Gait Recognition from a Single Image using a Phase-Aware Gait Cycle Reconstruction Network

Chi Xu[1,2], Yasushi Makihara[2], Xiang Li[1,2], Yasushi Yagi[2], and Jianfeng Lu[1]

[1] Nanjing University of Science and Technology, Nanjing 210094, China
[2] ISIR, Osaka University, Osaka 567-0047, Japan
{xu,makihara,li,yagi}@am.sanken.osaka-u.ac.jp, lujf@njust.edu.cn

Abstract. We propose a method of gait recognition just from a single image for the first time, which enables latency-free gait recognition. To mitigate large intra-subject variations caused by a phase (gait pose) difference between a matching pair of input single images, we first reconstruct full gait cycles of image sequences from the single images using an auto-encoder framework, and then feed them into a state-of-the-art gait recognition network for matching. Specifically, a phase estimation network is introduced for the input single image, and the gait cycle reconstruction network exploits the estimated phase to mitigate the dependence of an encoded feature on the phase of that single image. This is called phase-aware gait cycle reconstructor (PA-GCR). In the training phase, the PA-GCR and recognition network are simultaneously optimized to achieve a good trade-off between reconstruction and recognition accuracies. Experiments on three gait datasets demonstrate the significant performance improvement of this method.

Keywords: Gait cycle reconstruction, gait recognition, single image

1 Introduction

Gait is a common biometric modality used to identify a person. Gait has unique advantages compared with other biometrics such as DNA, fingerprints, the iris, and the face. For example, it can be authenticated at a long distance even without subject cooperation. Gait recognition has therefore received great attention for applications such as surveillance, forensics, and criminal investigation with CCTV footage [6, 19, 25].

Extensive studies on gait recognition have mainly used a silhouette sequence itself [41, 40, 7] or gait features extracted from a gait cycle of silhouette sequences [47, 37, 42, 38, 45, 48, 14]. However, capturing a video containing a certain time length or a full gait cycle usually requires waiting for some time (e.g., about one second for a full gait cycle), which is undesirable for real-time online applications. An extreme way to reduce latency in capturing is to try identifying a subject using just a single image, which has not been especially targeted in prior work to our knowledge. Besides latency, gait recognition from a single

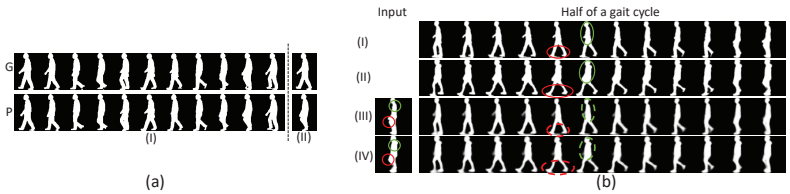

Fig. 1. Examples of different scenarios of gait recognition. (a) A matching pair (G and P) from the same subject for different problem settings. Left (I): gait recognition using a full gait cycle; right (II): gait recognition from a single image. There are significant differences between the pair of single images owing to the phase difference (i.e., double vs. single support). (b) A matching pair from different subject. (I) and (II) are the real silhouette images of a half gait cycles (i.e., ground truth). (III) and (IV) are the corresponding reconstructed half cycles from the single input image using our method. Clear motion differences (e.g., stoop by green circles and stride by red circles) are observed between real cycles (I) and (II), and also the reconstructions (III) and (IV), which means our method can successfully reconstruct the individual gait motion patterns to some extent.

image is also applicable to a case of temporal partial occlusion, which is another challenging factor in the real world. For example, in crowded scenes, a subject may be heavily occluded for most of the frames and those frames are useless, whereas the single-image gait recognition can still work once a single frame without occlusion is obtained.

Gait recognition from a single image, however, is quite challenging because gait phase differences (e.g., single vs. double support) introduce great intra-subject variations, as shown in Fig. 1(a). This largely degrades the performances of existing gait recognition methods such as the state-of-the-art network GaitSet [7] and conventional techniques using gait features such as a gait energy image (GEI) [13] (i.e., averaged silhouette over a full gait cycle).

On the other hand, a snapshot in an action video has proven to imply dynamic information that can predict past/future motions (i.e., implied motion) [21], which has been applied to video synthesis and action recognition [9, 33]. Similarly, a single gait image captured from a gait sequence, also intimates pose sequences before or after the frame while keeping the individuality of his/her gait thanks to temporally continuous variables (e.g., knee joint and back bending angles) [16], and hence provides the possibility of gait recognition from a single image. For example, a subject bending his/her back in the single-support is likely to bend his/her back in the double-support too (see Fig. 1(b)(I)), and a single-support phase with a greater knee flexion probably results in a double-support with a larger stride (see Fig. 1(b)(II)). Motivated by these facts, we tackle single-image gait recognition by first reconstructing a gait cycle of a silhouette sequence from the single image, which contains all the phases, before exploiting the subsequent matcher.

Recovering a silhouette sequence of a full gait cycle or a gait feature to be extracted from it, actually, is also often done in gait recognition from videos with

low frame-rates [1, 2, 4]. However, most of these methods does not work well for a very limited number of input frames (i.e., a single frame), and moreover, these approaches optimize only gait cycle reconstruction quality, and hence cannot guarantee the optimal recognition accuracy essentially.

We therefore propose a unified framework of a phase-aware gait cycle reconstruction network (PA-GCRNet) for gait recognition from a single image. This consists of a phase-aware gait cycle reconstructor (PA-GCR) module and a subsequent recognition network. Instead of simply minimizing the gait cycle reconstruction error, the proposed PA-GCRNet learns an appropriate gait cycle reconstruction, where the reconstruction quality is well maintained while ensuring optimal recognition performance simultaneously. The contributions of this work are threefold:

1. The first work aiming at gait recognition from a single image.

To our knowledge, this is the first work specially aimed at gait recognition from a single image. Compared with most existing gait recognition studies, which require acquisition of a gait video containing a certain time length (e.g., a gait cycle), single-image gait recognition is more suitable for real-time online applications because the result can be obtained once a single image is captured (i.e., without latency).

