End-to-end Model-based Gait Recognition using Synchronized Multi-view Pose Constraint

Xiang Li[1] Yasushi Makihara[1] Chi Xu[1] Yasushi Yagi[1]
[1] Osaka University, Osaka, Japan
{li, makihara, xu, yagi}@am.sanken.osaka-u.ac.jp

Abstract

We propose an end-to-end model-based cross-view gait recognition which employs pose sequences and shapes extracted by human model fitting. Specifically, we consider a problem setting where gait sequences from single different views are given as a pair to match in a test phase, while asynchronous multi-view gait sequences are given for each subject in a training phase. This work exploits multi-view constraint in the training phase to extract more consistent pose sequences from any views in the test phase, unlike the existing methods do not consider them. For this purpose, given asynchronous multi-view gait sequences, we introduce a phase synchronization step in the training phase so that we can impose pose consistency at each synchronized phase in a temporally up-sampled phase domain. We then train our network by minimizing a loss function based on the synchronized multi-view pose constraint as well as shape consistency, temporal pose smoothness, recognition accuracy, etc in an end-to-end manner. We also introduce the synchronization step in a test phase to reduce intra-subject variations caused by asynchronous pose features. Experimental results on the OU-MVLP and CASIA-B datasets show that the proposed method achieves the state-of-the-art performance for both gait identification and verification scenarios, especially a great improvement in terms of the pose representations.

1. Introduction

Gait recognition [36, 31, 7] aims to identity a walking person through his/her way of walking. Compared with other biometrics, it is more advantageous in some real-world applications such as surveillance, forensics [4, 30], and criminal investigation [18], because of the remote availability and non-contact properties.

Previous approaches to gait recognition are mainly divided into appearance-based and model-based ones. The appearance-based approach usually uses silhouette-based representations (e.g., gait energy image (GEI) [13]) and has been popular in the last two decades due to its simplicity and effectiveness. The appearance-based approach is, how-

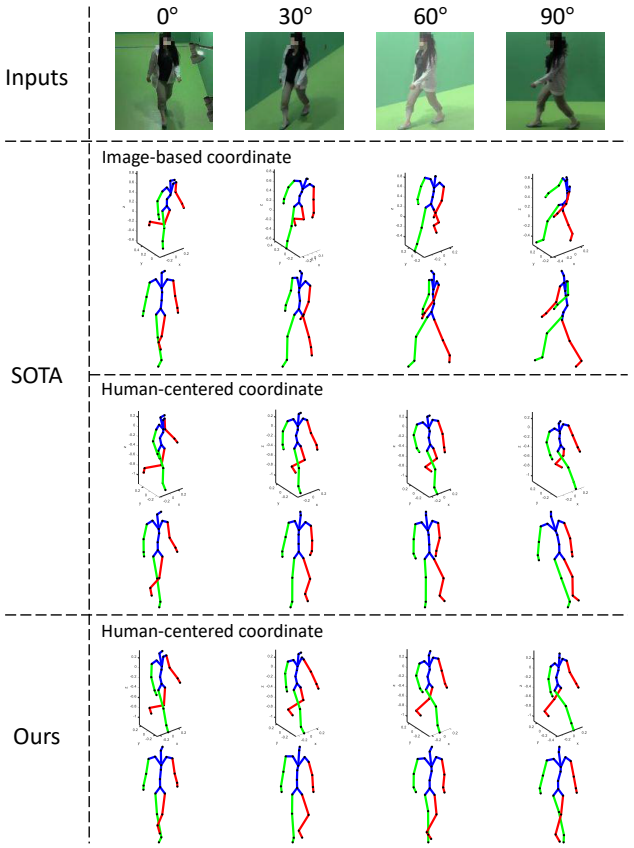

Figure 1. Examples of pose estimation results. Given multi-view input images with almost the same phase (gait stance), a state-of-the-art method (i.e., ModelGait [24]) independently estimates the 3D and 2D body joints in the image-based coordinate, which is subject to the view variation. Even after transforming them to a unified human-centered coordinate, the body joints still have certain differences among the multiple views. On the other hand, our method could narrow the differences and get more similar body joints by the synchronized multi-view pose constraint.

ever, easily affected by various covariates (e.g., view angles, carrying status, and clothing).

On the other hand, the model-based approach is generally more invariant to the covariates, particularly for view variation when using a 3D human model. Moreover, the model-based gait recognition accuracy has been improving for the last few years [10, 26, 27] thanks to recent advancement of deep learning-based human model fitting [5, 19]. Especially, a recent model-based gait recognition approach [24] (denoted as "ModelGait") which uses a skinned multi-person linear (SMPL) model [29] as a human model and integrates the human model estimation and recognition in an end-to-end framework, outperformed the state-of-the-art cross-view appearance-based methods.

The existing model-based cross-view gait recognition methods do, however, not make the most of multi-view gait sequences in a training phase. More specifically, in a problem setting of cross-view gait recognition, while gait sequences from single different views are given as a pair to match in a test phase, multi-view gait sequences are usually given for each subject in a training phase. The existing approaches independently treat a gait sequence from each different view in the training phase, and hence estimated pose sequences from different views cannot be necessarily similar to each other (see SOTA in Fig. 1).

In this study, we make the most of the multi-view pose constraint in the training phase so as to infer a more view-consistent pose sequence (i.e., with less intra-subject variation) even from a gait sequence of a single arbitrary view. One of challenges to realize this is that multi-view gait sequences in the training phase are not necessarily synchronous and hence existing multi-view approaches to human pose and shape reconstruction [25, 40] cannot be directly applied. We therefore introduce a synchronization process to cope with asynchronous multi-view gait sequences. Contributions of this study are threefold.

(1) A view-consistent pose estimator with a synchronized multi-view pose constraint.

We propose a training framework to estimate a more view-consistent pose sequence by making the most of asynchronous multi-view gait sequences in the training phase (denoted as "MvModelGait"). Unlike the existing methods that independently treat each of the asynchronous gait sequences in the image-based coordinate, the proposed method synchronizes pose sequences in a unified human-centered coordinate to impose multi-view consistency at the same synchronized phase. As such, we can infer a more view-consistent pose sequence even from a single arbitrary-view gait sequence in a test phase (see Ours in Fig. 1).

(2) Better recognition with synchronized pose sequences.

Unlike existing model-based approaches that use asynchronous pose sequences (e.g., pose sequences starting with different gait stances) for recognition, we introduce the phase synchronization step to the test phase too. This leads to better recognition accuracy since we can reduce the intra-subject variations induced by asynchronous gait sequences.

(3) State-of-the-art performance.

We achieve the state-of-the-art cross-view gait recognition accuracy by the ensemble of a body shape feature and the synchronized pose feature on the largest gait dataset with view variations (i.e., OU-MVLP [43]), and the most frequently used gait dataset with view, carrying and clothing variations (i.e., CASIA-B [53]).

