Real-Time Gait-Based Age Estimation and Gender Classification from a Single Image

Chi Xu[1,2] Yasushi Makihara[2] Ruochen Liao[2] Hirotaka Niitsuma[2] Xiang Li[1,2]
Yasushi Yagi[2] Jianfeng Lu[1]
[1] Nanjing University of Science and Technology, Nanjing, China [2] Osaka University, Osaka, Japan
{xu, makihara, liao, niitsuma, li, yagi}@am.sanken.osaka-u.ac.jp lujf@mail.njust.edu.cn

Abstract

In this paper, we propose a unified real-time framework for gait-based age estimation and gender classification that uses just a single image, which reduces the latency in video capturing compared with the existing methods based on a gait cycle. To cope with the problem of lacking motion information in the input single image, we first reconstruct a gait cycle of a silhouette sequence from the input image via a gait cycle reconstruction network. The reconstructed gait cycle is then fed into a state-of-the-art gait recognition network for feature representation learning, which is further used to obtain the class of the gender and the estimated probability distribution of integer age labels. Unlike the existing methods focusing on the gait sequences captured from the side view, the proposed method is applicable to the gait images from an arbitrary view with a single trained model, which is more suitable for real-world application scenarios (e.g., automatic access control). Stand-alone and client-server online systems were implemented based on the proposed method, which validates the real-time/online property in actual scenes. The experiments on the world's largest multi-view gait dataset demonstrate the effectiveness of the proposed method, which achieves performance improvement compared with the benchmark algorithms.

1. Introduction

Gait is a behavioral biometric that has unique advantages in the following aspects: it can be recognized without subject cooperation; it is effective even at a large distance from a camera with low-resolution images. Gait recognition has therefore been applied to surveillance, forensics, and criminal investigation by taking advantage of the CCTVs now widely installed in public areas [6, 15, 30].

Currently, most studies on gait analysis aim at person authentication and identification [38, 43, 11, 33, 48, 34, 21, 47, 7, 49, 23], i.e., hard biometrics. On the other hand, the soft biometrics, such as age estimation and gender classification, also own great application potential [51]. In addition to help with the surveillance applications as an additional cue (e.g., finding lost children/elderly), gait-based age estimation and gender classification can also be used for some commercial applications, for examples, an automatic access control for specific places or systems with age/gender limit, and a dialogue robot providing guidance and recommendation services in a shopping mall.

Most existing studies on gait-based age estimation [32, 27, 26, 28, 42, 41] and gender classification [13, 24, 55, 8, 36, 29, 56, 16] utilize the gait feature extracted from a gait cycle of a silhouette sequence, such as the gait energy image (GEI) [11]. Relying on a gait feature from a full gait cycle, however, results in the latency of capturing a video with the required time length (e.g., around one sec.), which is unfavorable for real-time online applications.

An direct solution to this problem is to reduce the required video length, such as estimation from just a single image, which has not been investigated to the best of knowledge. In this case, the time occupied by data capturing is greatly reduced, which is more suitable for the real-time systems. For example, a dialogue robot can know the age and gender of a customer in a shopping mall before he/she approaches, and hence can quickly change the guidance/recommendation mode and contents according to the estimation results. In the case of simultaneous processing of multiple subjects, which is often considered in the automatic access control, single image-based age estimation and gender classification may reduce the processing time by times, as well as solving the temporal partial occlusion possibly occurs when multiple subjects walking towards, which is a real-world challenging factor that affects the performance of gait cycle-based approaches.

Nevertheless, estimation of the age and gender using just a single image is pretty challenging because of lack of individual gait motion information. In fact, the age and gender-related gait patterns are reflected in terms of both body shape (e.g., larger head-to-body ratio for children, and

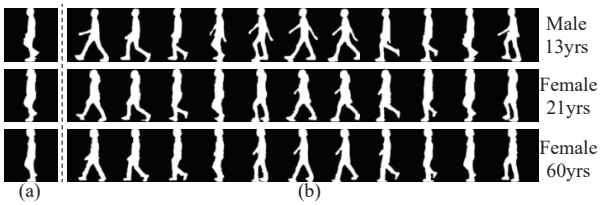

Figure 1. Comparison of age estimation and gender classification from (a) a single image, and (b) a full gait cycle. Each row shows the single image and gait cycle of the same subject, with his/her actual age and gender shown in the right. The single image does not contain any motion information, which makes age estimation and gender classification very difficult, whereas the age/gender-related characteristics appear in the full gait cycle (e.g., larger stride and arm swing for the first young male subject, and much smaller stride for the last elderly female subject) to help with the estimation.

middle-age spread and stoop for elderly) and temporal pose changes (e.g., larger stride length for youth, relatively small stride and arm swing for females compared with males) [32, 51], which are both contained in the gait cycle-based features, such as GEI. By contrast, only the body shape is visible in the captured single image, which may significantly affect the performance if we estimates the age and gender directly from the single image, as shown in Fig. 1(a). Since the motion patterns (e.g., stride) can never be observed in the input single-support phases, and the body shapes in these frames do not appear obvious age-related patterns (e.g., large head-to-body ratio, large stoop), it is very possible to obtain large estimation errors for these subjects if only this frame is considered, whereas a full gait cycle provides more motion characteristics for easier estimation (e.g., larger stride and arm swing for young male subject, and much smaller stride for elderly female subject), as shown in Fig. 1(b).

Therefore, it is reasonable to consider improving the performance of age estimation and gender classification by reconstructing a full gait cycle of a silhouette sequence from the given single image first, which has already been investigated in the field of single image-based gait recognition [50], and extremely low frame-rate gait recognition [2, 1, 3, 31]. The rationale behind the gait cycle reconstruction from a single image is that a snapshot of a single gait image in a natural gait sequence implies the motion (i.e., pose sequences) before or after this frame while maintaining the gait individuality [50]. Thus, the reconstructed gait cycle increases the individual gait motion information that can be used for the subsequent procedures.