2. Gait cycle reconstruction from an arbitrary input phase.

While most existing works focus on generating future action frames from an initial frame [9, 33], the proposed PA-GCR can reconstruct a gait cycle including future and past frames from an arbitrary input phase. To reduce the intra-subject variations in the reconstructed gait cycles caused by input phase difference, an phase estimation network is incorporated to mitigate the dependence of an encoded feature on the input phase. The proposed PA-GCR is further combined with the state-of-the-art sequence-based gait recognition network GaitSet [7], in an end-to-end training manner to achieve a good trade-off between gait cycle reconstruction performance and recognition accuracy, unlike traditional low frame-rate gait recognition methods that just focus on reconstruction quality instead of recognition performance.

3. State-of-the-art performance on three publicly available datasets.

The proposed method was evaluated on three publicly available gait datasets: the OU-ISIR Gait Database, Multi-View Large Population Dataset (OU-MVLP) [39], CASIA Gait Database, Dataset B (CASIA-B) [46], and OU-ISIR Treadmill Dataset D (OUTD-D) [26]. The proposed method yields significantly improved recognition performance on all three datasets compared with pure sequence-based GaitSet [7] and other state-of-the-art approaches to low frame-rate gait recognition.

2 Related Work

2.1 Gait recognition from low frame-rate videos

Temporal interpolation and super-resolution-based approaches. Temporal interpolation and super-resolution-based approaches were developed to

increase the number of frames to cope with the low frame-rate. Al-Huseiny et al. [3] proposed level-set morphing for temporal interpolation, and Prismall et al. [34] used linear interpolation for moment descriptors. These are, however, not applicable for very low frame-rates.

Makihara et al. [27] proposed reconstructing a gait period with a high frame-rate using phase registration data among multiple periods from a sequence with a low frame-rate and a manifold expressing a periodic temporal super-resolution (TSR) image sequence via energy minimization. Akae et al. [1] later used an exemplar of a gait image sequence with a high frame-rate to overcome the wagon wheel effect in [27]. A unified example-based and reconstruction-based periodic TSR was proposed in [2] to further solve the stroboscopic problem with good reconstruction even when the sequence has such a low frame-rate as to appear nearly still.

However, these methods only ensure optimal reconstruction quality rather than recognition accuracy, which is the main goal of gait recognition.

Metric learning-based approach. Unlike temporal interpolation and super-resolution-based approaches, metric learning-based approach directly applies a metric learning technique to videos with low frame-rates. Guan et al. [12] first extracted a gait feature just by averaging a sequence with a low frame-rate and applied a random subspace method (RSM) to reduce the generalization errors caused by the low frame-rate. However, this does not work well for extremely low frame-rates (e.g., 1 fps) because it is difficult to find robust subspaces with good generalizations for very few input frames.

Direct gait feature reconstruction-based approach. Recently, a direct gait feature reconstruction-based approach [4] was proposed that takes the average of a low frame-rate sequence as an input feature (e.g., an incomplete GEI) and reconstructs a GEI to be progressively extracted from a full gait cycle using a fully convolutional neural network. Similarly to traditional temporal interpolation and super-resolution-based approaches, this network only optimizes the reconstruction performance of the GEI, and still works poorly for very low frame-rates.

2.2 Gait representation

Most studies on gait recognition cope with normal frame-rates under various covariates, such as view [28, 42], walking speed [11, 43], clothing [17, 10], and carried objects [29, 23]. Rather than directly using silhouette image sequences, most traditional methods exploit feature templates (e.g., GEI and frequency-domain features [28]) extracted from the full gait cycle of a silhouette sequence for further processing such as feature transformation [28, 22] and spatial metric learning [11, 29, 5]. Recently, gait recognition performance has been greatly improved by introducing the convolutional neural network (CNN) framework, where the GEI is mainly used as the network input [47, 37, 42, 38, 45, 48]. Additionally, a few CNN-based approaches directly handle silhouette sequences [41, 40, 7] to use more temporal information than a single GEI template. Among them, GaitSet [7] achieves remarkable state-of-the-art recognition performance; hence, we combine

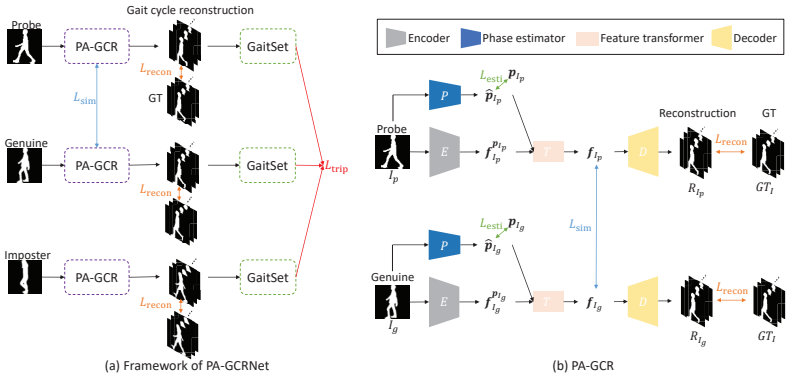

Fig. 2. Overview of the proposed method. GT denotes the ground truth. (a) shows the whole framework of the PA-GCRNet, which contains a PA-GCR for full gait cycle reconstruction and GaitSet [7] as the subsequent recognition network for discrimination learning. A triplet of samples (probe, genuine, and imposter) is fed into the PA-GCRNet in the training stage, where the network parameters are shared among each stream. (b) illustrates the details of the PA-GCR module, which consists of four components: an encoder, a phase estimator, a feature transformer, and a decoder. The imposter is omitted here because the similarity loss L_{sim} is only defined for the genuine pair.

the proposed PA-GCR with GaitSet for further feature discrimination learning from the reconstructed full gait cycles.

