Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.050018

ARTICLE

A Novel 3D Gait Model for Subject Identification Robust against Carrying and Dressing Variations

Jian Luo[1,*], Bo Xu[1], Tardi Tjahjadi[2] and Jian Yi[1]

[1]College of Information Science and Engineering, Hunan Normal University, Changsha, 410000, China

[2]School of Engineering, University of Warwick, Coventry, CV4 7AL, UK

*Corresponding Author: Jian Luo. Email: luojian@hunnu.edu.cn

Received: 25 January 2024 Accepted: 11 June 2024 Published: 18 July 2024

ABSTRACT

Subject identification via the subject's gait is challenging due to variations in the subject's carrying and dressing conditions in real-life scenes. This paper proposes a novel targeted 3-dimensional (3D) gait model (3D*Gait*) represented by a set of interpretable 3D*Gait* descriptors based on a 3D parametric body model. The 3D*Gait* descriptors are utilised as invariant gait features in the 3D*Gait* recognition method to address object carrying and dressing. The 3D*Gait* recognition method involves 2-dimensional (2D) to 3D*Gait* data learning based on 3D virtual samples, a semantic gait parameter estimation Long Short Time Memory (LSTM) network (3D-SGPE-LSTM), a feature fusion deep model based on a multi-set canonical correlation analysis, and SoftMax recognition network. First, a sensory experiment based on 3D body shape and pose deformation with 3D virtual dressing is used to fit 3D*Gait* onto the given 2D gait images. 3D interpretable semantic parameters control the 3D morphing and dressing involved. Similarity degree measurement determines the semantic descriptors of 2D gait images of subjects with various shapes, poses and styles. Second, using the 2D gait images as input and the subjects' corresponding 3D semantic descriptors as output, an end-to-end 3D-SGPE-LSTM is constructed and trained. Third, body shape, pose and external gait factors (3D-*eFactors*) are estimated using the 3D-SGPE-LSTM model to create a set of interpretable gait descriptors to represent the 3D*Gait* Model, i.e., 3D intrinsic semantic shape descriptor (3D-*Shape*); 3D skeleton-based gait pose descriptor (3D-*Pose*) and 3D dressing with other 3D-*eFators*. Finally, the 3D-*Shape* and 3D-*Pose* descriptors are coupled to a unified pattern space by learning prior knowledge from the 3D-*eFators*. Practical research on CASIA B, CMU MoBo, TUM GAID and GPJATK databases shows that 3D*Gait* is robust against object carrying and dressing variations, especially under multi-cross variations.

KEYWORDS

Gait recognition; human identification; three-dimensional gait; canonical correlation analysis.

1 Introduction

Gait as a means for remote biometric recognition and subject identification has shown practical potential deployments in many domains, e.g., access monitoring and intelligent security [1]. Several critical advantages of gait recognition for subject identification and authentication include dispensing with the subject's cooperation, being a long distance away, and being non-contact [2]. However, the challenges in natural scenes include changes in carrying conditions, wearing, hairstyles, viewing

angles, stride velocity and complex tracking environments [3,4]. Gait recognition is still not sufficiently accurate for real-world applications, e.g., due to challenges posed by outdoor environments [5]. Most advanced gait recognition approaches achieve good results or provide specific solutions when applied to publicly available gait datasets. However, changes in silhouettes of human gait can cause a dramatic reduction in recognition rate, e.g., in cross-view gait recognition [6].

Most gait recognition algorithms extract or use gait features from mask-generated gait silhouettes (i.e., binary gait silhouettes) to realise two-dimensional (2D) vision-based or joint-based gait recognition [7]. 2D masked-generated gait silhouettes help reduce the interferences due to the colour of clothes and style of shoes, etc. However, many details of the subject's body and motion information are lost, e.g., the contour of the hand or leg due to occlusion or overlap with the body trunk. Also, the left and right symmetric gait postures and the head and neck are hardly distinguishable. Gait is unique if the body structure and movements of the whole-body parts (which include the upper head, neck and hairstyle) are considered. Many gait recognition approaches either focus on Global Feature Representation (GFR) using an appearance-based method or extract Local Features Representation (LFR) from local gait parts. LFR focuses less on upper body features, and relations among local parts are neglected. In GFR, more attention should be paid to local details of human postures, which can help mitigate the adverse effects of object-carrying conditions [8].

Gait recognition researches are primarily classified into two approaches: appearance-based and model-based. However, few methods are related to 3-dimensional (3D), i.e., 3D volumetric gait [9] and pose-based gait recognition (*Pose*Gait) [10]. Most 3D models are based on 3D voxel, 3D body mesh, 3D poses or 3D joints. The voxel-based 3D model is unstructured with a high-density cluttered point cloud, which is not easy to deal with. Pose-based or joint-based models are poor in extracting body surface features, which help discriminate similar inter-class distances of movement of joints. In current 3D body mesh models, the clothing meshes are not considered or combined with body meshes. If the 3D body model and 3D clothing model cannot be separated, the model's advantages in dealing with clothing or carrying variations are significantly reduced. Thus, the model cannot distinguish the intrinsic body shape underneath the clothes. It is also not so feasible to realise other 3D processes which can be helpful against different variations, i.e., virtual dressing or virtual instance creation.

In this paper, we define a 3D*Gait* model, 3D*Gait* model, to address the problems mentioned above and provide a more targeted 3D*Gait* model. We also propose a novel 3D*Gait* recognition method via interpretable 3D*Gait* descriptors, an end-to-end 3D*Gait* descriptors estimation network and a feature fusion deep model based on a multi-set canonical correlation analysis. The paper makes the following noteworthy contributions.

First, an interpretable and more targeted 3D*Gait* model for subject identification is proposed. Unlike the 3D body model or 3D pose model, the 3D*Gait* model is represented by blending 3D intrinsic shape features, 3D*Gait* pose features and 3D external gait features (e.g., clothes, hairstyles, object carrying, and views) in a uniform 3D parametric gait model. Specifically, a 3D hair synthesising process is introduced to deal with hair styling for more targeted modelling. Based on the 3D*Gait* model, virtual dressing and virtual sample generation processes are introduced to significantly and logically extend the gait data under various walking conditions for training. The targeted 3D*Gait* model enables both pose-based and surface-based gait recognition methods to be robust against multiple walking conditions in real-life scenes.

Second, a gait sensory experiment based on body shape Latin Hypercube design is proposed. It is a practical and efficient method for labelling the semantic values of 3D*Gait* features from 2D gait data. By minimising the dissimilarity between 2D representations and 3D morphed counterparts,

described by 3D*Gait* descriptors, the transformation of 2D gait image features to 3D*Gait* descriptors is well estimated. The labelled 2D to 3D*Gait* data enable 3D*Gait* descriptors to be interpretable and used to train the 3D semantic gait parameter estimation Long Short Time Memory (LSTM) network (3D-SGPE-LSTM).

Third, a novel end-to-end 2D to 3D semantic body parameters estimation network 3D-SGPE-LSTM is proposed. A 3D virtual sample generation strategy is introduced to extend the training data for 3D-SGPE-LSTM by conducting virtual dressing, 3D rotation process, etc. The 3D process in the 3D*Gait* model makes 3D-SGPE-LSTM more robust to various walking conditions.

Finally, a deep learning architecture utilising multi-set canonical correlation analysis has been proposed to achieve efficient feature-level fusion. The approach turns the gait patterns into a unified representation space, enabling it to perform more effectively under diverse walking statuses.

