Performance Evaluation of Model-Based Gait on Multi-View Very Large Population Database With Pose Sequences

Weizhi An, Shiqi Yu, *Member, IEEE*, Yasushi Makihara, Xinhui Wu, Chi Xu, Yang Yu, Rijun Liao, and Yasushi Yagi, *Member, IEEE*

***Abstract*—Model-based gait recognition is considered to be promising due to the robustness against some variations, such as clothing and baggage carried. Although model-based gait recognition has not been fully explored due to the difficulty of human body model fitting and the lack of a large-scale gait database, recent progress in deep learning-based approaches to human body model fitting and human pose estimation is mitigating the difficulty. In this paper, we, therefore, address the remaining issue by presenting a large-scale human pose-based gait database, OUMVLP-Pose, which is based on a publicly available multi-view large-scale gait database, OUMVLP. OUMVLP-Pose has many unique advantages compared with other public databases. First, OUMVLP-Pose is the first gait database that provides two datasets of human pose sequences extracted by two standard deep learning-based pose estimation algorithms, OpenPose and AlphaPose. Second, it contains multi-view large-scale data, i.e., over 10,000 subjects and 14 views for each subject. In addition, we also provide benchmarks in which different kinds of gait recognition methods, including model-based methods and appearance-based methods, have been evaluated comprehensively. The model-based gait recognition methods have shown promising performances. We believe this database, OUMVLP-Pose, will greatly promote model-based gait recognition in the next few years.**

***Index Terms*—Gait database, benchmark, gait recognition, human body pose.**

I. Introduction

GAIT is one of the most popular behavioral biometrics in the world because it has unique advantages compared with face, iris, palm print, etc. Gait features can be captured at a long distance and are hard to disguise, and consequently, gait recognition technology has been added to the repertoire of tools available for crime prevention and forensic identification.

The gait recognition methods mainly fall into two groups: appearance-based methods and model-based methods. The appearance-based methods directly extract gait features from an image sequence to encode the spatial and temporal information. Most of the appearance-based methods usually extract silhouettes from raw videos first and then extract gait features, e.g., gait energy image (GEI) [1], which is created by averaging the pixels of silhouettes in a gait cycle, chrono gait image (CGI) [2], and gait flow image (GFI) [3]. Due to the simplicity of feature extraction and high performance in recognition accuracy, appearance-based methods have been more popular than model-based methods for more than a decade. However, the challenges caused by variations in view, speed, clothing, carrying status, etc. can affect the accuracy of the appearance-based gait recognition methods.

The model-based methods extract features by fitting a human articulated model to an image and by extracting kinematic information such as a sequence of joint positions or joint angles [4], [5], [6], [7], [8]. The model-based methods can be robust against appearance changes due to clothing and carrying status variations since extracted joint positions/angles are less affected by clothing and carrying status variations. However, human model fitting, a key procedure of the model-based approaches, has been thought to be error-prone, computationally exhaustive, and demands high image resolution. As a result, the model-based methods have been less employed in the video-based gait analysis community for more than a decade.

Situations surrounding the human model fitting (or human pose estimation) have, however, been drastically changing for these years. One such seminal work is a training-based approach to pose estimation with a depth sensor (e.g., Kinect) [9]. For example, Kastaniotis *et al.* [10] used skeleton data from a single Kinect sensor instead of a setup of multiple synchronous cameras in [11]. This shows that the body joints from Kinect can contribute to gait recognition, i.e., the feasibility of model-based gait recognition. The commonly used cameras in video surveillance are, however, not depth sensors such as Kinect but conventional cameras (e.g., color cameras or monochrome cameras).

Thereafter, deep learning-based approaches significantly advanced state-of-the-art human pose estimation, and standard

Manuscript received February 9, 2020; revised June 4, 2020; accepted June 27, 2020. Date of publication July 13, 2020; date of current version September 23, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61976144, and in part by the Japan Society for the Promotion of Science Grants-in-Aid for Scientific Research (A) under Grant JP18H04115. This article was recommended for publication by Associate Editor Z. Sun upon evaluation of the reviewers' comments.
(Corresponding author: Shiqi Yu.)

Weizhi An and Xinhui Wu are with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China.

Shiqi Yu is with the Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China (e-mail: yusq@sustech.edu.cn).

Yasushi Makihara, Chi Xu, Yang Yu, and Yasushi Yagi are with the Institute of Scientific and Industrial Research, Osaka University, Suita 565-0871, Japan.

Rijun Liao is with the Department of Computer Science and Electrical Engineering, University of Missouri–Kansas City, Kansas City, MO 64110 USA.

Digital Object Identifier 10.1109/TBIOM.2020.3008862

techniques such as OpenPose [12] and AlphaPose [13] have been widely used in many research fields, which indicates the possibility of model-based gait recognition with conventional cameras in visual surveillance scenarios. For example, Liao *et al.* [14] proposed a pose-based temporal-spatial network (PTSN) that takes a sequence of estimated human poses as input and showed its effectiveness on cross-view gait recognition with a publicly available gait database, i.e., CASIA B [15]. Although CASIA B contains large view variations (eleven views) from 0° to 180°, the number of subjects is still limited to 124, which is insufficient to fully demonstrate the possibility of model-based gait recognition in this deep learning era.

We, therefore, built the world's largest multi-view gait pose database named "OU-ISIR Gait Database, Multi-View Large Population Database with Pose Sequence (OUMVLP-Pose)"[1] to further advance state-of-the-art model-based gait recognition. The contributions of this study are three-fold.

- **The first gait database with pose sequences extracted by deep learning-based pose estimators.** While the existing gait databases were released in the form of images (e.g., GEI, silhouette sequences or RGB image sequences), we construct the gait database with the pose sequences obtained by two state-of-the-art pose estimation algorithms for the first time to the best of our knowledge. The constructive database is beneficial for the gait analysis community to revisit the model-based approaches.
- **Large-scale and multi-view database.** Since the database was built upon the large-scale multi-view gait database, i.e., OUMVLP [16], it contains 10,307 subjects with a wide range of views (14 views, 0°–90°, 180°–270° at 15° interval), which results in the world's largest gait database with pose sequence. As deep learning-based methods require massive samples for sufficient training and reliable evaluation, the constructed database is suitable for evaluating model-based gait recognition with deep neural networks such as [14].
- **Performance evaluation of the model-based gait recognition.** We conducted a set of evaluation experiments with a variety of model-based approaches ranging from a traditional method [4] to recent deep learning-based methods [14]. This can be a good milestone for future studies of model-based gait recognition. We also show the significant improvement from the traditional model-based approach to the current deep learning-based approach and show a still remaining gap of performances between the model-based and appearance-based approaches with deep learning frameworks.

The rest of the paper is organized as follows. Section II presents the existing gait databases and related work on pose estimation methods. Section III introduces the construction of our pose sequence database, and Section IV presents the performance evaluation results with the constructed database. Section V concludes this work.

[1]OUMVLP-Pose is available at http://www.am.sanken.osaka-u.ac.jp/BiometricDB/GaitLPPose.html.