We therefore propose a real-time end-to-end CNN framework for gait-based age estimation and gender classification from a single image, which first reconstructs a gait cycle with continuous pose changes from the input single image before the final estimation and classification tasks. More specifically, given a single gait image, a gait cycle reconstruction network first reconstructs a silhouette sequence of a gait cycle, which is further fed into a sequence-based gait recognition network for discriminative feature learning. Finally, the learned feature is used to obtain the estimated gender class and age label probability distribution. The contributions of this work are three-fold.

1. A single CNN model applicable to a gait image from an arbitrary view.

Unlike most studies on gait-based age estimation and gender classification focusing on the side view [32, 27, 26, 28, 42, 41, 24, 55], and a few works on gender classification that considered multiple views by applying view-dependent models [36, 10], the proposed method handles the gait image from an arbitrary view using a single CNN model without requiring the view information. This is more suitable for the real application scenarios as there are no restrictions on the observation view angle of the captured subjects.

2. Simultaneous age estimation and gender classification from a single gait image for the first time.

The proposed method uses a single image to conduct age estimation and gender classification simultaneously, which has not been studied to our knowledge. It enables the acquisition of estimated age and gender information without latency in gait video capture, and hence is more suitable for the real-time online system. The effectiveness of the proposed method is demonstrated through the experiments on the world's largest multi-view gait dataset, i.e., the OU-ISIR Gait Database, Multi-View Large Population Dataset (OU-MVLP) [39], which yields superior performance than the state-of-the-art approaches.

3. Implementation of online systems.

We implemented two online systems: a stand-alone system and a client-server system, based on the proposed method using a single video camera, which directly obtains the estimation results during the subjects' walking without extra requirements. The computational time of the stand-alone system was evaluated in an actual online environment, which validated its real-time property for possible application to automatic access control systems or dialogue robots. We also demonstrated that a client computer without a general-purpose GPU (e.g., a tablet computer) successfully ran our application online with a web API.

2. Related work

2.1. Gait-based age and gender analysis

Age estimation. Apart from some early works focusing on age group classification, most researches tackle age estimation, where approaches using different kinds of machine learning techniques and CNN frameworks were applied on appearance-based features (e.g., GEI). Compared with the traditional methods [32, 51, 28, 26, 20] that employed typical regression methods (e.g., Gaussian process regression [32]) with incorporated manifold analysis [28, 20], CNN-based

methods [40, 59, 22] achieved significantly promoted performance. For examples, Sakata et al. [40] utilized DenseNet for age value regression, and Zhu et al. [59] proposed a global and local CNN combining with an ordinal distribution regression.

Unlike above-mentioned methods that estimate a single age value, Sakata et al. [41] proposed a label distribution learning framework, which estimates a probability distribution (i.e., uncertainty) of the estimated age. This is more favorable for some real applications since it helps reduce the risk caused by estimation errors, for example, the robot can show broader recommendations for the target customer if the uncertainty is large.

However, most of the existing studies on gait-based age estimation focused on the gait images captured from the side view, which may limit their use in real-world applications.

Gender classification. Instead of fitting a human model by the model-based approaches [19, 53, 13, 45], appearance-based methods [24, 55, 12, 29] directly used the appearance-based features (e.g., GEI) without model fitting, which reduces the fitting errors and computational time. Several works [13, 8, 56, 9] fused the features from multiple views to gain more gait information for robust classification, and achieved better performance than aforementioned single view-based studies. To handle temporal partial occlusion in a gait cycle, Isaac et al. [14] classified the gender for each frame independently, and obtained the gender class of a sequence via majority voting. In [10], a real-time gender classification method was proposed using part of the gait cycle (i.e., setting as 15 frames), as well as a view-dependent classifier to reduce the effects of different views. Using part of the gait cycle, however, still requires the latency in video capturing, and the view estimation error may also affect the performance of view-dependent classification.

Multiple tasks. A few studies [37, 57, 58, 42] tackle age estimation, gender classification, and/or hard biometrics simultaneously. In [37], identity, gender and age were simultaneously output from a multi-task CNN, which adopts optical flow sequences as the input, and in [58], body shape parameters (i.e., BMI) was also obtained by inputting both GEI and silhouette images into a joint CNN. While Zhang et al. [57] proposed a deep CNN with multi-task learning for age estimation, which also integrates the gender information for performance improvement, Sakata et al. [42] proposed a multi-stage CNN to incorporate the tasks of gender and age group classification before final age regression.

However, the aforementioned single-task and multi-task researches all relied on a gait cycle or a sequence with a certain time length, which results in latency for data capture.

2.2. Gait cycle reconstruction

Reconstructing a gait cycle from a single image or a few images is often considered in the fields of gait recognition from a single image [50] and low frame-rate videos [31, 1, 2, 4, 5]. While a few works explored the direct reconstruction of a single GEI template [4, 5], most works chose to recover a silhouette sequence of a gait cycle instead [3, 31, 1, 2, 50], which contains more temporal information than a single GEI. Unlike the traditional manifold learning-based approaches [31, 1, 2] that only optimized the performance of reconstruction, an end-to-end CNN-based method [50] achieved superior recognition performance by simultaneous optimization of the phase-aware gait cycle reconstructor (PA-GCR) and following recognition network, which ensures the balance between reconstruction quality and recognition accuracy. We therefore employ PA-GCR as the module responsible for gait cycle reconstruction in our whole framework.

3. Proposed method

3.1. Overview

The proposed method automatically estimates the age and classifies the gender of a subject during his/her walking, as outlined in Fig. 2. After capturing a single gait image using a normal video camera, the silhouette image is first extracted by background subtraction-based graph-cut segmentation [35]. The size-normalized and gravity center-registered silhouette is then obtained [33] to be used as input for the proposed method.