The structure of this research is outlined below. Section 2 examines prior research, Section 3 elaborates on the 3D*Gait* methodology, and Section 4 showcases our experimental findings. Section 5 brings a conclusion.

2 Related Work

Considerable efforts have been made to tackle challenges in gait recognition posed by noise, clothing, object carrying, viewing angles, walking speed, etc. One appearance-based method averages gait silhouettes to gait energy image (GEI) [11]. Another method finds vital frames from a gait sequence based on silhouette matching and using the gait key frame images (GKIs) for recognition [12]. These straightforward visual representations are beneficial in standard walking scenarios. However, they are prone to variations in appearance resulting from clothing or carrying changes. Shape-based gait descriptors have been proposed to eliminate influences emphasising dynamic information representation, e.g., frame difference energy image. Shape-based gait descriptors are theoretically effective in subtle changes in clothing. Still, they are poor with other variations [13], i.e., object carrying, viewing angles, and heavy coats, which easily influence the human gait contour. A Global and Local Feature Extractor (GLFE) was proposed in [1] to improve the gait feature representation ability. It takes advantage of global visual information and details about the local region.

Other appearance-based methods have been proposed to address clothing variations. In [14], a clothing-insensitive gait recognition approach is presented. It leverages part-specific dressing categorisation and a dynamic weighting adjustment mechanism. The human body is divided into eight sections. Each is assigned a different weight according to the factor affected by clothing variations. In [15], a statistical shape analysis method is introduced to decompose GEI into three independent shape segmentations. The higher-order statistical moments extracted from the pooled segmented features are more robust to changes in clothing. However, the inherent drawback of appearance-based methods is their dependency on the viewing angle and poor performance at different walking speeds. Variations in clothing and shoes can also influence the gait recognition rate. In [16], an approach is proposed to integrate characteristic descriptors from 2D, 3D, and audio data to enhance recognition rates on the TUM-GAID database with varying footwear.

Model-based or fusion approaches have advantages in dealing with appearance variation. The underlying trajectories of time-varying gait parameters are captured using the five-link biped walking model to generate a gait dynamics graph for identification [17]. However, the model-based methods based on skeletons can also be affected by various factors, namely object carrying and occlusions. In [6], dynamic characteristics are integrated for gait identification, incorporating determinate learning

to account for clothing and carrying conditions. In [5], a versatile and practical framework named OpenGait has been devised. A straightforward and practically resilient benchmark model, GaitBase, is introduced by integrating cutting-edge gait recognition techniques.

Object-carrying conditions can dramatically influence the contour appearance of the body due to the merging of the body and the object in the gait silhouette. Few approaches ignore the body parts near the carried item during feature extraction. In [18], a consolidated framework for joint intensity adjustments is proposed for identifying gait, which is robust against various carrying conditions. In [19], a novel methodology named Pose-Based Temporal-Spatial Network is introduced. It takes into account changes in carrying and dressing.

Besides clothing and object carrying, identification of gait patterns across multi-views or cross-views is also a challenge, especially using gait videos captured from a few views for training. In contrast, the testing utilises a single-unit camera. Two approaches deal with the view variation, i.e., 2D-based and 3D-related. The view transformation model has been introduced in [20]. It translates 2D gait characteristics from one visual angle to a different one. A focusing mechanism is integrated to selectively attend to significant recurrently acquired partial representations from gait convolutional energy maps, utilised as attributes for viewpoint-agnostic gait identification [21]. In [22], multi-task generative adversarial networks (MGANs) are proposed to gain an understanding of unique feature depictions tailored to specific views to address variation in view angles. A 2D GEI-driven approach for extracting consistent features is proposed in [23]. It uses one uniform deep model to reconfigure the gait data from multiple viewpoints into a single designated view. A non-linear view transformation model is proposed for transforming the gait characteristics from diverse angles into a single standardised view for view-invariant gait representation, which is robust against cross-view variation [24].

3D-based view-invariant gait recognition methods use motion or visual descriptors of 3D body models, such as 3D joints and 3D volume images. In [9], the foremost method utilising 3D modelling techniques is introduced to retrieve gait markers from the 3D data stream. A structural framework is developed to simulate the human lower extremities, incorporating flexible segments at every joint. In [25], a 3D*Gait* identification method uses latent canonical covariates consisting of gait features. The study uses the 3D*Gait* dataset captured by 12 infrared cameras and with 41 retro-reflective markers on humans. Multiple cameras are usually needed to capture the 3D*Gait* data, and extra markers are used to model 3D joint data accurately for 3D*Gait* recognition.

The latest gait recognition methods typically deliver robust outcomes. However, the human gait is 3D and is influenced by various elements. The robustness of a gait identification method must still be enhanced using a more targeted 3D*Gait* model [17]. However, in 3D parametric gait modelling, a productive approach to extracting high-level semantic gait descriptors is lacking. Thus, to tackle the abovementioned problems, 3D*Gait* is put forward in this paper. It aids not only in extracting interpretable and semantic parameters of gait but also combines the various other factors, i.e., apparel, objects and hairstyles, for more targeted 3D*Gait* modelling.

To overcome these problems, the paper introduces the novel 3D*Gait* for gait recognition against variations in carrying and dressing. This study makes two significant contributions. The first contribution is that we introduce an innovative 3D*Gait* model, which is represented by a set of interpretable 3D*Gait* descriptors—the 3D*Gait* model built upon a standard 3D human, where the skeleton is embedded. Unlike the single skeleton or appearance-based 3D*Gait*, 3D*Gait* takes advantage of both approaches. To make full use of 3D*Gait*, virtual clothes and hair stylings are also introduced to 3D*Gait* as 3D-eFactors. As a result, using our interpretable 3D*Gait* descriptors (i.e., 3D-*Shape*, 3D-*Pose* and 3D-*eFactors*), the alteration is applied straight to the standard 3D body by shape, pose, hair and clothes

morphing. Our novel targeted 3D*Gait* provides a new approach for gait recognition against object carrying and clothing changes. The second contribution is that our 3D*Gait* model provides a virtual sample synthesis method to extend the gait dataset with diverse body contours, postures, apparel, hairstyles and object carryings. It simulates gait recognition under variant scenarios for research purposes.

3 Proposed Method: 3D*Gait*

3.1 Overview

Fig. 1 illustrates the framework of 3D*Gait*, which comprises three parts. The first is learning 2D to 3D*Gait* data for 3D-*Shape*, 3D-*Pose* and 3D-*eFactors* data. The other is the 3D*Gait* parameter estimation LSTM network (3D-SGPE-LSTM) using spatial-temporal 2D gait images. The third is the MCCA-DNet for multi-set feature fusion. The 3D-SGPE-LSTM network is trained by learned shape, pose and external gait factors derived from a mixed gait dataset with various walking conditions, i.e., view changes, ball and bag carrying, and variation in dressing. Using mixed gait datasets makes the 3D estimation model robust to different variations. We propose a set of view-invariant descriptors to represent the 3D*Gait*: 3D-*Shape*, 3D-*Pose* and 3D-*eFators*. Since the frames in a gait cycle are different due to walking conditions, ten key frames are chosen for constructing the 3D-*Pose*. As 3D-*Shape* and 3D-*Pose* descriptors have different physical meanings and subtle perturbation under different walking variations, the MCCA-DNet is introduced to transform them to a unified pattern space against various walking conditions.

