Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

View and Clothing Invariant Gait Recognition Via 3D Human Semantic Folding

Jian Luo[1,*], Tardi Tjahjadi[2]

[1] Hunan Provincial Key Laboratory of Intelligent Computing and Language Information Processing, Hunan Normal University, Changsha, Hunan 410000, China;

[2] School of Engineering, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, United Kingdom.

Corresponding author: Jian Luo (E-mail: luojian@hunnu.edu.cn).

This work was supported in part by the National Natural Science Foundation of China under Grant 61701179, in part by the Natural Science Foundation of Hunan Province, China under Grant 2019JJ50363, in part by the China Scholarship Council under Grant 201808430285, and in part by the Hunan Provincial Science and Technology Project Foundation, China, under Grant 2018TP1018 and Grant 2018RS3065.

ABSTRACT A novel 3-dimensional (3D) human semantic folding is introduced to provide a robust and efficient gait recognition method which is invariant to camera view and clothing style. The proposed gait recognition method comprises three modules: (1) 3D body pose, shape and viewing data estimation network (3D-BPSVeNet); (2) gait semantic parameter folding model; and (3) gait semantic feature refining network. First, 3D-BPSVeNet is constructed based on a convolution gated recurrent unit (ConvGRU) to extract 2-dimensional (2D) to 3D body pose and shape semantic descriptors (2D-3D-BPSDs) from a sequence of gait parsed RGB images. A 3D gait model with virtual dressing is then constructed by morphing the template of 3D body model using the estimated 2D-3D-BPSDs and the recognized clothing styles. The more accurate 2D-3D-BPSDs without clothes are then obtained by using the silhouette similarity function when updating the 3D body model to fit the 2D gait. Second, the intrinsic 2D-3D-BPSDs without interference from clothes are encoded by sparse distributed representation (SDR) to gain the binary gait semantic image (SD-BGSI) in a topographical semantic space. By averaging the SD-BGSIs in a gait cycle, a gait semantic folding image (GSFI) is obtained to give a high-level representation of gait. Third, a gait semantic feature refining network is trained to refine the semantic feature extracted directly from GSFI using three types of prior knowledge, i.e., viewing angles, clothing styles and carrying condition. Experimental analyses on CMU MoBo, CASIA B, KY4D, OU-MVLP and OU-ISIR datasets show a significant performance gain in gait recognition in terms of accuracy and robustness.

INDEX TERMS Gait recognition, Human identification, Three-dimensional gait, Virtual Gait

I. INTRODUCTION

Gait recognition and understanding (GRU) has a wide range of applications in the field of anti-terrorism, intelligent monitoring, access control, criminal investigation, pedestrian behaviour analysis, medical studies and reality mining (e.g., [1]). The advantages of GRU, e.g., without requiring subjects' cooperation, difficult to disguise gait, and gait is easily observed in low-resolution video, make it particularly attractive for subject identification and behaviour analysis (e.g., [2]). However, to successfully implement a GRU method for practical applications, several important issues must be overcome. One of these is the change in camera view when the human subject walks at different data capture sessions. It is also challenging for GRU to realize view-invariant or cross-view gait recognition from different cameras with changes in both camera azimuth and elevation angles. In most cases, only changes in azimuth view changes are considered. If only a few views of gait sequences are available for training, and a single camera is used in testing in the presence of changes in both azimuth and elevation angles, then it is expected that the recognition rate will be significantly reduced.

There are many other covariate factors that affect the accuracy of GRU, e.g., occlusion, the integrity of the gait image segmentation, and variations in clothing styles, carrying items, scene illumination, and walking speed [3-4]. Clothing variation is one of the most significant. Experiment results in [5] show that the gait recognition rate when

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2997814, IEEE Access

wearing a coat is much lower than when carrying a bag due to the large area of the subject's silhouette affected. This influence affects many appearance-based gait recognition methods. Thus, some gait recognition methods incorporate gait data of subjects with various clothing styles, or eliminate their influence by extracting dynamic joint features or body parts that are less affected. However, it is difficult to collect sufficient training data with various clothing styles under different views for every subject, and thus clothing variation remains an important issue in gait recognition. Compared with algorithms for 2-dimensional (2D) gait recognition, the 3-dimensional (3D) approach provides more flexibility to deal with clothing variations, i.e., by using virtual dressing and 3D clothes. But there are only few related studies due to the complexity of 3D modelling and virtual dressing.

It is still a challenge to explore a GRU system involving a large population as most publicly available gait databases are limited to hundreds of subjects. However, it is worth noting that gait datasets involving large populations under different walking conditions have been published recently by Osaka University, i.e., OU-MVLP [6] with 14 views and 10,307 subjects, and OU-LP-Bag [7] with various carrying conditions and 62,528 subjects of all age ranges. As gait datasets involve larger populations, an emerging challenge is that the number of gait frames to be processed is typically enormous, requiring much processing time and storage space. The much larger gait datasets also mean more subjects are involved, and it becomes difficult to publish them due to privacy issues. The datasets are more likely to be published in the form of binary silhouette or gait energy image (GEI), limiting the development of gait feature extraction from RGB images. Without the RGB sequences, it is difficult to detect the detailed clothing styles and carried items. Thus, how to convert the high-dimensional gait sequences into high-level feature representation of structured data while retaining their semantic meaning has important research significance. Most gait feature representation methods, e.g., GEI [4], and data dimensionality reduction methods, e.g., principal component analysis (PCA), address the above problems, but the effect of dimensionality reduction often depends on the number of specific samples. The data after dimensionality reduction is difficult to describe by semantics, i.e., they are usually considered a 'black box'.

Based on the above, a View and Clothing Invariant Gait Recognition via 3D Human Semantic Folding (VCIGR-3DHSF) is proposed in this paper. The method converts raw gait images into high-level semantic description based on 3D parametric body model. The 3D human body semantic folding is introduced to represent the feature in high-level pattern space. By converting image signals into semantic descriptors, gait visual features are both effectively represented in a new semantic space as structured data, and the dimensionality of the gait features reduced under instance and semantic level.

The novelties of VCIGR-3DHSF are as follows. First, by incorporating convolution gated recurrent units (ConvGRU), an instance-level body parsing network, a clothing recognition network and virtual dressing method, the 2D to 3D body pose and shape semantic descriptors (2D-3D-BPSDs), and an estimation and optimizing framework are proposed. Second, by making full use of the extracted 3D gait semantic parameters and semantic folding, 2D gait images are transformed to a description in a new semantic pattern space. It converts the unstructured raw gait data into structured data called gait semantic images. Third a SoftMax classifier with top-down refining mechanism is proposed to deal with gait recognition under various view and clothing conditions. The refining mechanism using a priori knowledge adjusts the gait semantic patterns to achieve even better performances under various scenarios.

The rest of this paper is organized as follows. Section II presents the related work. Section III presents the implementation of VCIGR-3DHSF. Section IV presents the experimental results and Section V concludes the paper.

II. RELATED WORK

GRU is divided into model-free and model-based methods according to whether a relevant body model is constructed. A model-free GRU method extracts the statistical data of gait contours in a gait cycle and matches known gait contours with similar shape and motion characteristics. GEI [4,8], as a classical gait feature representation, has led to many energy images of related features, such as frame difference energy image [9], gait entropy image [10] and pose energy image (PEI) [11]. Gait energy maps have low computational complexity, and due to contour averaging have better suppression of image distribution noise.

A model based GRU method has more advantages for addressing covariate factors such as changes in camera view and clothing, occlusion and carried item due to its incorporation of body model parameters. However, it is necessary to estimate the parameters from the gait contour. The required image resolution is also higher than that of a model-free method. Most current gait models are based on 2D descriptions, ranging from skeleton to shape, e.g., 2D rod skeleton, hinged skeleton and ellipse shape descriptions [12-14]. Since the gait model is a 3D structure, it is important to study gait with a 3D modelling method [15-16]. However, in most cases multiple cameras or 3D camera are needed to construct 3D voxel or volume models. These generate unstructured with redundant point cloud data, and without embedded skeleton the data cannot be used to morph pose or deform the body shape.

Gait recognition methods with variable views or multiple-views can be classified into two categories, i.e., model-free or model-based. In model-free approach, view transformation model (VTM) as cross-view gait recognition is widely used by transforming gait features from one viewing perspective to another [17-18]. View-invariant gait

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2997814, IEEE Access

features are extracted for multi-view gait recognition, i.e., based on uncorrelated multilinear sparse local discriminant canonical correlation analysis [19], deterministic learning [20], complete canonical correlation analysis [21], and view-invariant feature selectors [22]. In recent years, the deep learning network-based methods, i.e., convolution neural networks (CNNs), have been proposed to directly extract multi-view gait features from GEIs for gait recognition [1, 23], or transform the multi-view gait feature to one specific view using one uniform deep model [24]. For model-based methods, the view-invariant gait recognition is achieved by 3D, 2.5-dimensional (2.5D) or 2D modelling of the human body, extracting the relevant features of the model, such as joint angles based on skeleton model [14], walking posture parameters [25-26], etc. 3D gait entropy volume (3D-GEnV) [15] requires multiple views of a subject in order to construct the 3D volume model.

To address clothing variations, more attentions are given to certain body parts that are less sensitive to clothing styles [27], i.e., legs, using adaptive weight control strategy. In [13], lower limb joint angles are chosen as gait dynamic feature which is robust to clothing styles, and deterministic learning is used for recognition. A statistical shape analysis approach addresses various dressing by parsing GEI into three shape sections for feature extraction, i.e., horizontal, vertical and grid resolution [3]. The drawback of this approach is its dependency on the viewing angles. In [28], the combination of RGB, depth and audio features, are used to improve the robustness against dressing conditions including shoes changes. In [19], a fusion strategy combines the spatial-temporal and kinematic features for gait recognition, using deterministic learning to address dressing conditions. In [29] a time-based long short-term memory (LSTM) graph model is discussed for gait recognition, and a gait skeleton graph which is less sensitive to dressing is used for feature representation.

Most successful GRU methods have good results in fixed scenarios with limited conditions. Since human walking and body movement posture are affected by various factors as already mentioned, the generalization and recognition rate of a gait behaviour recognition algorithm still need to be greatly improved [30]. Especially in 3D gait recognition, little research has exploited 3D parametric body model and virtual dressing, which resulted in a lack of an effective way to describe gait using semantic descriptors. In order to facilitate 3D gait research and overcome the above-mentioned problems, VCIGR-3DHSF is proposed to extract semantic parameters of gait using ConvGRU-based 2D to 3D body parameters estimation network and a clothing recognition network. The semantic gait features are represented in 3D semantic pattern space by semantic folding. To improve the gait recognition accuracy the feature refining mechanism uses a priori knowledge of walking conditions to adjust the gait semantic folding image (GSFI) features before input to a SoftMax classifier.

III. PROPOSED METHOD: VCIGR-3DHSF

A. OVERVIEW

Fig.1 shows the overview of the proposed VCIGR-3DHSF. VCIGR-3DHSF is composed of three schemes. The first scheme extracts 2-dimensional (2D) to 3D body pose and shape semantic descriptors without clothes (2D-3D-BPSDs) from 2D gait images. It is based on our end to end 3D body pose, shape and viewing data estimation network (3D-BPSVeNet), and an optimizing process based on virtual dressing. The second is the 3D human semantic folding which encodes a sequence of scalar 2D-3D-BPSDs into visible GSFI based on sparse distributed representations (SDRs). The third is the view and clothing style invariant GSFI feature refinement based on GSFI refining network (GSFI-RNet) for better performance using a priori knowledge. This involves body parsing and clothing recognition network.

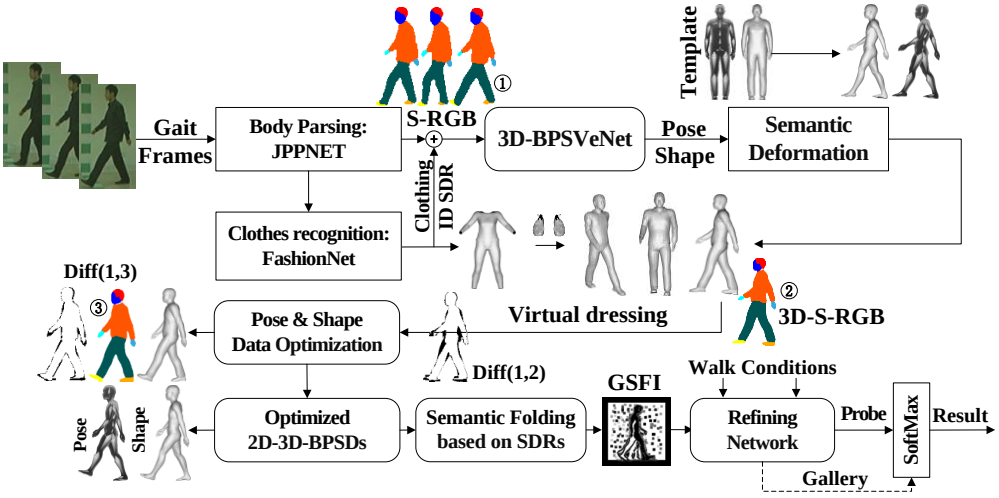


FIGURE 1. Overview of VCIGR-3DHSF.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2997814, IEEE Access

TABLE 1. Semantic parameters of human body shape and pose.