TABLE I
EXISTING MULTI-VIEW GAIT DATABASES

Database	#Subjects	#Views	Range of views
CMU Mobo [17]	25	6	$0° - 360°$
Soton small [18]	12	4	-
Soton multimodal [19]	>400	12	-
CASIA A [20]	20	3	$0° - 90°$
CASIA B [15]	124	11	$0° - 180°$
AVA [21]	20	6	-
WOSG [22]	155	8	-
KY4D [23]	42	16	$0° - 360°$
OU-TD C [24]	200	25	$0° - 360°$
OU-LP [25]	4,016	4	$55° - 85°$
OU-MVLP [16]	10,307	14	$0° - 90°$, $180° - 270°$

II. Related Work

A. Gait Databases

The existing major multi-view gait databases are shown in Table I. The CMU Mobo database [17] contains 25 individuals walking in four different walk patterns: slow walk, fast walk, incline walk and walking with a ball. All subjects are captured under six views from 0°-360°. The Soton Multimodal database [19] contains over 400 multimodal subjects involving gait, face and ear. The gaits of all the subjects are captured under 12 views. The CASIA A database [20] includes 20 subjects and four sequences per view per subject. They are captured at a rate of 25 frames per second and includes a total of 240 sequences under three views: 0°, 45°, and 90°. The CASIA B database [15] contains 124 subjects with large view variations from 0° 180° with 18° intervals. It includes 6 normal sequences, 2 carrying bag sequences and 2 clothing sequences. The AVA database [21] includes 20 subjects with different body sizes under six view angles. The WOSG database [22] contains 155 subjects with 8 views. The KY4D gait database [23] contains 42 subjects of three-dimensional volume data which constructed multi-view images captured by 16 cameras. The OU-ISIR Treadmill Database C [24] contains 200 subjects with 25 views, which includes 12 views with 30° intervals, 2 tilt views and 1 top view. The OU-ISIR LP [25] contains 4,016 subjects with 55°, 65°, 75°, and 85° views.

While the above mentioned gait databases lack either or both aspects of the number of subjects (less than 1,000) or the view variations, OUMVLP [16] contains both a large number of subjects and wide view variations, i.e., 10,307 subjects from 0° to 90°, 180° to 270° with 15° intervals as shown in Fig. 1. The original images[2] are captured with the image size of 1,280 × 980 pixels at the frame-rate of 25 fps by seven network cameras at intervals of 15° azimuth angles along a quarter of a circle whose center coincides with the center of the walking course. OUMVLP provides its data in the format of GEIs as well as silhouette sequences. Although the GEI is the most widely used gait feature in the video-based gait analysis community, a large-scale multi-view gait database with pose sequences is demanded since it enables us to more

[2]While silhouette sequences and GEIs are open to the public, the original images will not be released due to privacy issues.

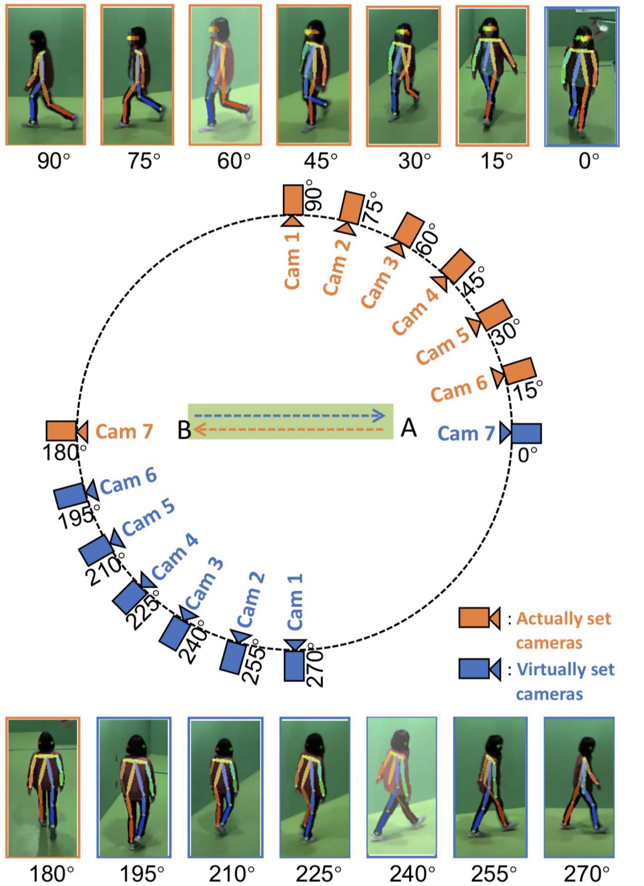


Fig. 1. Capturing setup of OUMVLP and examples of extracted pose from multiple views.

purely analyze gait, i.e., motion pattern when free from the body shape.

B. Gait Recognition

Gait recognition is a challenging task since there are a lot of variations. Many recent works are focusing on developing methods to extract robust gait feature to the variations. Some recent survey papers [26], [27] gave comprehensive analysis on gait recognition. Here we just list some recent works. Ben *et al.* [28] propose a coupled patch alignment (CPA) algorithm that effectively matches a pair of gaits across different views. Ben *et al.* [29], [30] proposed another two cross-view gait recognition methods respectively based on matrix and tensor, which are suitable for reducing the small sample size problem in discriminative subspace selection. Some other cross-view gait recognition methods can be found in [31], [32], [33]. To extract invariant gait feature, generative adversarial networks (GAN) are also employed in [34], [35]. In [36] the authors innovatively combined silhouette segmentation and gait recognition and proved that the combination can improve gait recognition obviously.

The previously mentioned methods are all appearance-based ones. In [6] the authors introduced a typical model-based gait recognition method. They used pendular motion to describe the thigh and lower leg motion, and studied on different walking styles, walking and running. Recently Liao *et al.* [14] took the advantage of deep learning and used a pose estimation by deep learning to recover human skeleton models. They also converted 2D pose data to 3D for view invariant feature extraction in [37]. The models by deep learning are much better than those by traditional methods. They also took a temporal-spatial neural network for gait recognition. We believe that model-based methods will be promoted greatly by deep learning. But we need a large gait database to advance gait recognition on model-based methods.

C. Pose Estimation

A human pose skeleton represents a person by a set of connections of human joints. It is a set of coordinates that can be connected to describe the pose of the person. Human pose estimation is challenging for computer vision. With the development of deep learning, human body pose estimation has achieved great progress in recent years. The pose estimation approaches are grouped into bottom-up DeepCut [38], OpenPose [12] and top-down approaches AlphaPose [13], and Mask-RCNN [39]. The top-down approaches detect the person first, followed by estimating the body parts. The bottom-up approaches detect all parts of every person, then group the parts belonging to distinct persons [40]. Bottom-up methods are more robust to occlusion and complex poses. However, most bottom-up methods do not directly benefit from human body structural information leading to many false positives. Top-down methods utilize global contexts and strong structural information, but they cannot handle complex poses. Moreover, the performance of top-down models is closely related to person detection results.

Cao *et al.* [12] proposed using deep learning to create accurate human models called "OpenPose", which can jointly detect human body joints including hands, feet, elbows and others. The method can handle multiple persons in an image. It can predict vector fields named part affinity fields (PAFs), which can directly expose the association between anatomical parts in an image. They designed an architecture to jointly learn part locations and their association, in which a set of 2D vector fields encodes the location and orientation of limbs over the image domain. These fields and joint confidence maps are jointly learned and predicted by CNN. Fang *et al.* [13] proposed a novel regional multiperson pose estimation framework to facilitate pose estimation, AlphaPose, in the presence of inaccurate human bounding boxes. The framework follows the top-down framework, which consists of three components: a symmetric spatial transformer network, parametric pose nonmaximum-suppression, and a pose-guided proposal generator. It significantly outperforms the state-of-the-art methods for multiperson human pose estimation in terms of accuracy and efficiency. AlphaPose is an accurate real-time multiperson pose estimation system, which can achieve 72.3 mean average precision (mAP) on the COCO dataset and an 82.1 mAP on the MPII dataset. In addition, the source code of AlphaPose is provided.