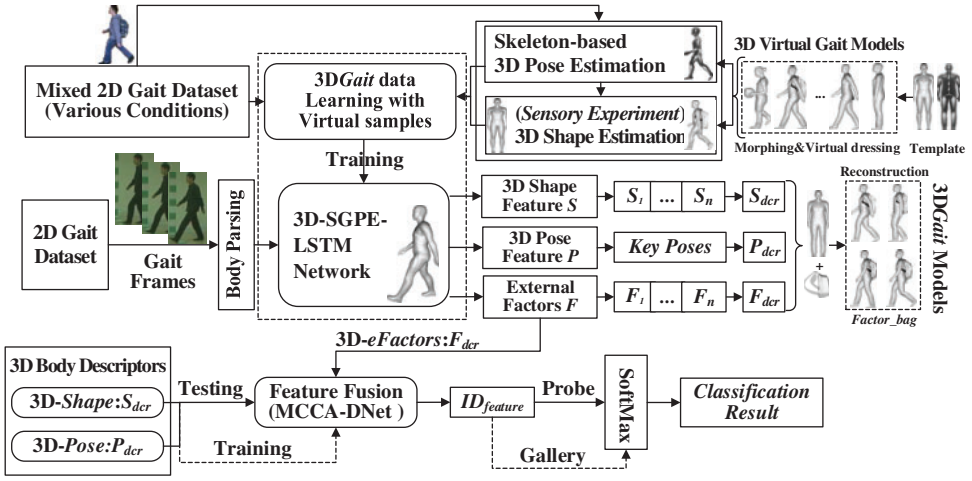


Figure 1: Overview of 3D*Gait* framework

3.2 Targeted 3D Parametric Gait Model with Virtual Hair Styling

In our previous work [26], a standard 3D human body is utilised for deforming the pose and projecting the pose to 2D space to fit onto the posture of 2D gait. The 3D parametric body model describes and constructs the corresponding 3D body model by semantic body labels. Further details of the 3D parametric gait model construction are provided in [26,27].

Besides the skeleton data of gait, shape descriptors are also crucial in gait recognition. Unlike posture, most semantic shape features are high-level physical parameters, i.e., body height, weight, and head size. A 3D body scanner for a static pose can measure these parameters. However, most

gait images are captured by 2D surveillance cameras as it is not practical to obtain the body shape data directly from 3D sensors. In our previous research, the shape deformations were conducted on the 3D human body to meet the 2D gait silhouettes for estimating the shape parameters. The various 3D-*eFactors*, especially object carrying, clothing and hairstyles, greatly influence the extracted gait silhouettes. As a result, the intrinsic body shape characteristics hidden underneath the clothing are sometimes entirely dissimilar from the ground truth.

To address the intrinsic body shape estimation under various clothing and gait factors, a sensory experiment with human evaluators is designed to quickly estimate the body shape parameters without conducting extensive and time-consuming 3D human body measurements. Table 1 demonstrates the interpretable parameters used to construct the 3D-*Shape* descriptor. The main classification consists of two groups: global features and the details of the body (i.e., head and neck size, arm thickness, etc.) The number of parameters is limited to reduce computational time and effort.

Table 1: Main interpretable parameters of *3D-Shape* descriptor

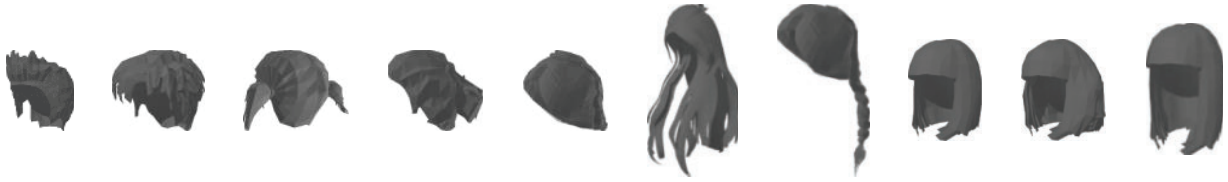
Category	Parameters	Classification	Parameters
Global	Gender Age	Head	Head scale horizontally Head scale vertically
	Height	Neck	Neck scale
	Weight	Torso	Torso scale
Arms	Length of arm Arm muscle thickness Height of leg		Size of chest Stomach scale Hip circumference
Legs	Leg muscle thickness	Feet	Feet dimension

In Table 1, the head and neck are used as the elements of the 3D-*Shape* due to their contribution to gait recognition. However, they can sometimes be obscured by various hairstyles. To make our parametric body model more targeted, 3D hair styling and virtual dressing are introduced in the sensory experiment.

A 3D standard human body is a template model that morphs human parts utilising shapes and pose parameters. The hair is usually not considered in the 3D*Gait* model. In our 3D*Gait* recognition framework, the head size and shape are essential physical features that must be estimated against various hair styling. Fig. 2 shows the multiple hairstyles of different subjects or the same subject at different time stamps.

Since there are only a few publicly available 3D hair datasets, we created several 3D models of hairstyles using 3D CAD software according to some critical semantic hair parameters for the 3D parametric models. Fig. 3 shows some 3D hairstyles, i.e., crew cuts, pixie cuts, ponytails, double pigtails, capes and straight hair. The same hairstyle can have two different parameters, i.e., length and thickness. In our 3D hair synthesising process, the standard style is set to 1, and an additional delta value is provided for a change in style.

Figure 2: Different hairstyles of gait



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

Figure 3: 3D hair models of different hairstyles with multiple lengths and loose degrees: (a) crew cuts; (b) pixie cuts; (c) double pigtails; (d) bun; (e) braid bun; (f) capes; (g) long ponytails; (h) short straight hair; (i) short and thick straight hair; (j) long straight hair

3.3 Learning Gait Data Using Targeted 3DGait Model by Sensory Experiment

1) 3D-*Pose* estimation and formalization of the 3D*Gait* data

Unlike most 3D body or pose data, our proposed 3D*Gait* is not only represented by dynamic pose data, i.e., 3D-*Pose*, but also by intrinsic body data, i.e., 3D-*Shape*, and 3D-*eFactors*. In 3D*Gait* framework, the end-to-end semantic gait parameter estimation 3D-SGPE-LSTM network is trained by the 2D gait images and their 3D*Gait* descriptors. The training data are derived from the mixed 2D gait datasets under different walking conditions. Thus, the 3D*Gait* descriptors should be accurately labelled before training. This paper the 2D to 3D*Gait* data are learned using the proposed 3D*Gait* model via a sensory experiment.

First, the 3D-*Pose* data are estimated before the body shape sensory experiment. The experiment aims to estimate 3D semantic shape values from 2D gait data by minimising the dissimilarity between 2D silhouettes and 3D body models with the same posture. The virtual 3D*Gait* model is then morphed using specific body shape values. In the experiment, the gait poses of 3D parametric body models in a gait cycle must be consistent with their 2D gait data and kept unchanged. Since the 3D body model is in the initial pose (I pose), the 3D*Gait* poses must be first estimated from their corresponding gait images using our previous method [26].