Category	Parameters	Category	Parameters	Category	Parameters
Shape-Global	Gender	Shape-Head	Head fat	Pose-Head-joints	neck
	Age		H-horizontal scale	Pose-Arms-joints	Left-shoulder
	Muscle		H-vertical scale		Right-shoulder
	Weight	Shape-Neck	Neck fat		Left-elbow
	Height		N-vertical scale		Right-elbow
	Proportions	Shape-Torso	Torso depth scale		Left-wrist
Shape-Arms	Arm length		T-horizontal scale		Right-wrist
	Arm thickness		T-vertical scale	Pose-Legs-joints	Left-hip
	Hand scale		Breast scale		Right-hip
Shape-Legs	Leg length		Stomach scale		Left-knee
	Leg thickness		Hip depth scale		Right-knee
	Foot scale		Buttocks volume		Left-ankle
Pose-Torso-joints	root	Pose-Torso-joints	chest		Right-ankle

B. 3D PARAMETRIC BODY MODEL WITH VIRTUAL CLOTHING

We refer the parameterized body model as the structured body mesh described by semantic body parameters. The deformation relationships between semantic body parameters and 3D mesh vertices are based on the statistical learning algorithms provided in the 3D body dataset. Table 1 shows the semantic body shape and pose parameters used in the proposed method. The shape descriptors are manually selected from around a hundred body shape parameters according to their sensitivity in gait recognition. Their values are normalized to the range [0 1], i.e., 0.5 is the average value. The pose joints are based on the skeleton of CMU mocap, and each joint has three degrees of freedom (DOF). The skeleton is embedded, and the 3D parametric model can be deformed both in shape and pose according to the given body parameters as shown in Fig. 1. To effectively extract the semantic gait features, the proposed method uses the 3D instances from the makehuman system [31], and the body parametric modelling method of our previous work [32].

We proposed a 2D-3D-BPSDs estimation method via a measuring function based on their silhouette difference as in [32], where binary 2D gait silhouettes are used for 3D body estimation. However due to the absence of RGB information, the estimation accuracy still needs to be improved, e.g., in 2D binary images it is not possible to distinguish a right foot from a left foot. If the two feet or hands overlap or self-occlusion occurs, then the precise position of them cannot be located. Furthermore, the speed of the required iterative computing is influenced by the initial 3D pose, i.e., the closer it is to the 2D gait, the smaller is the computational cost.

In order to improve the efficiency and the accuracy of the 2D-3D-BPSDs estimation, a sequence of gait silhouettes is utilized to estimate the semantic parameters of the 3D body model. We introduce the instance-level body parsing to obtain colour gait silhouettes for the estimation. The body parsing simultaneously segments the body from 2D images and parses each instance into finer grained body parts (i.e., hair, head, neck, left/right-hand, left/right-leg, foot, etc.). With more detailed 2D body parsed gait images, different body parts can be located more easily. By introducing a clothing recognition network, the clothing style is determined and used in the 3D body modelling by virtual dressing as shown in Fig. 1. The network eliminates the clothing influences and helps to improve the accuracy of the shape parameters estimation.

3D parametric body model, as a structured and parameters-controlled model, can morph to various 3D body using different body shape and pose parameters. The clothing is separated from the body model and virtual dressing is used to dress the body. Unlike modelling 3D parametric body, we introduced several 3D clothing models and slightly modified by 3D CAD software according to the key clothing styles in public gait datasets. Table 2 shows the list of clothing models for virtual dressing, where S-skirt, M-skirt and L-dress respectively denote short-skirt, medium skirt and long-dress. Fig. 2 illustrates some of them in details, i.e., shirt, coat, pants and skirt. The clothing models are constructed from the clothing categories introduced in [27] and DeepFashion [33] except for cap, bag and shoes.

TABLE 2. List of parametric clothing models for virtual dressing.

Category	Sub	Category	Sub class	Category	Sub
Tops	Tank	Pants	Leggings	Coat	Regular coat
	T-shirt		Regular pants		Medium coat
	Full shirt		Baggy pants		Long coat
	Sweater		Short pants		Raincoat
	Hoodie	Skirt/Dress	S-skirt	Others	Robe
	Blazer		M- skirt		Handbag
Hat	cap		L- dress		Backpack

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2997814, IEEE Access

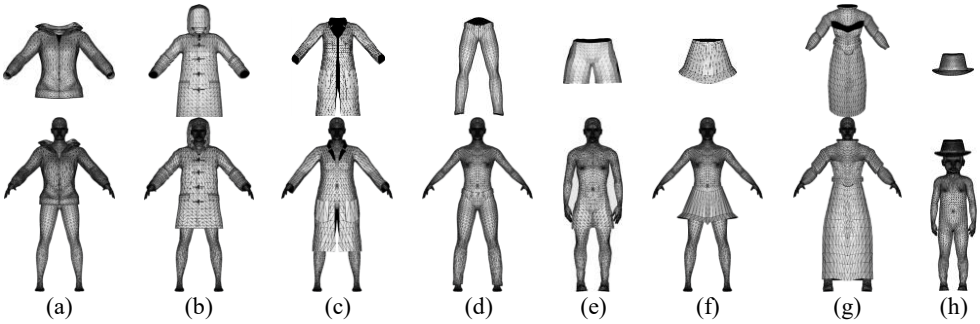


FIGURE 2. 3D parametric clothing models and virtual dressing: (a) Regular coat; (b) medium coat; (c) long coat; (d) regular pants; (e) short pants; (f) short skirt; (g) long dress; and (h) cap on kid.

C. 3D BODY POSE AND SHAPE DATA ESTIMATION NETWORK AGAINST VARIOUS CLOTHING CONDITIONS

In our proposed method, gait silhouette segmentation is achieved using a state-of-art joint body parsing and pose estimation network SS-JPPNet [34]. SS-JPPNet is trained on a dataset comprising over 50,000 annotated images with 19 semantic part labels, captured from a broad range of viewpoints, occlusions and scenarios. Its outputs are of three image formats, i.e., RGB body contour, body parsed image and binary silhouette.

Following the gait silhouette segmentation, an estimate of the initial 2D-3D-BPSDs including 3D joints data, shape parameter values and viewing data is made. In order to achieve view-invariant gait recognition, both azimuth and elevation angles must be considered. When a subject is walking from a far distance to the camera, the view between the body and camera changes continuously. In most gait recognition methods, these changes are ignored, especially in model free algorithms. However, camera views can influence the gait recognition accuracy especially if the subject walks in a big curve path. In order to obtain a better 3D initial gait model, an end to end 3D body pose, shape and viewing data estimation network (3D-BPSVeNet) is proposed. It is built upon three sub networks, i.e., the state-of-art DeeplabV3+ model [35] (a feature extractor using encoding), ConvGRU (a temporal feature encoder) and body parsing. The body parsing sub-network estimates the 3D joints and viewing angles in accordance with the extracted 2D features. The schematic diagram of the proposed network is shown in Fig. 3.

Fig. 3 shows several frames of body semantic parsing of RGB silhouettes with clothes ID embedded (SC-RGB) used as the inputs of 3D-BPSVeNet. SC-RGB, and 2D Gait RGB silhouettes with clothes ID embedded (GC-RGB) are directly used for training. Let the input gait sequence frames be denoted by $I_n, n = 1,2,3 \ldots, N$. First, deepLabV3+ is applied to the input gait silhouette I, i.e., SC-RGB or GC-RGB, to extract 2D gait feature $F = \mathcal{N}_{feature}(I)$. Then M consecutive frames of gait features are fed to ConvGRU to encode their spatial-temporal information, i.e., $\tilde{F} = convGRU(F_{k-m}, \ldots, F_{k-1}, F_k), m \in [1, M]$. ConvGRU exploits both CNNs and GRU. As a recurrent neural network, there are two important gates in a GRU unit [36], the updated gate z_t and the reset gate r_t. Compared with LSTM the state of the cell is removed, and the hidden state is used for information exchange which makes it efficient. The 3D-BPSVeNet outputs the joints and shape data of 3D body together with viewing data, i.e., the joints are encoded as delta values to the standard I pose. They are based on the skeleton structure of CMU mocap [37] and encoded in biovision hierarchical (BVH) format. Each joint has three DOF with its local coordinate.

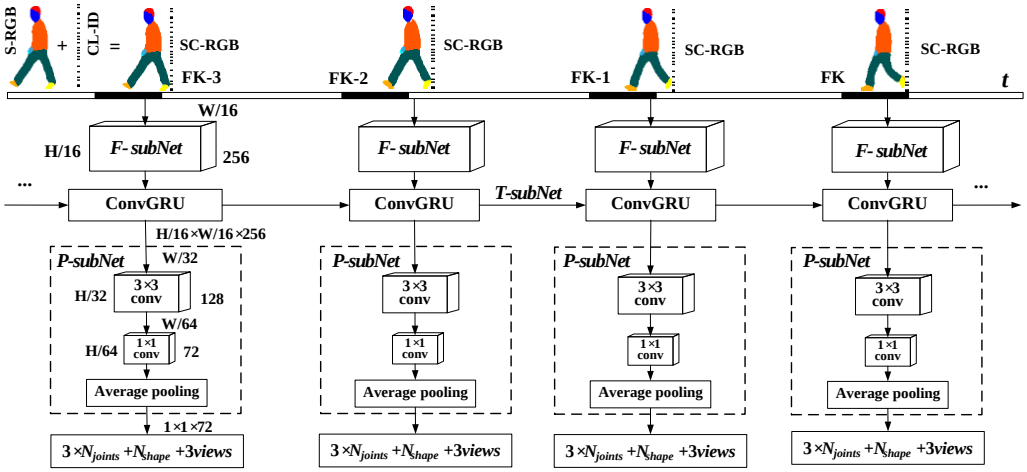


FIGURE 3. The schematic of 3D-BPSVeNet.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2997814, IEEE Access

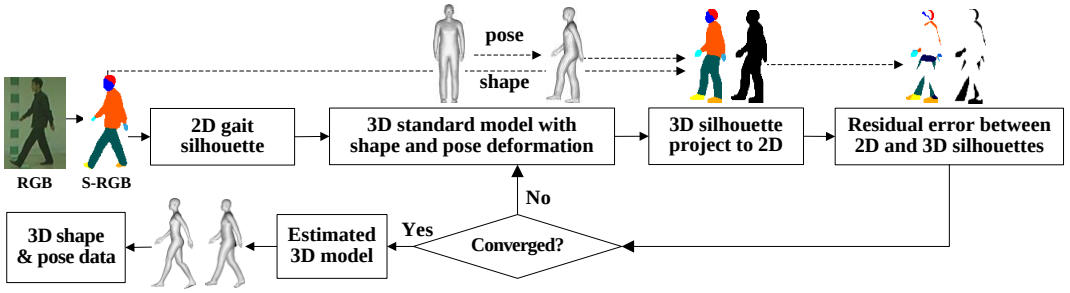


FIGURE 4. Extraction of 3D pose ground truth data.

In the F-subNet and T-subNet, the data have the same shape (stride of 16, 256 channels). In the subsequent parsing sub-network, a 3×3 convolutional layer and 1×1 convolutional layer with stride of 2 are designed to reduce the feature channels to the size of $\ell = (3N_j + N_s + 3)$, where N_j denotes the number of 3D joints, and each joint has 3 elements, i.e., $j = \Delta(x, y, z)$. N_s defines the number of 3D body shape parameters. After an average pooling, the ℓ size data is mapped to 3D body pose and shape parameters with the additional data on viewing data, i.e., azimuth and elevation angels.

To train the 3D-BPSVeNet, the L_2 based loss function is defined as

$$\mathcal{L} = \sum_{n=1}^{N} \left\| v_n \cdot (\mathcal{J}_n^{gt} - \mathcal{J}_n) \right\|_2^2 + \sum_{n=1}^{N} \left\| (\mathcal{S}_n^{gt} - \mathcal{S}_n) \right\|_2^2 + \sum_{n=1}^{N} \left\| (\gamma_n^{gt} - \gamma_n) \right\|_2^2 , \quad (1)$$

where N denotes the number of training samples. $\mathcal{J}_n^{gt} \in \mathbb{R}^{3N_j}$ is the normalized vector comprising all the ground truth 3D body joints data with three DOF, and $\mathcal{J}_n$ comprises the estimated joints data. $v_n \in \mathbb{R}^{3N_j}$ is the indicator vector denoting the status for each joint, i.e., visible or not (caused by self-occlusion). $\mathcal{S}_n^{gt} \in \mathbb{R}^{N_s}$ is the normalized vector comprising ground truth body shape values, and $\mathcal{S}_n$ comprises the estimated shape values. $\gamma_n^{gt} \in \mathbb{R}^3$ *is* the normalized vector comprising the ground truth data of viewing, and γ_n corresponds to the estimated data vector. To train the 3D-BPSVeNet, sufficient ground truth 2D to 3D estimated data is essential. To the best of our knowledge, there are no labelled 2D to 3D body parameters estimation data, especially for gait. To undertake the training, a semi-automatic method is introduced to construct the virtual ground truth data of 2D-3D-BPSDs.