2. Related work

2.1. Appearance-based gait recognition

Appearance-based gait recognition methods mainly use silhouette-based representations as gait features, e.g., GEI [13], frequency-domain feature [32], chrono-gait image [45], and even silhouette themselves [38]. Because various real-life covariates can easily affect these representations, many efforts have been made to improve their discriminative capability, such as traditional metric learning-based techniques [50, 11, 33], transformation-based techniques [32, 34, 35], deep convolutional neural networks [41, 47, 43, 48, 54, 6, 21, 55, 9, 23, 16, 49, 28] and generative adversarial networks [52, 14, 22]. Among them, the deep learning-based methods have become the main research direction and achieved the state-of-the-art performance. For example, Chao et al. [6] proposed the GaitSet network which regards gait silhouettes as a set. Zhang et al. [55] and Li et al. [23] proposed disentanglement networks that separate identity related and unrelated features. Fan et al. [9] proposed the GaitPart network which makes full use of part features. Hou et al. [16] proposed the gait lateral network (GLN) to learn discriminative and compact features.

2.2. Model-based gait recognition

Model-based gait recognition methods mainly consider pose sequences and body shape parameters obtained by human model fitting as gait features. Early approaches [44, 51, 3, 8, 2] faced difficulties in robustly and accurately fitting a human model, and hence got less satisfactory results. On the other hand, recent studies [10, 26, 27, 24] overcome the difficulties by using state-of-the-art human pose estimation methods (e.g., OpenPose [5] and human mesh recovery (HMR) network [19]), and hence are competitive with appearance-based approaches. For example, Liao et al. [26] first used OpenPose to extract 2D body joints as the pose feature, and then fed it to the pose-based temporal-spatial network. Because the 2D body joints are not invariant to the view variation, Liao et al. [27] further extended the work by estimating 3D joints from the 2D ones. Moreover, Li et al. [24] exploited body shape parameters as well as pose sequences extracted in a form of an SMPL model by using the

HMR network, and optimize the HMR network jointly with recognition network in an end-to-end manner.

These methods, however, do not make the most of multi-view gait sequences in the training phase as discussed in the introduction section unlike the proposed method which employs the synchronized multi-view pose constraint.

2.3. 3D human pose and shape estimation

Studies on 3D human pose and shape estimation often infer parametric 3D human body models (e.g., SMPL [29]) from 2D images. Most of them [37, 19] use learning-based regression models that focus on single-view images. However, estimation from a single-view image is often subject to ambiguity. For example, if a person walks towards a camera (i.e., observed from frontal view), estimated forward-backward motion (e.g., stride length) gets more ambiguous than that observed from side-view. Naturally, the 3D human body models could be more accurately estimated by using multi-view images. Liang and Lin [25] proposed a multi-view multi-stage framework, which iteratively transfers estimated pose and shape across views while the estimated camera calibration parameters across iteration stages. Shin and Halilaj [40] proposed a learnable volumetric aggregation approach to reconstruct 3D human body pose and shape from calibrated multi-view images.

Note that the above-mentioned multi-view approaches require multi-view images in the test phase, while the proposed method tackles a cross-view gait recognition scenario where multi-view gait sequences are given only in the training phase and gait sequences from single different views are given in the test phase.

3. MvModelGait

3.1. Overview

We build the proposed MvModelGait upon a state-of-the-art model-based gait recognition, ModelGait [24] as a backbone. An overall framework is shown in Fig. 2. Asynchronous multi-view RGB sequences are first independently fed into the backbone ModelGait to extract pose (3D and 2D body joints) and shape features. The pose features are then transformed from image-based coordinate to the unified human-centered coordinate and further synchronized based on estimated phases. The synchronized pose features from multiple views are forced to be consistent by the synchronized multi-view pose constraint $L_{\text{mv\_pose}}$ as well as fed into a CNN to extract more discriminative pose features for recognition.

3.2. ModelGait to extract pose and shape features

We will briefly explain our backbone ModelGait so that this paper can be self-contained. Readers may refer to [24] for more details. Given a cropped RGB sequence, ModelGait extracts pose $\boldsymbol{\theta} \in \mathbb{R}^{72}$ and shape $\boldsymbol{\beta} \in \mathbb{R}^{10}$ parameters of the SMPL [29] as well as weak-perspective camera parameters $\boldsymbol{k} \in \mathbb{R}^3$, based on the HMR network [19]. The SMPL can generate a triangulated body mesh which is composed of 6,890 vertices and is differentiable with respect to the SMPL parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. The silhouettes derived from the SMPL parameters can be rendered by a differentiable renderer [15]. Moreover, the pose parameter $\boldsymbol{\beta}$ is converted into 3D and 2D body joint positions as more intuitively understandable pose features, while the shape parameters are averaged over frames within a sequence so as to produce a common shape parameter.

Finally, a loss function L_{MG} for ModelGait is shown as

$$L_{\text{MG}} = \lambda_{\text{inner}} L_{\text{inner}} + \lambda_{\text{recont}} L_{\text{recont}} + \lambda_{\text{joints}} L_{\text{joints}}, \quad (1)$$

where the inner loss L_{inner} ensures temporal continuity of pose and camera parameters as well as consistency of shape parameters within sequences, the reconstruction loss L_{recont} keeps the consistency between renderer silhouettes by the SMPL parameters and ground-truth silhouettes so as to preserve more identity information, the joints loss L_{joints} avoids the pose corruption by over-fitting to the ground truth silhouettes, and coefficients λ_{inner}, λ_{recont}, and λ_{joints} are the weight parameters for each corresponding term.

Since the shape features are essentially independent of phase (gait stance), we can simply introduce multi-view consistent shape constraint $L_{\text{mv\_shape}}$ just by minimizing variance of the shape parameters over views regardless of phases (see Fig. 2), which is defined as

$$L_{\text{mv\_shape}} = \frac{2}{N_V(N_V-1)} \sum_{m=1}^{N_V} \sum_{n>m}^{N_V} \| \boldsymbol{f}^m_{\text{shape}} - \boldsymbol{f}^n_{\text{shape}} \|_2^2, \quad (2)$$

where N_V is the number of gait sequences from different views for a certain subject, and $\boldsymbol{f}_{\text{shape}}$ is the corresponding shape feature for a sequence.