Skeleton from the Carnegie Mellon University motion capture dataset is embedded in our parametric body model. All joints are described using the Biovision Hierarchical data style. A silhouette cost function for dissimilarity measure between 2D gait and 3D morphed gait silhouettes is proposed in [26]. However, only 2D binary gait images are used in our previous work. Due to overlapping and self-occlusion, the estimation of somebody's joints, i.e., two hands, can be distorted and inaccurate. To improve the performance of 3D*Gait* estimation, we modify the silhouette dissimilarity cost function by adding a penalty term for critical joints matching as

$$E = \sum_{\theta\in\Phi}\sum_{i=1}^{I}\left\|\Gamma_i\left(\Upsilon_\theta\left(P,S\right)\right)-\Gamma_i\left(\mathcal{P}_\theta^{2D}\right)\right\|^2+\sum_{\theta\in\Phi}\sum_{n=1}^{N}\left\|Mark_n\left(\Upsilon_\theta\left(P,S\right)\right)-Mark_n\left(\mathcal{P}_\theta^{2D}\right)\right\|^2, \quad (1)$$

where Φ is a gait view set. $\Gamma_i(\cdot)$ defines the *ith* silhouette contour marker extracted from the given 2D gait silhouette. I is the maximum number of markers. Each marker is based on the centroid of the gait silhouette as its original coordinate, given by $\Gamma_i(\cdot) = x_i i + y_i j$. $\mathcal{P}_\theta^{2D}$ is the 2D gait silhouettes at θ view and $\Upsilon_\theta\left(P_j, S\right)$ is the projected gait silhouette from the virtual 3D*Gait* model onto 2D space at θ view, where P is the joint data of body skeleton and S is the body shape parameter. $Mark_n(\cdot)$ defines the nth vital 2D body joint estimated from the 2D gait silhouette. Our method uses the joints estimation algorithm in [28], and a certain amount of vital joints are used in Eq. (1). By minimising the silhouette dissimilarity cost function, 2D gait images approximate the optimal pose of 3D*Gait*.

After obtaining the gait poses using the enhanced 2D to 3D*Gait* pose estimation method, each 2D gait frame in the gait cycle is paired with its corresponding 3D*Gait* skeleton with different joints data. The shape sensory experiment is then conducted using the scheme with the formalized concepts and data as follows. Let $G = \left\{G^1, G^2, \ldots, G^N\right\}$ be a set of mixed gait samples under various walking conditions, and $G^N = \left\{g_1^N, g_2^N, \ldots, g_M^N\right\}$ denotes each sample comprises M different gait frames in a cycle. Let $S = [s^1, s^2, \ldots, s^{K_s}]^T \in \mathbb{R}^{K_s}$ defines the K_s semantic body shape parameters controlling the 3D body shape deformation. $P = [p^1, p^2, \ldots, p^{K_j}]^T \in \mathbb{R}^{3K_j}$ defines the vector of K_j body joint data based on human skeleton which can be estimated according to Eq. (1). $F = [f^1, f^2, \ldots, f^{K_f}]^T \in \mathbb{R}^{K_f}$ represents the flags for K_f external gait factors, i.e., clothing variation, object carrying, and hairstyles. In order to distinguish the shape dissimilarity between ith 2D gait sample and the corresponding 3D sample, a dissimilarity descriptor is defined as $Ds^i = \langle Ks^i, Vs^i \rangle$, where Ks^i denote the dissimilarity of frames under the same views, and Vs^i represent dissimilarity among different views with similar poses if multi-view 2D gait data existed. The frames are selected according to the gait pose, i.e., posture with minimum self-occlusions, and with the distinct shape feature. The multi-view dissimilarity Vs^i make the estimation more accurate. In this paper, the scale of Ks^n and Vs^n ranges from 0 to 4. Each score defines a meaning, i.e., 4–very different, 3–different, 2–medium, 1–similar, and 0–very similar.

Let $In^i = [In^i_{m+1}, In^i_{m+2}, \ldots, In^i_{m+t}]$ be the normalized input based on the 2D gait images with L key frames, where $m \in [1M]$ and $m + t \leq M$. This means several gait frames are used to estimate the body shape parameters instead of a single gait image. Let $Out^i = [S^{iT}, P^{iT}, F^{iT}]$ be composed of three category elements, i.e., K_s body shape parameters, K_f external gait factors and K_j joints data of body skeleton. In^i and Out^i are respectively the input and output data for training the 3D-SGPE-LSTM network. In this paper, $K_s = 16$ body shape parameters are used, i.e., as shown in Table 1, and several walking variations are introduced, i.e., gait views, clothing styles, object carrying, hairstyles, etc.

2) 3D-*Shape* sensory experiment using Latin Hypercube design

The 3D human body is controlled by all the shape parameters, pose and walking conditions. The shape features, i.e., 3D-*Shape*, are denoted as $S \in \mathbb{R}^{K_s}$. An interactive learning framework is proposed to obtain the physical body shape parameters for all the gait samples from 2D gait images. The scheme iteratively updates and generates the new values of S based on their former data until most semantic values minimise the dissimilarity between the 2D gait appearance and the 3D synthesised gait model.

To make full use of human knowledge and reduce subjective errors, five evaluators are trained to evaluate the dissimilarity between a 2D gait sample and the corresponding 3D synthesised virtual gait model. In the body shape sensory experiment, the 3D*Gait* model is morphed by the parameters of 3D-*Shape* descriptors, i.e., S, with the pre-estimated 3D-*Pose* and 3D-*eFactors* descriptor fixed. After evaluation, the average dissimilarity for body shape parameters is determined and represented as $Ds^n = \langle Ks^n, Vs^n \rangle$ for sample S^n, where Ks^n denotes the dissimilarity using gait frames in the same view while Vs^n denotes the multi-view evaluation using frames in different views.

Before the shape evaluation, a set of 3D body models with evenly distributed values of body parameters are given. This necessitates an optimal set of K_s dimensional points of body shape parameters S to be generated first. The points in the set must be distributed as uniformly as possible in their experiment domain. The process is known as the uniform design or space-filling design. One popular approach to generating good space-filling points is the Audze-Eglais optimal Latin Hypercube design (AE-OLHD) [29]. We let the experiment domain of body shape be denoted in the unit cube $C^{K_s} = [0, 1]^{K_s}$. In Latin Hypercube design (LHD), a K_s-dimensional LHD of N points is a set of n sampling data in the experiment domain, i.e., $S = [S_1, S_2, \ldots, S_N]$, where $S_i = (s_i^1, s_i^2, \ldots, s_i^{K_s})$. AE-OLHD [29] is efficient in scattering points uniformly over the experiment domain. It generates evenly spread body shape points by minimizing the objective $U = \sum_{i=1}^{N} \sum_{j=i+1}^{N} d(S^i, S^j)^{-2}$ where $d(S^i, S^j)$ defines the Euclidean distance. By generating the evenly spread N body shape points, denoted as $S = [S_1, \ldots, S_n, \ldots, S_N]$, the virtual 3D body models are deformed with different shape parameter S_n.

The tth frame of the 3D body model with shape data S_n^t and pose data P^t is then morphed from standard body model X_{std}, and represented by $\mathcal{M}_n^t = \mathcal{P}(P^t) \cdot \mathcal{S}(S_n) \cdot X_{std}$, where $\mathcal{P}(\cdot)$ is pose deformation and $\mathcal{S}(\cdot)$ is shape deformation [26]. After virtual dressing, the final models of gait under various gait conditions (as demonstrated in Fig. 4) are denoted by $\mathcal{Y}_n^t = \mathcal{F}(\mathcal{M}_n^t, F)$ where F is the data of 3D-*eFactors*, i.e., clothing, hairstyles and carrying flags. $\mathcal{F}(\cdot)$ represents the virtual dressing and carrying condition, i.e., clothing, ball holding, bag carrying and dressing different hairstyles. The 2D gait image and the corresponding virtually generated 3D*Gait* model with different body shape S_n are then compared directly by evaluators as shown in Fig. 4, where their evaluated dissimilarity at same view and multi-view are denoted by $Ds = \{\langle Ks^1, Vs^1\rangle, \ldots, \langle Ks^N, Vs^N\rangle\}$.