The semi-automatic method was developed in our previous work in [16] and [32]. In [32] 3D gait pose data are estimated by observing the silhouette difference between 2D gait contour and 3D projected body under the same view using a silhouette similarity degree function for binary images. Using a binary image to estimate 2D-3D-BPSDs has its disadvantages. For example, the left and right hands (or legs) are often difficult to distinguish due to the lack of RGB information. To overcome this problem, the RGB body parsed images are introduced instead of binary images. The process is illustrated in Fig. 4. First, a 3D body model similar to the current gait posture is initialized. Then, the selected 3D body model is rotated to the view consistent with the 2D gait and projected onto the 2D space to form a reference template. Finally, the residual error between the 2D and 3D-2D projected body parsed silhouettes is determined. If the residual error is large than the set threshold or the maximum number of iterations has not been reached, thus the 3D body model will undergo further pose deformation by updating the pose parameters. The synthesized 3D body model will fit the 2D gait better, and the residual error is updated until the residual error is less than or equal to the set threshold.

In this paper, the residual error measuring function defined in Eq. (2)-(4) is a real-valued function of a fixed number of 2D-3D-BPSDs as inputs. However, the function is a continuous but complex function without an underlying mathematical definition. To simplify the problem and facilitate the realization, Powell's conjugate direction method is introduced as the basic optimization method to extract the 2D-3D-BPSDs truth data as illustrated in Fig. 4. By using the Powell's method, the function need not be differentiable, and no derivatives are taken. It is useful to calculate the local minimum of such a function. In the real application, the values of shape parameters are first fixed and minimized using Eq. (2) to obtain the optimal values of pose parameters. When the pose parameters are refined, they are then fixed to gain the optimal values for shape. The experimental results show that the accuracy of the estimated data of 2D-3D-BPSDs are greatly improved by the clothes recognition, virtual dressing process and multi-view data.

The silhouette similarity degree function for measuring the residual error at a given view α is

$$\mathcal{L}_\alpha = \frac{1}{2m \times n} \sum_{i=1}^{m \times n} w_b \left\| \left(g_i^{2D,\alpha} - g_i^{3D,\alpha} \right) \right\|_2^2 + \frac{1}{2m \times n} \sum_{d=1}^{D} \sum_{i=1}^{m \times n} w_d \left\| \left(c_{d,i}^{2D,\alpha} - c_{d,i}^{3D,\alpha} \right) \right\|_2^2, \quad (2)$$

where m and n are respectively the height and width of the normalized gait images, and i is the index of pixels in gait images. Let $g_i^{2D,\alpha}$ be the pixel value in 2D body parsed image

$$P_g^{2D,\alpha} = J_{PPNET}(\mathcal{B}^\alpha) = \{ g_i^{2D,\alpha}, i = 1 \dots m \times n \}, \quad (3)$$

obtained from 2D RGB gait $\mathcal{B}^\alpha$ using SS-JPPNET. $g_i^{3D,\alpha}$ defines the pixel value corresponding to body parsed image of 3D projected image. The 3D projected gait image is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2997814, IEEE Access

denoted by $\mathcal{P}_\alpha(\mathcal{J}, \mathcal{S}, C_p)$. Its corresponding 3D model comprises S as the body shape parameters, $\mathcal{J}$ as the parameters of joints and C_p as the clothing parameter of p type. The body parsed image of $\mathcal{P}_\alpha(\mathcal{J}, \mathcal{S}, C_p)$ is

$$\begin{aligned} P_g^{3D,\alpha} &= J_{PPNET}(\mathcal{P}_\alpha(\mathcal{J}, \mathcal{S}, C_p)) \\ &= \{g_i^{3D,\alpha}, i = 1 \dots m \times n\}. \end{aligned} \tag{4}$$

Let D be the number of parsed body parts of interest, i.e., head, leg and hand (displayed in different colour in Fig. 5), $c_{d,i}^{2D,\alpha}$ is the pixel value of body part d in $P_g^{2D,\alpha}$, and $c_{d,i}^{3D,\alpha}$ is the pixel value of body part d in $P_g^{3D,\alpha}$. w_b is the weight which determines the global fitness of two different gait silhouettes, and w_d are the weights that overcome the sub-optimal decisions when significant part of the body is lost.

(a) (b) (c) (d) (e) (f)

FIGURE 5. (a) & (c) RGB gait images; (e) 3D projected image after texture mapping; and (b), (d) and (f): the corresponding body parsed images.

By minimizing the silhouette similarity degree function of Eq. (2) the 2D-3D-BPSDs are estimated and denoted by $\mathcal{J}_{opt} = \{\Delta(x_i, y_i, z_i),\ i \in [1 \dots N_j]\}$ and $\mathcal{S}_{opt} = \{s_j,\ j \in [1 \dots N_s]\}$. N_j and N_s respectively denote the number of joint and shape parameters as listed in Table 1. If multi-view data are considered for more accurate estimation, the total residual error can be redefined by $\mathcal{L} = \sum_{\alpha \in \Phi} \mathcal{L}_\alpha$, where Φ is a view set. Before iterating, the initial viewing data γ, i.e., elevation angle, is manually assigned according to the dataset. The gait images from CASIA B dataset with different views are used to construct the virtual ground truth dataset of 2D-3D-BPSDs. We manually check the final optional results and adjust the pose and shape to get the best ground truth data for each subject. Using the semi-automatic method, 2D-3D-BPSDs are estimated from the input 2D images, and the additional check with manual modification ensures the data to be more accurate.

The data from CASIA B is insufficient to train the 3D-BPSVeNet. To enlarge the training data, we morph the 3D body models with virtual random body shape parameters $\mathcal{S}_{vir}$ and clothing parameter $\mathcal{C}_{p_{vir}}$. They are projected onto 2D space to obtain the 2D virtual gait image with pose data $\hat{\mathcal{J}}$, i.e., $\mathcal{B}_{vir,\hat{\mathcal{J}}}^{\alpha} = \mathcal{P}_\alpha(\hat{\mathcal{J}}, \mathcal{S}_{vir}, \mathcal{C}_{p_{vir}})$. Let $\mathcal{B}^{i,\alpha} = \{\mathcal{B}_1^{i,\alpha}, \dots \mathcal{B}_m^{i,\alpha}, \dots, \mathcal{B}_M^{i,\alpha}\}$ be a given gait set where $\mathcal{B}_m^{i,\alpha}$ denotes the mth 2D RGB gait frame of ith sample at view α. M is the maximum number of frames in a gait cycle. For $\mathcal{B}^{i,\alpha}$, the M corresponding 3D pose data are denoted as $\mathcal{J}_{set} = \{\mathcal{J}_1, \dots, \mathcal{J}_M\}$ and the shape data set as $\mathcal{S}_{set} = \{\mathcal{S}_1, \dots, \mathcal{S}_M\}$. Virtual generated samples are based on the extension of $\mathcal{J}_{set}$ and $\mathcal{S}_{set}$. The $\mathcal{S}_{set}$ can be enlarged by uniformly synthesizing N_s^{vir}, new virtual shape data set, i.e., $\mathcal{S}_{set}^{vir} = \{\mathcal{S}_1^{vir}, \dots, \mathcal{S}_{N_S^{vir}}^{vir}\}$, and the mixed data set is $\mathcal{S}_{set}^{mixed} = \mathcal{S}_{set} \cup \mathcal{S}_{set}^{vir}$. The $\mathcal{J}_{set}$ is enlarged by T times linear interpolation based on joints data in a cycle, and $\mathcal{J}_{set}^{vir} = \{\mathcal{J}_1^{vir}, \dots, \mathcal{J}_{T \times M}^{vir}\}$. The corresponding virtual generated gait set is $\mathcal{B}_{vir}^{i,\alpha} = \{\mathcal{B}_{vir,1}^{i,\alpha}, \dots \mathcal{B}_{vir,m}^{i,\alpha}, \dots, \mathcal{B}_{vir,T \times M}^{i,\alpha}\}$, which is $T \times N_s^{vir} \times M$ times larger than the original $\mathcal{B}^{i,\alpha}$. By using the estimated 2D-3D-BPSDs from $\mathcal{B}^{i,\alpha}$, and the virtual generated data, the sequence training dataset is constructed. Let $In^i = \left(In_{m+1}^i, In_{m+2}^i, \dots, In_{m+t}^i\right)^T$ be the ith sequence based input gait comprising t consecutive frames in a gait cycle where $m + t \leq M$ and $m \in [1\ M]$. The output is

$$\begin{aligned} Out^i &= \left(Out_1^i, Out_2^i, \dots, Out_K^i, \dots, Out_{K+3}^i\right)^T \\ &= \left(\mathcal{J}_{opt}^i, \mathcal{S}_{opt}^i, \gamma^i\right)^T, \end{aligned} \tag{5}$$

where $K = 3N_j + N_s$ and $\gamma^i \in \mathbb{R}^3$. γ^i denote the views, i.e., azimuth and elevation angels. $\mathcal{J}_{opt}^i$ are the 3D pose parameters corresponding to last gait frame In_{m+t}^i, and S_{opt}^i are the average shape values of t input gait frames. The 3D-BPSVeNet can be adequately trained using batches of the input In^i and output Out^i.

D. 3D GAIT SEMANTIC DATA OPTIMIZATION

Using the 3D-BPSVeNet, the 3D pose parameters $\mathcal{J}_{opt,0}$, shape parameters $\mathcal{S}_{opt,0}$ and views are estimated. However, due to the limited availability of ground truth data for real 2D-3D-BPSDs, the training samples are still less than satisfactory. Thus, the estimated 3D body data, especially from 2D gait images under various conditions, need to be optimized. The optimization of 2D-3D-BPSDs comprises the following three steps. First, recognize the 2D clothing styles and virtual dress the 3D body with clothing. Second, adjust the shape parameters to optimize the pose parameters using semantic parsed gait image. Finally, adjust the pose and update the body shape parameters.

FashionNet [33] is introduced to recognize clothes. It is based on the clothes dataset DeepFashion which consists of 800K clothing items with comprehensive annotations. It can predict clothing category, attribute and landmarks, that help to determine the length of clothes. According to the basic category of clothing, the prior designed virtual clothes are selected to dress (using virtual dressing [38]) the 3D body before shape deformation.

After virtual dressing, the initialized 3D model is refined using an algorithm similar to that shown in Fig. 4 by minimizing Eq. (2). The data corresponding to moving parts, i.e., hands and legs, are assigned larger weights, i.e., set to 0.6, due to their importance in motion. If there is a significant loss of this data, the larger weights ensure that moving parts do not lose their total energy quickly so as not to be trapped in local optimum. The other static body parts, i.e., head and trunk, are assigned smaller weights, thus ensuring the lost

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2997814, IEEE Access

data have less effects on the global optimum. Since body pose and shape parameters have different physical meanings, we first fix the values of shape parameters and minimize Eq. (2) to obtain the optimal pose parameter $\hat{\mathcal{J}}_{opt}$. This is followed by determining the optimal shape parameter $\hat{\mathcal{S}}_{opt}$. The final optimal body semantic parameters for input sample i are denoted by $\mathcal{P}_b = \{\hat{\mathcal{J}}^i_{opt}, \hat{\mathcal{S}}^i_{opt}, \hat{\gamma}^i\}$.

E. GAIT SEMANTIC FOLDING

Gait semantic folding comprises two steps as illustrated in Fig. 6: gait semantic sparse distributed representation (GS-SDR); and folding. GS-SDR is the process of encoding an unstructured gait images to a Sparse Distributed Binary Gait Semantic Image (SD-BGSI) using a topographical semantic space based on 3D body semantic parameters. By averaging a sequence of SD-BGSIs, a GSFI is obtained. The GSFI is used as the basic gait semantic feature for further gait recognition against various walking conditions.

By using 3D-BPSVeNet and the refining process, the body semantic parameters as listed in Table 1 are estimated as $\mathcal{P}_b = \{\mathcal{J}_{opt}, \mathcal{S}_{opt}, \gamma\}$. Motivated by the efficiency of GEI and to exploit sparse distributed representations (SDRs), which is the fundamental form of pattern representation in our brain [39], we encode the scalar body semantic data to binary GS-SDR. SDRs are robust to noise and usually in the form of a binary sequence. According to the brain-like HTM theory [39], the bits correspond to neurons in the brain, where a one denotes a relatively active neuron and a zero a relatively inactive neuron. Our GS-SDR shares the same conceptual foundation with the HTM theory.

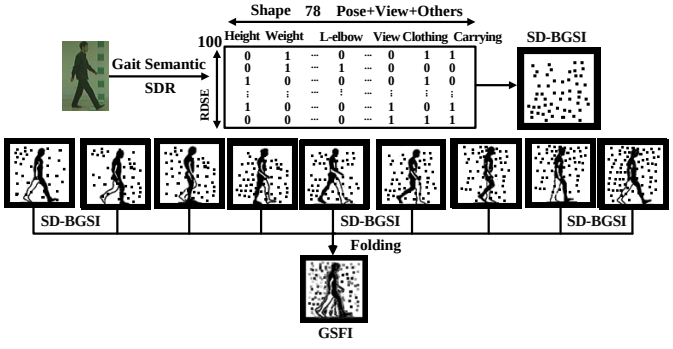


FIGURE 6. Generation of SD-BGSI and SD-GSFI.