In the training phase, we input a pair of silhouettes from the same subject with different phases (i.e., poses) to mitigate the possible intra-subject variations existed in the reconstructed gait cycles and the final estimation results. The parameters are shared among this parallel networks. A full gait cycle is first reconstructed from the input image via the PA-GCR [50], which is further fed into the state-of-the-art sequence-based gait recognition network (i.e., GaitSet [7]) to learn a discriminative feature representation. The learned feature is finally used to obtain the probability distribution of the gender labels and integer age labels, as well as the expectation calculated from the estimated age label distribution, which is regarded as the estimated age value. In the test phase, only a single image is required to output the gender class and estimated age.

3.2. Gait cycle reconstruction via PA-GCR

The PA-GCR is a gait cycle reconstruction module that takes a single silhouette image as the input and output a full gait cycle with specified phases. Four parts are contained in the PA-GCR, i.e., encoder, phase estimator, feature transformer, and decoder. During the training process, the PA-GCR is jointly optimized with the following feature learning module in an end-to-end manner, which aims to achieve the trade-off between the performance of reconstruction, and final age estimation and gender classification. We will briefly

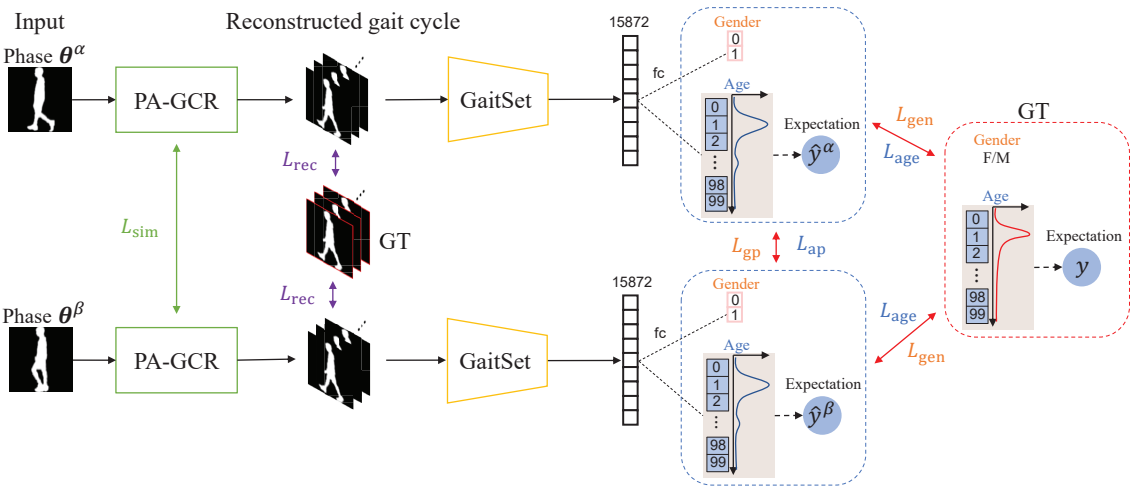

Figure 2. Overview of the proposed method. The digits represent the dimension of the features. GT and fc denote the ground truth and fully connected layer, respectively. In the training phase, a pair of single images from the same subject is used as the inputs, and the network parameters are shared; in the test phase, only a single image is fed into the network to output the estimated gender class and age label distribution, and the expectation of the label distribution is computed as the final estimated age value.

introduce it and readers may refer to [50] for more details.

Following [50], the phase of a silhouette is represented as a 2D vector using cosine and sine functions of a cyclic angle, and the ground truth of the gait cycle is set to a fixed number of frames with synchronized phases among different subjects. To mitigate the dependence of the encoded feature on the input phase, a similarity loss is defined to minimize the difference between the phase-independent features obtained by the feature transformer for an input pair of the same subject as

$$L_{\text{sim}} = \frac{1}{N}\sum_{n=1}^{N} \|\boldsymbol{f}_n^{\alpha} - \boldsymbol{f}_n^{\beta}\|_2^2, \tag{1}$$

where $\boldsymbol{f}_n^{\alpha}$ and $\boldsymbol{f}_n^{\beta}$ are the transformed features of n-th input pair $(n = 1, \ldots, N)$ with phase $\boldsymbol{\theta}^{\alpha}$ and $\boldsymbol{\theta}^{\beta}$ $(\alpha, \beta = 1, \ldots, T)$, respectively. N is the number of training pairs, and T is the number of frames in a gait cycle, which is set to 20 experimentally.

The reconstructed cycle is then generated by the decoder using the transformed phase-independent feature, which is constrained by a reconstruction loss as

$$L_{\text{rec}} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t\in\{\alpha,\beta\}} \|RC_n^t - GTC_n\|_2^2, \tag{2}$$

where RC_n^t $(t \in \{\alpha, \beta\})$ is the reconstructed gait cycle for n-th input with phase $\boldsymbol{\theta}^t$, and GTC_n is the corresponding ground truth gait cycle.

3.3. Feature extraction via GaitSet

Unlike most gait recognition networks using GEI as an input [44, 48, 47, 54, 21], GaitSet [7] directly inputs a silhouette sequence, which is treated as a set of images, and yields the state-of-the-art performance in the gait recognition community. We therefore employ GaitSet to extract discriminative gait feature from the reconstructed gait cycle.

Each of the silhouette in the reconstructed gait cycle is first fed into a CNN to extract frame-level feature independently, which is then aggregated into a single set-level feature using set pooling. A discriminative representation is further obtained from the set-level feature by horizontal pyramid mapping, where features on different scales with different spatial locations are combined.

The feature learned by GaitSet is finally used to obtain the gender class and estimated age via a fully connected layer with Softmax normalization, respectively. As mentioned in [41] and Section 2.1, the uncertainty of the estimated age is quite beneficial for real applications, we therefore output a probability distribution of discrete age labels instead of a single age value; hence, the dimension of the output vector for gender and age is set to two and Y, respectively, and Y is set to 100 in our implementation.