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

Figure 4: 2D gait images and their corresponding 3D virtually dressed gait models under different gait factors: (a) 2D gait carrying a knapsack; (b) 3D model with virtual backpack carrying; (c) texture mapping effect of (b); (d) silhouette difference between (a) and (c); (e) change in hairstyle based on (b); (f) with long coat but without backpack based on (b); (g) 2D gait with a bag; (h) synthesised 3D model with a virtual bag carrying; and (i) texture mapping effect of (h)

Using the aforementioned process, the evaluation on the N sets of shape parameters are obtained, and the optimal set of body shape values are determined according to the minimum observation result in Ds. The final step is to refine the values of body parameters in an even smaller changes by the evaluators. Finally, the optimal estimated body shape is obtained and denoted by $S_{opt} = \tilde{S}_{max} = \left[s_{opt}^1, s_{opt}^2, \ldots, s_{opt}^{K_s}\right]^T$. The output learning data of 3D-SGPE-LSTM network including K_s body shape parameters S, K_j joints data P and K_f walking condition flags F are then compacted as

$$Out^i = \left[out_1^i, \ldots, out_{K_s+K_j}^i, \ldots, out_{K_s+K_j+K_c}^i\right] = \left[S^{iT}, P^{iT}, F^{iT}\right]. \quad (2)$$

The input sample i is denoted by $In^i = \left[In^i_{m+1}, In^i_{m+2}, \ldots, In^i_{m+t}\right] = [G^i_1, G^i_2, \ldots, G^i_t]$ which comprises the corresponding 2D gait images with L consecutive key frames. The body joint values P^i_j in Out^i corresponding to the pose of last gait frame is In^i, i.e., gait image G^i_t.

3) 3D-SGPE-LSTM network

In this paper a set of 3D*Gait* descriptors derived from the output of 3D-SGPE-LSTM network is proposed. The first is 3D-*Shape*, based on K_s body shape values and denoted by $S_{dcr} = S_{opt} = \left[s^1_{opt}, \ldots, s^{K_s}_{opt}\right]^T$. The second is 3D-*Pose* which is composed of N_k gait pose data in a walking cycle, i.e., $P_{dcr} = \left(P^1, P^2, \ldots, P^{N_k}\right)$, where $P^n = [P^{n,1}, P^{n,2}, \ldots, P^{n,K_j}]^T$ denotes the K_j joints data including the three degrees of freedom based on body skeleton. The third is 3D-*eFactors*, based on N_f different gait factors, and denoted by $F_{dcr} = (f^1, f^2, \ldots, f^{N_f})$. Each input sample In^i with t gait frames is used to estimate the pose data based on body skeleton corresponding to the last gait frame, and the key pose data are chosen from all the pose data in a walking cycle. The detailed method used to choose the fixed number of key walking frames based on gait postures is given in our previous work [26].

3D-SGPE-LSTM network is a sequential model in the 3D*Gait* framework and aims to estimate the 3D*Gait* parameters, i.e., shape, pose data and external gait factors, based on spatial-temporal 2D gait images. The network comprises three typical layers, i.e., CNN layer, fully connection (FC) layer and LSTM part, as illustrated in Fig. 5. A Residual Network 50 without the top layer is introduced to operate as CNN layers. The ith input of the network is denoted by $In^i = \left(g^i_1, g^i_2, \ldots, g^i_t\right)$, i.e., it consists of t consecutive frame-by-frame 2D gait images. As aforementioned, the output data are coded into three gait-related data types (i.e., body shape parameters S, body pose data P and external gait factor parameters F), and denoted by $Out^i = [S^{i^T}, P^{i^T}, F^{i^T}]$. To simplify the training process, the 3D-*eFactors* estimation model of 3D-SGPE-LSTM framework is divided into several sub 3D-*eFactors* networks according to the region of interest (ROI) regions. They are trained separately against various walking conditions and different tasks.

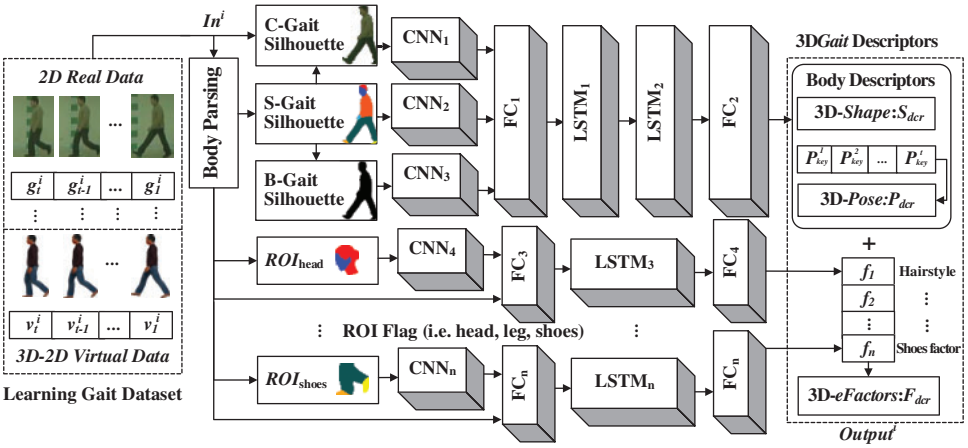


Figure 5: 3D-SGPE-LSTM Network

Fig. 5 shows there are several CNNs which operate as feature-extracting layers to transfer the gait static characteristics and ROI characteristics corresponding to different types of gait images, i.e., colour gait (C-Gait), body semantic segmentation gait (S-Gait), binary gait (B-Gait) silhouettes and ROI regions. The parameters of 3D-*eFactors* are split and trained in different models, as shown in Fig. 5.

The construction of the four types of gait images is as follows. First, the subject is segmented in the original image with a bounding box using a person detection model, i.e., You Only Look Once version 5 (YoLov5). Second, the segmented RGB gait images are passed to a novel pose-extracting model, introduced in [28], to obtain accurate S-Gait silhouettes with six body parts, thirteen clothing categories and small-scale accessories (i.e., hat, pants, shoes, jumpsuit, coat, scarf, etc.). Third, using the S-Gait silhouettes, the human body is parsed with the background removed, and the B-Gait silhouettes are easily obtained by setting all the parsed body regions in S-Gait silhouettes to black. Fourth, using the B-Gait silhouettes as masks, the C-Gait silhouettes are segmented from the RGB gait images. Finally, in the S-Gait silhouettes, parsed body parts, clothing categories and small-scale accessories are selected according to ROI region to estimate 3D-*eFactors* using the proposed CNN-LSTM network. The detailed analysis of 3D-eFactors is helpful for the virtual dressing process of the 3D*Gait* model. Our method uses the face and hair body parts for hairstyle estimation. The upper clothing, arm, pant and leg regions are used to analyse the detailed clothing classifications. The ROIs corresponding to left and right lower legs and shoes are used to estimate the shoe category.

Compared with most of the popular gait features extracted only from B-Gait silhouettes, our method takes advantage of different types of gait images. They carry more helpful information than B-Gait images, which make dressing conditions, carrying items and body parts more easily determined. The first full connection (FC_1) following the three CNNs deals with the feature fusion of three types of gait silhouettes. Direct learning gait features from C-Gait silhouettes must overcome the problem of colour variation, mainly caused by colourful clothing styles and variations in illumination and view angles. To make gait recognition robust, the chosen B-Gait images ignore the colours in most gait recognition methods. However, some helpful information has also been removed, thus weakening the ability of the technique to address invariant gait recognition issues under diverse walking conditions.