The gait semantic folding is based on the 2D-3D-BPSDs estimated by 3D-BPSVeNet with a refining process. As in Section III.C, let $\mathcal{J}^k_{opt} = \{j^k_i = \Delta(x_i, y_i, z_i) | i \in [1, N_j]\}$, $\mathcal{S}^k_{opt} = \{s^k_j | j \in [1, N_s]\}$ and $\gamma^k \in \mathbb{R}^3$ respectively denote the refined 3D semantic body joints, shape and viewing parameters. N_j and N_s respectively denote the maximum number of joint and shape parameters. j^k_i denotes the ith 3D joint data of k frames in a gait cycle, and γ^k is viewing data. s^k_j denotes the jth shape parameters of the k frames in a gait cycle. The length of 2D-3D-BPSDs is $\ell = (3N_j + N_s + 2)$. Additional clothing and carrying conditions with six parameters are added to 2D-3D-BPSDs.

The generation of SD-BGSI is illustrated in Fig. 6, where each column represents a single gait semantic parameter. The numeric value of the semantic parameter is encoded as a spare binary column vector using the sparse distributed scalar encoder (SDSE) introduced in [39]. In SDSE encoding, w is defined as the number of ON-bits that are set to encode a single value, and n is the number of bits in the output which must be greater than w. A radius and a resolution are also defined, i.e., two values separated by greater than the radius have non-overlap, and two values separated by greater than the resolution have different representations. According to the SDSE, $resolution = radius/w$ and $n = w * range/radius$. The input data range is normalized to [0 1] in this paper and the w is set to 11, which should be an odd number. The resolution is set to 0.01 and the number of bits in the output n is determined to be 100. The SDSE maps a scalar value into an array of bits, i.e., ON-bits are significantly less than the zero-bits. The similarity of two SDSE vectors is given by the overlap score. If x and y are two SDSE vectors with length n, the overlap between them is defined as their dot product, i.e.,

$$overlap(x, y) \equiv x \cdot y \,. \tag{6}$$

It simply computes the number of ON (i.e., 1) bits between the two SDSE vectors at the same locations. Several columns of SDSE vectors are constructed to form an SDR matrix, which is the SD-BGSI after visualization.

A match between two SD-BGSIs is then defined by $match(x, y|\theta) \equiv overlap(x, y) \geq \theta$. The match is inexact as in fuzzy theory if $\theta < w$, where w is defined to assume that the two SD-BGSIs have the same cardinality w. If $\theta = w$, an exact match is determined. The inexact representation is one of the significant properties of SD-BGSIs, which makes the processing of SD-BGSIs more robust to noise and changes the input. Thus, the match of two SD-BGSIs is determined by checking if they overlap sufficiently [39], which can be directly undertaken with the semantic meaning using the logical "AND" or "OR" operation.

In a gait cycle, there are several SD-BGSIs, i.e., each gait frame corresponds to a SD-BGSI. To obtain a more efficient gait feature representation, GSFI is calculated based on the principle of GEI, i.e., averaging the SD-BGSIs in a gait cycle. As aforementioned, the 3D body parameters are normalized to [0 1] range. The average value of each semantic pixel in GSFI denotes the probability of ON-bit. For the purpose of visible display, they are re-normalized to [0 255] for each pixel. Unlike averaging the scalar values, the GSFI is more similar to the statistical representation of GEI. But it is essentially not the same as GEI is derived from raw binary gait images, and GSFI is based on 3D body semantic pattern space, i.e., pose and shape. It is the structural gait feature descriptor and is less sensitive to various walking conditions.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2997814, IEEE Access

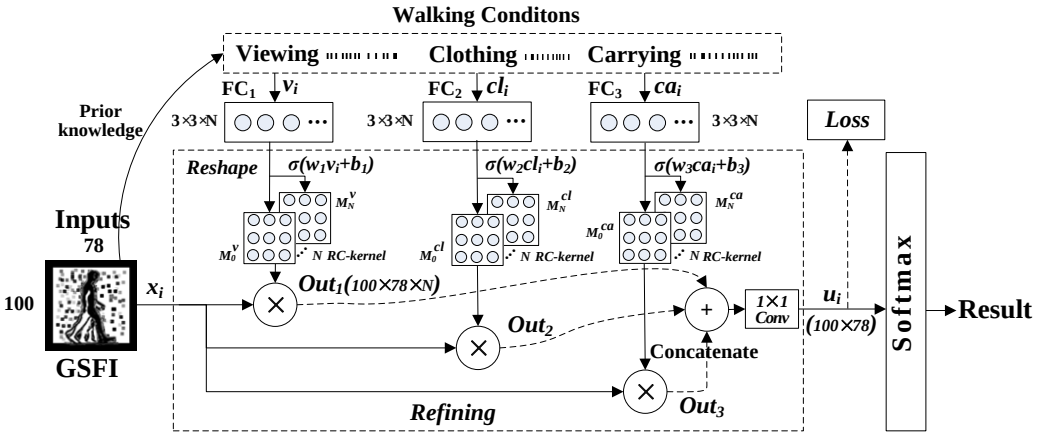


FIGURE 7. Refining structure of GSFI for view and clothing invariant gait recognition.

F. GSFI REFINEMENT FOR VIEW AND CLOTHING INVARIANT GAIT RECOGNITION

Fig. 7 illustrates the proposed refining method using GSFI as input and SoftMax as the classifier. The method comprises two phases, i.e., refining and recognition. The feature refining is motivated by the fact that the a priori knowledge about walking conditions can be used to construct a feature adjustor. In fact, our GSFI is view-invariant gait feature descriptor, i.e., the shape parameter is less sensitive to views. The 3D dynamic joint data are also view-invariant, i.e., the motion information of joints is encoded by values relative to the data of standard template using BVH (Biovision hierarchical data) format which makes it also robust to views. However, the estimation of 2D-3D-BPSDs for the same subject may sometimes be slightly different under different walking conditions. The refining mechanism uses the statistics of different views, clothing and carrying items to adjust GSFI features before classification. For example, carrying a ball influences the dynamic data of two hands, and the refining mechanism assigns small weight to the hand joint data using the knowledge learned from normal walking.

Let $X = \{x_i = I^i_{GSFI} \in \mathbb{R}^{\ell\times\ell}, i = 1,\dots,I\}$ denotes the set of GSFIs with I samples. Three types of walking conditions., i.e., viewing, clothing and carrying, are introduced for refinement as shown in Fig. 7. The refining $3 \times 3 \times N$ convolutional kernels (RC-Kernels) are generated according to the walking conditions. The refining process is achieved via the convolution of GSFI and the RC-Kernels.

As shown in Fig. 7, the connection networks FC_1 to FC_3 are used to directly connect the input data of three walking conditions that are represented in the form of SDRs vector as discussed in Section III.E. The input viewing data is denoted as $v_i = \{(v_{azimuth}, v_{elevation}) \in \mathbb{R}^2\}$. The clothing style is composed of upper, down and additional dressing, and denoted by $cl_i = \{(cl_{upper}, cl_{lower}, cl_{addition}) \in \mathbb{R}^3\}$. The carrying condition is described by three variables, i.e., object carrying style and the (*x*,*y*) location of the corresponding body part. It is defined as $ca_i = \{(ca_{style}, ca_x, ca_y) \in \mathbb{R}^3\}$. The sigmoid activate function is introduced to normalize the outputs of FC_1 to FC_3 within the range [0 1]. They are then reshaped to form the RC-Kernels for convolution on GSFI. The three outputs of the convolution are $Out_1 = Conv(GFSI, RC_v\_Kernals)$, $Out_2 = Conv(GFSI, RC_{Cl}\_Kernals)$ and $Out_3 = Conv(GFSI, RC_{Ca}\_Kernals)$. These have a dimension of $100 \times 78 \times N$ and are concatenated for fusion. A 1×1 convolution operation and followed by a sigmoid activate function are then applied. The final output, i.e., the refined GSFI, is $u_i = GSFI\_MNet(x_i)$ which has the same size as the input GSFI.

The refining network and the SoftMax classification network are trained separately. The refining network adjusts the higher-level features extracted directly from GSFIs and makes the features more invariant to viewing angles, clothing styles and carrying items. Its loss function is

$$\mathcal{L}^u_{oss} = \sum_{i=1}^{I} \sum_{\tilde{u}_p \in U^i_{pos}} \left\| u_i - \tilde{u}_p \right\|_2^2 , \quad (7)$$

where U^i_{pos} denotes N_{pos} positive outputs set based on the anchor sample x_i, i.e., the positive output $\tilde{u}_p$ is from the same subject anchor but under different view, clothing and carrying conditions. After feature refinement for gallery GSFIs, the gallery feature set are denoted by $In_{gallery} = \{u^{gal}_i, i \in [1\ N_g]\}$ and is used as input data to train the SoftMax classifier for recognition. The SoftMax classifier has two important functions, i.e., a score function and the cross-entropy loss function. The score function, i.e., $S(x_i; W; b) = Wx_i + b$, maps each input $x_i = u^{gal}_i$ to the scores of each category. The cross-entropy loss function then converts the classification scores into its probability distribution by using one-hot encoding vector as final output. The cross-entropy loss function for all the samples in the training dataset is defined as

$$\mathcal{L}^{ce}_{oss} = -\frac{1}{N_g} \sum_{i\in[1\ N_g]} \log\left(\frac{e^{S_{y_i}}}{\sum_j e^{S_j}}\right), \quad (8)$$

where S_j represents the score value of the jth class in the score function vector S, y_i is the correct classification label

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2997814, IEEE Access

information of the input x_i, S_{y_i} denotes the target class score of x_i, and N_g denotes the total number of samples used in the training. After training using the gallery data, the samples in probe dataset are applied for testing. The kth input of the test samples is denoted as x_k^{probe} and its output is the probability distribution of all categories, i.e., ID labels. The classification result is determined by the category with the highest probability value.

IV. EXPERIMENT

To evaluate our VCIGR-3DHSF, the datasets CMU MoBo, CASIA B and KY4D with clothing variation, object carrying, occlusion, etc., were selected for experiments. The clothing related OU-ISIR dataset B, and the multi-view gait dataset OU-MVLP with binary gait silhouettes from large population were also used.

To train our 3D-BPSVeNet and GSFI-RNet, we chose 24 subjects in CASIA B, i.e., ID-001 to ID-024, and estimated their ground truth 2D-3D-BPSDs using the semi-automatic approach involving the loss function in Section III.C. Three walking conditions, i.e., normal, carrying a bag, and wearing a coat, and 11 views were included. We also used the virtual sample generation method in Section III.C to increase the number of samples as follows. The number was first doubled by morphing to the virtually generated 100 sets of typical shape parameters. It was further increased by twice using linear interpolation of poses derived from subjects of ID-001 to ID-024, and doubled by random dressing with 3D virtual clothes from the clothing dataset. In addition, two elevation angle changes were added by rotating the 3D gait models, i.e., ±8°. 20% of the total samples were duplicated and randomly added with horizontal or vertical bar located at 5% to 30% height of a gait image. The total number of 2D-3D-BPSDs gait sequence patterns used was 22,800, sufficient to train the networks.

A. EXPERIMENTS ON CMU MOTION OF BODY DATASET

The CMU MoBo [40] consists of six image sequences for each of the twenty-five subjects walking on a treadmill captured by a number of cameras. Each subject undertook four different walking conditions: slow, fast walking, inclined walking and walking with a ball. In order to demonstrate the robustness of our method against incomplete gait silhouettes, missing data was simulated by adding horizontal or vertical bar to the gallery silhouettes. Using the settings in [9], a horizontal or vertical bar was introduced as interference to gait silhouettes with the probability varying from 10% to 100%. The width of a vertical bar varies from 20 to 50 pixels with 10 pixels as step size, and the horizontal bar varies from 40 to 100 pixels. Unlike the situation in [9], RGB images with equally distributed bars that simulate potential occlusions were used in our experiments.

TABLE 3. Rank-1 recognition rates (%) with horizontal and vertical bar occlusions.