3.4. Loss functions

Age estimation. To supervise the output age label distribution, we adopt JensenShannon (JS) divergence [25] to measure the similarity between the estimated probability distribution and a ground truth probability distribution. Compared with the KullbackLeibler (KL) divergence used in [41], the JS divergence is an extended version that considers symmetry, and hence is more suitable for the proposed method which takes a pair of single images as the network input. Given two probability distributions $P(x_i)$ and $Q(x_i)$ for discrete variables $\{x_i\}$ $(i = 1, \ldots, I)$, the JS divergence

between them is defined as

$$\mathrm{JS}(P\|Q)=\frac{1}{2}\mathrm{KL}\left(P\left\|\frac{P+Q}{2}\right.\right)+\frac{1}{2}\mathrm{KL}\left(Q\left\|\frac{P+Q}{2}\right.\right), \quad (3)$$

where $\mathrm{KL}(\cdot)$ is the KL divergence [18] between two distributions, which is defined as $\mathrm{KL}(P\|Q) = \sum_{i=1}^{I} P(x_i)(\log P(x_i) - \log Q(x_i))$.

Denote the estimated age label distribution for the n-th input pair with phase $\boldsymbol{\theta}^t$ $(t \in \{\alpha, \beta\})$ as $\hat{\boldsymbol{a}}_n^t = [\hat{a}_n^t(1), \ldots, \hat{a}_n^t(Y)]^T \in \mathbb{R}^Y$, where $\hat{a}_n^t(y)$ $(y = 1, \ldots, Y)$ is the estimated probability for the interger age y. The corresponding ground truth probability distribution is denoted as $\boldsymbol{a}_n^t = [a_n^t(1), \ldots, a_n^t(Y)]^T$, which is set to a Gaussian distribution with the mean and standard deviation of the ground truth age y_n and 1, respectively. Similar as [41], we adopt the single estimated age value as the expectation of the estimated probability distribution, which is computed as $\hat{y}_n^t = \sum_{y=1}^{Y} y\hat{a}_n^t(y)$. We therefore define a loss function to minimize both the JS divergence between the estimated and ground truth age label distributions, and the L1 distance between the expected age and ground truth age as

$$L_{\mathrm{age}} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t\in\{\alpha,\beta\}}(\mathrm{JS}(\hat{\boldsymbol{a}}_n^t\|\boldsymbol{a}_n^t) + \|\hat{y}_n^t - y_n\|_1). \quad (4)$$

Because the inputs from the same subject with different phases may cause the intra-subject differences in the estimated age, we additionally define a loss function to force the age estimation results to be similar between the input same subject pair, which is computed as

$$L_{\mathrm{ap}} = \frac{1}{N}\sum_{n=1}^{N}(\mathrm{JS}(\hat{\boldsymbol{a}}_n^{\alpha}\|\hat{\boldsymbol{a}}_n^{\beta}) + \|\hat{y}_n^{\alpha} - \hat{y}_n^{\beta}\|_1). \quad (5)$$

Gender classification. Let the output vector for gender be $\hat{\boldsymbol{g}}_n^t = [\hat{g}_n^t, 1-\hat{g}_n^t]^T \in \mathbb{R}^2$, where $\hat{g}_n^t$ and $1-\hat{g}_n^t$ indicates the predicted probability for female and male, respectively. A binary cross-entropy loss is defined between the predicted probability and ground truth gender class g_n as

$$L_{\mathrm{gen}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{t\in\{\alpha,\beta\}}(g_n\log\hat{g}_n^t + (1-g_n)\log(1-\hat{g}_n^t)). \quad (6)$$

Similarly, a loss function to minimize the difference in output gender probability between the input training pair is computed as

$$\begin{aligned} L_{\mathrm{gp}} = -\frac{1}{N}\sum_{n=1}^{N}&(\hat{g}_n^{\beta}\log\hat{g}_n^{\alpha} + (1-\hat{g}_n^{\beta})\log(1-\hat{g}_n^{\alpha}) \\ &+ \hat{g}_n^{\alpha}\log\hat{g}_n^{\beta} + (1-\hat{g}_n^{\alpha})\log(1-\hat{g}_n^{\beta})). \end{aligned} \quad (7)$$

Joint losses. We finally combine all loss functions to simultaneously optimize the whole framework in an end-to-end manner, which ensures a well balance between the reconstruction quality and accuracies of gender classification and age estimation. The joint loss function is computed as

$$\begin{aligned} L_{\mathrm{joint}} = {} & w_{\mathrm{sim}}L_{\mathrm{sim}} + w_{\mathrm{rec}}L_{\mathrm{rec}} + w_{\mathrm{age}}(L_{\mathrm{age}} + w_{\mathrm{ap}}L_{\mathrm{ap}}) \\ & + w_{\mathrm{gen}}(L_{\mathrm{gen}} + w_{\mathrm{gp}}L_{\mathrm{gp}}), \end{aligned} \quad (8)$$

where w_{sim}, w_{rec}, w_{age}, w_{ap}, w_{gen}, and w_{gp} are the corresponding weights for each single loss function.

4. Online systems

We implemented two online systems: one is a stand-alone system and the other is a client-server system where the client and server communicate via a web API each other. While a computer for the stand-alone system and the server is equipped with a general-purpose GPU, i.e., NVIDIA GeForce RTX 2080 Ti, that for the client does not need it and a conventional tablet computer, Surface Go 2, was used as the client.

We employed Microsoft Kinect v2 as an input device and extract walking persons' silhouettes from each captured single depth image by a background subtraction. We then register and size-normalize each person's silhouette into an image with the size of 64×64 pixels. The above mentioned pre-preprocessing part was implemented by C++ and OpenCV. On the other hand, we implemented a deep learning module composed of gait cycle reconstruction and age/gender estimation by Python and PyTorch and converted them by Boost.Python so as to be called from C++.

The stand-alone system was implemented by simply combining the above-mentioned pre-processing and deep learning modules. Hence, the stand-alone system can show the estimated gender and age for each subject in each captured single frame, as shown in Fig. 5.