3 Gait Recognition using PA-GCRNet

3.1 Overview

An overview of the proposed PA-GCRNet is shown in Fig. 2(a). Given a captured gait image, a silhouette can be first extracted using graph-cut segmentation based on background subtraction [30], or recent state-of-the-art semantic segmentation methods such as RefineNet [24] based on deep learning. A normalized silhouette is then obtained via height normalization and registration using the center of gravity [28], after which it is used as an input for the proposed method.

The proposed PA-GCRNet consists of two parts: PA-GCR and the recognition network (GaitSet in our implementation). Similarly to GaitSet [7], the proposed PA-GCRNet takes a triplet of inputs in the training stage, where the network parameters are shared among the three inputs. The PA-GCR tries fully reconstructing a gait cycle of a silhouette sequence that contains a fixed number of frames with corresponding phases (e.g., left-leg-forward double-support in the first frame) by considering the phase of a single input silhouette. The reconstructed full gait cycle of silhouettes is then fed into the subsequent recognition network to learn more discriminative features for gait recognition. In the testing stage, the dissimilarity between a matching pair is computed as the L2 distance between the discriminative features learnt by the recognition network.

3.2 PA-GCR

The proposed PA-GCR has four components: an encoder, a phase estimator, a feature transformer, and a decoder, as shown in Fig. 2(b). One potential issue is that the phase difference among input silhouettes may affect gait cycle reconstruction results. For example, a double-support silhouette reconstructed from a single-support input silhouette may be different from that reconstructed from a double-support silhouette, because the poses in dynamic parts such as legs and arms in the double-support phase are not observed in the single-support phase. There may be large intra-subject variations in the reconstructed gait cycle of silhouettes if we directly use encoded features from input silhouettes with various phases (kinds of phase-dependent encoded features). We therefore introduce the feature transformer to transform the phase-dependent encoded features into more phase-independent features by taking the estimated phase information into account for the following decoder. This is more advantageous because it reduces intra-subject variations in the reconstructed gait cycle of silhouettes.

Phase representation Considering the periodicity of human gait, it is necessary to represent the phase (gait stance) using a periodically continuous label. The phase can be defined by a cyclic angle representation with the domain $[0, 2\pi)$ similarly to a general periodic variable. The cyclic angle representation, however, is discontinuous from 2π to 0; hence we use a redundant two-dimensional vector representation without discontinuity consisting of sine and cosine functions. Assuming that a gait cycle has T frames and that the phase evolves linearly over the frames. The phase vector $\boldsymbol{p}_t \in \mathbb{R}^2$ at the t-th frame is expressed as

$$\boldsymbol{p}_t = [\cos\theta_t, \sin\theta_t]^T, \tag{1}$$

where $\theta_t = \theta_0 + 2\pi\frac{t}{T}$, and θ_0 is a phase shift.

Gait cycle of a silhouette sequence for training We need full gait cycles of silhouette sequences from multiple training subjects (i.e., ground truth) to train the proposed network, and the cycles should be phase-synchronized among the training subjects to mitigate the impact of phase inconsistency on reconstruction performance in the training data. However, the real gait cycle needs to be first interpolated into the common gait cycle (e.g., 100 frames per cycle) because the number of frames of a real cycle might be different among the training subjects (e.g., 25 frames per cycle for subject A, 32 frames per cycle for subject B). We therefore apply a geometric transformation based on free-form deformation (FFD) [36] to interpolate intermediate frames between original frames and to generate the silhouette sequence with the common gait cycle because the FFD is suitable for expressing the transformation of a non-rigid human body and helps preserve gait individuality in the transformed image [8, 44].

After obtaining the silhouette sequences with the common gait cycle, we synchronize them among the training subjects using a baseline algorithm [35].

More specifically, we first choose a subject as the standard, and then compute the sum of silhouette differences over the common gait cycle between another subject and this standard. We compute the sum for each shift amount of the starting frame, and we adopt the silhouette sequence with the shift amount that minimizes the summed difference as the training data for the reconstructor.

Networks ***1) Encoder.*** The encoder E first extracts a low-dimensional feature from the input silhouette, which somewhat depends on the phase of the input silhouette. Given the input silhouette I, the obtained low-dimensional feature from the encoder is denoted as

$$\boldsymbol{f}_I^{\boldsymbol{p}_I} = E(I), \tag{2}$$

where $\boldsymbol{p}_I$ is the phase of input I.

The encoder is designed as a CNN with an input size of $1 \times 64 \times 64$. Four convolutional layers are used with a filter size of 4×4 and stride of two, and the number of filters is increased from 64 to 512 in successive doubling steps. We apply a batch-normalization layer [18] and the rectified linear unit (ReLU) activation function [31] after each convolutional layer. Finally, a 100-dimensional feature is obtained through a fully connected layer.

2) Phase estimator. A phase estimator P is used to estimate the phase label of the input silhouette to make the phase-dependent encoded feature more phase-independent in the next step. The phase estimator is represented as

$$\hat{\boldsymbol{p}}_I = P(I) \in \mathbb{R}^2. \tag{3}$$

The phase estimator has a structure similar to that of the encoder, but with one more fully connected layer to regress the 2D phase label. A normalization layer is used to ensure that $\|\hat{\boldsymbol{p}}_I\|_2 = 1$, which is a characteristic of the sine and cosine functions. The output phase label is compared with the ground truth label $\boldsymbol{p}_I$ to compute an estimation loss as

$$L_{\text{esti}} = \|\hat{\boldsymbol{p}}_I - \boldsymbol{p}_I\|_2^2. \tag{4}$$

3) Feature transformer. A feature transformer T is inserted between the encoder and the decoder to reduce the reconstruction difference caused by the input phase difference for the same subject. The feature transformer transforms the phase-dependent encoded feature $\boldsymbol{f}_I^{\boldsymbol{p}_I}$ into the phase-independent feature $\boldsymbol{f}_I$, which is formulated as

$$\boldsymbol{f}_I = T(\text{cat}(\boldsymbol{f}_{\text{I}}^{\boldsymbol{p}_{\text{I}}}, \hat{\boldsymbol{p}}_{\text{I}})), \tag{5}$$

where cat indicates a concatenation.