Instead of ignoring colours, we introduce three schemes to deal with the effects of colour by generating 3D-2D virtual dressing data, adding noise to C-Gait images, and constructing a feature fusion layer. Virtual sample generation extends our training dataset by dressing different colour clothing on its 3D*Gait* model and projecting it to 2D space. Typically, the input sample In^i and its output Out^i are coupled, i.e., with one to one mapping. However, to make our 3D-SGPE-LSTM network more robust, virtually generated samples are introduced to multiple inputs and mapped to the same output. The 3D-2D virtually generated samples to $In^i = \left(g_1^i, g_2^i, \ldots, g_t^i\right)$ are denoted by $In_{v,d}^i = \left(v_{1,d}^i, v_{2,d}^i, \ldots, v_{t,d}^i\right)$, where $v_{t,d}^i$ defines the 3D-2D mapping gait image with dth clothing styles, i.e., different clothing styles have varied colours as shown in Fig. 6. The shape parameters and 3D skeleton pose are the same as the gait frame t obtained in our sensory experiment. As a result, multiple input samples including virtually generated samples in set $\left\{In^i, In_{v,d}^i\right\}, d \in [1 \ldots D]$ have the same output Out^i mapping, which significantly helps to progress the generalisation capability of our model.

The three types of gait CNNs feature fusion process are based on a FC1 layer. Provided in input, three 2D gait features $\mathcal{F}_{CNN}^i$ extracted by ResNet-50, i.e., $\mathcal{F}_{CNN}^i = \mathcal{N}_{feature}(G^i) \in \mathbb{R}^d$, $G^i \in \{Img_{C\_Gait}, Img_{S\_Gait}, Img_{B\_Gait}\}$, $i \in [1 \ldots 3]$, are fused by the mapping function such that

$$FC_{neuron}^i = ReLU\left[\left(\sum_{n=1}^{3} \delta^n \mathcal{F}_{CNN}^n\right)^T \cdot w_i + b_{ias}^i\right], \tag{3}$$

where FC_{neuron}^i denotes the ith Rectified Linear Unit (ReLU) neuron in fully connected layer with maximum I_{max} neurons in total. $w_i \in \mathbb{R}^d$ and b_{ias}^i is a bias. $\sum_{n=1}^{3} \delta^n = 1$ and $ReLU(\cdot)$ is the activation function defined as $ReLU(x) = max(0, x)$. As shown in Fig. 5, the S-Gait silhouette plays an essential role in the feature fusion, and the segmentation accuracy of S-Gait directly influences the efficiency of

the model. The mean squared error cost function is introduced in the 3D-SGPE-LSTM network for estimating the 3D-*Shape* and 3D-*Pose* parameters, i.e.,

$$\mathcal{L}_{oss} = \frac{1}{N}\sum_{i=1}^{N}\left(\left\|S_{dcr}^{i} - \hat{S}_{dcr}^{i}\right\|^{2} + \left\|P_{dcr}^{i} - \hat{P}_{dcr}^{i}\right\|^{2}\right), \tag{4}$$

where N defines the total input samples. The i*th* input sample of the network is $In^{i} = \left(g_1^i, g_2^i, \ldots, g_4^i\right)$, i.e., it consists of four consecutive frame-by-frame 2D gait images. S_{dcr}^{i} are the ground truth body shape data of i*th* sample, and $\hat{S}_{dcr}^{i}$ are the predicted data from In^{i}. P_{dcr}^{i} are the ground truth body pose data of the last gait image g_4^i in i*th* sample, and $\hat{P}_{dcr}^{i}$ are the predicted data. Table 2 shows the 3D-SGPE-LSTM main training hyperparameter settings. The ResNet50 is introduced as a base CNN network for fine-tuning, and the last three layers are set as trainable.

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

Figure 6: 3D-2D virtual gait data with varied clothing styles: (a), (d), and (g)–2D gait of different poses; and (b), (c), (e), (f), (h), and (i)–render effects of two different virtual dressing corresponding to their gait poses

Table 2: 3D-SGPE-LSTM main training hyperparameter settings

No.	Parameter	Parameter value	Explain
1	Base network	ResNet50	CNN feature extractor
2	Normalised gait image size	120 × 80 × 3	Height 120 pixels, width 80 pixels
3	Learning_rate	0.001	The learning rate for training
4	Lstm units	128	Number of hidden nodes in LSTM
5	Lstm time steps	4	4 Gait frames for an input sample
6	train_epochs	200	Number of training sessions
7	train_batch_size	32	Number of samples per training
8	dropout_rate	0.1	Parameters to prevent overfitting
9	Optimizer	Adam	Adam optimiser
10	Trainable params	1,905,694	The last three layers of ResNet50 are trainable

In our 3D-SGPE-LSTM network, two LSTM layers are introduced to get the gait characteristics using sequential gait data, i.e., the fused features from FC1. Compared with one static gait image, using several gait frames in a walking cycle to extract the 3D-*Shape* and 3D-*Pose* data is more efficient and robust by exploiting contextual information. More data usually means more info and inherent features, which help the network to attain better performance. LSTM, a type of RNN with a long short-term sequence memory function, is appropriate for temporal feature extraction and is widely

used in text and speech recognition. Compared with RNN, LSTM introduces the concept of "gate" to discard unwanted data and record meaningful information. As shown in Fig. 5, after all t gait frames are input to the 3D-SGPE-LSTM network, the final estimated values are output using connection layer FC_2, which works as a regression mapping. Varying gait images from datasets are utilised to create a combined training set. Our sensory data is also leveraged for 3D-SGPE-LSTM training.

In our experiments, each input sample to the 3D-SGPE-LSTM model is of size $3 \times 4 \times 120 \times 80 \times 3$ in tensor. It comprises three types of four frame-by-frame gait images, i.e., four colour gait images, four binary gait images and four body semantic segmentation gait images. The gait image size is normalised to 120×80, i.e., a 3-channel RGB image. An experiment was undertaken to evaluate the model inference speed. The average speed per sample is less than 60 ms, as shown in Table 3. The experiment was undertaken on a 3.8 GHz Intel Core i7-10700K computer, with 16 GB RAM, RTX 2070 8G GPU, in a Python 3.8 environment.

Table 3: Evaluation of the inference speed of 3D-SGPE-LSTM

No.	Sample no.	Batch size	Total time (ms)	Time per sample (ms)
1	16	8	177.05	41.07
2	32	16	691.9	51.62
3	32	32	289.86	39.06
4	64	32	566.08	38.85
5	128	32	1126.08	38.80

4) 3D*Gait* recognition using invariant fusion feature from MCCA-DNet

The fusion of static 3D-*Shape* and dynamic 3D-*Pose* under different 3D-*eFactors* is motivated by two factors. First, the two descriptors complement each other for gait recognition under different walking conditions. 3D-*Shape* is the intrinsic trait of the body and is robust to viewing angles, walking speed and object carrying. 3D-*Pose*, conversely, is less sensitive to static factors, i.e., variations in clothing and hairstyles and occlusions on static body parts. Second, the fusion of the two features under different 3D-*eFactors* using MCCA projects them to a uniform pattern space. Thus, a novel gait recognition framework against various walking variations is applied. Fig. 7 illustrates the semantic feature fusion model, which comprises two phases: semantic feature concatenation and invariant feature projection by MCCA.