3.3. Synchronized multi-view pose constraint

Coordinate transformation. As shown in Fig. 1, the initial pose features (body joints) estimated from ModelGait are defined in the image coordinate, which depends on observation views and hence induces inconsistency among different views even for the same pose in 3D. More specifically, the first three dimensions of the 72-dimensional pose parameter $\boldsymbol{\theta}$ correspond to the 3D root rotation which rotates the remaining 69-dimensional pose parameters defined in a default coordinate (call it a human-centered coordinate later) into those in the input image coordinate. We therefore simply replace the 3D root rotation with a pre-defined fixed one (typically, no rotation) to get pose features in the unified human-centered coordinate (see Fig. 1), which are

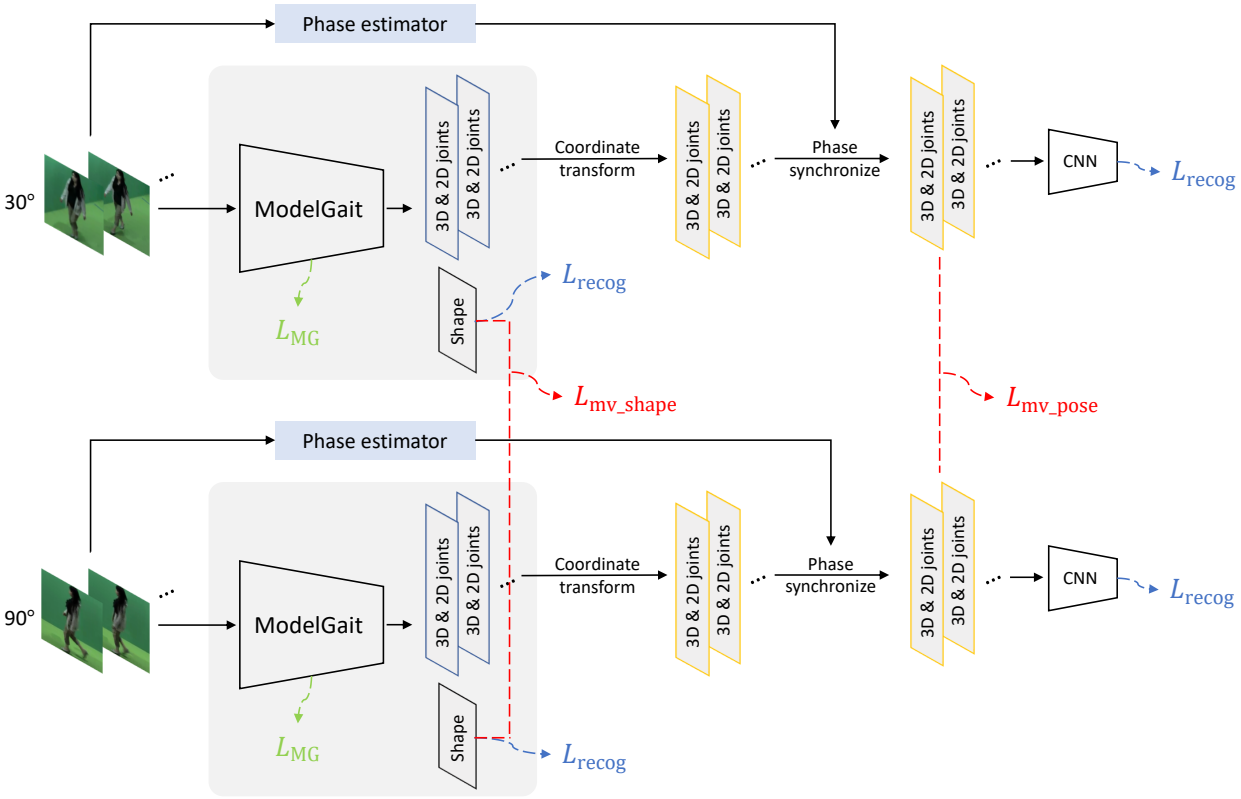

Figure 2. Overall framework of the proposed MvModelGait in the training. The gray-background part indicates the backbone network, ModelGait, which independently extracts initial pose (body joints) and shape features for multi-view RGB input sequences. The blue-background part indicates the phase estimator, which estimates the phase information of the input sequences for the latter synchronization. L_{MG}, $L_{\text{mv\_pose}}$ and $L_{\text{mv\_shape}}$ are the losses to ensure the unified pose and shape feature extraction from multi-view sequences. L_{recog} is the recognition loss computed based on the shape/pose features. Networks parameters are shared for different view angles.

unaffected by the view variation and therefore more suitable for the view consistency.

Phase estimation and synchronization. Because we cannot directly impose multi-view consistency on asynchronous multi-view gait sequences, we first apply a phase synchronization step. Since the ground-truth phase labels of the training set are generally unavailable and manual phase annotation is costly, we employ a baseline algorithm [38] for phase estimation as a pre-processing step in both training and test stages. In addition, we assume that each subject has silhouette sequence whose length is the same as his/her gait cycle.

At first, a certain subject in the training set is chosen as a standard one for phase estimation, and a silhouette sequence whose gait cycle is T_{std} frames, is up-sampled so that a new gait cycle has more frames (let it be $T_{\text{up}}(> T_{\text{std}})$) by interpolation with free-from deformation-based geometric transformation [39]. We assign a phase label i to the i-th up-sampled silhouette $I_{\text{up}}(i)$, and define a set of phase labels as $\mathcal{P}_{\text{up}} = \{0, 1, \ldots, T_{\text{up}} - 1\}$.

We then assume another non-standard subject whose gait period is T frames ($T < T_{\text{up}}$) and whose j-th silhouette is $I(j)$. If a phase (gait stance) of the initial frame ($j = 0$) of the non-standard subject corresponds to the phase s of the standard subject in the up-sampled domain, a phase of the j-th frame of the non-standard subject corresponds to the phase $(s + \lfloor dj \rfloor) \bmod T_{\text{up}}$ of the standard subject, where $d = T_{\text{up}}/T$. Based on the baseline algorithm [38], we obtain the optimal phase s^* by minimizing the sum of silhouettes differences over frames as

$$s^* = \arg\min_{s} \sum_{j=0}^{T-1} ||I_{\text{up}}((s + \lfloor dj \rfloor) \bmod T_{\text{up}}) - I(j)||^2. \tag{3}$$

After obtaining the phase label of the input sequences, we synchronize all body joints by linear interpolation to make them temporally consistent with a pre-defined canonical phase label with T frames in the up-sampled domain (i.e., a set of labels $\{\lfloor dj \rfloor \mid j = 0, \ldots, T-1\}$).