We implement the feature transformation using a fully connected layer to obtain the transformed 100D feature $\boldsymbol{f}_I \in \mathbb{R}^{100}$ from the 102D concatenated vector of the encoded feature $\boldsymbol{f}_I^{\boldsymbol{p}_I} \in \mathbb{R}^{100}$ and the estimated phase $\hat{\boldsymbol{p}}_I \in \mathbb{R}^2$. We expect the transformed feature $\boldsymbol{f}_I$ to be more independent of the input phase than the encoded feature $\boldsymbol{f}_I^{\boldsymbol{p}_I}$, i.e., those for the same subjects are more similar

to each other among different phases of the input silhouettes. Therefore, we minimize the similarity loss for the phase-independent feature $\boldsymbol{f}_I$ via

$$L_{\text{sim}} = \|\boldsymbol{f}_{I_p} - \boldsymbol{f}_{I_g}\|_2^2, \tag{6}$$

where I_p and I_g denote the probe and genuine in a training triplet sample, respectively.

4) Decoder. The output feature of the feature transformer is then fed into the decoder D to fully reconstruct a gait cycle with a predefined number of frames M, and the decoding process is formulated as

$$R_I = D(\boldsymbol{f}_I), \tag{7}$$

where R_I denotes the gait cycle of M silhouettes reconstructed from the input silhouette I.

The structure of the decoder is symmetrical to that of the encoder. A fully connected layer along with reshaping is first used to convert the input 100D feature into the same size as the feature output from the last convolutional layer in the encoder, and then four deconvolutional layers are used for up-sampling. A sigmoid activation function is applied after the last deconvolutional layer that outputs the reconstructions with a size of $M \times 64 \times 64$, where each channel indicates a reconstructed image at a specific phase common to all subjects. A reconstruction loss is computed to ensure the reconstructed gait cycle is similar to the corresponding ground truth (training data) GT_I, which is defined as

$$L_{\text{recon}} = \|R_I - GT_I\|_2^2. \tag{8}$$

3.3 Combining PA-GCR with GaitSet

Next, the reconstructed gait cycle from the PA-GCR is fed into GaitSet to obtain a more discriminative feature.

GaitSet [7] is a set-based gait recognition network that takes a set of silhouettes as an input. After obtaining features from each input silhouette independently using a CNN, set pooling is applied to aggregate features over frames into a set-level feature. The set-level feature is then used for discrimination learning via horizontal pyramid mapping, which extracts features of different spatial locations on different scales. The feature output from GaitSet G for the reconstructed gait cycle of silhouettes R_I is formulated as

$$\boldsymbol{h}_I = G(R_I). \tag{9}$$

For a batch size of $S \times K$ in the training stage, where S is the number of subjects and K is the number of samples per subject, the batch all triplet loss is [15]

$$L_{\text{trip}} = \frac{1}{N} \sum_{i=1}^{S} \sum_{a=1}^{K} \sum_{\substack{s=1 \\ s \neq a}}^{K} \sum_{\substack{j=1 \\ j \neq i}}^{S} \sum_{n=1}^{K} \max(\text{margin} + d_{i,s}^{i,a} - d_{j,n}^{i,a}, 0), \tag{10}$$

where $N = SK(SK - K)(K - 1)$ is the number of all triplets in a batch, $d_{i,s}^{i,a} = \|\boldsymbol{h}_{I_{i,s}} - \boldsymbol{h}_{I_{i,a}}\|_2^2$ is the dissimilarity score of the genuine pair, and $d_{j,n}^{i,a} = \|\boldsymbol{h}_{I_{j,n}} - \boldsymbol{h}_{I_{i,a}}\|_2^2$ is the dissimilarity score of the imposter pair.

3.4 Unified loss function

Because the phase estimator directly works on the input image, we train it separately from the main pipeline (i.e., the encoder, the feature transformer, the decoder, and GaitSet). We therefore define a unified loss function to optimize the whole main pipeline jointly to achieve a trade-off between reconstruction and recognition accuracy. The unified loss function is calculated as the weighted sum of the three aforementioned loss functions:

$$L_{\text{uni}} = w_{\text{sim}} L_{\text{sim}} + w_{\text{recon}} L_{\text{recon}} + w_{\text{trip}} L_{\text{trip}}, \tag{11}$$

where w_{sim}, w_{recon}, and w_{trip} are the respective weights for the three losses.

4 Experiments

4.1 Datasets

We evaluated the proposed method on three publicly available datasets: OU-MVLP [39], CASIA-B [46], and OUTD-D [26].

OU-MVLP contains image sequences of 10,307 subjects captured from 14 views at a frame rate of 25 fps, and is the largest gait dataset with a wide view variation in the world. We only focused on the side view (90°) to investigate the recognition performance using a single frame without other covariates. According to the original protocol [39], 5,153 subjects were used for training and the other disjoint 5,154 subjects were used for testing, with one probe sequence and one gallery sequence for each subject. We used this as the main dataset for the following experiments because of its high statistical reliability.

CASIA-B is one of the most widely used gait datasets and consists of gait sequences of 124 subjects captured at 25 fps. Each subject has six normal walking sequences for each of the 11 views. Similarly to the OU-MVLP experiment, only sequences at 90° were used for our evaluation (Section 4.4). We adopted the same challenging protocol as in [4], where the first 24 subjects were used for training and the last 100 were used for testing, with one gallery sequence (NM #01) and five probe sequences (NM #02-06).

OUTD-D is a dataset that focuses on gait fluctuations (i.e., silhouette differences of the same phase) over several periods. Hence, it includes a larger number of frames, 360 in each sequence. 185 subjects with two sequences (probe and gallery) for each subject were captured at 60 fps from the side view in this dataset. Using the same protocol as in [1, 2], we used 85 subjects for training and the other 100 for testing (Section 4.4).

We randomly selected a single frame from a sequence as the input for evaluation for all the datasets.