As for the client-server system, once the client sends a capture depth image to the server with the web API, the server processes it with the pre-processing and deep learning modules and returns the estimation results to the client again via the web API. The latency due to communication over network naturally arises for the client-server system, and hence the real-time processing at standard video rate (e.g., 30 fps) is difficult. The system can, however, still work well thanks to the proposed single frame-based processing under the latency (e.g., a few frames per second), unlike the conventional video-based gait analysis assumes inputs are captured at the standard video rate.

5. Experiments

5.1. Dataset

We trained and evaluated the proposed method on OU-MVLP [46], which is the world's largest gait dataset with a wide view variation. It contains 10,307 subjects (5,114

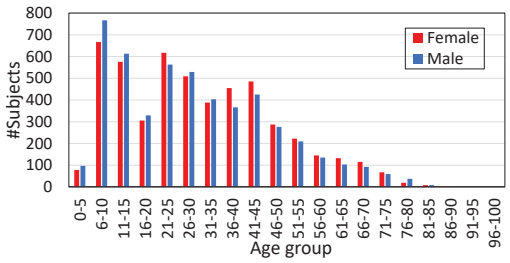

Figure 3. Distribution of subjects' genders and ages in the OU-MVLP dataset.

males and 5,193 females) with ages ranging from 2 to 87 years old. The distribution of subjects' genders and ages is shown in Fig. 3. Each subject was captured from 14 views, ranging 0°–90° and 180°–270° in 15° interval, with two sequences ('00' and '01') captured for each view. Following the original protocol, we used 5,153 subjects for training and the other disjoint 5,154 subjects for testing. In the training phase, the images from both '00' and '01' sequences in all views were simultaneously used to train a single CNN model; in the test phase, the performance was evaluated for each view independently using only '00' sequence, from which a single frame was randomly chosen as the network input.

5.2. Training details

The network was trained using Adam [17] with the batch size of 8×16, which indicates 8 subjects with 16 samples for each were chosen to constitute a mini-batch. Note that only the same subject pairs in a batch were used for training process. We followed the training strategy in [50], i.e., first trained PA-GCR only for the first 30K iterations with an learning rate of 10^{-4}, and then included the following GaitSet to train more 250K iterations with learning rates of 10^{-5} and 10^{-4} for PA-GCR and GaitSet, respectively, which were both reduced by 0.1 for the final 100K iterations. The weight parameters w_{sim}, w_{rec}, w_{age}, w_{ap}, w_{gen}, and w_{gp} in Eq. 8 were experimentally set to 0.0005, 10, 1, 1, 2, and 0.01, respectively.

5.3. Evaluation metrics

We evaluated the accuracy of gender classification using mean correct classification rate (CCR), and evaluated that of the estimated age using mean absolute error (MAE) and cumulative score (CS) [32]. Let $\hat{y}_i$ and y_i be the estimated age and ground truth age of the i-th test sample, the MAE is computed as $\mathrm{MAE} = \frac{1}{N_s}\sum_{i=1}^{N_s}|\hat{y}_i - y_i|$, where N_s is the number of test samples. The CS for y-year absolute error tolerance is computed as $CS(y) = N_s(y)/N_s$, where $N_s(y)$ is the number of samples whose estimation absolute error is within y years.

To measure the performance of estimated probability distribution, we computed a mean cross-entropy (MCE) between the estimated and ground truth distribution as

Table 1. MAE [year], CSs [%], MCE, and mean CCR [%] over all samples from all 14 views for the proposed method, GEINet [41] and GaitSet [7]. Bold indicates the best results.

Method	MAE	CS(1)	CS(5)	CS(10)	MCE	CCR
GEINet	9.11	15.70	46.73	65.82	-13.60	N/A
GaitSet	9.02	**16.39**	47.59	66.01	-5.04	92.72
Ours	**8.39**	15.84	**48.00**	**68.40**	**-4.27**	**94.27**

$$\mathrm{MCE} = \frac{1}{N_s}\sum_{i=1}^{N_s}\sum_{y=1}^{100} a_i(y)\log(\hat{a}_i(y)), \tag{9}$$

where $a_i(y)$ and $\hat{a}_i(y)$ are the ground truth and estimated probability of age label y for i-th test sample. We further defined the ground truth distribution as a delta function of the ground truth age y_i, and hence Eq. 9 is converted to

$$\mathrm{MCE} = \frac{1}{N_s}\sum_{i=1}^{N_s}\log(\hat{a}_i(y_i)), \tag{10}$$

which is a log likelihood for the estimated age label distribution that gets larger when the estimated probability for the ground truth age is larger.

5.4. Comparison with benchmarks

Because this is the first work tackling single gait image-based age estimation and gender classification, we compared the proposed method with a baseline, i.e., pure GaitSet [7] directly using a single image without gait cycle reconstruction, and a state-of-the-art age estimation method, i.e., GEINet that estimates an age label distribution using KL divergence-based loss function [41] in Table 1. Note that both the GaitSet and GEINet were trained and tested using the same samples as ours.

Compared with the GEINet with a simple network structure, the sequence-based GaitSet yields slightly better MAE and CSs, and significantly better MCE, which indicates that GaitSet still extracts more effective features even if only a single image is used. All three methods gain similar CS for 1-year absolute error, which is easier to be achieved for children and teenagers because the body shape differences (e.g., head-to-body ratio) are obviously observed even in a single frame. On the other hand, age estimation for adults is much more difficult if only a single frame is used, which only contains the shape information; hence, the proposed method obtains much better performance for adults by reconstructing a gait cycle, which is illustrated by the CS for larger absolute error tolerance, and also the scatter plots in the supplementary material. Therefore, the proposed method achieves the best performance for both age estimation and gender classification in general.

We also report the MAE, MCE and CCR of each view for the proposed method in Table 2. The performance of

Table 2. MAE [year], MCE and CCR [%] of each view for the proposed method.