Loss function. Suppose that we have gait sequences from N_V different views with T frames for a certain subject, and that their corresponding synchronized body joint from n-th view at the j-th frame is $\boldsymbol{p}_j^n \in \mathbb{R}^M (j = 0, \ldots, T -$

$1; n = 1, \ldots, N_V)$, where $\boldsymbol{p}_j^n$ is an M-dimensional concatenated 1D vector of synchronized 3D and 2D body joint locations, the synchronized multi-view pose constraint is defined as the following loss function

$$L_{\text{mv\_pose}} = \frac{2}{N_V(N_V-1)T} \sum_{m=1}^{N_V} \sum_{n>m}^{N_V} \sum_{i=0}^{T-1} \|\boldsymbol{p}_j^m - \boldsymbol{p}_j^n\|_2^2. \tag{4}$$

3.4. Recognition

Given a mini-batch of sequences that consists of P subjects and each subject has K sequences from different view angles (i.e., PK sequences in total), we first extract a 10-dimensional SMPL shape parameter $\boldsymbol{f}_{\text{shape}}^i = \bar{\boldsymbol{\beta}}^i \in \mathbb{R}^{10}$ averaged over frames of the i-th gait sequence ($i = 1, \ldots, PK$) as a shape feature. We then extract a synchronized pose sequence $P^i = [\boldsymbol{p}_0^i, \ldots, \boldsymbol{p}_{T-1}^i] \in \mathbb{R}^{M \times T}$ from the i-th gait sequence. The synchronized pose sequence P^i is further fed into a CNN to extract more discriminative spatio-temporal feature in the same way as [27, 24]. The CNN architecture is the same as that used in [24], which is also shown in Table 1. After going through the CNN, a 52-dimensional output vector is used as a final pose feature $\boldsymbol{f}_{\text{pose}}^i \in \mathbb{R}^{52}$ for recognition.

Gait recognition has two different tasks: gait identification and verification, thus we use different loss functions $L_{\text{recog}} \in \{L_{\text{trip}}, L_{\text{cont}}\}$ for different tasks similar to [43]. Regarding gait identification, all triplet sequences $(Query, Genuine, Imposter)$ from the mini-batch are selected, and their corresponding features are fed into a triplet loss function [46] , which is defined as,

$$L_{\text{trip}} = \frac{1}{N} \sum_{n=1}^{N} \max(m - D_{\text{same}}^n + D_{\text{diff}}^n, 0), \tag{5}$$

where m is a pre-defined margin, $D_{\text{same}}^n = \|\boldsymbol{f}_Q^n - \boldsymbol{f}_G^n\|_2^2$ and $D_{\text{same}}^n = \|\boldsymbol{f}_Q^n - \boldsymbol{f}_I^n\|_2^2$ are the squared L2 distances of the same and different subject pairs in the n-th triplet, respectively.

Regarding gait verification, all pair sequences $(Probe, Gallery)$ from the mini-batch are selected, and their corresponding features and labels are fed into a contrastive loss function [12] , which is defined as,

$$L_{\text{cont}} = \frac{1}{N_s} \sum_{i=1}^{N_{\text{pair}}} y_n D^n + \frac{1}{N_d} \sum_{i=1}^{N_{\text{pair}}} (1 - y_n) \max(m - D^n, 0), \tag{6}$$

where m is a pre-defined margin, N_s, N_d are the number of same and different subject pairs, and $D^n = \|\boldsymbol{f}_P^n - \boldsymbol{f}_G^n\|_2^2$ is the squared L2 distances of the n-th pair, y_n is the corresponding label (if the pair is from the same subject, $y_n = 1$; otherwise, $y_n = 0$).

Table 1. The architecture of the CNN for pose features. Each convolutional layer is followed by a batch normalization layer and ReLU activation function.

Layers	Filters	Stride	Output Size
Conv1	64×3×3	(2, 1)	$\frac{M}{2} \times T$
Conv2	128×3×3	(2, 1)	$\frac{M}{4} \times T$
Conv3	256×3×3	(2, 1)	$\frac{M}{8} \times T$
FC	52	-	52

3.5. Training and inferring

The whole network is trained in an end-to-end manner with a weighted linear sum of the aforementioned losses as

$$\begin{aligned} L_{\text{total}} = L_{\text{MG}} &+ \lambda_{\text{mv\_shape}} L_{\text{mv\_shape}} + \lambda_{\text{mv\_pose}} L_{\text{mv\_pose}} \\ &+ \lambda_{\text{recog}} L_{\text{recog}}, \end{aligned} \tag{7}$$

where $\lambda_{\text{mv\_shape}}$, $\lambda_{\text{mv\_pose}}$, and λ_{recog} are the weight parameters, and L_{recog} is chosen based on the recognition tasks.

When inferring for a test case, according to the feature (pose or shape) and tasks to be evaluated, we apply the corresponding trained model to extract the required feature $\boldsymbol{f}$ for the input RGB sequences, then compute the L2 distance between $\boldsymbol{f}$ for two sequences as a dissimilarity score for matching.

4. Experiment

4.1. Datasets

We evaluated the proposed method with two publicly available datasets: OU-MVLP and CASIA-B.

OU-MVLP [43] is the largest database with various view variations at present. It contains 10,307 subjects with 14 view angles (0°, 15°, …, 90°; 180°, 195°, …, 270°). Each subject has two sequences with normal walking status per view. We followed the protocol as [43], and set 5,153 subjects for training, while the rest 5,154 subjects for testing.

CASIA-B [53] is one of the most frequently used gait databases. It contains 124 subjects with 11 view angles (0°, 18°, …, 180°). Each subject has 10 sequences with three walking conditions per view, which are divided into six normal walking (NM), two bag carrying (BG) and two coat wearing (CL) sequences. We followed the one of the protocols as [47], and set the first 74 subjects for training, while the rest 50 subjects for testing. For evaluation, NM #1–4 were assigned as galleries, and the other six were divided into three probes: NM #5–6, BG #1–2, and CL #1–2.

4.2. Implementation details

Input data. The cropped RGB sequences and their corresponding silhouette sequences in a gait period were the required input data for training. We followed the methods

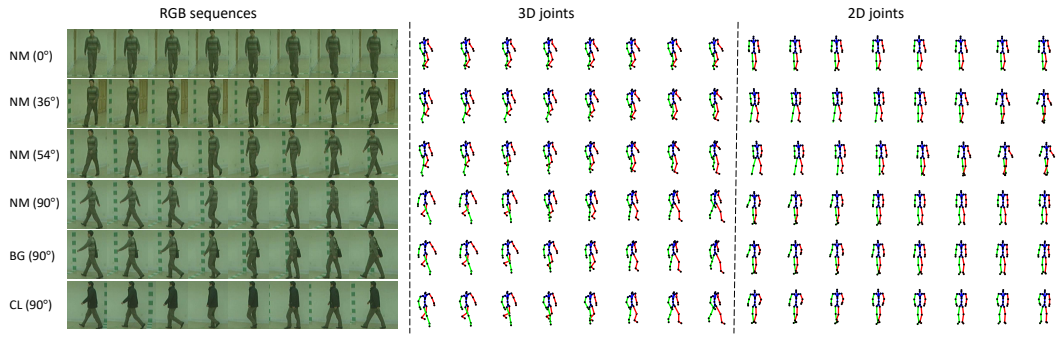

Figure 3. Visualization of pose estimation results under different views and walking conditions on CASIA-B. The left part shows the synchronized RGB input sequences of the half gait period, the middle and right parts are the estimated synchronized 3D and 2D joints in the human-centered coordinate, respectively.

in [24] to obtain these data. The RGB and silhouette sequences were then scaled to 224×224 and 64×64 for HMR regression and differentiable renderer, respectively. The temporal up-sampled frame number T_{up} was set to 150 for fine-grained phase estimation. The frame number T in a period was set to 15 to save the memory. Thus, we generally chose 15 frames at equal intervals from the real gait period, while ignoring those with less than 15 frames. The dimension M for the pose feature was set to 120 as we chose 24 joints (23 joints + 1 root joint) and each of them has 5D (3D + 2D) locations.