4.2 Implementation details

We trained the proposed network using the Adam optimizer [20] with a batch size of $S \times K = 8 \times 16$. We used the same number of channels in GaitSet for

OU-MVLP and CASIA-B as in [7], and used the same number for OUTD-D as that for CASIA-B. The margin in Eq. 10 was set to 0.2 for all three datasets. We first prepared the ground truth of a full gait cycle containing 100 frames for the phase synchronization introduced in Section 3.2. Considering the computational complexity and memory size needed to train the network, we set the number of frames in the reconstructions as $M = 25$, and evenly down-sampled 25 frames from the original 100 frames to correspond to the training ground truth.

The weights in Eq. 11 were set as $w_{\text{sim}} = 0.0005$ and $w_{\text{recon}} = w_{\text{trip}} = 1$. To first achieve stable reconstruction results, L_{trip} was excluded (i.e., only PA-GCR was included) for the first 30K training iterations with an initial learning rate of 10^{-4} for OU-MVLP, and for the first 20K iterations with an initial learning rate of 10^{-5} for CASIA-B and OUTD-D. We then involved L_{trip} with a learning rate of 10^{-4} for GaitSet while reducing the learning rate for PA-GCR by 0.1. The whole network was trained with 50K more iterations for CASIA-B and OUTD-D and 250K more iterations for OU-MVLP, where the learning rates for both PA-GCR and GaitSet were again reduced by 0.1 for the last 100K iterations.

The recognition performance was evaluated using rank-1 identification rate and equal error rate (EER) [32].

4.3 Visualizing gait cycle reconstruction

We first visualize gait cycles reconstructed by PA-GCRNet using a **test** example of a genuine pair. We choose the challenging case of a large phase difference between the matching pair, where the input probe and gallery image are in the single-support and double-support phase, respectively. Fig. 3 shows that the reconstruction results are similar to the corresponding ground truths. We also give the mean squared error (MSE) as a measure of the difference between the reconstruction result and ground truth [2]:

$$\text{MSE} = \frac{1}{MWH} \sum_{m=1}^{M} \sum_{i=1}^{W} \sum_{j=1}^{H} \|R_I(i,j,m) - GT_I(i,j,m)\|_2^2, \tag{12}$$

where W and H are the image width and height, respectively, and M is the number of images in the gait cycle (25 in this case). The figure also shows the mean squared L2 distance between the ground truth pair and reconstruction pair to illustrate the difference between the matching pair.

Using the single input silhouette, the proposed method successfully reconstructs a natural gait cycle with a continuous phase change. Although there is still a reconstruction error, the body shapes and poses in the reconstruction are similar to those in the ground truth. The reconstructed gallery and probe pair are also quite similar (see Figs. 3(d) and (e)), which demonstrates that the PA-GCRNet reconstructions are independent of the phase of the input silhouettes to some extent.

On the other hand, the ground truth pair in Fig. 3(b) and (c) has a larger difference because there may be pose variations even for the same phase. This

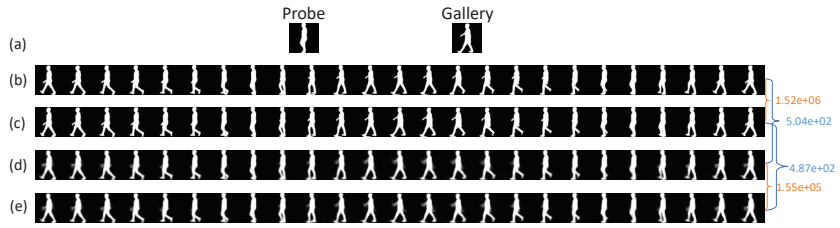

Fig. 3. Examples of gait cycle reconstruction. (a) Input silhouette; left: probe; right: gallery. (b) Ground truth of probe gait cycle. (c) Ground truth of gallery gait cycle. (d) Reconstructed probe gait cycle. (e) Reconstructed gallery gait cycle. Blue digits indicate the errors between the reconstruction and corresponding ground truth, and orange digits indicate the mean squared L2 distances between the corresponding probe and gallery pairs.

implies that the proposed network not only forces the reconstruction results to be similar to the ground truth, but also reduces the intra-subject variation between the reconstructed same subject pair, which is more beneficial for matching. This is because the end-to-end training includes both reconstruction and recognition, which makes the PA-GCRNet achieve a good trade-off between reconstruction quality and recognition performance. More reconstruction examples are shown in the supplementary material.

In addition, taking a look at the case that a pair of inputs are in the same phase but from different **test** subjects (see Figs. 1(b)(III) and (IV)), the pose differences (e.g., different back bending and stride) are continuously observed between the reconstructed gait cycle pair, which demonstrates the proposed network can keep the gait pose individuality to some extent. That is, an individual gait pose sequence is able to be reconstructed from a single gait image by the network, which provides more gait characteristics of a specific subject, and hence possibly helps improve the recognition accuracy.

4.4 Comparison with state-of-the-art methods

OU-MVLP In this section, the proposed method is compared with state-of-the-art methods for OU-MVLP. There is no existing work on this topic, and only GaitSet [7] has been used to test performance based on a single input image. Therefore, we compare our method with GaitSet[3] and the baseline, i.e., direct matching (DM) between the selected single probe and gallery image pair, as shown in Table 1.

The proposed method significantly outperforms the benchmarks. For example, the rank-1 identification rate is over five times higher than that of the benchmarks, which makes it possible to achieve gait recognition from a single image. The proposed method completes a matching task and obtains the dissimilarity score

[3] The results were obtained by using their model on our test set (selected single image from a sequence).

Table 1. Rank-1 identification rate [%] (denoted as Rank-1) and EER [%] of the proposed method and other benchmarks for OU-MVLP. Bold and bold italic indicate the best and second-best results, respectively. This font convention is used to indicate performance throughout this paper.

Method	Rank-1	EER
DM	4.4	41.3
Gaitset [7]	***14.0***	***19.6***
PA-GCRNet (proposed)	**80.3**	**1.3**

between a matching image pair in 5 milliseconds using a Quadro RTX 6000 GPU, demonstrating its real-time executability.