III. OUMVLP-Pose Database

OUMVLP-Pose was built upon OUMVLP [16]. OUMVLP contains 10,307 subjects of round-trip walking sequences captured by seven network cameras at intervals of 15° (this

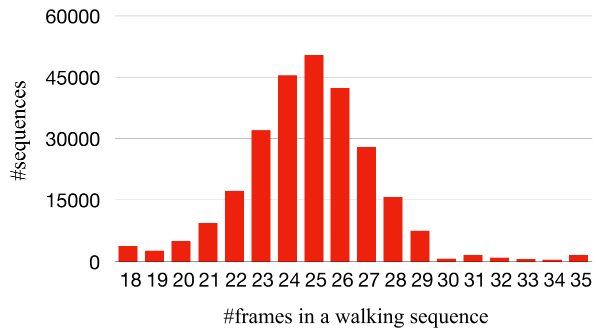


Fig. 2. Statistics of the number of frames in a sequence.

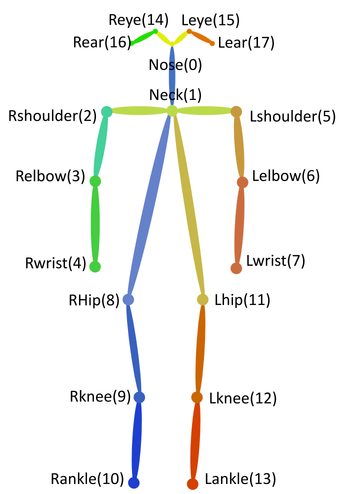


Fig. 3. The human skeleton model with 18 joints.

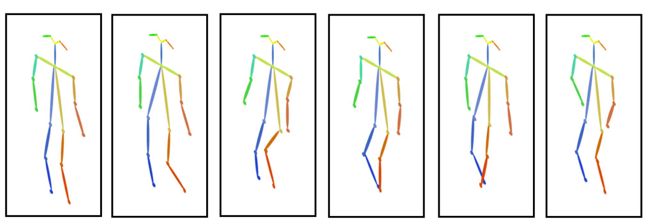


Fig. 4. A pose sequence from the 45° view.

sums to 14 views by considering the round trip on the same walking course) with an image size of 1,280×980 pixels and a frame-rate of 25 fps. The video capturing setup is shown in Fig. 1. The statistics of the number of frames per sequence is shown in Fig. 2. The number of frames in a sequence is from 18 to 35, and most of the sequences contain approximately 25 frames. More details about the database can be found in [16].

We then extracted pose sequences from RGB images of OUMVLP. More specifically, we employed pretrained versions of OpenPose [12] and AlphaPose [13] to extract human joint information. As shown in Fig. 3, the estimated results include 18 joints in total: Nose, Neck, RShoulder (right shoulder, the following names named similarly), RElbow, RWrist, LShoulder, LElbow, LWrist, RHip, RKnee, RAnkle, LHip, LKnee, LAnkle, Reye, LEye, REar, and LEar. We show some samples from multiple views in Fig. 1, and some extracted pose sequences in Fig. 4.

Two datasets in OUMVLP-Pose were created. One was created using the OpenPose method, and another was by the AlphaPose method. The two datasets contain the same number of subjects and the same parameters. The only difference is the pose accuracy of the two pose estimation methods. After the acceptance of the paper, we will release OUMVLP-Pose to the research community.

IV. Performance Evaluation

First, we evaluate the performances of two existing model-based approaches on the constructed database: one is a method using Fourier transform analysis on leg movement [4], which was proposed in the early stage of gait recognition, and the other is a recent deep learning-based approach. Second, we compare the model-based approach with widely used appearance-based approaches.

A. Model-Based Benchmarks

1) Fourier Transform Analysis on Legs Movement: Model-based approaches have been actively studied mainly in the early stage of gait recognition studies. We chose a method from such model-based approaches for comparison with the recent deep learning-based approaches. More specifically, we chose a Fourier transform-based approach to gait recognition proposed by Cunado *et al.* [4]. The method extracts two angles from legs, the thigh angle and the knee angle. The angle values from a sequence can be put into a vector which will be the input of Fourier transform. The length of the angle vector in our experiments was set to 20 as the other experiments. The phase-weighted Fourier magnitude spectra is the feature vector for classification. We implemented the algorithm as described in [4]. The classifier we used is NN (nearest neighbor). Compared with the CNN-based methods described in the following part of the paper, the Fourier method only uses 2 joint angles for gait recognition, not all joints as other methods.

2) CNN for Feature Extraction: Considering the recent progress of deep learning approaches on many computer vision and biometric authentication tasks, it is natural to employ the deep learning-based approaches for the model-based method of gait recognition.

For this purpose, we first apply a normalization procedure to the pose sequences because the size of a human body changes according to the distance between the subject and the camera, which is undesirable for recognition purposes. In this study, we used the distance $d_{\text{neck}-\text{hip}}$ between the neck and the middle point of the hip (computed as the center of RHip and LHip) as a normalization factor. We normalize the position so that the neck joint $\vec{p}_{\text{neck}}$ is located at the origin and the neck-hip distance $d_{\text{neck}-\text{hip}}$ is unity. Specifically, the position of the i-th body joint $\vec{p}_i$ is normalized to a new position $\vec{p}_i'$ in the normalized coordinate as

$$\vec{p}_i' = \frac{\vec{p}_i - \vec{p}_{\text{neck}}}{d_{\text{neck}-\text{hip}}}. \tag{1}$$

Next, we apply deep learning-based approaches to the normalized pose sequences. As one of the most standard methods, we apply a convolutional neural network to the sequence of

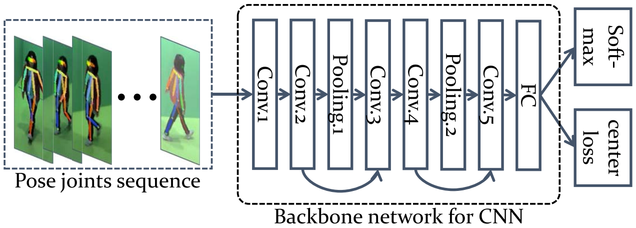


Fig. 5. The network structure for the CNN-Pose method. Two losses are used to train the network.

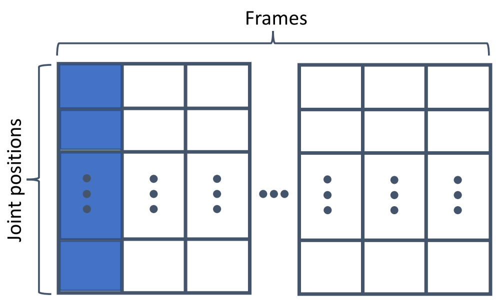


Fig. 6. The positions of 18 joints are stored in a column vector. N vectors from the N consecutive frames are concatenated to a matrix of size $36 \times N$.

a pose (i.e., the normalized positions of the joints). More specifically, we first construct a matrix whose row and column correspond to the normalized position of the joints and frames. Since we have a two-dimensional position (x, y) for each of the 18 joints for N frames, the size of the matrix is 36 $\times N$. The data structure sent to the CNN is illustrated in Fig. 6. Given the matrix as input, we then apply two-dimensional convolution layers, pooling layers, and a full connection layer, as shown in Fig. 5. The network is similar to that in [37], but with fewer layers, and it is easier to train. To normalize in the temporal domain, the frame with the largest distance between two feet is selected as the first frame for the input data. If the frames after the selected first frame are not enough to N frames, the frames before the selected first frame will be padded to the end.