Method	Horizontal bar width				Vertical bar width			
	40	60	80	100	20	30	40	50
IDTW[41]	64.0	60.2	62.4	63.2	66.2	67.3	65.8	66.4
GEI[4]	79.6	80.6	81.0	79.6	81.0	82.0	82.0	80.6
GHI[42]	54.4	54.4	57.8	53.4	52.8	54.6	56.0	56.2
GMI[43]	46.0	46.4	46.4	39.6	48.8	50.8	46.4	48.4
FD-GEI[9]	79.5	81.4	80.3	80.3	83.4	83.2	82.2	81.4
V-3DHSF	92.2	91.2	90.4	86.8	90.2	90.8	90.0	88.4
V-3DHSF-V	95.2	94.2	94.6	92.8	94.8	94.2	93.6	92.2

For the CMU MoBo dataset, the gait data of fast walk were used as gallery while the slow walk data as probe. The comparison with other data-driven or model-based methods of the lateral-view gait recognition results is shown in Table 3. The results for our VCIGR-3DHSF-V (denoted by V-3DHSF-V, i.e., where virtual samples with added bars were used to train the 3D-BPSVeNet), show good performance. By using the virtual sample generation process, 2D-3D-BPSDs were estimated to mitigate the effect of imperfect silhouettes. The results for VCIGR-3DHSF (denoted by V-3DHSF, i.e., gait recognition without using virtual noise samples) shows the recognition rate is slightly reduced. Nevertheless, they both represent performances significantly better than the other methods. This is because instead of using a static binary image, sequences of 2D RGB gait images were used in our framework to estimate 2D-3D-BPSDs. The influence of incomplete gait semantic data caused by occlusion or missing data are mitigated by neighbouring frames. In order to illustrate the performance of VCIGR-3DHSF under other walking variations, further experiments as shown in Table 4 were conducted. Unlike some methods, e.g., FSVB [44] STM-SPP [45], WBP [46], SGRVDL [47] and PEI [11], in our experiments the SC-RGB gait images were used instead of binary gait images to give more information of gait.

TABLE 4. Twelve experiments on CMU MoBo gait dataset (in lateral view).

Exp.	Gallery set	Probe set	Gallery/Probe
A	Slow walk	Fast walk	25×3×4
B	Slow walk	Ball-carrying walk	25×3×4
C	Slow walk	inclined walk	25×3×4
D	Fast walk	Slow walk	25×3×4
E	Fast walk	Ball-carrying walk	25×3×4
F	Fast walk	Inclined walk	25×3×4
G	Inclined walk	Slow walk	25×3×4
H	Inclined walk	Fast walk	25×3×4
I	Inclined walk	Ball-carrying walk	25×3×4
J	Ball-carrying walk	Slow walk	25×3×4
K	Ball-carrying walk	Fast walk	25×3×4
L	Ball-carrying walk	Inclined walk	25×3×4

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2997814, IEEE Access

TABLE 5. Recognition results (%) on Mobo data set.

Exp.	FSVB	WBP	STM-SPP	SGRVDL	Method [8]	PEI	VCIGR-3DHSF
A	82	92	94	96	92	100	100
B	77	73	93	87	-	92	96
C	-	-	-	-	-	60	94
D	80	92	91	92	92	88	96
E	61	61	84	88	-	60	95
F	-	-	-	-	-	72	96
G	-	-	-	-	-	76	93
H	-	-	-	-	-	80	95
I	-	-	-	-	-	48	94
J	89	75	82	87	-	92	96
K	73	63	82	88	-	84	94
L	-	-	-	-	-	76	94

Table 5 shows VCIGR-3DHSF outperforms the other methods especially for ball-carrying condition (Exp. B, E and I) and inclined walk (Exp. C, F and L). Other experimental results that are not presented in the original papers have been left blank. The table shows that when the gait data are under normal conditions (e.g., Exp. A and D), the existing methods show high recognition results as well. However, most methods are not robust to abnormal changes (e.g., carrying a ball and inclined walk). This is because the 2D binary gait silhouettes are more easily degraded by various walking conditions especially by heavy coat and carrying items. In contrast, the VCIGR-3DHSF shows satisfactory recognition results across all types of conditions. When faced with the carrying conditions, the body parsing network removes the ball, and the carrying refining matrix for GSFI assigns small weight to the joints of hands. In most cases, the carrying condition makes the hand joints unchanged. When training the GSFI-RNet, virtual samples with the hand joints data unchanged are generated to make GSFI-RNet robust against carrying conditions.

In our framework, the body parsing SS-JPPNET [34] is introduced to parse the human body, and the clothing recognition network FashionNet [33] helps the recognition of clothing styles. The gait semantically parsed images without background, e.g., Fig. 8(b), are used to estimate the initial 2D-3D-BPSDs by our 3D-BPSVeNet. The 2D-3D-BPSDs is then optimized by virtual dressing, e.g., Fig8(d)-(f), for better performance.

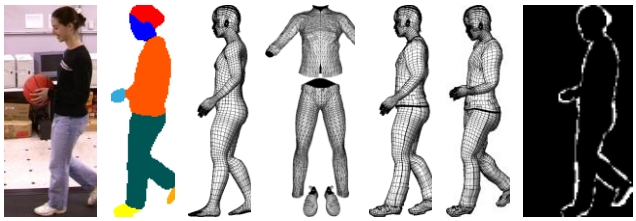


FIGURE 8. Refining 2D-3D-BPSDs by virtual dressing: (a) walk with a ball; (b) body parsing image of (a); (c) estimated 3D gait model; (d) 3D clothes; (e) dressing on model; (f) after refining; and (f) silhouette difference between (b) and (f).

B. GAIT RECOGNITION UNDER NORMAL CONDITION ON CASIA B DATASET

CASIA Database B is a multi-view gait dataset with two variations, i.e., clothing changes and object carrying. The dataset contains video sequences of 124 subjects captured from 11 views in the range [0° 180°] with an interval of 18°. Each view of a subject comprises 10 video sequences: 6 sequences for normal walking, and 4 sequences under two variations, e.g., wearing a coat, and carrying a bag, a knapsack, or a handbag [6].

The view-invariant performance of VCIGR-3DHSF was evaluated using the CASIA Dataset B. We excluded 24 subjects for 3D-BPSVeNet training, and the rest of the hundred normal walking subjects, i.e., ID025-ID124, were chosen for evaluation. Similar to the settings in [24], they were assigned to two groups. Two normal sequences, i.e., nm05 and nm06, out of six were selected on each view for probe data and the rest for gallery. At each time, only one probe view was used for testing, and the gallery views ranged from 18° to 162° except for the probe view. Fig. 9 compares the rank-1 recognition rate of different methods, i.e., GEI-SVD [48], GFI-CCA [49], Gabor-CMMF [50], C3A [21], ViFS-LDA [22], SPAE-NN [24], and ours with gallery views from 18° to 162°. Gabor-CMMF extracts Gabor features from GEIs and uses coupled multi-linear marginal fisher criterion for feature encoding. For GaborSD-CMMF, only the cross-view recognition result under the 54° probe is reported and for C3A [21], 108° probe is not reported.

The results show that VCIGR-3DHSF performs well especially when large view change occurs. There are several reasons for this. The first is due to our GSFI which is derived from two types of view invariant body semantic data, i.e., body shape data and dynamic joint data. The second is that the GSFI refining network helps to overcome the value deviation issue in 2D to 3D semantic parameter estimation. In our framework, a single view gait data supplemented with a few of other views in the refining process are used to extract 2D-3D-BPSDs. Due to the occurrences of different self-occlusions, the 2D-3D-BPSDs from two different views might differ even for the same pose of the same subject. Thus, semantic feature refining is introduced to address this.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2997814, IEEE Access

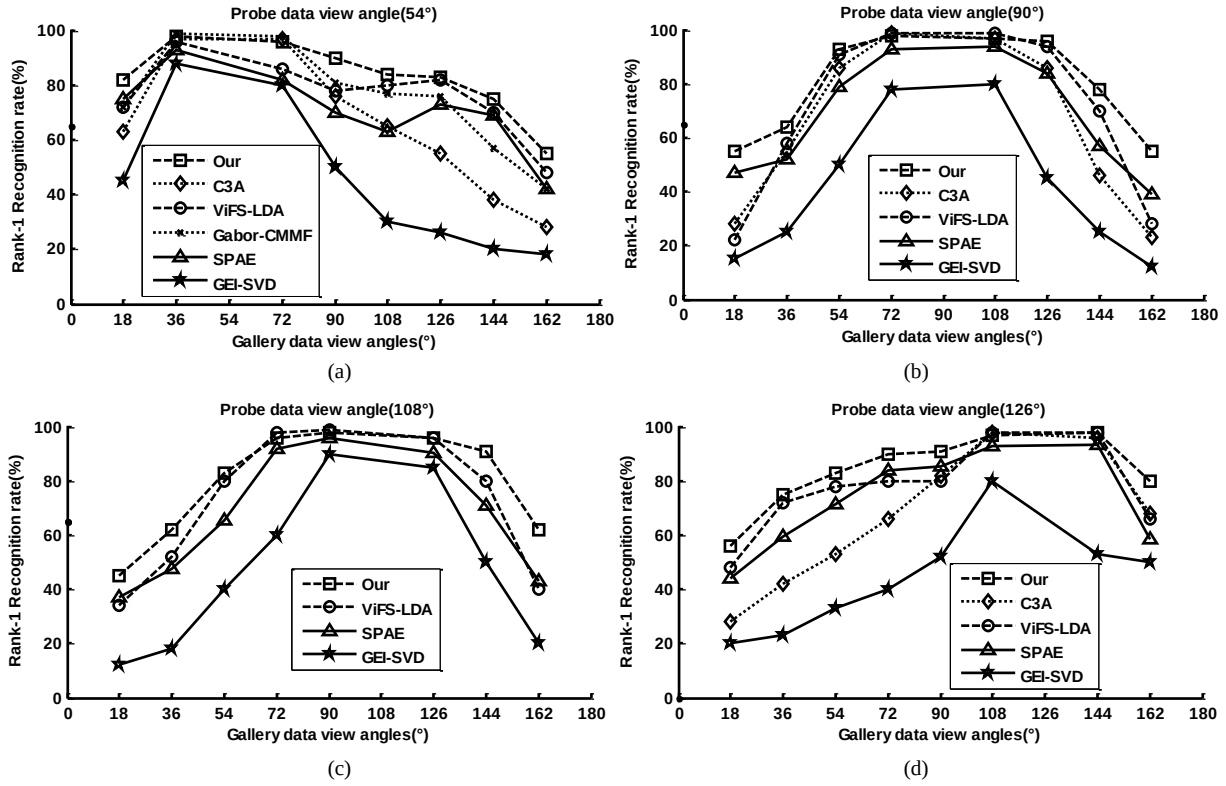


Fig. 9. Rank-1 recognition rates of different methods.

By using the data of gait views and the knowledge learned by GSFI refining network, both the azimuth and elevation angle refining matrices help to improve the GSFI for better performance. In fact, the elevation angle refining matrices greatly help in cross elevation view gait recognition.

C. GAIT RECOGNITION UNDER VARIOUS CONDITIONS ON CASIA B DATASET

To further evaluate the performance of our VCIGR-3DHSF against various walking conditions, CASIA Dataset B was used. First, normal sequences of 100 subjects were selected on each view for gallery data. The coat wearing and bag carrying data were for probe. At each time, one gallery view was used for training and testing the probe data under the same view. The rank-1 recognition results of our VCIGR-3DHSF outperforms GEI-GaitSet [2], GFI-CCA [49], GPSM [16] as shown in Fig. 10 under views from 18° to 162°. The GFI-CCA method which takes GFI as a gait feature only reported results under 36° to 144° views.

In the second experiment, we set the probe views to 54°, 90° and 126° with two walking conditions. The gallery data were chosen from normal walking sequences under views of 36°, 72°, 108°, 126° and 144°. Tables 6 to 8 show the performances of our method, GEI-NN [6], MGANs [52], SPAE-NN [24], GFI-CCA [49], RLTDA [53], and Deep-CNNs [1]. Theses tables show our VCIGR-3DHSF performs best, especially with bag and clothing conditions with large view changes. It is robust and less sensitive to various dressing conditions and object carrying.

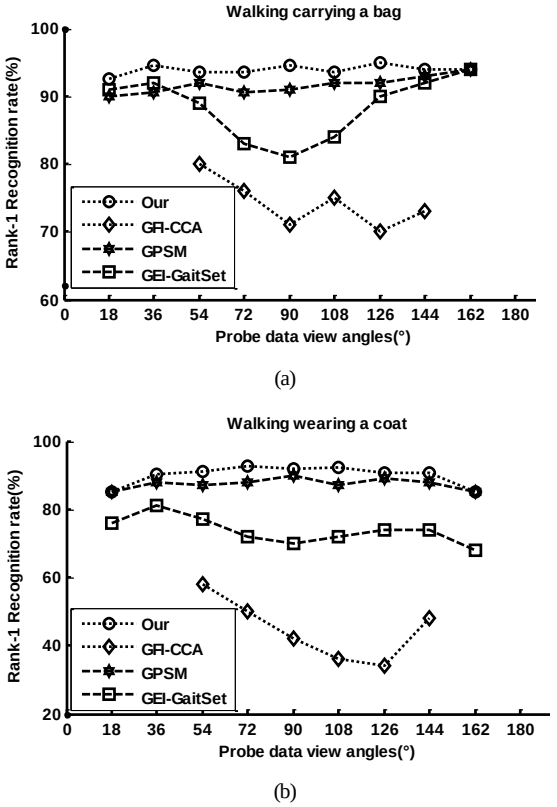


FIGURE 10. Recognition results of VCIGR-3DHSF, AVGR-BPRS, VI-MGR and GFI-CCA under various variations

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2997814, IEEE Access

Table 6. Rank-1 cross-view gait recognition (%) with probe under 54°.

