through results presented in Fig. 10, it is possible to acknowledge the performance similarities in normal conditions with relation to the other approaches; however, it outperforms them when the gallery set contains sequences with appearance changes, which is an attractive aspect for real-time and on-field intelligent surveillance systems.

V. Conclusion

In this paper, we propose to use a modified GAN based on a U-Net architecture to overcome appearance variations due to changes of clothing, carrying conditions, and view angle. Since gait recognition is a complex task where the amount of data is limited, we propose to use a combined approach between generative and discriminative models. First, our generative model is trained to convert images from any condition and any viewpoint to images in normal conditions at side view. Then, the generated images are classified by using LDA. Through experiments, we have seen that this configuration improved the gait recognition accuracy.

The proposed framework achieves a superior performance among the reviewed approaches which focus on the same issue, especially in the case of carrying-bag and clothing-coat variations. This feature makes our approach suitable for practical applications such as intelligent surveillance systems.

In the future, we will extend this model to deal more challenging covariates, such as time variations, prove the model in other benchmarking large databases since the use of more subjects is important to have reliable results, and improve the accuracy in the case of large view changes between the gallery and prove set. Furthermore, we are going to implement more complex and powerful models to deal with cross-view recognition.

References

[1] P. Karampelas and T. Bourlai, *Surveillance in Action*. 2018.

[2] 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 Secur.*, vol. 7, no. 5, pp. 1511–1521, 2012.

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

[4] M. Tariq and M. A. Shah, "Review of Model-Free Gait Recognition in Biometric Systems," 2017.

[5] G. V. Veres, L. Gordon, J. N. Carter, and M. S. Nixon, "What image information is important in silhouette-based gait recognition?," *Proc. 2004 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition, 2004. CVPR 2004.*, vol. 2, pp. 776–782, 2004.

[6] D. Tao, X. Li, X. Wu, and S. J. Maybank, "General tensor discriminant analysis and Gabor features for gait recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1700–1715, 2007.

[7] I. Rida, S. Almaadeed, and A. Bouridane, "Gait recognition based on modified phase-only correlation," *Signal, Image Video Process.*, vol. 10, no. 3, pp. 463–470, 2016.

[8] I. Rida, X. Jiang, and G. L. Marcialis, "Human body part selection by group lasso of motion for model-free gait recognition," *IEEE Signal Process. Lett.*, vol. 23, no. 1, pp. 154–158, 2016.

[9] I. Rida, S. Al-Maadeed, and A. Bouridane, "Unsupervised Feature Selection Method for Improved Human Gait Recognition," pp. 1133–1137, 2015.

[10] I. Rida, L. Boubchir, N. Al-Maadeed, S. Al-Maadeed, and A. Bouridane, "Robust model-free gait recognition by statistical dependency feature selection and Globality-Locality Preserving Projections," *2016 39th Int. Conf. Telecommun. Signal Process. TSP 2016*, pp. 652–655, 2016.

[11] S. Yu, H. Chen, Q. Wang, L. Shen, and Y. Huang, "Invariant feature extraction for gait recognition using only one uniform model," *Neurocomputing*, vol. 239, pp. 81–93, 2017.

[12] D. S. Matovski, M. S. Nixon, S. Mahmoodi, and J. N. Carter, "The effect of time on gait recognition performance," *IEEE Trans. Inf. Forensics Secur.*, vol. 7, no. 2, pp. 543–552, 2012.

[13] S. Yu, H. Chen, S. Engineering, E. B. Garc, N. Poh, and U. Kingdom, "GaitGAN : Invariant Gait Feature Extraction Using Generative Adversarial Networks," *Cvpr*, 2017.

[14] S. Yu, Q. Wang, L. Shen, and Y. Huang, "View invariant gait recognition using only one uniform model," *Proc. - Int. Conf. Pattern Recognit.*, pp. 889–894, 2017.

[15] B. Khalid, T. Xiang, and S. Gong, "Gait recognition using gait entropy image," *3rd Int. Conf. Imaging Crime Detect. Prev. (ICDP 2009)*, pp. P2–P2, 2009.

[16] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "GEINet: View-invariant gait recognition using a convolutional neural network," *2016 Int. Conf. Biometrics, ICB 2016*, 2016.

[17] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer, "The humanID gait challenge problem: Data sets, performance, and analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 162–177, 2005.

[18] Y. Zhang, Y. Huang, L. Wang, and S. Yu, "A comprehensive study on gait biometrics using a joint CNN-based method," *Pattern Recognit.*, vol. 93, pp. 228–236, Sep. 2019.

[19] Y. Dupuis, X. Savatier, and P. Vasseur, "Feature subset selection applied to model-free gait recognition," *Image Vis. Comput.*, vol. 31, no. 8, pp. 580–591, 2013.

[20] J.-H. Yoo, D. Hwang, K.-Y. Moon, and M. S. Nixon, "Automated Human Recognition by Gait using Neural Network," *1st Work. Image Process. Theory, Tools Appl. (IPTA 2008)*, pp. 1–6, 2008.

[21] M. Alotaibi and A. Mahmood, "Automatic Real Time Gait Recognition based on Spatiotemporal Templates," 2015.

[22] M. Alotaibi and A. Mahmood, "Improved Gait recognition based on specialized deep convolutional neural networks," *2015 IEEE Appl. Imag. Pattern Recognit. Work.*, pp. 1–7, 2015.

[23] T. Yeoh, A. E. Hernan, and K. Tanaka, "Clothing-invariant Gait Recognition Using Convolutional Neural Network," *2016 Int. Symp. Intell. Signal Process. Commun. Syst.*, pp. 1–5, 2016.

[24] S. Yu *et al.*, "GaitGANv2: Invariant gait feature extraction using generative adversarial networks," *Pattern Recognit.*, vol. 87, pp. 179–189, 2019.

[25] R. Zhang, D. Yin, Z. Zhou, Z. Cao, F. Meng, and B. Hu, "Improving Cross-View Gait Recognition With Generative Adversarial Networks Rui," 2019, no. Npsc, pp. 43–47.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation."

[27] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 5967–5976, 2017.

[28] I. R. Tinini Alvarez and G. Sahonero-Alvarez, "Gait Recognition Based on Modified Gait Energy Image," in *2018 IEEE Sciences and Humanities International Research Conference (SHIRCON)*, 2018, vol. 1, pp. 1–4.

[29] X. Shi, Z. Guo, F. Nie, L. Yang, J. You, and D. Tao, "Two-Dimensional Whitening Reconstruction for Enhancing Robustness of Principal Component Analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2130–2136, 2016.

[30] T. Chau, "A review of analytical techniques for gait data. Part 1: Fuzzy, statistical and fractal methods," *Gait Posture*, vol. 13, no. 1, pp. 49–66, 2001.

[31] T. Chau, "A review of analytical techniques for gait data. Part 2: neural network and wavelet methods," *Gait Posture*, vol. 13, no. 2, pp. 102–120, 2001.

[32] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," *Proc. - Int. Conf. Pattern Recognit.*, vol. 4, pp. 441–444, 2006.