For feature extraction in gait recognition, it is crucial to reduce the intra-class variation and enlarge the inter-class variation, and hence, the multiloss strategy is employed to optimize the network. As in [37], we employed two losses: cross-entropy loss based on softmax and a center loss. The cross-entropy loss with softmax can be used for classifying the input into multiple different classes while the center loss learns a center for deep features of each class and gives a penalty for the distances between the deep features. With joint supervision, we can simultaneously enlarge the inter-class differences and reduce the intra-class differences. We call the above mentioned CNN network architecture CNN-Pose throughout this paper.

After we trained a model with the AlphaPose data, we analyzed the distributions of the intra-class distances and inter-class distances of the extracted gait feature vectors. All samples from the same subject were used to compute the intra-class variation, and samples from different subjects were for the inter-class variation. The histograms of the two variations

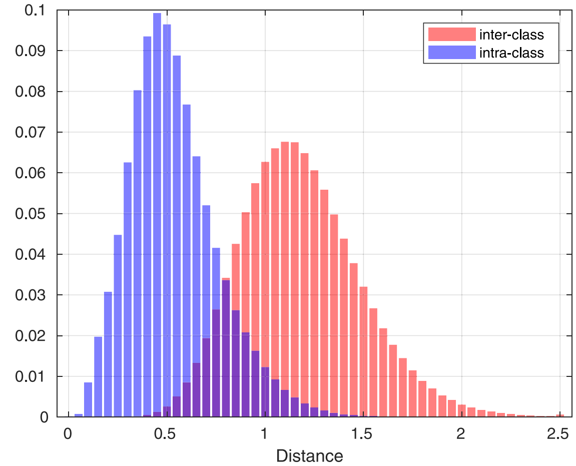


Fig. 7. The distributions of the inter-class distance and the intra-class distance on the test set.

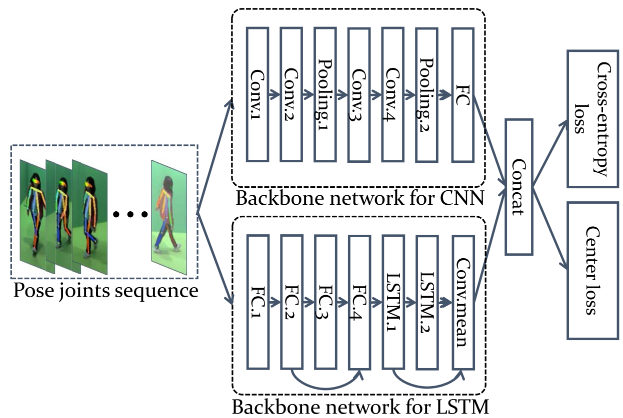


Fig. 8. The PTSN architecture, which contain two pipelines: CNN and LSTM.

are shown in Fig. 7. From the figure we can find that the extracted gait feature can distinguish different subjects even there is still an overlap between the two distributions.

3) PTSN by Combining CNN and LSTM: In addition, we introduce another popular architecture to encode temporal information from pose sequences, i.e., long short-term memory (LSTM), as shown in Fig. 8, which is from the PTSN method in [14]. Two types of features extracted through CNN and LSTM are combined to capture the dynamic-static information from gait poses, which has a powerful representation capacity to extract invariant features from different gaits. We call the above mentioned CNN network architecture PTSN throughout this paper.

B. Appearance-Based Benchmarks

To evaluate the performance of the model-based features, some appearance-based features should also be involved and compared. Therefore, we employ the following typical appearance-based benchmarks, which are designed for cross-view gait recognition ranging from classical linear algebraic methods to recent deep learning-based methods.

- The VTM method [41] acquires the VTM with the training data of multiple subjects from multiple view angles.

TABLE II
EXPERIMENTAL DESIGN OF THE OUMVLP-POSE DATABASE

Training	Test	
	Gallery Set	Probe Set
ID: 1-5153 Seq: 00, 01	ID: 5154-10307 Seq: 00	ID: 5154-10307 Seqs: 01

In a recognition phase, the VTM transforms gallery gait features into the same view angle as that of an input feature, and the features match under the same view.

- Linear discriminant analysis (LDA) [42] is adopted as a baseline in OUMVLP. Principal component analysis (PCA) is first applied to an unfolded feature vector of GEI to reduce dimension and, subsequently, LDA is applied to obtain the discriminant features.
- GEINet [43] is based on one of the simplest CNNs where one input GEI is fed, and the number of nodes in the final layer (fc4) is equal to the number of training subjects. A softmax value calculated from the output of the final layer is regarded as the probability of matching a corresponding subject.
- LB (local at the bottom) [44] is one of the state-of-the-art gait recognition networks that takes a pair of GEIs as the input. Paired convolutional filters are used to compute the pixelwise weighted sum of the pair on the first layer, which simulates the differences (i.e., matching) between a probe and a gallery image. The cross-entropy loss is adopted for training, where the two softmax values return the probability that the input pair belongs to the same subject or different subjects.

C. Experimental Design and Evaluation Criteria

There are 10,307 subjects in the database. We divided them into two sets. The first one, which contains 5,153 subjects, is the training set, and the second, which contains the remaining 5,154 subjects, is the test set. The test set is separated into a gallery set and a probe set. Since each subject roughly owns 2 sequences, we put the sequence "00" in the gallery set and the sequence "01" in the probe set. The experimental design is also shown in Table II.

In our training phase, we set the training batch as 1024, and the learning rate as 0.001. The learning rate decreased 10 times every 300 iterations. The size of input data is $B \times N \times 36$ as shown in Fig. 6, where B is the batch size in training and N is the number of frames in a sequence. In our training phase we choose 20 for N. The 20 continuous frames which start from the frame with the largest distance between feet will be taken as the input sequence. If the selected frames are less than 20, we will select the remaining frames from the start of the original sequence and pad them to the end of the selected.

Two evaluation criteria were employed to evaluate different recognition accuracies: the rand-1 recognition rate, and the equal error rate (EER). The results and analysis are listed in the following subsections.

D. Performance of Benchmarks

First, we evaluated the recognition accuracy of CNN with the rank-1 recognition rates on the two datasets, AlphaPose and OpenPose. Due to the evaluation in the OU-ISIR MVLP database, the recognition rate of the 0-90° gallery vs. the 0-90° probe is similar to that of the 0-90° gallery vs. the 180-270° probe, the 180-270° gallery vs. the 0-90° probe, and the 180-270° gallery vs. the 180-270° probe. We also adopted the same evaluation criteria to focus on four typical view angles 0°, 30°, 60°, and 90°. The specific CNN network is shown in Table III and Fig. 5.

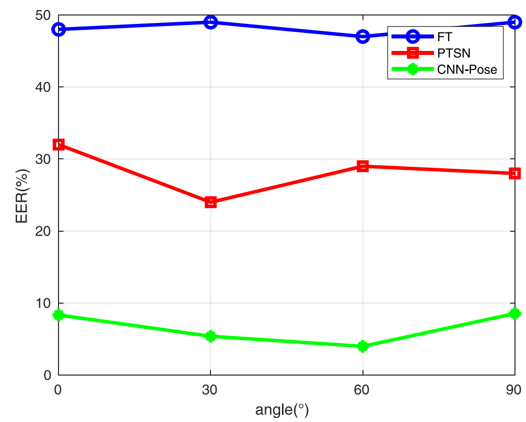


Fig. 9. The EERs of three model-based methods on the AlphaPose dataset where the probe angle is the same as the gallery angle.

Table IV and Table V show the rank-1 recognition rates on two datasets with the CNN network. From the two tables, we find that the recognition rate will be relatively high when the probe angle is the same as the gallery angle. View variation can greatly reduce the recognition rate. The average rate on the ApahaPose data is 20.42% and greater than 14.76% on the OpenPose data. In Table VI and Table VII, the EERs are listed. A lower EER value means a better recognition rate. From the four tables, it is obvious that a better quality pose estimation can lead to a better recognition rate.