Training detail and parameters. The proposed MvModelGait has the same number of network parameters as ModelGait [24], and shares a similar training strategy. First, the HMR was initialized with a pre-trained model in [19], while other layers was initialized with default values. Then, the optimizer Adam [20] was initialized with the learning rate 10^{-4}. After a certain number of iterations, the learning rate was reduced by 10 times. For OU-MVLP, our model ran for a total of 60K iterations and the learning rate was reduced at 30K. For CASIA-B, our model ran for a total of 15K iterations and the learning rate was reduced at 10K.

In a mini-batch, (P, K) was set to (8, 8). The margin m in Eqs. (5) and (6) was set to 0.2. The hyper-parameters in Eqs. (1) and (7) were basically set to 1, except for $\lambda_{\text{joints}} = 100$ and $\lambda_{\text{mv\_pose}} = 0.01$.

Evaluation metrics. The rank-1 identification rate (denoted as Rank-1) and the equal error rate (denoted as EER) were used for performance evaluation on identification and verification tasks, respectively.

4.3. Pose visualization

Figure 1 shows the comparison of pose estimation results between the backbone alone (i.e., ModelGait [24]) and the proposed method. For RGB input sequences from four view angles with a similar pose, ModelGait estimates the 3D and 2D body joints in the image-based coordinate, which is variant to the view variation and causes a lot of difficulties for recognition. Even if they are transformed to the unified human-centered coordinate, there still exists relatively large differences. On the other hand, the proposed method could make the most of the multi-view pose constraint so as to narrow the differences and generate more consistent joints.

In addition, we visualize the cases under different views and walking conditions in Fig. 3. The RGB input sequences are coarsely selected frames from nearly a half gait period (8 frames). The 3D and 2D joints are synchronized to be temporally consistent with a pre-defined canonical phase label, which starts with a double-support phase. Note that there might exist discretization error between actual phases for the RGB input sequences and assigned discrete phases for the 3D/2D joints (chosen from T_{up} discrete phases). This figure shows the proposed method could well handle different walking conditions (NM, BG or CL) and estimate consistent body joints of the same subject. Regarding different views, the body joints are successfully transformed into the unified human-centered coordinate, which are more suitable for the pose constraint and recognition.

Regarding the difficult case (0°vs. 90°), the estimated 3D joints show some bias, especially for the stride length. This may because the initial 3D pose estimation of the proposed method is still dependent on the input view to some extent, thus the ill-pose problem in the 3D pose estimation cannot be completely solved only by the proposed multi-view pose constraint. Despite this, the 2D joints are much more similar and may occupy more weights for recognition.

4.4. Comparison with the state-of-the arts

OU-MVLP. Table 2 shows the comparison of the proposed method and the baseline ModelGait [24] based on the angular differences of four typical views (i.e., 0°, 30°, 60°, 90°). We also introduce an ensemble version of pose and shape features by averaging their dissimilarity scores.

From the results, the proposed method shows large improvement on the pose feature, i.e., 3.8% higher Rank-1 identification rate and 0.12% lower EER, and the performance degradation from 0°to 90°angular difference is also reduced from 23.8% to 16.4%, which indicates the effec-

Table 2. Rank-1 identification rates and EERs of our method and ModelGait [24] on OU-MVLP based on the angular differences. Models are trained using all view angles, while tested based on the angular differences of four typical views (i.e., 0°, 30°, 60°, 90°). Bold indicates the best accuracy. This convention is consistent throughout this paper.

Methods	Rank-1 [%]					EER [%]				
	Angular difference					Angular difference				
	0°	30°	60°	90°	Mean	0°	30°	60°	90°	Mean
ModelGait (pose) [24]	98.3	91.9	81.9	74.5	88.8	0.19	0.30	0.41	0.58	0.33
Ours (pose)	**98.9**	**94.7**	**88.0**	**82.5**	**92.6**	**0.13**	**0.20**	**0.25**	**0.31**	**0.21**
ModelGait (shape) [24]	**99.6**	97.7	95.0	91.4	96.7	0.12	0.19	0.23	**0.29**	0.19
Ours (shape)	99.5	**97.8**	**95.3**	**91.8**	**96.8**	**0.11**	**0.17**	**0.21**	0.31	**0.18**
ModelGait (ensemble) [24]	99.5	**98.2**	95.2	91.3	96.9	0.12	0.19	0.14	0.28	0.17
Ours (ensemble)	**99.6**	95.8	**99.0**	**93.4**	**97.3**	**0.10**	**0.14**	**0.11**	**0.19**	**0.13**

Table 3. Rank-1 identification rates and EERs for each probe view averaged over the 14 gallery views on OU-MVLP, where the identical view is excluded. "-" means not provided. The upper and lower blocks of Rank-1 identification rates are the results without and with non-enrolled probes, respectively.