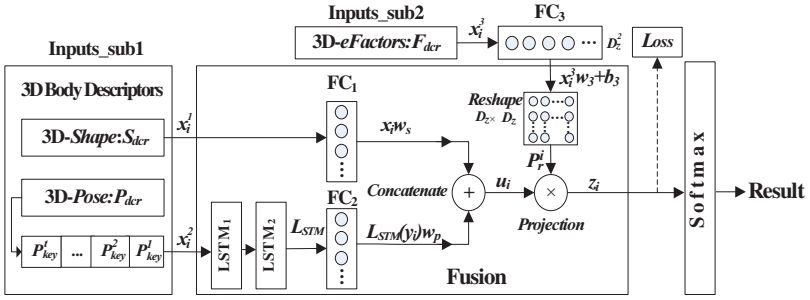


Figure 7: 3D*Gait* Semantic feature fusion deep network based on MCCA (MCCA-DNet)

The system has two sub-input styles: 3D body descriptors and 3D-*eFactors*. The 3D body descriptors, i.e., 3D-*Shape* and 3D-*Pose*, are first concatenated in feature-level fusion. The concatenation model consists of two LSTM networks and two full connection layers of FC_1 and FC_2. The model maps 3D-*Shape* and 3D-*Pose*, with different dimensionalities and physical meanings, into a shared space with the same dimension. The two mapped features are then concatenated and further processed using the MCCA-based feature projecting matrix controlled by 3D-*eFactors*. After these two steps, 3D-*Shape* and 3D-*Pose* are mapped into a novel pattern representation, considering diverse 3D-*eFactors*.

Let $X = \{x_i, i = 1, \ldots, I\}$ and $x_i = \left[x_i^1, x_i^2, x_i^3\right]$ is the input of the feature fusion network and I is the sample numbers. $x_i^1 = S_{dcr}^i \in \mathbb{R}^{K_s}$ denotes 3D-*Shape*, and $x_i^2 = P_{dcr}^i \in \mathbb{R}^{K_j \times t}$ denotes 3D-*Pose*. P_{dcr}^i are t frames of gait pose features. $x_i^3 = F^i \in \mathbb{R}^{K_f}$ defines the vector of 3D-*eFactors*, i.e., comprising viewing angels, object carrying conditions and clothing variation. First, x_i^1 and x_i^2 are mapped into the same dimension by two projecting matrices: $\omega_s \in \mathbb{R}^{K_s \times (D_z/2)}$ for shape and $\omega_p \in \mathbb{R}^{D_L \times (D_z/2)}$ for pose. $(D_z/2)$ denotes the dimensionality of trait within the common space after projection and D_L is the dimensionality of LSTM encoding output. The concatenation feature of subject i is defined as $u_i \in \mathbb{R}^{D_z}$ as shown in Fig. 7. After feature concatenation, the output trait set $U = [u_1 \ldots u_i \ldots u_I] \in \mathbb{R}^{D_z \times I}$ is processed by an MCCA-based invariant feature projecting matrix P_r on the basis of different walking conditions. The invariant characteristic set of ith subject after feature projection is denoted by ${z_i}^T = {u_i}^T P_r^i$, where $P_r^i = [p_1^i p_2^i \ldots p_{D_z}^i] \in \mathbb{R}^{D_z \times D_z}$. The newly defined pattern space is explained by the characteristic set $Z = \left\{z_i \in \mathbb{R}^{D_z}\right\}$ where $i \in [1\ I]$. After feature fusion, the subjects are expected to have different walking conditions in the unified space. The projection process is based on MCCA, which computes multi-sets of correlation projection matrices to separate the multiple (more than two) sets of variables, i.e., with different walking conditions, in new subspace. Unlike canonical correlation analysis which analyzes the maximum correlation between two variables, MCCA applies an objective function of the covariance matrixes from multiple vectors to compute maximum correlation of their canonical variables. In this paper, the training of MCCA-DNet is to learn the optimal parameters that maximize the correlation of the data sets with different walking conditions.

To describe the multi-set correlations, we classify the samples into multi-sets according to walking conditions. Let $X_k \in X, k = [1\,K]$ constitute the subcollection of X where every instance shares identical walking conditions and K is the total number of combinations with different walking conditions. $U_k \in U$ denotes the sets obtained after merging features. Let $p_m, p_n \in \mathbb{R}^{D_z}$ denotes the correlation coefficient of two sets X_m and X_n, and calculated by their concatenation characteristic U_m and U_n, i.e., features U_m and U_n, using

$$\rho_{m,n} = \frac{p_m^T U_m {U_n}^T p_n}{\sqrt{(p_m^T U_m {U_m}^T p_m)(p_n^T U_n {U_n}^T p_n)}}. \tag{5}$$

The criterion used by our MCCA-based feature fusion model is

$$\underset{P_r}{max} \sum_{m=1}^{K} \sum_{n \in U_{pos}^m} \rho_{m,n}, \tag{6}$$

subject to $p_m^T U_m {U_m}^T p_m = p_n^T U_n {U_n}^T p_n = 1$ and $P_r = [p_1 p_2 \ldots p_{D_z}]$. U_{pos}^m denotes the items of wholly positive sets to X_m that indicates their subject matter is identical but different walking conditions. Thus, Eq. (6) is reformulated as

$$\underset{P_r}{arg\,min} \mathcal{J} = \underset{P_r}{arg\,min} \sum_{m=1}^{K} \sum_{n \in U_{pos}^m} \left\| U_m^T P_r^m - U_n^T P_r^n \right\|_2^2, \tag{7}$$

where P_r is a tensor of size $D_z \times D_z \times K$. $P_r = \tanh\left[\left(x_i^3\right)^T \omega_3 + b_3\right]$ which is under the control of the walking conditional vector x_i^3 over the FC_3. The computation of P_r is converted into learning process of FC_3 network. The loss function of our feature fusion model is defined as

$$\mathcal{L}_{oss} = \sum_{m=1}^{K} \sum_{n \in U_{pos}^m} \sum_{i=1}^{N} \left\| u_m^i P_r^m - u_n^i P_r^n \right\|_2^2, \tag{8}$$

where N stands for the sampled quantity of sets U_m and U_n. The MCCA-DNet converts the multi-set data into a consistent pattern space under different conditions. Thus, it undergoes training before the SoftMax classifier. After feature fusion by MCCA-DNet, the refined trait z from the training (i.e., gallery) set with the class label is fed into the SoftMax classifier. The refined features from the test (i.e., probe) sets are used for testing.

4 Experiments

We validate the functioning capacity of the mentioned 3D*Gait* on several widely used cross-variation gait datasets, i.e., CMU MoBo, CASIA-B, TUM-GAID and the Gait database of Polish-Japanese Academy of Information Technology (GPJATK), under different 3D-*eFactors*, with multi-view angles, ball and bag carrying, and variations in clothing, hairstyle and walking speed. The three networks in 3D*Gait* are trained separately, i.e., 3D-SGPE-LSTM for semantic gait parameter estimation, MCCA-DNet for unifying the descriptors and the SoftMax recognition network. A mixed gait dataset with various walking conditions was constructed to conduct the sensory experiment and train the network with the learned body data and variation flags, as shown in Table 4.