We then evaluated the recognition accuracy of all model-based benchmarks mentioned previously. The rank-1 recognition rates are shown in Table VIII, and the EERs are illustrated in Fig. 9. The probe angle of each experiment in Table VIII and Fig. 9 is the same as its gallery angle. From the results, it can be found that the FT method achieves an average recognition rate of 0.73%. The recognition rate for random guess is $1/5154 = 0.0194\%$. The FT method is about 37 times better than random guess. By taking account of the fact that the original paper reported 80% and 90% rank-1 identification rates on 10 galleries by kNN classifiers (k = 1 and 3, respectively), the obtained accuracy for the FT method on our database is reasonable. It shows that even one thigh angle and one knee angle can contribute to gait recognition obviously.

The CNN methods in Table VIII and Fig. 9 achieves much better performance than the FT method for more body joints and the CNN classifiers. We believe that there is still great potential for pose-based methods. The pose data are 2D data in the experiments in this paper. Some methods can convert 2D pose data to 3D as that in [37], which will obviously improve the robustness with view variation. Besides, the progress on pose estimation will also advance model-based gait recognition for their better accuracy on human pose estimation.

TABLE III
IMPLEMENTATION DETAILS OF THE CNN NETWORK

Layers	Number of filters	Filter size	Stride	Padding	Group	Activation function
Conv.1	32	3×3	1	0	Y	ReLU
Conv.2	64	3×3	1	0	N	ReLU
Pooling.1	-	2×2	2	0	N	-
Conv.3	64	3×3	1	1	Y	ReLU
Eltwise.1	Sum operation between conv. and pooling layer					
Conv.4	128	3×3	1	0	Y	ReLU
Pooling.2	-	2×2	2	0	N	-
Conv.5	128	3×3	1	1	Y	ReLU
Eltwise.2	Sum operation between conv. and pooling layer					
FC	512	-	-	-	N	-

TABLE IV
RANK-1 RECOGNITION RATES BY CNN NETWORK USING OPENPOSE DATASET FOR ALL COMBINATIONS OF VIEWS

Probe \ Gallery	**0°**	**15°**	**30°**	**45°**	**60°**	**75°**	**90°**	**180°**	**195°**	**210°**	**225°**	**240°**	**255°**	**270°**	**mean**
0°	31.98	18.73	11.16	9.14	6.35	4.3	2.11	5.6	6.36	4.53	5.27	4.38	2.46	2.18	8.18
15°	16.06	52.38	30.73	21.49	14.4	7.31	4.22	6.46	10.64	7.7	9.24	7.36	3.96	2.86	13.92
30°	9.77	30.31	53.39	45.28	27.98	15.34	7.73	5.76	10.3	10.62	14.78	11.44	6.19	4.74	18.12
45°	7.85	21.22	43.38	68.58	53.35	29.16	13.33	5.92	9.24	10.69	17.48	15.83	9.79	7.5	22.38
60°	4.85	13.52	26.43	51.53	69.07	41.8	18.22	4.49	7.86	9.19	16.28	16.69	10.46	7.72	21.29
75°	3.54	8.68	16.4	30	42.92	58.24	30.64	3.45	5.95	7.07	12.5	13.82	11.38	10.2	18.2
90°	2.06	3.78	7.29	12.19	16.89	28.73	37.93	2.34	3.25	4.45	7.04	8.99	8.73	8.89	10.9
180°	4.87	6.19	5.74	5.21	4.7	3.29	2.03	33.73	13.06	7.61	6.43	4.93	2.49	1.53	7.27
195°	6.31	12.02	11.74	10.94	9.11	5.5	3.52	16.07	50.02	22.1	20.06	12.19	5.6	3.29	13.46
210°	4.14	7.06	11.02	11.3	9.74	6.61	4.17	7.66	20.5	31.21	26.79	16.18	7.7	4.2	12.02
225°	4.99	10.21	16.52	21.48	20.65	15.56	9.42	8.49	19.82	27.67	61.89	42.27	18.63	9.52	20.51
240°	4.11	7.48	12.92	17.71	19.27	14.68	9.69	5.07	11.8	14.76	38.19	52.09	22.73	11.1	17.26
255°	2.73	5.38	8.16	11.88	14.54	14.21	11.19	2.79	6.5	8.34	19.5	25.33	40.9	20.76	13.73
270°	2.05	3.15	5.02	8.02	10.32	11.44	11.42	1.6	3.24	4.6	9.99	11.07	18.35	31.11	9.38
mean	7.52	14.29	18.56	23.20	22.81	18.30	11.83	7.82	12.75	12.18	18.96	17.33	12.10	8.97	14.76

TABLE V
RANK-1 RECOGNITION RATES BY CNN NETWORK USING ALPHAPOSE DATASET FOR ALL COMBINATIONS OF VIEWS

Probe \ Gallery	**0°**	**15°**	**30°**	**45°**	**60°**	**75°**	**90°**	**180°**	**195°**	**210°**	**225°**	**240°**	**255°**	**270°**	**mean**
0°	47.25	34.46	23.58	17.64	12.63	7.48	5.22	7.85	7.5	8.13	9.39	8.42	5.55	5.44	14.32
15°	28	64.53	51.11	37.8	25.99	15.75	9.7	8.16	11.48	12.97	15.97	14.69	8.7	7.31	22.3
30°	18.85	47.59	69.13	62.05	42.84	26	15.75	7.76	11.73	15	21.76	19.28	12.18	10.12	27.15
45°	12.89	35.37	62.32	76.4	62.52	39.36	22.18	7	10.32	15.82	24.7	23.29	15.56	12.71	30.03
60°	8.69	24.63	44.06	61.71	73.21	50.74	27.92	6.17	8.81	12.41	22.85	23.79	17.27	15.23	28.39
75°	5.99	14.68	27.13	38.84	52.46	62.67	39.2	4.65	5.95	9.95	16.53	19.62	16.93	16.44	23.64
90°	4.23	8.91	15.68	22.16	28.68	39.35	49.07	3.26	3.84	6.88	12.29	15.43	14.93	15.38	17.15
180°	5.68	7.6	7.57	6.18	5.39	4.03	3.22	30.8	12.97	7.64	7.5	6.05	3.39	2.72	7.91
195°	7.08	13.16	15.05	13.09	11.09	7.45	5.38	15.74	38.78	20.78	17.77	12.35	7.27	4.78	13.55
210°	6.43	13.08	17.67	17.1	15.21	10.93	7.55	8.48	19.44	34.14	30.25	19.47	11.31	7.16	15.59
225°	8.85	17.86	25.12	27.69	26.6	20.13	14.21	8.97	17.82	31.73	63.08	45.79	25.47	16.53	24.99
240°	7.47	15.93	22.43	25.06	28.07	22.03	16.76	6.01	11.28	20.34	44.55	60.6	35.22	21.25	24.07
255°	5.88	11.39	16.24	19.4	22.43	20.59	18.19	3.43	7.04	13.07	25.65	35.44	51.13	33.41	20.23
270°	4.53	8.08	13.45	15.7	19.83	19.81	18.54	2.86	4.48	8.27	16.13	22.49	31.9	44.93	16.5
mean	12.27	22.66	29.32	31.49	30.50	24.74	18.06	8.65	12.25	15.51	23.46	23.34	18.34	15.24	20.42

E. Comparison Benchmarks With Appearance-Based Methods

We compared the recognition rates with those by some appearance-based methods. The results of VTM, LDA, GEINet, and LB) are from the paper which introduces OUMVLP [16], and the results of GaitSet are from [31]. All comparisons are listed in Table IX. Different from the results in Table VIII, the results in Table IX are the averages on different probe angles with specific gallery angles 0°, 30°, 60° and 90°. The corresponding EERs are illustrated in Fig. 10. From the comparisons in Table IX and Fig. 10, we can find most appearance-based methods achieves better recognition rates than the model-based ones. This shows that the OUMVLP-Pose database is challenging because only the positions of the joints are included. There is no body shape or body appearance feature.