Gallery		Methods						
		GEI-NN	RLTDA	SPAE-NN	GFI-CCA	MGANs	Deep-CNNs	Our
36°	Bag	24	81	62	70	78	93	92
	Coat	17	69	42	50	50	50	85
72°	Bag	9	72	66	60	90	90	90
	Coat	8	58	37	22	56	62	82
126°	Bag	17	-	47	32	68	-	66
	Coat	4	-	29	28	35	-	63

Table 7. Rank-1 cross-view gait recognition (%) with probe under 90°.

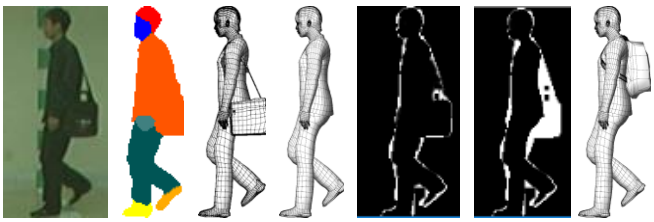
Gallery		Methods						
		GEI-NN	RLTDA	SPAE-NN	GFI-CCA	MGANs	Deep-CNNs	Our
72°	Bag	31	75	64	60	89	93	92
	Coat	22	63	38	35	55	78	84
108°	Bag	44	76	61	58	83	89	88
	Coat	28	72	40	42	50	76	80
144°	Bag	2	-	24	26	43	-	52
	Coat	2	-	18	28	38	-	50

Table 8. Rank-1 cross-view gait recognition (%) with probe under 126°.

Gallery		Methods						
		GEI-NN	RLTDA	SPAE-NN	GFI-CCA	MGANs	Deep-CNNs	Our
72°	Bag	11	-	24	25	80	-	72
	Coat	9	-	25	22	43	-	62
108°	Bag	23	66	56	45	83	93	91
	Coat	9	65	42	35	50	58	78
144°	Bag	32	72	57	50	78	86	85
	Coat	18	64	35	29	55	51	76

There are several reasons why VCIGR-3DHSF performs well. Using the clothing recognition network, a priori knowledge of dressing and object carrying conditions are determined first. Different clothing styles are chosen and the initial 3D human model is virtually dressed before the 2D-3D-BPSDs refining process. The virtual dressing ensures the predicted parameters of body shape with clothing are more accurate for heavy garments and skirt, or with bag carrying. For carrying conditions, to make the estimation more tolerant and robust, virtual data on different object carrying are used or manually synthesized when training the 3D-BPSVeNet as illustrated in Fig 11.

Fig. 10(a) shows that most methods achieve good performance when the views are close to 18° or 162°, and achieve poor performance near 90°. The latter is due to the large bag contours that influence the gait silhouettes segmentation at this view. The bag silhouettes merge with the gait contours when the gait silhouettes are extracted using traditional segmentation methods. We introduced JPPNET to accurately parse the body with output S-RGB, thus aiding to locate the hand position in carrying condition. This is not possible with 2D binary images due to the overlap of the carried item with other body parts or objects. By using the robust 2D-3D-BPSDs extraction method, the influence of the carrying condition is greatly reduced.



(a) (b) (c) (d) (e) (f) (g)

FIGURE 11. Generation of virtual bag carrying models: (a) 2D gait image with a bag; (b) semantic gait image of (a); (c) synthesized 3D mesh model with similar carrying condition of (a); (d) 3D body mesh of (c) without a bag; (e) silhouette difference between (b) and (c); (f) silhouette difference between (b) and (d); and (g) 3D virtual body mesh with a backpack.

D. EXPERIMENTS ON KY4D DATABASES WITH CURVED TRAJECTORIES

Kyushu University 4-dimensional (4D) Gait Database (KY4D) [54] is characterized by its 4D gait data comprising

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2997814, IEEE Access

a set of 3D visual hull models with 2D image sequences. The forty-two subjects involved in the dataset walked along four straight paths $\{t1, t2, t3, t4\}$ and two curved trajectories $\{t5, t6\}$. The 2D gait images were captured by 16 high-definition cameras, suitable for identifying subjects walking along curved trajectories. Since KY4D is a multi-view gait database, we exploited it in 2D-3D-BPSDs optimization using Eq. (2). The silhouette similarity measuring function based on multi-view is defined as

$$\mathcal{L} = \frac{1}{2m\times n}\sum_{\theta\in\Phi}\sum_{i=1}^{m\times n} w_b \left\|(g_i^{2D,\alpha} - g_i^{3D,\alpha})\right\|_2^2 + \frac{1}{2m\times n}\sum_{\theta\in\Phi}\sum_{d=1}^{D}\sum_{i=1}^{m\times n} w_d \left\|(c_{d,i}^{2D,\alpha} - c_{d,i}^{3D,\alpha})\right\|_2^2, \quad (9)$$

where Φ is a multi-view set determined by the number of cameras. The redefined cost function illustrates the union of the residual error from all gait views. By minimizing the multi-view silhouette similarity measuring function, accurate 3D human body pose and shape parameters are estimated.

In our experiment, only the straight path walking sequences were used as gallery for training and the curved trajectories for testing. Fig. 12 shows that our method outperforms the approaches by López [26], Iwashita [54], Castro [55] and Seely [56] for curved gait trajectories. The VCIGR-3DHSF works best in curved walking condition due to two reasons. First, our 3D-BPSVeNet estimates camera views by a sequence of 2D gait images, i.e., four frames in our experiment. The difference in walking directions correspond to camera view changes. Since the walking direction within four frames are similar to straight walk, it makes our body feature extraction of 2D-3D-BPSDs less influenced by the curved trajectories. Second, the information on changing walking views is embedded in our GSFI when averaging the different viewing data in SD-BGSIs. It takes into account the GSFI refining process, thus making our 2D-3D-BPSDs more robust to view changes regardless of self-occlusions.

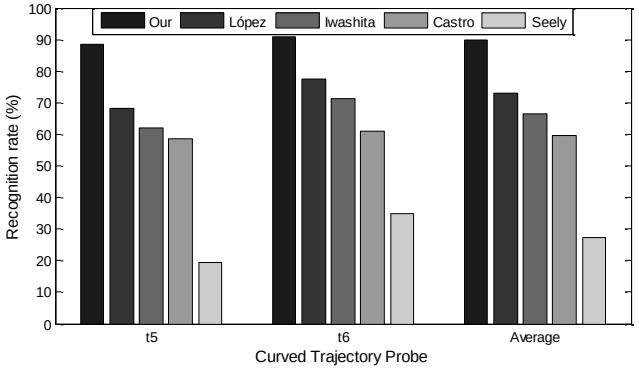


FIGURE 12. Gait recognition rates comparison on KY4D gait dataset.

E. EXPERIMENTS ON OU-MVLP DATASET

OU-MVLP [6] is multi-view gait dataset incorporating a large population (i.e., 10307 subjects), captured with 14 view angles ranging from 0°-90°, 180°-270° with 15° interval. Each view of a subject contains two video sequences with a resolution of 1280x980 pixels. It is helpful for evaluating algorithms for cross-view gait recognition under large population condition. We used the same criteria settings in [6] to evaluate our method under four typical view angles ranging from 0°-90°. In the baseline of 1in-GEINet, 10,307 subjects were divided into two disjoint groups, i.e., 5153 for training and 5154 for testing. The methods compared in our experiment are 1in-GEINet baseline [6], VTM [57], CNNs-LB [1] and CNNs-Siamese [58].

Since only binary gait images are published due to privacy reasons, body parsed S-RGB images cannot be used. Instead, we transformed all the S-RGB images to binary format when training the 3D-BPSVeNet. It makes our method less accurate in extracting the 2D-3D-BPSDs, and the clothing recognition network based on RGB images cannot be used. To address this problem, a clothing combination classification based on GEIs, as illustrated in Fig. 13, is introduced for coarse clothing recognition.

Twelve clothing combinations were used in the experiment as listed in Table 9. The ResNet-50 convolutional network [59] with SoftMax classifier was used for recognition and about 10,00 subjects in OU-MVLP were manually selected for training. The keys for different types of clothing in Table 9 are: FS - Full shirt; Hd - Hoodie; Br - Blazer; RC - Regular coat; MC - Medium coat; LC - Long coat; RC - Rain coat; Lg - Leggings; RP - regular pants; Ht - hat; SS - Short skirt; MS - Medium skirt; LD - Long dress; and Rb - Robe.

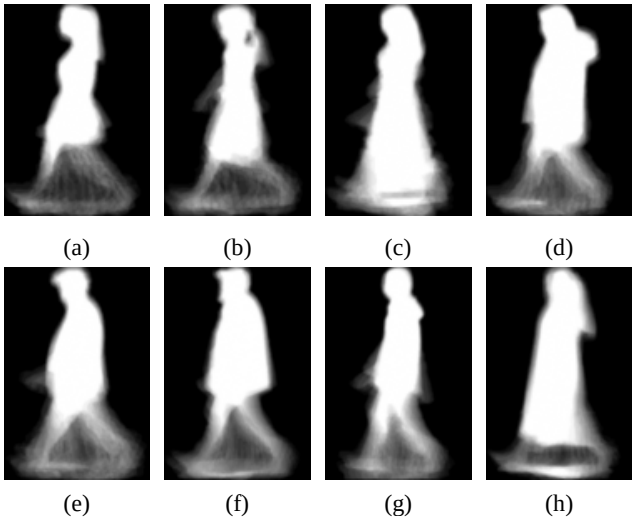


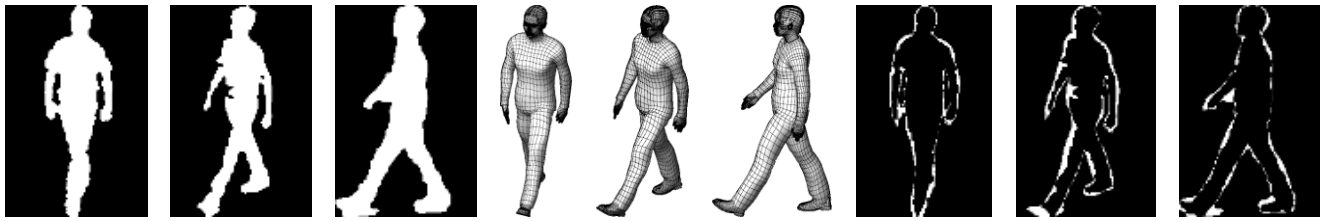
FIGURE 13. GEIs in OU-MVLP under different clothing conditions: (a) short skirt; (b) medium skirt; (c) long dress; (d) medium coat; (e) hat and blazer; (f) raincoat; (g) hoodie; and (h) robe.

Besides clothing recognition, the multi-view gait data were also used. According to Eq. (9), large multi-view gait data, i.e., 5153 subjects, help to obtain more accurate 2D-3D-BPSDs in the optimization as illustrated in Fig. 14. These data were added to train our 3D-BPSVeNet, which made it adapt to the new data in OU-MVLP.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2997814, IEEE Access

TABLE 9. Different clothing combinations used in the OU-ISIR B dataset.

Index	Upper	Lower	Addition	Index	Upper	Lower	Addition
1	FS	RP	-	8	LD	LD	-
2	FS	Lg	-	9	Hd	RP	-
3	RC	RP	-	10	Br	RP	-
4	MC	RP	-	11	RC	RP	-
5	LC	RP	-	12	Rb	RP	-
6	FS	SS	-	13	Br	RP	Ht
7	FS	MS	-	14	FS	RP	Ht



(a) (b) (c) (d) (e) (f) (g) (h) (i)

FIGURE 14. Refining 3D gait model using multi-view data: (a) 15° gait of ID-10 subject from OU-MVLP; (b)-(c) respectively 45° and 90° gait data for refining; (d) refined 3D model using (b) & (c); (e)-(f): the corresponding 3D gait of (b) & (c); and (g)-(i) silhouette difference between 2D gait silhouettes and their corresponding 3D gait.