	Methods	Probe view														
		0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	Mean
Rank-1 [%]	PTSN-O [1]	6.4	11.0	15.4	18.8	17.6	15.1	8.8	5.2	10.6	10.5	17.3	14.6	11.6	7.7	12.2
	PTSN-α [1]	11.8	19.0	23.9	26.5	24.9	20.6	14.7	6.1	11.6	14.2	22.1	21.3	17.9	14.3	17.8
	GaitSet [6]	84.7	93.6	96.7	96.7	93.6	95.3	94.2	86.9	92.8	96.0	96.1	93.0	94.5	92.8	93.3
	ACL [54]	74.0	88.3	94.6	95.4	88.0	91.3	90.0	76.7	89.5	95.0	94.9	88.0	90.8	89.8	89.0
	GaitPart [9]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	95.1
	GLN [16]	89.3	95.8	**97.9**	**97.8**	**96.0**	96.7	96.1	90.7	95.3	**97.7**	**97.5**	95.7	96.2	95.3	95.6
	ModelGait [24]	92.8	96.2	96.8	96.3	94.7	96.6	96.6	93.5	95.4	96.3	96.7	96.5	96.5	96.2	95.8
	Ours	**93.5**	**96.5**	97.1	96.9	95.7	**96.8**	**97.1**	**93.7**	**95.6**	96.6	97.0	**97.1**	**97.1**	**97.0**	**96.2**
	GaitSet [6]	79.5	87.9	89.9	90.2	88.1	88.7	87.8	81.7	86.7	89.0	89.3	87.2	87.8	86.2	87.1
	GaitPart [9]	82.6	88.9	90.8	91.0	89.7	89.9	89.5	85.2	88.1	90.0	90.1	89.0	89.1	88.2	88.7
	GLN [16]	83.8	**90.0**	91.0	**91.2**	**90.3**	90.0	89.4	85.3	89.1	**90.5**	**90.6**	89.6	89.3	88.5	89.2
	Ours	**87.7**	89.7	**91.1**	90.1	89.8	**90.3**	**90.3**	**88.1**	**89.4**	89.4	90.0	**90.8**	**90.0**	**89.7**	**89.7**
EER [%]	PTSN-O [1]	16.0	13.3	13.0	11.2	11.6	12.8	17.1	17.6	14.9	18.9	12.3	13.2	14.9	18.2	14.6
	PTSN-α [1]	15.1	12.5	11.9	11.1	11.2	12.5	14.8	22.2	17.8	21.3	11.8	11.9	13.0	14.7	14.4
	GaitSet [6]	1.45	0.93	0.76	0.75	0.99	0.79	0.86	2.80	1.61	1.53	2.20	1.83	1.15	1.00	1.33
	ModelGait [24]	0.34	0.34	0.20	0.18	0.31	0.26	0.17	0.28	0.28	0.36	0.34	0.21	0.20	0.20	0.26
	Ours	**0.29**	**0.29**	**0.18**	**0.14**	**0.24**	**0.23**	**0.15**	**0.24**	**0.22**	**0.27**	**0.24**	**0.18**	**0.17**	**0.17**	**0.21**

tiveness of the proposed synchronized multi-view pose constraint in solving the view covariate. For the shape feature, the proposed method achieves similar performance as the baseline because the proposed synchronized multi-view pose constraint has almost no effects on it. Whereas for the ensemble, the proposed method achieves better results owing to the improvement of the pose feature.

Table 3 shows the comparison of the proposed method and other state-of-the-arts. Because probe sets contain some subjects that are not in the gallery, we provide both results with and without non-enrolled probes for rank-1 identification rates. The results of ModelGait and the proposed method are from the ensemble ones, while those of individual pose and shape features are reported in the supplementary material. From the results, the proposed method achieves the best performance for both scenarios.

CASIA-B. Table 4 shows the comparison of the proposed method and other state-of-the-arts. For three different settings (i.e., NM, BG and CL), the proposed method was trained on three different training set including the corresponding probe and gallery sets. Similar to OU-MVLP, the results of ModelGait and the proposed method are from the ensemble ones, while those of individual pose and shape features are reported in the supplementary material. For NM setting, the proposed method achieves the best rank-1 identification rate of 98.1%. For BG and CL settings, the proposed method yields the second best rank-1 identification rates of 93.4% and 80.7%, respectively. This may because the influence of carrying and clothing changes on the SMPL model estimation is relatively greater than the normal walking status. Note that the proposed method is still comparable with the best benchmarks (e.g., 0.6% lower than GLN [16] for BG, 0.8% lower than MT3D [28] for CL), and the two benchmarks are worse than the proposed method for the other two settings. On average over the three settings, the proposed method achieves the best accuracy.

4.5. Ablation study

The proposed method has two important components: synchronized multi-view pose constraint in the training

Table 4. Rank-1 identification rates [%] of comparison methods on CASIA-B using the first 74 subjects for training. The mean result over all 10 gallery views for each probe view is given, where the identical view is excluded.

Probe	Methods	Probe view											
		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
NM #5–6	ViDP [17]	-	-	-	64.2	-	60.4	-	65.0	-	-	-	-
	CNN ensemble [47]	88.7	95.1	98.2	96.4	94.1	91.5	93.9	97.5	98.4	95.8	85.6	94.1
	Takemura's [43]	83.2	91.2	95.8	93.4	91.2	87.8	89.4	93.6	96.0	95.8	81.6	90.8
	PSTN [48]	87.0	93.8	96.2	94.4	92.2	91.8	92.0	95.0	96.0	96.4	84.8	92.7
	Song's GaitNet [42]	75.6	91.3	91.2	92.9	92.5	91.0	91.8	93.8	92.9	94.1	81.9	89.9
	Zhang's GaitNet [55]	91.2	92.0	90.5	95.6	86.9	92.6	93.5	96.0	90.9	88.8	89.0	91.6
	GaitSet [6]	90.8	97.9	99.4	96.9	93.6	91.7	95.0	97.8	98.9	96.8	85.8	95.0
	GaitPart [9]	94.1	98.6	99.3	98.5	94.0	92.3	95.9	98.4	99.2	97.8	90.4	96.2
	GLN [16]	93.2	**99.3**	99.5	98.7	96.1	95.6	97.2	98.1	**99.3**	**98.6**	90.1	96.9
	ACL [54]	92.0	98.5	**100.0**	**98.9**	95.7	91.5	94.5	97.7	98.4	96.7	91.9	96.0
	MT3D [28]	95.7	98.2	99.0	97.5	95.1	93.9	96.1	**98.6**	99.2	98.2	92.0	96.7
	PoseGait [27]	55.3	69.6	73.9	75.0	68.0	68.2	71.1	72.9	76.1	70.4	55.4	68.7
	ModelGait [24]	96.9	97.1	98.5	98.4	**97.7**	98.2	97.6	97.6	98.0	98.4	**98.6**	97.9
	Ours	**97.5**	97.6	98.6	98.8	**97.7**	**98.9**	**98.9**	97.3	97.6	97.8	97.9	**98.1**
BG #1–2	LB [47]	64.2	80.6	82.7	76.9	64.8	63.1	68.0	76.9	82.2	75.4	61.3	72.4
	Zhang's GaitNet [55]	83.0	87.8	88.3	93.3	82.6	74.8	89.5	91.0	86.1	81.2	85.6	85.7
	GaitSet [6]	83.8	91.2	91.8	88.8	83.3	81.0	84.1	90.0	92.2	94.4	79.0	87.2
	GaitPart [9]	89.1	94.8	96.7	95.1	88.3	**94.9**	89.0	93.5	96.1	93.8	85.8	91.5
	GLN [16]	91.1	**97.7**	**97.8**	**95.2**	92.5	91.2	92.4	**96.0**	**97.5**	95.0	88.1	**94.0**
	MT3D [28]	91.0	95.4	97.5	94.2	92.3	86.9	91.2	95.6	97.3	**96.4**	86.6	93.0
	PoseGait [27]	35.3	47.2	52.4	46.9	45.5	43.9	46.1	48.1	49.4	43.6	31.1	44.5
	ModelGait [24]	**94.8**	92.9	93.8	94.5	93.1	92.6	94.0	94.5	89.7	93.6	90.4	93.1
	Ours	93.9	92.5	92.9	94.1	**93.4**	93.4	**95.0**	94.7	92.9	93.1	**92.1**	93.4
CL #1–2	LB [47]	37.7	57.2	66.6	61.1	55.2	54.6	55.2	59.1	58.9	48.8	39.4	54.0
	Zhang's GaitNet [55]	42.1	58.2	65.1	70.7	68.0	70.6	65.3	69.4	51.5	50.1	36.6	58.9
	GaitSet [6]	61.4	75.4	80.7	77.3	72.1	70.1	71.5	73.5	73.5	68.4	50.0	70.4
	GaitPart [9]	70.7	85.5	86.9	83.3	77.1	72.5	76.9	82.2	83.8	80.2	66.5	78.7
	GLN [16]	70.6	82.4	85.2	82.7	79.2	76.4	76.2	78.9	77.9	78.7	64.3	77.5
	MT3D [28]	76.0	**87.6**	**89.8**	85.0	81.2	75.7	**81.0**	**84.5**	**85.4**	**82.2**	68.1	**81.5**
	PoseGait [27]	24.3	29.7	41.3	38.8	38.2	38.5	41.6	44.9	42.2	33.4	22.5	36.0
	ModelGait [24]	**78.2**	81.0	82.1	82.8	80.3	76.9	75.5	77.4	72.3	73.5	74.2	77.6
	Ours	77.0	80.0	83.5	**86.1**	**84.5**	**84.9**	80.6	80.4	77.4	76.6	**76.9**	80.7