Metrics	0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	Mean
MAE	8.91	8.83	8.27	8.48	8.35	8.08	7.83	9.24	8.86	8.61	8.48	8.11	7.88	7.74	8.39
MCE	-4.28	-4.36	-4.29	-4.34	-4.27	-4.21	-4.19	-4.42	-4.34	-4.30	-4.34	-4.19	-4.16	-4.16	-4.27
CCR	91.5	93.0	94.5	94.6	95.0	94.9	95.7	93.0	93.6	94.9	94.8	94.4	94.6	94.9	94.3

Table 3. Ablation experiments evaluated by MAE [year], MCE and mean CCR [%] over all 14 views.

Network	Age loss	Pair loss	#Training input frame	Test input	Results		
					MAE	MCE	CCR
GaitSet	JS + L1	√	1	1 frame	9.02	-5.04	92.72
Proposed	L1 (regression)	√	1	1 frame	8.72	N/A	93.87
Proposed	KL + L1	√	1	1 frame	8.61	-4.47	93.71
Proposed	JS + L1	×	1	1 frame	8.55	-4.77	94.07
Proposed	JS + L1	√	1	1 frame	**8.39**	**-4.27**	**94.27**
GaitSet (upper bound for reconstruction)	JS + L1	√	20	GT cycle	6.63	-4.75	96.04

both age estimation and gender classification is better for the views around the side view (e.g., 75°and 90°), and is worse for those around the front/back view (e.g., 0°and 15°). This is understandable because both the body shape (e.g., middle-age spread) and motion (e.g., stride) patterns that indicate age and gender are clearer near the side view, which is true for both a single image and a reconstructed gait cycle.

5.5. Ablation study

The results of ablation experiments for the proposed method are shown in Table 3. In the first row, the part of gait cycle reconstruction (i.e., PA-GCR) was removed from the proposed method, i.e., the GaitSet was trained and tested using a single frame. The second and third rows show the effects of loss function L_{age} by: changing the task of age label distribution estimation into single age value regression, i.e., using a fully connected layer with 1-dim output to regress a scalar value and computing a L1 loss between the output and ground truth age; replacing the JS divergence in Eq. 4 with KL divergence used in [41]. The fourth row shows the results of removing the loss functions defined between the input pair (i.e., L_{ap} and L_{gp}). The fifth row is the result of the proposed method. The upper bound of the gait cycle reconstruction framework is shown in the last row, which was done by testing with ground truth gait cycle using the GaitSet model pre-trained with 20 frames-input.

Compared with directly extracting features from a single image using GaitSet (first row), the performance of both age estimation and gender classification was clearly improved using the proposed method (fifth row), which shows the effectiveness of reconstructing a gait cycle before feature extraction. As shown in the second, third and fifth rows, the age label distribution estimation-based methods worked better than the regression-based method that outputs a single age value, because the former mitigated the estimation errors caused by the subjects with similar appearances but different actual ages (e.g., a middle-aged with slim body) by providing the uncertainty for different age labels [41]. Additionally, thanks to the symmetric property of JS divergence measure, the proposed method gained better results than that using the KL divergence measure. Including constraints on output between training pairs also improved the performance (comparing fourth and fifth rows), as it reduced the intra-subject differences in estimations resulted from the phase differences of the input frames. Therefore, all analyzed components contributed to the proposed method.

The GaitSet using the ground truth gait cycle (last row) obtained much better results for MAE but worse for MCE compared with the proposed method using reconstructed gait cycle. This means that the accuracy of estimated age probability distribution is somewhat sacrificed to achieve a relatively good estimation of a single age value during the training process, which easily happens when the weighted sum-based loss function is adopted for multiple tasks.

5.6. Visualizing age label distribution

We then qualitatively evaluated the estimated age label distribution by comparing the proposed method with the one frame-based GaitSet (i.e., first row in Table 3) and the ground truth cycle-based GaitSet (i.e., last row in Table 3). Two examples are shown in Fig. 4, where the first subject is an adult captured from 90°, and the second is an elderly captured from 0°. More examples are shown in the supplementary material.

Because the first subject shows a relatively larger head-to-body ratio in the given single frame due to the hair style, the one frame-based GaitSet under-estimated this subject as a child with a large probability (i.e., sharp peak in the distribution) based only on this shape characteristics. Similarly, due to the quite limited shape information captured from 0°, where age-related patterns (e.g., stoop) are difficult to be observed, the one frame-based GaitSet estimated the second subject as a youth, which results in a large estimation error.

By contrast, the reconstructed gait cycle reflects more in-

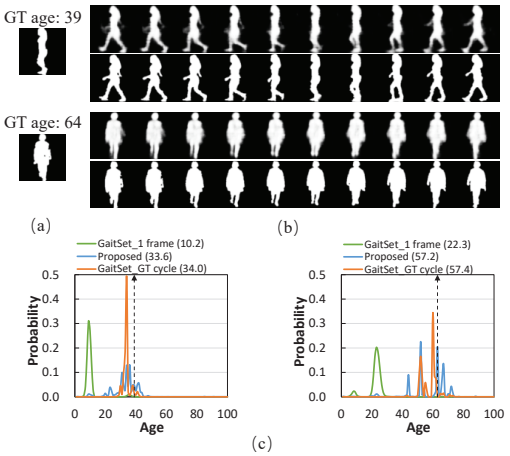

Figure 4. Two examples for visualizing age label distribution. (a) Input single image with the ground truth age. (b) For each example, first row is a half of the reconstructed gait cycle by the proposed method, and second row is the corresponding ground truth half cycle. (c) Estimated age label distribution by each method (left: first subject. right: second subject). The digits shown after each method is the corresponding estimated age, and the black dotted line indicates the ground truth age.

formative motion patterns, such as the stride-to-height ratio, which should be relatively smaller for the adults compared to the children [32, 52], and hence the proposed method estimated the first subject as an adult instead. Moreover, due to the limited physical changes, the gait differences among adults (e.g., 20–40 years old) are quite small, which is difficult to correctly judge an exact age value; hence, the proposed method showed the uncertainty on estimated age by assigning probabilities to a range of age labels, while the probabilities around the ground truth age is also relatively larger. Therefore, the proposed method returned a reasonable age label distribution that is consistent with our prior knowledge and insights from [41].