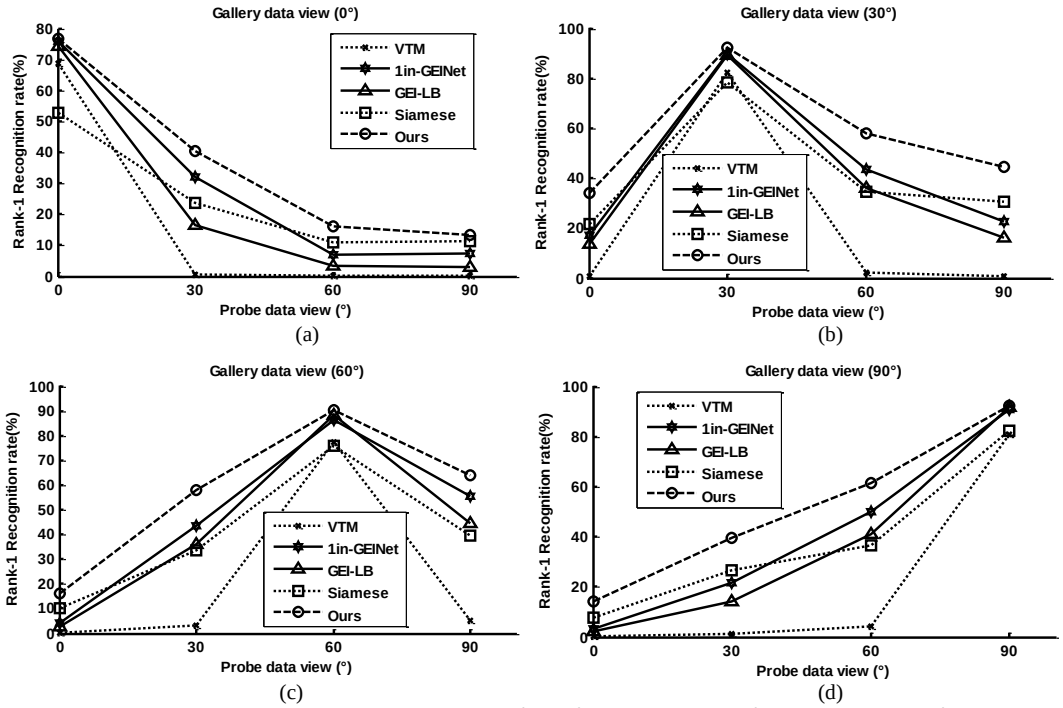


FIGURE 15. Recognition rates of different methods with probe view from 0° to 90°: (a) gallery view is 0°; (b) gallery view is 30°; (a) gallery view is 60°; and (a) gallery view is 90°.

The large population data when training our GSFI-RNet also greatly helped to overcome the value deviation problem in 2D to 3D semantic parameter estimation for adjusting the intrinsic semantic features for recognition. The comparisons results are shown in Fig. 15.

Fig. 15 shows that our VCIGR-3DHSF has advantages in cross-view recognition even when the population of the subjects is larger. Unlike VTM-based methods and most deep learning approaches that transform the feature of probe gait data to gallery viewing angle, or extract the view-invariant features that are unexplained and less semantic relevance, our method extracts the view-invariant body features directly by an end to end 3D-BPSVeNet with full semantic meaning. Also, the mismatched feature that often occurs in view transformation or extraction is avoided, especially with large view changes. The framework of VTM-based or data-driven based method, e.g., deep learning [60], requires large training samples to gain a more generic model, and better performance is achieved by learning from more gait samples. However, RGB gait images under various walking conditions from a large population are not easy to obtain. Also, the camera settings fixed in one scenario might

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2997814, IEEE Access

be different to real-world application scenarios due to the change of their elevation angle. Our parametric 3D body model with virtual dressing is greatly helped by virtual sample generation process. The 3D body knowledge with viewing angles are fully utilized which make our method performs well under large view changes.

F. EXPERIMENTS ON OU-ISIR DATASET B

The OU-ISIR dataset B [27] is focused on different clothing combinations and is useful for evaluating the robustness of gait recognition algorithm against clothing variations. It is composed of 68 subjects from side view with up to 32 combinations of different types of clothing. Since the OU-ISIR dataset only provides binary gait silhouettes we cannot use our clothing recognition network. However, all the clothing combinations are given in [27] and used as our a priori knowledge in our experiments. Table 10 shows the different clothing combinations used in the OU-ISIR B dataset, i.e., RP - Regular pants (Regular jeans); BP - Baggy pants (Chinos); SP - Short pants; Sk - Skirt (Medium skirt); CP - Casual pants (Chinos); HS - Half shirt; FS - Full shirt; LC - Long coat; Pk - Parker (Hoodie); DJ - Down jacket (Parka); CW - Casual wear (Full shirt); RC - Rain coat; Cs - Casquette cap (Hat); and Mf - Muffler.

TABLE 10. Different clothing combinations used in the OU-ISIR B dataset.

Exp.	S_1	S_2	S_3	Exp.	S_1	S_2	Exp.	S_1	S_2
3	RP	HS	Ht	0	CP	CW	F	CP	FS
4	RP	HS	Cs	2	RP	HS	G	CP	Pk
6	RP	LC	Mf	5	RP	LC	H	CP	Dj
7	RP	LC	Ht	9	RP	FS	I	BP	HS
8	RP	LC	Cs	A	RP	Pk	J	BP	LC
C	RP	DJ	Mf	B	RP	Dj	K	BP	FS
X	RP	FS	Ht	D	CP	HS	L	BP	Pk
Y	RP	FS	Cs	E	CP	LC	M	BP	DJ
N	SP	HS	-	P	SP	Pk	R	RC	-
S	Sk	HS	-	T	Sk	FS	U	Sk	PK
V	Sk	DJ	-	Z	SP	FS	-	-	-

We used the experiment settings in [51] to evaluate our VCIGR-3DHSF. The dataset was divided into three groups: (1) a training set comprising 446 sequences of 20 subjects with all types of clothing, used to train the GSFI-RNet; (2) a gallery set comprising sequences of the remaining 48 subjects with standard clothing; and (3) a probe set comprising 856 sequences for these 48 subjects with other types of clothing excluding the standard clothing. Fig. 16 shows the performances of our method and GEI, CI-SSA [3] and VI-MGR [51]. N.B. CI-SSA only reported recognition results in several clothing combination, i.e., Exp. 3, 5, 6, 7, 8, B, C, E and R.

Fig. 16 shows that our method significantly outperforms GEI, VI-MGR and CI-SSA, especially when the subjects wore heavy coat or skirt, i.e., clothing conditions C, J, M, U and V. Our VCIGR-3DHSF exploited 3D virtual dressing as illustrated in Fig. 17 and feature refining network, i.e., GSFI-RNet, using a priori knowledge of clothing for feature refinement.

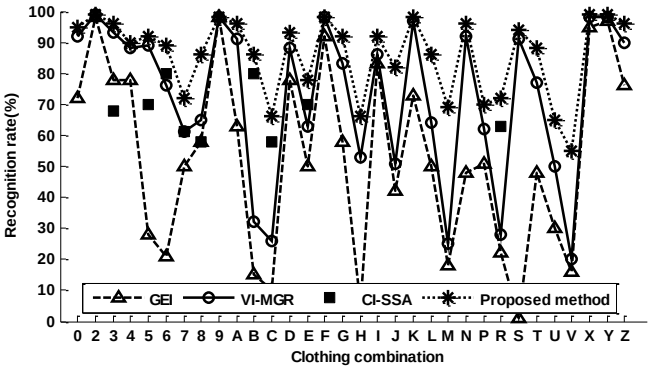


FIGURE 16. Recognition accuracy of various methods on OU-ISIR dataset B with different clothing combinations.

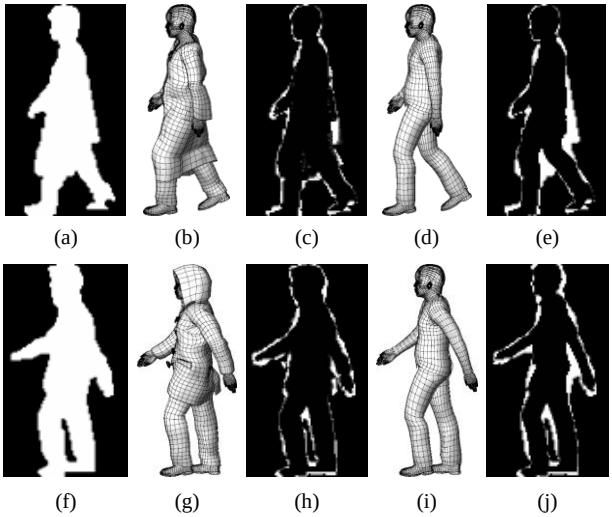


FIGURE 17. Refining 3D gait model using virtual dressing: (a) J combination of ID-3 subject from OU-ISIR; (b) refined 3D gait model with long coat; (c) difference between (a) and (b); (d) normal dressing of (b); (e) difference between (a) and (d); (f) R combination of ID-3 subject; (g) refined 3D gait model with raincoat; (h) difference between (f) and (g); (i) normal dressing of (g); and (j) difference between (f) and (i).

G. COMPUTATIONAL COMPLEXITY

In our proposed VCIGR-3DHSF method, the extraction of 2D-3D-BPSDs from 2D gait images is the time-consuming part of the gait recognition. Thus, we discuss the computational complexity of the 2D-3D-BPSDs extraction, and the minimum silhouette residual error search involved in optimizing the 2D-3D-BPSDs using Eq. (2). In the Powell's conjugate direction method, the number of iterations in Eq. (2) is greatly influenced by the initial data. To speed up the process, an end to end 3D-BPSVeNet is proposed to gain a better 3D initial gait model. A good set of global data values of 2D-3D-BPSDs greatly reduces the time in using Eq. (2). Another strategy is also introduced to speed up the computation. An extra penalty item is added to Eq. (2) to make the pose estimation results more reasonable by using body shape and motion knowledge, i.e.,

$$\mathcal{L}_{new} = \mathcal{L}_{\alpha} + \sum_{m\in[1\,M]} r_{ule_m}(\mathcal{J}) + \sum_{n\in[1\,N]} \hat{r}_{ule_n}(\mathcal{S}), \qquad (10)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2997814, IEEE Access

where $\{r_{ule_m}|j \in [1\ M]\}$ denotes a set of rules on joints with M items, and $\{\hat{r}_{ule_n}|n \in [1\ N]\}$ denotes a set of rules on body shape with N items. The rule item function $r_{ule}(\cdot)$ inputs the current joints data $\mathcal{J}$ or shape data $\mathcal{S}$ to check for any violation of the rules. It returns a large positive value when it violates the rule and zero otherwise. Since the physical variables of body shape are related to each other, i.e., the weight is highly related to height and can be estimated using Body Mass Index. As for pose data, the constraints for the maximum ranges of joints and the conditions for normal walking movement also aid to speed up the process. Table 11 shows the typical running time in optimizing 2D-3D-BPSDs using Powell's estimation method on a PC with an Intel Core i7(3.6GHz) CPU and 8GB RAM. The optimized strategy method has been discussed earlier and the original method is initialized with template I-pose without using the 3D-BPSVeNet, and no extra penalty item is added to Eq. (2). The computational complexity can be improved further by using Graphical Processing Units.

TABLE 11 Typical running time in optimizing 2D-3D-BPSDs.

Methods	Average time(seconds)
Optimized strategy method	6.8
Original method	82.5

V. CONCLUSION

In this paper, a view and clothing invariant gait recognition system based on semantic folding is presented. A novel gait feature descriptor, i.e., GSFI, and a semantic feature refining network are introduced. VCIGR-3DHSF converts unstructured gait image data to structured gait semantic image via 2D-3D body parameter estimation and semantic folding. By using the a priori knowledge of viewing angles, clothing styles and carried items, the proposed system is robust to various walking conditions that commonly occur in real application scenarios.

The method is based on the accurate extraction of the 2D-3D-BPSDs for semantic folding representation. In order to speed up the process, an end to end 3D-BPSVeNet is trained using mixed training samples, i.e., real data and virtual generated data. The process for accurate body parameters estimation is then conducted based on virtual dressing which greatly helps to overcome the effects of clothing variations. To make the semantic folding descriptor GSFI more effective for recognition, a semantic feature refining network is proposed. In addition, the method also exploits deep learning network, i.e., CNN, RCNN and GRU. Since a large dataset is normally required for adequate training, and this is a problem for 3D gait recognition, we exploited full use of the a priori knowledge to generate virtual samples, i.e., utilizing parametric body model and 3D clothing models. By introducing the clothing recognition network and body parsing network trained on a large dataset, we achieved accurate gait recognition against changing viewing angles and clothing. The other most important improvement is that RGB images are used for gait recognition. Compared with the traditional gait recognition methods based on binary gait images, more information is exploited in our method. The experimental results show that VCIGR-3DHSF is effective in view-invariant gait recognition against most walking conditions.

REFERENCES

[1] Z. Wu, Y. Huang, L. Wang, X. Wang, T. Tan, "A Comprehensive Study on Cross-View Gait Based Human Identification with Deep CNNs", *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol.39, no.2, pp.209-226, March 2016.

[2] H. Chao, Y. He, J. Zhang, J. Feng, "Gaitset: Regarding gait as a set for cross-view gait recognition", In Proceedings of the *AAAI Conference on Artificial Intelligence* 2019, Hawaii, USA, 27 Jan.-1 Feb. 2019; pp.1-8.

[3] A. Nandy, R. Chakraborty, Chakraborty P, "Cloth Invariant Gait Recognition using Pooled Segmented Statistical Features", *Neurocomputing*, vol.191, pp.117-140, Feb. 2016.

[4] J. Han, B. Bhanu, "Individual Recognition Using Gait Energy Image", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.28, no.2,316-322, Feb. 2006.

[5] X. Chen, J. Xu, "Uncooperative gait recognition: Re-ranking based on sparse coding and multi-view hypergraph learning", *Pattern Recognition*, vol.53, pp.116-129, Dec. 2016.

[6] 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. on Computer Vision and Applications*, vol.10, no.4, pp.1-14, Feb. 2018.