phase and temporally synchronized pose features in the test phase. We analysis the contributions of each component and provide their respective results in Table 5. The pose feature was used to conduct the ablation study since the effect of these two components is larger on the pose than on the shape. As results, it turns out that both components contribute to the proposed method, and the best accuracy is achieved when both of them are employed.

Table 5. Ablation study using pose feature under different walking conditions on CASIA-B. The mean Rank-1 rate [%] of all cross-view results is presented. "TSP" means temporally synchronized pose features.

Methods	NM	BG	CL
Ours (w/o $L_{\text{mv\_pose}}$)	91.8	86.2	60.7
Ours (w/o TSP)	91.5	86.8	61.8
Ours	**93.1**	**88.0**	**64.3**

5. Conclusion

This paper describes an end-to-end model-based gait recognition approach using the synchronized multi-view pose constraint. Specifically, we make the most of asynchronous multi-view gait sequences in the training phase so that we can generate more view-consistent pose sequence from arbitrary single-view gait sequence in the training phase. As a direct benefit from the synchronization, temporally aligned pose features are employed not only in the training phase but in the test phase to further improve the accuracy. Experimental results show that the proposed method outperforms other state-of-the-art methods.

Because pose and shape features are separately treated in the training and test phases in this work, the integration of both features is worth exploring in the future. We will also replace the baseline algorithm for phase synchronization with deep learning-based phase sequence estimator.

Acknowledgment This work was supported by JSPS KAKENHI Grant No. JP19H05692 and JP20H00607, MEXT "Innovation Platform for Society 5.0" Program Grant Number JPMXP0518071489.

References

[1] W. An, S. Yu, Y. Makihara, X. Wu, C. Xu, Y. Yu, R. Liao, and Y. Yagi. Performance evaluation of model-based gait on multi-view very large population database with pose sequences. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(4):421–430, 2020. 7

[2] G. Ariyanto and M.S. Nixon. Marionette mass-spring model for 3d gait biometrics. In *Proc. of the 5th IAPR International Conference on Biometrics*, pages 354–359, March 2012. 2

[3] A.F. Bobick and A.Y. Johnson. Gait recognition using static activity-specific parameters. In *CVPR*, volume 1, pages 423–430, 2001. 2

[4] I. Bouchrika, M. Goffredo, J. Carter, and M. Nixon. On using gait in forensic biometrics. *Journal of Forensic Sciences*, 56(4):882–889, 2011. 1

[5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*, 2018. 2

[6] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *AAAI*, 2019. 2, 7, 8

[7] Patrick Connor and Arun Ross. Biometric recognition by gait: A survey of modalities and features. *Computer Vision and Image Understanding*, 167:1–27, 2018. 1

[8] D. Cunado, M.S. Nixon, and J.N. Carter. Automatic extraction and description of human gait models for recognition purposes. *Computer Vision and Image Understanding*, 90(1):1–41, 2003. 2

[9] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. Gaitpart: Temporal part-based model for gait recognition. In *CVPR*, June 2020. 2, 7, 8

[10] Yang Feng, Yuncheng Li, and Jiebo Luo. Learning effective gait features using lstm. In *ICPR*, pages 325–330, 2016. 2

[11] Y. Guan, C. T. Li, and F. Roli. 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. 2

[12] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, volume 2, pages 1735–1742, 2006. 5

[13] J. Han and B. Bhanu. Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):316– 322, 2006. 1, 2

[14] Y. He, J. Zhang, H. Shan, and L. Wang. Multi-task gans for view-specific feature learning in gait recognition. *IEEE Transactions on Information Forensics and Security*, 14(1):102–113, Jan 2019. 2

[15] Yoshitaka Ushiku Hiroharu Kato and Tatsuya Harada. Neural 3d mesh renderer. In *CVPR*, 2018. 3

[16] Saihui Hou, Chunshui Cao, Xu Liu, and Yongzhen Huang. Gait lateral network: Learning discriminative and compact representations for gait recognition. In *ECCV*, pages 382–398, Cham, 2020. Springer International Publishing. 2, 7, 8

[17] M. Hu, Y. Wang, Z. Zhang, J. J. Little, and D. Huang. View-invariant discriminative projection for multi-view gait-based human identification. *IEEE Transactions on Information Forensics and Security*, 8(12):2034–2045, 2013. 8

[18] H Iwama, D. Muramatsu, Y. Makihara, and Y. Yagi. Gait verification system for criminal investigation. *IPSJ Transactions on Computer Vision and Applications*, 5:163–175, Oct. 2013. 1

[19] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *CVPR*, pages 7122–7131, 2018. 2, 3, 6