On the other hand, the ground truth cycle-based GaitSet also outputted a probability distribution for an appropriate range of age labels. While the proposed method assigned similar probabilities to a relatively wider age range (e.g., 30–40 years old for the first subject), the ground truth cycle-based GaitSet predicted a much larger probability for a specific age around the ground truth age (i.e., 34 years old for the first subject). This results in smaller estimation error for the final expected age from the distribution, whereas the probability for the ground truth age is relatively smaller compared with the proposed method, which is consistent with the results of quantitative evaluation shown in Table 3.

5.7. Processing time of online system

To validate the real-time processing capability of the online stand-alone system, we evaluated the running time of the system in an online environment. While a single/multiple subjects walking in the capture scene, the system automatically outputted the estimated age and gender for each of the subject frame-by-frame, as shown in Fig. 5. The processing time for different number of subjects detected from the captured image is shown in Table 4. Obviously, the proposed system meets the requirements of a real-time online system, since the computation time is less than/around the frame rate of Kinect (i.e., 30 fps) for even multiple subjects.

Figure 5. A snapshot of the online system in an actual online scene. Left: overview of the capture scene with a walking subject and the results shown in the system at the same time, right: an enlarged image of the system shown in the left. The system still works for a person from a back view unlike face biometrics. Video examples are found in the supplementary material.

Table 4. Processing time [msec.] for different number of subjects detected from the captured image.

#Detected subjects	0	1	2	3	4
Processing time	2–3	10–12	16–19	26–27	35

6. Conclusion

This paper presented a real-time CNN framework for gait-based age estimation and gender classification from a single image. A full gait cycle is first reconstructed from the input single image by the PA-GCR, and the reconstructed gait cycle is then used for feature leaning by the GaitSet. Finally, the estimated gender class and age label probability distribution are obtained from the proposed network simultaneously. We also implemented two online systems: the stand-alone system and the client-server system to demonstrate the proposed method works real-time/online.

An important future work is to evaluate the performance of the proposed method by capturing images in an actual online situation. Another possible extension is to directly output a continuous age distribution instead of discrete label distribution, and also include a regularizer for smoothing the estimated distribution. Additionally, since the system may receive more than one frame when the subjects walking towards, the fusion of the results obtained by multiple frames is also worth investigating, which can further improve the estimation accuracy and stability compared with the independent frame-by-frame version.

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.

References

[1] N. Akae, Y. Makihara, and Y. Yagi. Gait recognition using periodic temporal super resolution for low frame-rate videos. In *Proc. of the Int. Joint Conf. on Biometrics (IJCB2011)*, pages 1–7, Washington D.C., USA, Oct. 2011.

[2] N. Akae, A. Mansur, Y. Makihara, and Y. Yagi. 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)*, pages 1537–1543, Providence, RI, USA, Jun. 2012.

[3] Muayed S. Al-Huseiny, Sasan Mahmoodi, and Mark S. Nixon. Gait learning-based regenerative model: A level set approach. In *The 20th Int. Conf. on Pattern Recognition*, pages 2644–2647, Istanbul, Turkey, Aug. 2010.

[4] Maryam Babaee, Linwei Li, and Gerhard Rigoll. Person identification from partial gait cycle using fully convolutional neural networks. *Neurocomputing*, 338:116 – 125, 2019.

[5] M. Babaee, Y. Zhu, O. Kpkl, S. Hrmann, and G. Rigoll. Gait energy image restoration using generative adversarial networks. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2596–2600, 2019.

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

[7] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. 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] L. Chen, Y. Wang, and Y. Wang. Gender classification based on fusion of weighted multi-view gait component distance. In *2009 Chinese Conference on Pattern Recognition*, pages 1–5, 2009.

[9] Zhang De. Research on gait-based gender classification via fusion of multiple views. *International journal of database theory and application*, 8:39–50, 2015.

[10] Trung Dung Do, Van Huan Nguyen, and Hakil Kim. Real-time and robust multiple-view gender classification using gait features in video surveillance. *Pattern Analysis and Applications*, 23(1):399413, 2020.

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

[12] M. Hu and Y. Wang. A new approach for gender classification based on gait analysis. In *2009 Fifth International Conference on Image and Graphics*, pages 869–874, 2009.

[13] G. Huang and Y. Wang. Gender classification based on fusion of multi-view gait sequences. In *Proc. of the 8th Asian Conf. on Computer Vision*, volume 1, pages 462–471, Nov. 2007.

[14] Ebenezer R.H.P. Isaac, Susan Elias, Srinivasan Rajagopalan, and K.S. Easwarakumar. Multiview gait-based gender classification through pose-based voting. *Pattern Recognition Letters*, 126:41 – 50, 2019. Robustness, Security and Regulation Aspects in Current Biometric Systems.

[15] 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.

[16] C. Kalaiselvan and A. Sivanantha Raja. Robust gait-based gender classification for video surveillance applications. *Applied Mathematics and Information Sciences*, 11:1207–1215, 2017.

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint*, arXiv: 1412.6980 (2014), 2014.

[18] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 03 1951.

[19] L. Lee and W. Grimson. Gait analysis for recognition and classification. In *Proc. of the 5th IEEE Conf. on Face and Gesture Recognition*, volume 1, pages 155–161, 2002.

[20] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, and Mingwu Ren. Gait-based human age estimation using age group-dependent manifold learning and regression. *Multimedia Tools Appl.*, 77(21):2833328354, Nov. 2018.

[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*, 14(12):3102–3115, Dec 2019.

[22] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, and Mingwu Ren. Make the bag disappear: Carrying status-invariant gait-based human age estimation using parallel generative adversarial networks. In *Proc. of the IEEE 10th International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, Sept. 2019.