F. Impact on the Number of Training Subjects

The recognition rate of the CNN network changes with different quantities of training data. We set three different training sets for evaluation: 1,000, 3,000 and 5,153. The last 5,154 subjects are put into the test set. The impact of the different training subjects is shown in Table X on the AlphaPose dataset. In each of the experiments, the probe angle is the same as the gallery angle. For 00°, 30°, 60° and 90°, the recognition rate rises with an increased number of training subjects. We can

TABLE VI
EERs Using OpenPose Dataset for All Combinations of All Views

Probe \ Gallery	0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	mean
0°	8.74	10.55	12.83	12.64	14.17	16.75	20.86	17.35	15.22	20.37	13.95	15.76	17.44	20.75	15.53
15°	9.47	6.03	9.02	8.29	10.45	13.06	18.44	16.21	11.82	18.11	10.97	13.03	14.92	18.7	12.75
30°	11.93	8.93	8.38	7.7	9.61	11.52	17.29	16.81	13.25	17.38	10.92	12.53	13.82	17.7	12.7
45°	11.28	8.19	7.52	4.56	5.73	7.93	14.63	15.69	12.36	17.06	8.72	9.84	11.46	15.39	10.74
60°	12.66	9.89	9.29	5.4	4.75	7.01	13.95	16.7	12.99	17.71	9.14	9.69	10.79	15.07	11.07
75°	14.83	12.95	11.85	8.05	7.13	6.71	12.95	17.53	14.77	19.2	10.42	10.74	11.1	14.81	12.36
90°	20.09	18.17	17.4	14.36	13.65	12.76	13.44	21.92	19.44	21.73	15.28	15.76	15.74	16.57	16.88
180°	15.76	15.41	16.69	15.59	16.44	18.16	21.73	12.62	15.06	20.68	14.67	17.31	18.68	22.07	17.2
195°	14.64	12.23	13.9	12.56	13.98	15.61	20.06	15.16	8.7	16.25	10.49	13.22	15.5	19.54	14.42
210°	20.24	18.24	17.45	17.38	17.97	19.33	23.33	20.57	16.51	16.98	15.22	17.5	19.12	22.66	18.75
225°	13.87	11.45	11.54	9.15	10.09	11.59	16.27	15.53	11.04	16.02	5.11	7.44	10.15	15.51	11.77
240°	15.39	13.1	13.19	10.3	10.61	11.83	16.59	17.73	13.17	17.64	7.31	6.58	9.5	15.58	12.75
255°	17.69	16.02	15.28	12.55	12.13	12.93	17	20.08	16.26	19.81	10.47	9.53	8.69	14.46	14.49
270°	20.92	19.57	18.94	16.21	15.82	16.04	18.28	23.01	19.91	23.03	15.6	15.72	13.91	14.11	17.93
mean	14.82	12.91	13.09	11.05	11.61	12.95	17.49	17.64	14.32	18.71	11.31	12.48	13.63	17.35	14.24

TABLE VII
EERs Using AlphaPose Dataset for All Combinations of All Views

Probe \ Gallery	0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	mean
0°	8.34	8.31	9.89	10.6	11.62	14.32	17.41	21.16	18.44	22.06	13.46	14.76	16.26	17.98	14.61
15°	9.07	4.93	5.94	6.64	8.18	10.7	13.73	21.11	16.05	20.34	10.67	11.76	13.34	15.39	11.99
30°	10.99	6.68	5.39	5.42	6.57	8.81	12.48	21.32	16.09	19.72	9.91	10.57	12.12	14.12	11.44
45°	11.61	7.47	5.7	4.21	4.67	6.83	10.1	21.24	16.09	19.73	8.71	8.95	10.67	12.38	10.6
60°	12.98	8.46	6.82	4.64	3.99	5.48	8.55	22.01	16.83	20.04	9.3	9.1	9.81	11.32	10.67
75°	15.44	10.81	9	6.93	5.44	5.14	8.21	23.24	18.14	21.24	10.9	10.22	10.77	11.81	11.95
90°	18.27	14.3	12.09	10.58	9.18	8.42	8.55	24.69	20.56	22.72	13.36	12.27	12.64	13.29	14.35
180°	21.7	20.53	20.04	20.62	20.83	22.05	23.86	18.9	22.62	26.64	20.32	21.35	23.7	24.75	21.99
195°	18.77	15.39	15.54	15.87	15.95	17.7	19.99	22.14	14.05	20.94	14.87	15.68	18.32	20.53	17.55
210°	22.96	20.67	19.69	20.12	19.74	21.43	22.25	27.39	21.01	21.16	18.67	19.03	21.09	22.86	21.29
225°	13.99	10.51	9.12	8.46	8.26	9.93	11.89	20.79	14.65	18.14	5.25	6.59	9.12	11.81	11.32
240°	14.79	11.19	9.39	8.28	7.99	9.18	11.24	21.89	15.82	19.37	6.57	5.79	7.87	10.73	11.44
255°	16.38	13.25	11.08	9.51	9.1	9.76	11.52	23.36	18.17	20.64	8.91	7.58	7.02	9.86	12.58
270°	17.98	14.77	13.08	11.32	10.58	11	11.95	25.25	20.34	22.61	11.89	10.68	9.75	9.63	14.35
mean	15.23	11.95	10.91	10.23	10.15	11.48	13.70	22.46	17.78	21.10	11.63	11.74	13.03	14.75	14.01

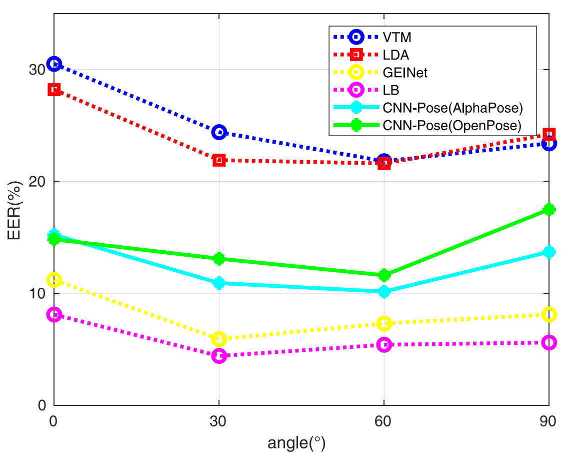


Fig. 10. The EERs of four appearance-based methods (VTM, LDA, GEINet and LB) and the pose-based CNN method.

TABLE VIII
The Rank-1 Recognition Rates of Three Model-Based Methods on the AlphaPose Dataset Where the Probe Angle Is the Same as the Gallery Angle

Methods	0°	30°	60°	90°	mean
Fourier transform analysis	0.33	0.76	0.96	0.87	0.73
PTSN	24.0	38.2	29.3	28.5	30.0
CNN-Pose	47.3	69.1	73.2	49.0	59.7

TABLE IX
The Rank-1 Recognition Rates of Four Appearance-Based Methods (VTM, LDA, GEINet, LB and GaitSet) and the Pose-Based CNN Method. The Rates Are the Averages on Different Probe Angles With a Specific Gallery Angle 0°, 30°, 60° and 90°

Methods	0°	30°	60°	90°	mean
VTM [41]	17.4	21.4	21.6	21.6	20.5
LDA [42]	18.4	26.2	28.1	24.8	24.4
GEINet [43]	30.6	43.3	47.3	41.5	40.7
LB [44]	24.3	38.8	43.0	37.3	35.9
GaitSet [31]	79.5	89.9	88.1	87.8	86.3
CNN-Pose(OpenPose)	7.5	18.6	22.8	11.8	18.0
CNN-Pose(AlphaPose)	12.3	29.3	30.5	18.1	22.5

expect the recognition rate can continue to increase with more training data.