[7] M.Z. Uddin, T.T. Ngo, Y. Makihara, N. Takemura, X. Li, D. Muramatsu, Y. Yagi, "The OU-ISIR Large Population Gait Database with Real-Life Carried Object and its performance evaluation", *IPSJ Trans. on Computer Vision and Applications*, vol.10, no.1, pp.1-11, May 2018.

[8] L. Wei, K. C.-C. Jay, J. Peng, "Gait recognition via GEI subspace projections and collaborative representation classification", *Neurocomputing*, vol.275, pp.1932-1945, Nov. 2018.

[9] C. Chen, J. Liang, H. Zhao, H. Hu, J. Tian, "Frame difference energy image for gait recognition with incomplete silhouettes", *Pattern Recognition Letters*, vol.30, no.11, pp. 977-984, Aug. 2009.

[10] K. Bashir, T. Xiang, S. Gong, "Gait recognition without subject cooperation", Pattern Recognition Letters, vol.31, no.13, pp.2052-2060, Oct. 2010.

[11] A. Roy, S. Sural, J. Mukherjee, "Gait recognition using Pose Kinematics and Pose Energy Image", *Signal Processing*, vol.92, no.3, pp.780-792, March 2012.

[12] M. Deng, C. Wang, F. Cheng, et al, "Fusion of spatial-temporal and kinematic features for gait recognition with deterministic learning", *Pattern Recognition*, vol.67:186-200, Feb. 2017.

[13] W. Zeng, C. Wang, Y. Li, "Model-Based Human Gait Recognition Via Deterministic Learning", *Neural Networks the Official Journal of the International Neural Network Society*, vol.35, no.2, pp.92-102, Aug. 2012.

[14] J. Kovač, P. Peer, "Human Skeleton Model Based Dynamic Features for Walking Speed Invariant Gait Recognition", *Mathematical Problems in Engineering*, vol.2014, pp.1-15, Jan. 2014.

[15] D. López-Fernández, F. J. Madrid-Cuevas, A. Carmona-Poyato, R. Muñoz-Salinas, R. Medina-Carnicer, "Entropy volumes for viewpoint-independent gait recognition", *Machine Vision and Applications*, vol.26, no.7-8, pp.1079-1094, Aug. 2015.

[16] T. Jin, L. Jian, T. Tjahjadi, F. Guo, "Robust Arbitrary-View Gait Recognition based on 3D Partial Similarity Matching", *IEEE Transactions on Image Processing*, vol.26, no.1, pp.7-23, Jan. 2017.

[17] W. Kusakunniran, Q. Wu, J. Zhang, et al, "Gait Recognition Under Various Viewing Angles Based on Correlated Motion Regression", *IEEE Transactions on Circuits & Systems for Video Technology*, vol.22, no.6, pp.966-980, June 2012.

[18] D. Muramatsu, A. Shiraishi, Y. Makihara, et al, "Gait-Based Person Recognition Using Arbitrary View Transformation Model", *IEEE Transactions on Image Processing*, vol.24, no.1, pp.140-154, Jan. 2015.

[19] H. Hu, "Multiview Gait Recognition Based on Patch Distribution Features and Uncorrelated Multilinear Sparse Local Discriminant Canonical Correlation Analysis", *IEEE Transactions on Circuits and Systems for Video Technology*, vol.24, pp.617-630, Apr. 2014.

[20] Z. Wei, W. Cong, "View-invariant gait recognition via deterministic learning", *Neurocomputing*, vol.175, pp.324-335, Jan. 2016.

[21] X. Xing, K. Wang, T. Yan, Z. Lv, "Complete canonical correlation analysis with application to multi-view gait recognition", *Pattern Recognition*, vol.50, pp.107-117, Feb. 2016.

[22] N. Jia N, V. Sanchez, C.T. Li C, "On view-invariant gait recognition: a feature selection solution", *IET biometrics*, vol.7, no.4, pp. 287-295, March 2018.

[23] N. Takemura, Y. Makihara, D. Muramatsu,T. Echigo, Y.Yagi, "On input/output architectures for convolutional neural network-based cross-view gait recognition", *IEEE Transactions on Circuits and Systems for Video Technology*, vol.29, no.9, pp.2708-2719, Sept. 2019.

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

[25] D. Muramatsu, Y. Makihara, Y. Yagi View, "Transformation Model Incorporating Quality Measures for Cross-View Gait Recognition", *IEEE Transactions on Cybernetics*, vol.46, no.7, pp.1602-1615, July 2016.

[26] D. López-Fernández, F.J. Madrid-Cuevas, A. Carmona-Poyato, et al, "A new approach for multi-view gait recognition on unconstrained paths", *Journal of Visual Communication and Image Representation*, vol.38, pp.396-406, March 2016.

[27] M.A. Hossain, Y. Makihara ,J. Wang , Y. Yagi, "Clothing-invariant gait identification using part-based clothing categorization and adaptive weight control", *Pattern Recognition*, vol.43, no.6, pp.2281-2291, June 2010.

[28] F.M. Castro ,M. Marín-Jiménez, N. Guil, "Multimodal features fusion for gait, gender and shoes recognition", *Machine Vision and Applications*, vol.27, no.8, pp.1213-1228, May 2016.

[29] F. Battistone, A. Petrosino, "TGLSTM: a Time based Graph Deep Learning Approach to Gait Recognition", *Pattern Recognition Letters*, vol.126, pp.132-138, Sept. 2019.

[30] Muqing D., Cong W., Tongjia Z, "Individual identification using a gait dynamics graph", *Pattern Recognition*, vol.83, pp.287-298, June 2018.

[31] M. Bastioni, Re Simone, "Ideas and methods for modeling 3D human figures: The principal algorithms used by MakeHuman and their implementation in a new approach to parametric modeling", The 1st *ACM Bangalore Annual Conferenc*e, *COMPUTE 2008*, Bangalore, India, January 18-20, 2008, pp.1-6.

[32] J. Luo, J. Tang, T. Tjahjadi, X. Xiao, "Robust arbitrary view gait recognition based on parametric 3D human body reconstruction and virtual posture synthesis", *Pattern Recognition*, vol.60, pp.361-377, Dec. 2016.

[33] Z. Liu, P. Luo, S. Qiu, X. Wang, X. Tang, "DeepFashion: Powering Robust Clothes Recognition and Retrieval With Rich Annotations", The *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR) 2016, Las Vegas, USA, 27-30 June 2016, pp.1096-1104.

[34] X. Liang, K. Gong,X. Shen, et al, "Look into Person: Joint Body Parsing & Pose Estimation Network and a New Benchmark", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.41, no.4, pp.871-885, April 2019.

[35] LC. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation", *European Conference on Computer Vision*– ECCV 2018, Munich, Germany, 8-14 Sept. 2018, pp.833-851.

[36] K. Cho, B.V. Merrienboer, C. Gulcehre, et al, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation", *Computer Science*, vol.2014, pp.1-15, June 2014.

[37] CMU, "Carnegie-Mellon Mocap Database", http://mocap.cs.cmu.edu, 2003.

[38] L. Liu, Z. Su, X. Fu, et al, "A data-driven editing framework for automatic 3D garment modeling", *Multimedia Tools and Applications*, vol.76, no.10, pp.12597-12626, Jan. 2017.

[39] C. Yuwei, A. Subutai, H. Jeff, "The HTM Spatial Pooler—A Neocortical Algorithm for Online Sparse Distributed Coding", *Frontiers in Computational Neuroscience*, vol.11, pp.1-15, Nov. 2017.

[40] R. Gross, J. Shi, "The CMU Motion of Body (MoBo) Database", *Technical Report CMU-RI-TR-01-18*, Robotics Institute, Carnegie Mellon University, pp.1-13, July 2001.

[41] S. Yu, D. Tan, K. Huang, T. Tan, "Reducing the Effect of Noise on Human Contour in Gait Recognition", *Advances in Biometrics, International Conference*, ICB 2007, Seoul, Korea, August 27-29, 2007, pp.338-346.

[42] J. Liu, N.Zheng, "Gait History Image: A Novel Temporal Template for Gait Recognition", *IEEE International Conference on Multimedia and Expo*, Beijing, China, 2-5 July 2007, pp.663-666.

[43] Q. Ma, S. Wang, D.Nie, J.Qiu, "Recognizing Humans Based on Gait Moment Image", Eighth A*cis International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/distributed Computing*, Qingdao, China, 30 July-1 Aug. 2007, pp.606-610.

[44] S. Lee, Y. Liu, R. Collins, "Shape variation-based frieze pattern for robust gait recognition", In: Proceedings of *IEEE Conference on CVPR*, Minneapolis, MN, USA, 18-23 June 2007, pp.1–8.

[45] D.S. Choudhury, T. Tjahjadi, "Silhouette-based gait recognition using Procrustes shape analysis and elliptic Fourier descriptors", *Pattern Recognition*, vol.45, no.9, pp.3414-3426, Sept. 2012.

[46] W. Kusakunniran, Q. Wu, H. Li, J. Zhang, "Automatic gait recognition using weighted binary pattern on video", In: Proceedings of the Sixth *IEEE International Conference on Advanced Video and Signal Based Surveillance*, Genova, Italy,2-4 Sept. 2009, pp.49-54.

[47] W. Zeng, C. Wang, "Silhouette-based gait recognition via deterministic learning", A*dvances in Brain Inspired Cognitive Systems*. 6th International Conference, Beijing, China, 9-11 June 2013, pp.1-10.

[48] W.Kusakunniran, Q.Wu, H.Li, J.Zhang, "Multiple views gait recognition using view transformation model based on optimized gait energy image", In Proc. *IEEE Int. Conf. Comput. Vision*, Kyoto, Japan , Sep.-Oct. 2009, pp.1058–1064.

[49] K. Bashir, T. Xiang, S. Gong, "Cross-view gait recognition using correlation strength", *British Machine Vision Conference*, Aberystwyth, UK, 31 Aug.-3 Sept. 2010, pp.1-11.

[50] X. Ben, Z. Peng, Z. Lai, R. Yan, Z. Zhai, W. Meng, "A general tensor representation framework for cross-view gait recognition", *Pattern Recognition*, vol.90, 87-98, June 2019.

[51] D.S. Choudhury, T. Tjahjadi, "Robust view-invariant multiscale gait recognition. *Pattern Recognition*, vol.48, no.3, pp.798-811, March 2015.

[52] H. Yiwei, Z. Junping, S. Hongming, L. Wang, "Multi-task GANs for View-Specific Feature Learning in Gait Recognition", *IEEE Transactions on Information Forensics and Security*, vol.14, no.1, 102-113, Jan. 2019.

[53] H. Hu, "Enhanced gabor feature based classification using a regularized locally tensor discriminant model for multiview gait recognition", *IEEE Transactions on Circuits and Systems for Video Technology*, vol.23 , no.7, pp.1274-1286, July 2013.

[54] Y. Iwashita, K. Ogawara, R. Kurazume, "Identification of people walking along curved trajectories", *Pattern Recognition Letters*, vol.48, pp.60-69, Oct. 2014.

[55] F.M. Castro, M.J. Marín-Jiménez, R.M. Carnicer, "Pyramidal fisher motion for multiview gait recognition", In: 22nd *International Conference on Pattern Recognition*, ICPR 2014, Stockholm, Sweden, August 24–28, 2014, pp.1692-1697.

[56] R. Seely, S. Samangooei, M. Lee, J. Carter, M. Nixon, "The University of Southampton Multi-Biometric Tunnel and introducing a novel 3D gait dataset", In: 2nd *IEEE International Conference on Biometrics: Theory, Applications and Systems*, Arlington, USA,29 Sept.-1 Oct. 2008, pp.1-6.

[57] Y.Makihara, R.Sagawa, Y.Mukaigawa, T.Echigo, Y.Yagi, "Gait recognition using a view transformation model in the frequency domain", In Proc. of the 9th *European Conf. on Computer Vision*, Graz, Austria, 7-13 May 2006, pp.151–163.

[58] C. Zhang, W. Liu, H. Ma, H. Fu, "Siamese neural network based gait recognition for human identification", In: 2016 *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP). IEEE, Shanghai, China, 20-25 March 2016, pp.2832-2836.

[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", The *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR) 2016, 27-30 June 2016, Las Vegas, USA pp770–778.

[60] F. Jiang, K. Wang, L. Dong, C. Pan, W. Xu, K. Yang, "Deep Learning Based Joint Resource Scheduling Algorithms for Hybrid MEC Networks", *IEEE Internet of Things Journal*. 2019, doi:10.1109/JIOT.2019.2954503.

Jian Luo received B.Sc. in communication engineering from Hunan Normal University, China and M.Sc. in electronic science and technology from Hunan University China, in 2007 and 2010, respectively, and Ph.D. in information science and engineering in 2016 from Central South University China. He has been a lecturer at Hunan Normal University China since 2017. His research interest is gait recognition.

Tardi Tjahjadi received B.Sc. in mechanical engineering from University College London in 1980, and M.Sc. in management sciences in 1981 and Ph.D. in total technology in 1984 from UMIST, U.K. He has been an associate professor at Warwick University since 2000 and a reader since 2014. His research interests include image processing and computer vision.