[20] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint*, 2014. 6

[21] X. Li, Y. Makihara, C. Xu, Y. Yagi, and M. Ren. Joint intensity transformer network for gait recognition robust against clothing and carrying status. *IEEE Transactions on Information Forensics and Security*, pages 1–1, 2019. 2

[22] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, and Mingwu Ren. Gait recognition invariant to carried objects using alpha blending generative adversarial networks. *Pattern Recognition*, 105:107376, 2020. 2

[23] X. Li, Y. Makihara, C. Xu, Y. Yagi, and M. Ren. Gait recognition via semi-supervised disentangled representation learning to identity and covariate features. In *CVPR*, June 2020. 2

[24] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, Shiqi Yu, and Mingwu Ren. End-to-end model-based gait recognition. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020. 1, 2, 3, 5, 6, 7, 8

[25] Junbang Liang and Ming Lin. Shape-aware human pose and shape reconstruction using multi-view images. In *ICCV*, 08 2019. 2, 3

[26] Rijun Liao, Chunshui Cao, Edel B. Garcia, Shiqi Yu, and Yongzhen Huang. Pose-based temporal-spatial network (ptsn) for gait recognition with carrying and clothing variations. In *Proceedings of the 12th Chinese Conference on Biometric Recognition*, pages 474–483, 2017. 2

[27] Rijun Liao, Shiqi Yu, Weizhi An, and Yongzhen Huang. A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognition*, 98:107069, 2020. 2, 5, 8

[28] Beibei Lin, Shunli Zhang, and Feng Bao. Gait recognition with multiple-temporal-scale 3d convolutional neural network. In *MM '20: The 28th ACM International Conference on Multimedia*, 2020. 2, 7, 8

[29] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 2, 3

[30] Niels Lynnerup and Peter Kastmand Larsen. Gait as evidence. *IET Biometrics*, 3(2):47–54, 6 2014. 1

[31] Yasushi Makihara, Darko S. Matovski, Mark S. Nixon, John N. Carter, and Yasushi Yagi. *Gait Recognition: Databases, Representations, and Applications*, pages 1–15. John Wiley & Sons, Inc., Jun. 2015. 1

[32] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi. Gait recognition using a view transformation model

in the frequency domain. In *ECCV*, pages 151–163, Graz, Austria, May 2006. 2

[33] Y. Makihara, A. Suzuki, D. Muramatsu, X. Li, and Y. Yagi. Joint intensity and spatial metric learning for robust gait recognition. In *CVPR*, pages 5705–5715, Jul. 2017. 2

[34] Yasushi Makihara, Akira Tsuji, and Yasushi Yagi. Silhouette transformation based on walking speed for gait identification. In *CVPR*, San Francisco, CA, USA, Jun 2010. 2

[35] D. Muramatsu, A. Shiraishi, Y. Makihara, M.Z. Uddin, and Y. Yagi. Gait-based person recognition using arbitrary view transformation model. *IEEE Trans. on Image Processing*, 24(1):140–154, Jan 2015. 2

[36] Mark S. Nixon, Tieniu N. Tan, and Rama Chellappa. *Human Identification Based on Gait*. Int. Series on Biometrics. Springer-Verlag, Dec. 2005. 1

[37] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *CVPR*, June 2018. 3

[38] S. Sarkar, J.P. Phillips, Z. Liu, I.R. Vega, P. Gro ther, and K.W. Bowyer. The humanid gait challenge problem: Data sets, performance, and analysis. *IEEE Trans. of Pattern Analysis and Machine Intelligence*, 27(2):162–177, Feb. 2005. 2, 4

[39] Thomas W. Sederberg and Scott R. Parry. Free-form deformation of solid geometric models. *SIGGRAPH Comput. Graph.*, 20(4):151–160, Aug. 1986. 4

[40] Soyong Shin and Eni Halilaj. Multi-view human pose and shape estimation using learnable volumetric aggregation. In *arXiv preprint arXiv:2011.13427*, 11 2020. 2, 3

[41] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi. Geinet: View-invariant gait recognition using a convolutional neural network. In *ICB*, Jun. 2016. 2

[42] Chunfeng Song, Yongzhen Huang, Yan Huang, Ning Jia, and Liang Wang. Gaitnet: An end-to-end network for gait based human identification. *Pattern Recognition*, 96:106988, 2019. 8

[43] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi. On input/output architectures for convolutional neural network-based cross-view gait recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(9):2708–2719, 2019. 2, 5, 8

[44] D.K. Wagg and M.S. Nixon. On automated model-based extraction and analysis of gait. In *Proc. of the 6th IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pages 11–16, 2004. 2

[45] Chen Wang, Junping Zhang, Jian Pu, Xiaoru Yuan, and Liang Wang. Chrono-gait image: A novel temporal template for gait recognition. In *Proc. of the 11th European Conf. on Computer Vision*, pages 257–270, Heraklion, Crete, Greece, 2010. 2

[46] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *CVPR*, June 2014. 5

[47] 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 Transactions on Pattern Analysis and Machine Intelligence*, 39(2):209–226, Feb 2017. 2, 5, 8

[48] C. Xu, Y. Makihara, X. Li, Y. Yagi, and J. Lu. Cross-view gait recognition using pairwise spatial transformer networks. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2020. 2, 8

[49] Chi Xu, Yasushi Makihara, Xiang Li, Yasushi Yagi, and Jianfeng Lu. Gait recognition from a single image using a phase-aware gait cycle reconstruction network. In *ECCV*, pages 386–403, Cham, 2020. Springer International Publishing. 2

[50] Dong Xu, Shuicheng Yan, Dacheng Tao, Lei Zhang, Xuelong Li, and Hong jiang Zhang. Human gait recognition with matrix representation. *IEEE Trans. Circuits Syst. Video Technol*, 16(7):896–903, 2006. 2

[51] C. Yam, M.S. Nixon, and J.N. Carter. Automated person recognition by walking and running via model-based approaches. *Pattern Recognition*, 37(5):1057–1072, 2004. 2

[52] Shiqi Yu, Rijun Liao, Weizhi An, Haifeng Chen, Edel B. Garcia, Yongzhen Huang, and Norman Poh. Gaitganv2: Invariant gait feature extraction using generative adversarial networks. *Pattern Recognition*, 87:179 – 189, 2019. 2

[53] S. Yu, D. Tan, and T. Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *ICPR*, volume 4, pages 441–444, Hong Kong, China, Aug. 2006. 2, 5

[54] Y. Zhang, Y. Huang, S. Yu, and L. Wang. Cross-view gait recognition by discriminative feature learning. *IEEE Transactions on Image Processing*, 29:1001–1015, 2020. 2, 7, 8

[55] Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Jian Wan, Nanxin Wang, and Xiaoming Liu. Gait recognition via disentangled representation learning. In *CVPR*, Long Beach, CA, June 2019. 2, 8