CASIA-B Considering the very limited number of training subjects in CASIA-B, we adopted two strategies for the proposed method: training the network from scratch only using 24 training subjects in CASIA-B, and fine-tuning the network from the model pre-trained for OU-MVLP. For the latter, we only fine-tuned the PA-GCR from the OU-MVLP while still training the GaitSet part from scratch because of the different settings of GaitSet for these two datasets [7]. Additionally, to validate the generalization capability of the proposed network, we also investigated the performance of cross-dataset testing, i.e., directly tested with CASIA-B using the pre-trained PA-GCRNet on OU-MVLP.

Table 2 (left) shows the results for GaitSet, DM and our method along with that of ITCNet, which was reported [4] using a single input image for this dataset. Note that the ITCNet protocol differed from ours by choosing 14 different frames for each probe and gallery sequence and then obtaining the result via fusion. The other benchmarks and our method used only a single image.

Because of the severe overfitting caused by the very limited training samples, the proposed method cannot perform well by training from scratch on this dataset but still gains a little improvement compared with GaitSet. The overfitting problem can be solved and the performance of the proposed method largely improved by fine-tuning the PA-GCR pre-trained via OU-MVLP with a better generalization capability. Although the result is not as good as for OU-MVLP, this is understandable because some low-quality silhouettes with segmentation errors are included in this dataset. These may affect the recognition performance if the selected matching image pair has different segmented silhouettes.

It is worth mentioning to that the cross-dataset testing achieves a quite good performance, which is only slightly worse than the fine-tuned model in terms of EER. That means, the proposed network trained on a large-scale dataset (i.e., OU-MVLP) has a good generalization to be directly used for another dataset, and hence has a chance to be directly applied in real application scenarios.

OUTD-D We finally compare the methods for OUTD-D. Table 2 (right) shows the results of our method using both training from scratch and fine-tuning, as well as cross-dataset testing, as was done for CASIA-B. The results for NoTSR [2] (direct matching between the averaged silhouettes over selected frames), Morph [3],

Table 2. Rank-1 identification rate [%] (denoted as Rank-1) and EER [%] of the proposed method and other benchmarks for CASIA-B and OUTD-D. Note that ITCNet [4], NoTSR [2], Morph [3], TSR [1], and Unified TSR [2] used different protocols from ours.

Dataset	CASIA-B		OUTD-D	
Method	Rank-1	EER	Rank-1	EER
DM	14.1	39.7	17	39.0
Gaitset [7]	33.3	17.7	42	13.0
PA-GCRNet (scratch)	39.4	14.7	73	6.4
PA-GCRNet (cross-dataset)	**74.7**	***9.9***	***75***	***4.1***
PA-GCRNet (fine-tune)	**74.7**	**8.1**	**91**	**3.5**
ITCNet [4]	50.0	22.8	-	-
NoTSR [2]	-	-	51	15.0
Morph [3]	-	-	52	14.0
TSR [1]	-	-	44	16.5
Unified TSR [2]	-	-	87	3.5

TSR [1], and Unified TSR [2] were obtained for 1 fps, which means six frames from one sequence were simultaneously used in those methods.

Compared with the benchmarks using the same protocol, the proposed method achieves much better results even for training from scratch. The performance of the proposed method can be further improved by fine-tuning the PA-GCR from the pre-trained model for OU-MVLP. This even outperforms the state-of-the-art method for low frame-rate gait recognition (Unified TSR [2]), which uses more than one frame for each sequence simultaneously. Again, the cross-dataset testing works well and gains a better result than the model trained from scratch, which further demonstrates the good generalization capability of the proposed method[4].

4.5 Ablation study

We analyzed the effects of each component of the proposed method on OU-MVLP, as shown in Table 3. The first row shows the result of the baseline GaitSet, which used the same settings as in the original paper [7]. One component of our method was removed for each row from the second to the fourth row. Specifically, we did the following: in the second row, the GaitSet was retrained using the same strategy (a single image) as the proposed method to fairly confirm the effectiveness of PA-GCR reconstruction; in the third row, PA-GCR and GaitSet were separately trained to verify the effectiveness of the proposed unified framework; in the fourth row, the phase estimator and feature transformer with the similarity loss (Eq. 6) were removed from the proposed network to validate the effects of using the phase information of the input image. The fifth row shows the result of our method. For reference, the sixth row reports the upper bound of the proposed phase-aware framework (i.e., using the ground truth phase label rather than the estimate by the phase estimator). The last row shows the upper bound of reconstruction (i.e., test results for the ground truth of the gait cycle using the pre-trained GaitSet model in the first row).

[4] Reconstruction results for cross-dataset testing are shown in the supplementary material.

Table 3. Ablation experiments evaluated using rank-1 identification rate [%] (denoted as Rank-1) on OU-MVLP. Ground truth is denoted as GT. "×" indicates phase information is not used.

Model	Removed component	Phase info.	#Training input frame	Test input	Rank-1
GaitSet [7]	-	×	30	1 real frame	14.0
GaitSet	PA-GCR	×	1	1 real frame	34.4
PA-GCR + GaitSet	Unified training	Estimated	1	1 real frame	50.5
GCRNet	Use of phase info.	×	1	1 real frame	***77.4***
PA-GCRNet (proposed)	-	Estimated	1	1 real frame	**80.3**
PA-GCRNet (upper bound of phase-aware framework)	-	GT	1	1 real frame	80.6
GaitSet (upper bound of reconstruction)	-	×	30	GT gait cycle (25 frames)	97.7

Comparing the results in the second and fifth rows, it is obvious that removing the proposed PA-GCR significantly reduces the recognition performance. This demonstrates the need to involve gait cycle reconstruction in this task. The proposed unified framework performs much better than the separated training strategy in the third row because it achieves a trade-off between reconstruction and recognition accuracy through its unified optimization. The effectiveness of using input phase information is also confirmed by comparing the fourth and fifth rows. Additionally, the phase estimator yields an error of only 0.02, which is the mean squared L2 distance between the estimated and ground truth phases[5]. Therefore, the difference in performance between the real test (i.e., using the estimated phase) and the upper bound of using the ground truth phase is quite small for the proposed method.