[23] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, and Mingwu Ren. Gait recognition via semi-supervised disentangled representation learning to identity and covariate features. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[24] X. Li, S.J. Maybank, S. Yan, D. Tao, and D. Xu. Gait components and their application to gender recognition. *Trans. on Systems, Man, and Cybernetics, Part C*, 38(2):145–155, Mar. 2008.

[25] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

[26] Jiwen Lu and Yap-Peng Tan. Gait-based human age estimation. *IEEE Trans. on Information Forensics and Security*, 5(4):761–770, Dec. 2010.

[27] Jiwen Lu and Yap-Peng Tan. Ordinary preserving manifold analysis for human age estimation. In *IEEE Computer Society and IEEE Biometrics Council Workshop on Biometrics 2010*, pages 1–6, San Francisco, CA, USA, Jun. 2010.

[28] J. Lu and Y. P. Tan. Ordinary preserving manifold analysis for human age and head pose estimation. *IEEE Transactions on Human-Machine Systems*, 43(2):249–258, March 2013.

[29] Jiwen Lu, Gang Wang, and Thomas S. Huang. Gait-based gender classification in unconstrained environments. In *Proc of the 21st International Conference on Pattern Recognition*, pages 3284–3287, 2012.

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

[31] Y. Makihara, A. Mori, and Y. Yagi. Temporal super resolution from a single quasi-periodic image sequence based on phase registration. In *Proc. of the 10th Asian Conf. on*

Computer Vision, pages 107–120, Queenstown, New Zealand, Nov. 2010.

[32] Y. Makihara, M. Okumura, H. Iwama, and Y. Yagi. Gait-based age estimation using a whole-generation gait database. In *Proc. of the Int. Joint Conf. on Biometrics (IJCB2011)*, pages 1–6, Washington D.C., USA, Oct. 2011.

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

[34] Y. Makihara, A. Suzuki, D. Muramatsu, X. Li, and Y. Yagi. Joint intensity and spatial metric learning for robust gait recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6786–6796, July 2017.

[35] Yasushi Makihara and Yasushi Yagi. 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.

[36] H. Mannami, Y. Makihara, and Y. Yagi. Gait analysis of gender and age using a large-scale multi-view gait database. In *Proc. of the 10th Asian Conf. on Computer Vision*, pages 975–986, Queenstown, New Zealand, Nov. 2010.

[37] M. J. Marn-Jimnez, F. M. Castro, N. Guil, F. de la Torre, and R. Medina-Carnicer. Deep multi-task learning for gait-based biometrics. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 106–110, 2017.

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

[39] Takemura Noriko, Makihara Yasushi, Muramatsu Daigo, Echigo Tomio, and Yagi Yasushi. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ Transactions on Computer Vision and Applications*, 10(1):4, Feb 2018.

[40] Atsuya Sakata, Yasushi Makihara, Noriko Takemura, Daigo Muramatsu, and Yasushi Yagi. Gait-based age estimation using a densenet. In *ACCV Workshops*, pages 55–63, 2018.

[41] Atsuya Sakata, Yasushi Makihara, Noriko Takemura, Daigo Muramatsu, and Yasushi Yagi. How confident are you in your estimate of a human age? uncertainty-aware gait-based age estimation by label distribution learning. In *IEEE International Joint Conference on Biometrics*, Sep. 2020.

[42] Atsuya Sakata, Noriko Takemura, and Yasushi Yagi. Gait-based age estimation using multi-stage convolutional neural network. *IPSJ Transactions on Computer Vision and Applications*, 11(4):1–10, 2019.

[43] 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 Transactions of Pattern Analysis and Machine Intelligence*, 27(2):162–177, 2005.

[44] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi. Geinet: View-invariant gait recognition using a convolutional neural network. In *2016 International Conference on Biometrics (ICB)*, pages 1–8, 2016.

[45] L.R Sudha and Dr. R Bhavani. Gait based gender identification using statistical pattern classifiers. *International Journal of Computer Applications*, 40(8):30–35, February 2012.

[46] Noriko Takemura, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. 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.

[47] 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*, pages 1–1, 2018.

[48] 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, 2017.

[49] 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.

[50] 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 *The 16th European Conference on Computer Vision (ECCV)*, August 2020.

[51] Chi Xu, Yasushi Makihara, Gakuto Ogi, Xiang Li, Yasushi Yagi, and Jianfeng Lu. The ou-isir gait database comprising the large population dataset with age and performance evaluation of age estimation. *IPSJ Transactions on Computer Vision and Applications*, 9(1):24, Dec 2017.

[52] Chi Xu, Yasushi Makihara, Yasushi Yagi, and Jianfeng Lu. 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.

[53] J. Yoo, D. Hwang, and M.S. Nixon. Gender classification in human gait using support vector machine. In *Advanced Concepts For Intelligent Vision Systems*, pages 138–145, 2006.

[54] S. Yu, H. Chen, E. B. G. Reyes, and N. Poh. Gaitgan: Invariant gait feature extraction using generative adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 532–539, July 2017.

[55] S. Yu, T. Tan, K. Huang, K. Jia, and X. Wu. A study on gait-based gender classification. *IEEE Trans. on Image Processing*, 18(8):1905–1910, Aug. 2009.

[56] D. Zhang and Y. Wang. Using multiple views for gait-based gender classification. In *The 26th Chinese Control and Decision Conference (2014 CCDC)*, pages 2194–2197, 2014.

[57] S. Zhang, Y. Wang, and A. Li. Gait-based age estimation with deep convolutional neural network. In *2019 International Conference on Biometrics (ICB)*, pages 1–8, 2019.

[58] Yuqi Zhang, Yongzhen Huang, Liang Wang, and Shiqi Yu. A comprehensive study on gait biometrics using a joint cnn-based method. *Pattern Recognition*, 93:228 – 236, 2019.

[59] Haiping Zhu, Yuheng Zhang, Guohao Li, Junping Zhang, and Hongming Shan. Ordinal distribution regression for gait-based age estimation. *Science China Information Sciences*, 63(2), 2020.