V. Conclusion

A large population pose database is introduced in this paper. It is a large database with multiple view angles and 10,307 subjects. The pose data were extracted from the RGB videos in the OU-ISIR multi-view large population database (OUMVLP) using deep learning-based pose estimation methods. Two datasets, the OpenPose dataset and AlphaPose dataset were created using two methods, OpenPose and AlphaPose, respectively. In addition to the body pose data, we also provide benchmarks and analysis on the database.

TABLE X
THE RANK-1 RECOGNITION RATES ON THE ALPHAPOSE SKELETON DATA WITH DIFFERENT NUMBERS OF TRAINING SAMPLES WHERE THE PROBE ANGLE IS THE SAME AS THE GALLERY ANGLE

#Training subjects	**0°**	**30°**	**60°**	**90°**	**mean**
1,000	11.9	32.2	31.9	10.4	21.6
3,000	45.8	67.3	72.9	48.8	58.7
5,153	47.3	69.1	73.2	49.0	59.7

With progress in human body modeling, we believe that model-based gait recognition should be investigated further. The proposed benchmark method, CNN-Pose, is relatively simple. However, it achieved encouraging results. The pose data are in a 2D dimension and are not robust to view variation. Therefore, in the future, some 3D human models can be built for gait recognition. Since the model data is in a 3D space, we can rotate the model in the 3D space and extract view-invariant gait features. In addition, the model-based feature should be more robust to clothing and carrying condition changes.

ACKNOWLEDGMENT

The authors are deeply grateful to Mr. Xiang Li for his help with the computer resource setup.

REFERENCES

[1] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, Feb. 2006.
[2] C. Wang, J. Zhang, L. Wang, J. Pu, and X. Yuan, "Human identification using temporal information preserving gait template," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2164–2176, Nov. 2012.
[3] T. H. Lam, K. H. Cheung, and J. N. Liu, "Gait flow image: A silhouette-based gait representation for human identification," *Pattern Recognit.*, vol. 44, no. 4, pp. 973–987, 2011.
[4] D. Cunado, M. S. Nixon, and J. N. Carter, "Using gait as a biometric, via phase-weighted magnitude spectra," in *Proc. Int. Conf. Audio VideoBased Biometric Pers. Authentication*, 1997, pp. 93–102.
[5] L. Wang, H. Ning, T. Tan, and W. Hu, "Fusion of static and dynamic body biometrics for gait recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 2, pp. 149–158, Feb. 2004.
[6] C. Yam, M. S. Nixon, and J. N. Carter, "Automated person recognition by walking and running via model-based approaches," *Pattern Recognit.*, vol. 37, no. 5, pp. 1057–1072, 2004.
[7] R. Urtasun and P. Fua, "3D tracking for gait characterization and recognition," in *Proc. 6th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Seoul, South Korea, 2004, pp. 17–22.
[8] G. Ariyanto and M. S. Nixon, "Marionette mass-spring model for 3D gait biometrics," in *Proc. 5th IAPR Int. Conf. Biometrics (ICB)*, New Delhi, India, 2012, pp. 354–359.
[9] J. Shotton *et al.*, "Real-time human pose recognition in parts from single depth images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Providence, RI, USA, 2011, pp. 1297–1304.
[10] D. Kastaniotis, I. Theodorakopoulos, and S. Fotopoulos, "Pose-based gait recognition with local gradient descriptors and hierarchically aggregated residuals," *J. Electron. Imag.*, vol. 25, no. 6, 2016, Art. no. 063019.
[11] R. Tanawongsuwan and A. Bobick, "Gait recognition from time-normalized joint-angle trajectories in the walking plane," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2. Kauai, HI, USA, 2001, pp. 726–731.
[12] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 7291–7299.
[13] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, 2017, pp. 2334–2343.
[14] R. Liao, C. Cao, E. B. Garcia, S. Yu, and Y. Huang, "Pose-based temporal-spatial network (PTSN) for gait recognition with carrying and clothing variations," in *Proc. Chin. Conf. Biometric Recognit. (CCBR)*, 2017, pp. 474–483.
[15] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, vol. 4. Hong Kong, China, 2006, pp. 441–444.
[16] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition," *IPSJ Trans. Comput. Vis. Appl.*, vol. 10, no. 1, p. 4, Feb. 2018.
[17] R. Gross and J. Shi, "The CMU motion of body (MOBO) database," Dept. Robotics Inst., Carnegie Mellon Univ., Pittsburgh, PA, USA, Rep. CMU-RI-TR-01-18, Jun. 2001.
[18] M. S. Nixon, J. N. Carter, J. M. Nash, P. S. Huang, D. Cunado, and S. V. Stevenage, "Automatic gait recognition," *Biometrics*, vol. 7, no. 02, p. 3, 1999.
[19] S. Samangooei, J. Bustard, M. S. Nixon, and J. N. Carter, "On acquisition and analysis of a dataset comprising of gait, ear and semantic data," in *Multibiometrics for Human Identification*. Cambridge, U.K.: Cambridge Univ. Press, 2011, pp. 277–301.
[20] L. Wang, H. Ning, W. Hu, and T. Tan, "Gait recognition based on procrustes shape analysis," in *Proc. Int. Conf. Image Process. (ICIP)*, vol. 3. Rochester, NY, USA, 2002, pp. 433–436.
[21] D. López-Fernández, F. J. Madrid-Cuevas, Á. Carmona-Poyato, M. J. Marín-Jiménez, and R. Muñoz-Salinas, "The AVA multi-view dataset for gait recognition," in *Proc. Int. Workshop Activ. Monitor. Multiple Distrib. Sens.*, 2014, pp. 26–39.
[22] B. DeCann, A. Ross, and J. Dawson, "Investigating gait recognition in the short-wave infrared (SWIR) spectrum: Dataset and challenges," in *Proc. SPIE Biometric Surveillance Technol. Human Activ. Identif. X*, vol. 8712, 2013, pp. 101–116.
[23] Y. Iwashita, R. Baba, K. Ogawara, and R. Kurazume, "Person identification from spatio-temporal 3D gait," in *Proc. Int. Conf. Emerg. Security Technol.*, Canterbury, U.K., 2010, pp. 30–35.
[24] Y. Makihara *et al.*, "The OU-ISIR gait database comprising the treadmill dataset," *IPSJ Trans. Comput. Vis. Appl.*, vol. 4, no. 1, pp. 53–62, 2012.
[25] H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi, "The OU-ISIR gait database comprising the large population dataset and performance evaluation of gait recognition," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 5, pp. 1511–1521, Oct. 2012.
[26] P. Connor and A. Ross, "Biometric recognition by gait: A survey of modalities and features," *Comput. Vis. Image Understand.*, vol. 167, pp. 1–27, Feb. 2018.
[27] J. P. Singh, S. Jain, S. Arora, and U. P. Singh, "Vision-based gait recognition: A survey," *IEEE Access*, vol. 6, pp. 70497–70527, 2018.
[28] X. Ben, C. Gong, P. Zhang, X. Jia, Q. Wu, and W. Meng, "Coupled patch alignment for matching cross-view gaits," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 3142–3157, Jun. 2019.
[29] X. Ben, C. Gong, P. Zhang, R. Yan, Q. Wu, and W. Meng, "Coupled bilinear discriminant projection for cross-view gait recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 3, pp. 734–747, Mar. 2020.
[30] X. Ben, P. Zhang, Z. Lai, R. Yan, X. Zhai, and W. Meng, "A general tensor representation framework for cross-view gait recognition," *Pattern Recognit.*, vol. 90, pp. 87–98, Jun. 2019.
[31] H. Chao, Y. He, J. Zhang, and J. Feng, "GaitSet: Regarding gait as a set for cross-view gait recognition," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 33, 2019, pp. 8126–8133.
[32] Y. Zhang, Y. Huang, S. Yu, and L. Wang, "Cross-view gait recognition by discriminative feature learning," *IEEE Trans. Image Process.*, vol. 29, pp. 1001–1015, 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8759096
[33] C. Xu, Y. Makihara, X. Li, Y. Yagi, and J. Lu, "Cross-view gait recognition using pairwise spatial transformer networks," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Feb. 21, 2020, doi: 10.1109/TCSVT.2020.2975671.
[34] S. Yu *et al.*, "GaitGANv2: Invariant gait feature extraction using generative adversarial networks," *Pattern Recognit.*, vol. 87, pp. 179–189, Mar. 2019.
[35] X. Li, Y. Makihara, C. Xu, Y. Yagi, and M. Ren, "Gait recognition invariant to carried objects using alpha blending generative adversarial networks," *Pattern Recognit.*, vol. 105, Sep. 2020, Art. no. 107376.
[36] C. Song, Y. Huang, Y. Huang, N. Jia, and L. Wang, "GaitNet: An end-to-end network for gait based human identification," *Pattern Recognit.*, vol. 96, Dec. 2019, Art. no. 106988.