5 Conclusion

This paper presented PA-GCRNet for gait recognition from a single image. Given a single input image, the PA-GCR fully reconstructs the gait cycle of a silhouette in conjunction with the phase estimator and then feeds the reconstruction into a subsequent recognition network like GaitSet for matching. This method achieved significantly higher recognition performance on three publicly available gait datasets.

One future research goal is to extend the proposed PA-GCR to accept multiple input images for low frame-rate gait recognition. Additionally, the proposed method can also potentially be extended to general actions to generate both future and past frames from an arbitrary frame, which is beneficial for video synthesis and action recognition, and this remains future work.

Acknowledgment. This work was supported by JSPS KAKENHI Grant No. JP18H04115, JP19H05692, and JP20H00607, Jiangsu Provincial Science and Technology Support Program (No. BE2014714), the 111 Project (No. B13022), and the Priority Academic Program Development of Jiangsu Higher Education Institutions.

[5] A mean squared L2 distance of 0.02 is equivalent to approximately 9° of the circumference, i.e., less than one phase for a gait cycle containing 25 phases.

References

1. Akae, N., Makihara, Y., Yagi, Y.: Gait recognition using periodic temporal super resolution for low frame-rate videos. In: Proc. of the Int. Joint Conf. on Biometrics (IJCB2011). pp. 1–7. Washington D.C., USA (Oct 2011)
2. Akae, N., Mansur, A., Makihara, Y., Yagi, Y.: Video from nearly still: an application to low frame-rate gait recognition. In: Proc. of the 25th IEEE Conf. on Computer Vision and Pattern Recognition (CVPR2012). pp. 1537–1543. Providence, RI, USA (Jun 2012)
3. Al-Huseiny, M.S., Mahmoodi, S., Nixon, M.S.: Gait learning-based regenerative model: A level set approach. In: The 20th Int. Conf. on Pattern Recognition. pp. 2644–2647. Istanbul, Turkey (Aug 2010)
4. Babaee, M., Li, L., Rigoll, G.: Person identification from partial gait cycle using fully convolutional neural networks. Neurocomputing **338**, 116 – 125 (2019)
5. Bashir, K., Xiang, T., Gong, S.: Cross view gait recognition using correlation strength. In: BMVC (2010)
6. Bouchrika, I., Goffredo, M., Carter, J., Nixon, M.: On using gait in forensic biometrics. Journal of Forensic Sciences **56**(4), 882–889 (2011)
7. Chao, H., He, Y., Zhang, J., Feng, J.: Gaitset: Regarding gait as a set for cross-view gait recognition. In: Proc. of the 33th AAAI Conference on Artificial Intelligence (AAAI 2019) (2019)
8. El-Alfy, H., Xu, C., Makihara, Y., Muramatsu, D., Yagi, Y.: A geometric view transformation model using free-form deformation for cross-view gait recognition. In: Proc. of the 4th Asian Conf. on Pattern Recognition (ACPR 2017). IEEE (Nov 2017)
9. Gao, R., Xiong, B., Grauman, K.: Im2flow: Motion hallucination from static images for action recognition. In: CVPR (2018)
10. Guan, Y., Li, C., Roli, F.: On reducing the effect of covariate factors in gait recognition: A classifier ensemble method. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(7), 1521–1528 (July 2015)
11. Guan, Y., Li, C.T.: A robust speed-invariant gait recognition system for walker and runner identification. In: Proc. of the 6th IAPR International Conference on Biometrics. pp. 1–8 (2013)
12. Guan, Y., Li, C.T., Choudhury, S.: Robust gait recognition from extremely low frame-rate videos. In: Biometrics and Forensics (IWBF), 2013 International Workshop on. pp. 1–4 (April 2013). https://doi.org/10.1109/IWBF.2013.6547319
13. Han, J., Bhanu, B.: Individual recognition using gait energy image. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(2), 316– 322 (2006)
14. He, Y., Zhang, J., Shan, H., Wang, L.: Multi-task gans for view-specific feature learning in gait recognition. IEEE Transactions on Information Forensics and Security **14**(1), 102–113 (Jan 2019). https://doi.org/10.1109/TIFS.2018.2844819
15. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. CoRR **abs/1703.07737** (2017), http://arxiv.org/abs/1703.07737
16. Horst, F., Lapuschkin, S., Samek, W.and Müller, K., Schöllhorn, W.: Explaining the unique nature of individual gait patterns with deep learning. Scientific Reports **9**, 2391 (02 2019). https://doi.org/10.1038/s41598-019-38748-8
17. Hossain, M.A., Makihara, Y., Wang, J., Yagi, Y.: Clothing-invariant gait identification using part-based clothing categorization and adaptive weight control. Pattern Recognition **43**(6), 2281–2291 (Jun 2010)

18. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. CoRR **abs/1502.03167** (2015), http://arxiv.org/abs/1502.03167
19. Iwama, H., Muramatsu, D., Makihara, Y., Yagi, Y.: Gait verification system for criminal investigation. IPSJ Transactions on Computer Vision and Applications **5**, 163–175 (Oct 2013)
20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint **arXiv: 1412.6980 (2014)** (2014)
21. Kourtzi, Z., Kanwisher, N.: Activation in human mt/mst by static images with implied motion. Journal of cognitive neuroscience **12**, 48–55 (02 2000). https://doi.org/10.1162/08989290051137594
22. Kusakunniran, W., Wu, Q., Zhang, J., Li, H.: Support vector regression for multi-view gait recognition based on local motion feature selection. In: Proc. of IEEE computer society conference on Computer Vision and Pattern Recognition 2010. pp. 1–8. San Francisco, CA, USA (Jun 2010)
23. Li, X., Makihara, Y., Xu, C., Yagi, Y., Ren, M.: Joint intensity transformer network for gait recognition robust against clothing and carrying status. IEEE Transactions on Information Forensics and Security **14**(12), 3102–3115 (Dec 2019)
24. Lin, G., Milan, A., Shen, C., Reid, I.: RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In: CVPR (Jul 2017)
25. Lynnerup, N., Larsen, P.: Gait as evidence. IET Biometrics **3**(2), 47–54 (6 2014). https://doi.org/10.1049/iet-bmt.2013.0090
26. Makihara, Y., Mannami, H., Tsuji, A., Hossain, M., Sugiura, K., Mori, A., Yagi, Y.: The ou-isir gait database comprising the treadmill dataset. IPSJ Transactions on Computer Vision and Applications **4**, 53–62 (Apr 2012)
27. Makihara, Y., Mori, A., Yagi, Y.: Temporal super resolution from a single quasi-periodic image sequence based on phase registration. In: Proc. of the 10th Asian Conf. on Computer Vision. pp. 107–120. Queenstown, New Zealand (Nov 2010)
28. Makihara, Y., Sagawa, R., Mukaigawa, Y., Echigo, T., Yagi, Y.: Gait recognition using a view transformation model in the frequency domain. In: Proc. of the 9th European Conference on Computer Vision. pp. 151–163. Graz, Austria (May 2006)
29. Makihara, Y., Suzuki, A., Muramatsu, D., Li, X., Yagi, Y.: Joint intensity and spatial metric learning for robust gait recognition. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6786–6796 (July 2017). https://doi.org/10.1109/CVPR.2017.718
30. Makihara, Y., Yagi, Y.: Silhouette extraction based on iterative spatio-temporal local color transformation and graph-cut segmentation. In: Proc. of the 19th International Conference on Pattern Recognition. Tampa, Florida USA (Dec 2008)
31. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on International Conference on Machine Learning. pp. 807–814. ICML'10, Omnipress, USA (2010), http://dl.acm.org/citation.cfm?id=3104322.3104425
32. Phillips, P., Moon, H., Rizvi, S., Rauss, P.: The feret evaluation methodology for face-recognition algorithms. IEEE Transactions of Pattern Analysis and Machine Intelligence **22**(10), 1090–1104 (2000)
33. Pintea, S.L., Gemert, J.C., Smeulders, A.W.M.: Déjà vu: - motion prediction in static images. In: ECCV (2014)
34. Prismall, S.P., Nixon, M.S., Carter, J.N.: Novel temporal views of moving objects for gait biometrics. In: Kittler, J., Nixon, M.S. (eds.) Audio- and Video-Based Biometric Person Authentication. pp. 725–733. Springer Berlin Heidelberg, Berlin, Heidelberg (2003)

35. Sarkar, S., Phillips, P.J., Liu, Z., Vega, I.R., Grother, P., Bowyer, K.W.: The humanid gait challenge problem: data sets, performance, and analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(2), 162–177 (Feb 2005). https://doi.org/10.1109/TPAMI.2005.39
36. Sederberg, T.W., Parry, S.R.: Free-form deformation of solid geometric models. SIGGRAPH Comput. Graph. **20**(4), 151–160 (Aug 1986). https://doi.org/10.1145/15886.15903, http://doi.acm.org/10.1145/15886.15903
37. Shiraga, K., Makihara, Y., Muramatsu, D., Echigo, T., Yagi, Y.: Geinet: View-invariant gait recognition using a convolutional neural network. In: 2016 International Conference on Biometrics (ICB). pp. 1–8 (2016)
38. Takemura, N., Makihara, Y., Muramatsu, D., Echigo, T., Yagi, Y.: On input/output architectures for convolutional neural network-based cross-view gait recognition. IEEE Transactions on Circuits and Systems for Video Technology pp. 1–1 (2018). https://doi.org/10.1109/TCSVT.2017.2760835
39. Takemura, N., Makihara, Y., Muramatsu, D., Echigo, T., Yagi, Y.: Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. IPSJ Trans. on Computer Vision and Applications **10**(4), 1–14 (2018)
40. Wolf, T., Babaee, M., Rigoll, G.: Multi-view gait recognition using 3d convolutional neural networks. In: 2016 IEEE International Conference on Image Processing (ICIP). pp. 4165–4169 (2016)
41. Wu, Z., Huang, Y., Wang, L.: Learning representative deep features for image set analysis. IEEE Transactions on Multimedia **17**(11), 1960–1968 (Nov 2015). https://doi.org/10.1109/TMM.2015.2477681
42. Wu, Z., Huang, Y., Wang, L., Wang, X., Tan, T.: A comprehensive study on cross-view gait based human identification with deep cnns. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(2), 209–226 (2017)
43. Xu, C., Makihara, Y., Li, X., Yagi, Y., Lu, J.: Speed invariance vs. stability: Cross-speed gait recognition using single-support gait energy image. In: Proc. of the 13th Asian Conf. on Computer Vision (ACCV 2016). pp. 52–67. Taipei, Taiwan (Nov 2016)
44. Xu, C., Makihara, Y., Yagi, Y., Lu, J.: Gait-based age progression/regression: a baseline and performance evaluation by age group classification and cross-age gait identification. Machine Vision and Applications **30**(4), 629–644 (Jun 2019). https://doi.org/10.1007/s00138-019-01015-x, https://doi.org/10.1007/s00138-019-01015-x
45. Yu, S., Chen, H., Reyes, E.B.G., Poh, N.: Gaitgan: Invariant gait feature extraction using generative adversarial networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 532–539 (July 2017). https://doi.org/10.1109/CVPRW.2017.80
46. Yu, S., Tan, D., Tan, T.: A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In: Proc. of the 18th Int. Conf. on Pattern Recognition. vol. 4, pp. 441–444. Hong Kong, China (Aug 2006)
47. Zhang, C., Liu, W., Ma, H., Fu, H.: Siamese neural network based gait recognition for human identification. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2832–2836 (2016)
48. Zhang, K., Luo, W., Ma, L., Liu, W., Li, H.: Learning joint gait representation via quintuplet loss minimization. In: 2019 conference on computer vision and pattern recognition (CVPR 2019) (2019)