[37] R. Liao, S. Yu, W. An, and Y. Huang, “A model-based gait recognition method with body pose and human prior knowledge,” *Pattern Recognit.*, vol. 98, Feb. 2020, Art. no. 107069.

[38] L. Pishchulin *et al.*, “DeepCut: Joint subset partition and labeling for multi person pose estimation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 4929–4937.

[39] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, 2017, pp. 2961–2969.

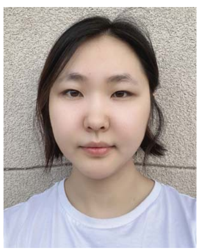
[40] S. Jin *et al.*, “Towards multi-person pose tracking : Bottom-up and top-down methods,” in *Proc. ICCV PoseTrack Workshop*, vol. 2, 2017, p. 7.

[41] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi, “Gait recognition using a view transformation model in the frequency domain,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2006, pp. 151–163.

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

[43] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, “GEINet: View-invariant gait recognition using a convolutional neural network,” in *Proc. Int. Conf. Biometrics (ICB)*, Halmstad, Sweden, 2016, pp. 1–8.

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



Weizhi An received the M.S. and B.S. degrees from the College of Computer Science and Software Engineering, Shenzhen University, China, in 2019 and 2016, respectively. She is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, University of Texas at Arlington, USA. Her research interests include computer vision and deep learning.

Shiqi Yu (Member, IEEE) received the B.E. degree in computer science and engineering from the Chu Kochen Honors College, Zhejiang University in 2002, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences in 2007. He is currently an Associate Professor with the Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China. He worked as an Assistant Professor and an Associate Professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences from 2007 to 2010, and Shenzhen University from 2010 to 2019. His research interests include computer vision, pattern recognition, and artificial intelligence.

Yasushi Makihara received the B.S., M.S., and Ph.D. degrees in engineering from Osaka University in 2001, 2002, and 2005, respectively. He was appointed as a specially appointed Assistant Professor (full-time), an Assistant Professor, and an Associate Professor with the Institute of Scientific and Industrial Research, Osaka University, in 2005, 2006, and 2014, where he is currently a Professor with the Institute for Advanced Co-Creation Studies. His research interests are computer vision, pattern recognition, and image processing, including gait recognition, pedestrian detection, morphing, and temporal super resolution. He has obtained several honors and awards, including the 2nd International Workshop on Biometrics and Forensics in 2014, the IAPR Best Paper Award, the 9th IAPR International Conference on Biometrics in 2016, the Honorable Mention Paper Award, and the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology, Prizes for Science and Technology, Research Category in 2014. He has served as an Associate Editor in Chief of the *IEICE Transactions on Information and Systems*, an Associate Editor of the *IPSJ Transactions on Computer Vision and Applications*, a Program Co-Chair of the 4th Asian Conference on Pattern Recognition in 2017, a Area Chair of ICCV 2019, CVPR 2020, and ECCV 2020. He is a member of IPSJ, IEICE, RSJ, and JSME.

Xinhui Wu received the B.S. degree from Southeast University, China in 2017, and the M.S. degree from the College of Computer Science and Software Engineering, Shenzhen University, China, in 2020. Her research interests include computer vision and deep learning.

Chi Xu received the B.S. degree in computer science and technology from the Nanjing University of Science and Technology, China, in 2012, where she is currently pursuing the Ph.D. degree in pattern recognition and intelligent system. Since January 2016, she has been with the Institute of Scientific and Industrial Research, Osaka University, Japan, as a Visiting Researcher. Her research interests are gait recognition, machine learning, and image processing.

Yang Yu revived the B.S. degree from the School of Electronic and Information, Dalian University of Technology, China, in 2011, and the M.S. degree from the School of Electronic and Information, Harbin Institute of Technology, China, in 2013. He is currently pursuing the Ph.D. degree with the Department of Information Science and Technology, Osaka University, Japan. His research interests are multiple object tracking and segmentation, human pose estimation, and machine learning.

Rijun Liao received the M.S. degree and B.S. degree from the College of Computer Science and Software Engineering, Shenzhen University, China, in 2018 and 2015 respectively. His research interests include computer vision and deep learning.

Yasushi Yagi (Member, IEEE) received the Ph.D. degree from Osaka University in 1991, where he is a Professor with the Institute of Scientific and Industrial Research. In 1985, he joined the Product Development Laboratory, Mitsubishi Electric Corporation, where he worked on robotics and inspections. He became a Research Associate in 1990, a Lecturer in 1993, an Associate Professor in 1996, and a Professor in 2003 with Osaka University, where he was also the Director of the Institute of Scientific and Industrial Research from 2012 to 2015, and the Executive Vice President from 2015 to 2019. His research interests are computer vision, medical engineering and robotics. He was awarded the ACM VRST2003 Honorable Mention Award, the IEEE ROBIO2006 Finalist of T. J. Tan Best Paper in Robotics, the IEEE ICRA’2008 Finalist for Best Vision Paper, the MIRU’2008 Nagao Award, and the PSIVT’2010 Best Paper Award. International conferences for which he has served as Chair include: FG’1998 (Financial Chair), OM-INVIS’2003 (Organizing Chair), ROBIO’2006 (Program Co-Chair), ACCV’2007 (Program Chair), PSVIT’2009 (Financial Chair), ICRA’2009 (Technical Visit Chair), ACCV’2009 (General Chair), ACPR’2011 (Program Co-Chair), and ACPR’2013 (General Chair). He has also served as the Editor of IEEE ICRA Conference Editorial Board from 2007 to 2011. He is the Editorial Member of *International Journal of Computer Vision* and the Editor-in-Chief of the *IPSJ Transactions on Computer Vision and Applications*. He is a fellow of IPSJ and a member of IEICE and RSJ.