Table 4: Elements of the mixed gait dataset

Variations	Dataset	Sub. no.	Describes	Pattern no.	3D-*eFactor* variables encoding
Speed	CMU-MoBo	5	2-level speed	10	Speed variable $f^1 \in [1 \ldots 2]$
Views	CASIA-B	24	11 views	264	Azimuth angle variable $f^2 \in [0 \ldots 360]$; elevation angle variable $f^3 \in [0 \ldots 90]$
Clothing	CASIA-B TUM-GAID	24 10	Coat (11 views) Coat	264 10	Styles (normal, coat, etc.) $f^4 \in [1 \ldots 10]$; tight/loose, long/short $f^5, f^6 \in [0 \ldots 10]$
Carrying	CASIA-B CMU-MoBo	24 5	Bag (11 views) Ball carrying	264 5	Styles (bag, ball etc.) $f^7 \in [0 \ldots 10]$; position $f^8, f^9 \in [0 \ldots 10]$; shape $f^{10} \in [0 \ldots 10]$; size $f^{11} \in [0 \ldots 10]$
Hairstyles	CASIA-B TUM-GAID	10 10	11 views Lateral view	110 10	Styles (crewcut, ponytail, straight hair, etc.) $f^{12} \in [0 \ldots 10]$; length $f^{13} \in [0 \ldots 10]$; thickness $f^{14} \in [0 \ldots 10]$

The semantic feature fusion network is also trained using the mixed gait dataset. The recognition network is then trained according to each gait dataset for different tests. In Table 4, sub. no. denotes the number of subjects from the given dataset and f^k encodes the variations. The walking condition vector is denoted by $F = [f^1, f^2, \ldots, f^{K_f}]^T \in \mathbb{R}^{K_f}$.

4.1 Experiments Based on CMU MoBo Database

The CMU MoBo [30] comprises six camera recordings of twenty-five individuals. Although the number of subjects is small, all subjects were under four varied walking situations. They are slow walking, quick walking, inclining walking and walking with a ball. To evaluate the effectiveness of 3D*Gai*t when the walking conditions are changed, i.e., due to speed variation and object carrying, the experiments are designed by the settings in [31], as shown in Table 5. Twenty-five subjects in CMU MoBo are used in experiments, i.e., twenty-five subjects of the slow walk are in the gallery set, and their fast walk is involved in the probe group. Twenty-five subjects of the quick walk are in the gallery group, and ball-taking is involved in the probe group. The gallery and probe groups have equal group sizes of $25 \times 3 \times 4$.

Table 5: The experiments on CMU MoBo dataset for robustness test

Experiment	Gallery walk group	Probe walk group
SvsQ	Slowly	Quickly
SvsB	Slowly	Ball-taking
SvsI	Slowly	Inclining
QsS	Quickly	Slowly
QvsB	Quickly	Ball-taking
QvsI	Quickly	Inclining
IvsS	Inclining	Slowly
IvsQ	Inclining	Quickly
IvsB	Inclining	Ball-taking
BvsS	Ball-taking	Slowly
BvsQ	Ball-taking	Quickly
BvsI	Ball-taking	Inclining

In the training process of our 3D-SGPE-LSTM network, virtual data comprising inclining and ball-taking samples with different body shapes are generated. They are added to enhance the model's resilience against diverse disruptions. In this task, we trained the 3D-SGPE-LSTM network using the data from the mixed dataset, as shown in Table 4, together with the virtually generated samples, as shown in Fig. 8. The input virtually generated samples to 3D-SGPE-LSTM are denoted by $In^{i}_{v,ball\_inlcine_n} = \left(v^{i}_{1,ball\_inlcining_n}, v^{i}_{2,ball\_inlcining_n}, \ldots, v^{i}_{t,ball\_inlcining_n}\right)$, where $v^{i}_{t,ball\_inlcine_n}$ denotes the tth 3D to 2D gait image under external factors of ball taking or inclining walking with nth virtually generated 3D-*Shape* descriptors of ith input sample. The corresponding 3D-*Pose* and 3D-*eFactors* output of $v^{i}_{n,ball\_inlcine_n}$ are the same as In^{i}, but with the different virtually generated 3D-*Shape* descriptor. The 3D-SGPE-LSTM network is separately trained using the SoftMax recognition network with the mixed dataset. In the recognition process, only the gait from the gallery group is used to train the SoftMax classifier. The probe data are used for the test. The feature fusion network is trained with the mixed dataset and CMU-MoBo examples at two-speed levels.

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

Figure 8: Illustration of synthetic sample generation: (a) 2D body with a ball taking; (b) 3D pose and shape estimate from (a); (c) 3D virtual model with a ball taking based on (b); (d) texture mapping effect of (c); (e) 3D body model with 3° inclining rotation of (c); (f) 3D body model with 36° horizontal rotation of (c); (g) 3D body model with different hairstyle of (c); (h) 3D body model with different hairstyle and clothes of (c); and (i) 3D body model with different carrying, hairstyle and 6° inclining rotation of (c)

Unlike methods that only consider the lateral view data, our experiments included three views, i.e., captured by cameras vr03_7, vr05_7 and vr07_7. The object carrying in these views is seen, i.e., taking a basketball, which influences the silhouette shapes and the dynamic skeleton representation. The rank-1 results of our proposed approach are compared with radial basis function (RBF) [32], Pyramidal Fisher Motion (PFM) [33], gait dynamics graph (GDG) [17], deep convolutional and recurrent neural network (CRNN) [34], GEI subspace projections (GSP) [13], and Pose Energy Image (PEI) [31] as shown in Table 6. The data of the probe view are derived from cameras vr03_7.

Table 6: Recognition results (%) comparison on mobo dataset for robustness test to cross-variation

Exp.	GDG	CRNN	RBF	PFM	GSP-CRC	PEI	PoseMapGait [35]	3D*Gait*
SvsQ	92	92	96	92	85	100	86	98
SvsB	–	–	87	100	93	92	62	95
SvsI	–	–	–	–	–	60	82	90
QvsS	91	92	92	92	82	88	–	96
QvsB	–	–	88	83	84	64	–	89
QvsI	–	–	–	–	–	72	–	92
IvsS	–	–	–	–	–	60	–	88
IvsQ	–	–	–	–	–	80	–	91
IvsB	–	–	–	–	–	32	–	85
BvsS	–	–	87	48	91	92	–	94
BvsQ	–	–	88	48	85	84	–	93
BvsI	–	–	–	–	–	60	–	86

Table 6 shows some experiments involving inclining walking, i.e., Exp. SvsI, QvsI, BvsI, IvsS, IvsQ and IvsB, but only the method in [31] reported the recognition result under this walking condition. The recognition rates are pretty low when the inclining factor is involved in the cross-factor experiment, i.e., 60% in the experiments: slow *vs*. inclining walk, inclining *vs*. slow and ball *vs*. inclining. This is because

the silhouette in inclining walking significantly differs from normal walking. Most 2D appearance-based gait recognition methods cannot deal with this variation. Since our proposed approach is based on a parametric 3D human template, it can exploit any view rotation in 3D space. The multi-view virtual samples, including both vertical and horizontal, are virtually generated by rotating the 3D models and added into the training dataset for the 3D-SGPE-LSTM network, as demonstrated in Fig. 8. The worst accuracy of 32% is for an inclining walk and ball-taking walk. That is due to the variation of combined external gait factors, i.e., ball taking and inclining walking. However, 3D*Gait* has advantages against various walking conditions, even when the variations are combined. This is mainly due to the targeted 3D*Gait* model having significant potential in 3D virtual dressing and virtual sample generation process, which cannot be performed on 2D gait images, as illustrated in Fig. 8.

4.2 Experiments Based on the CASIA B Database

CASIA B is the most widely used gait database under several walking conditions, i.e., view changes, object carrying and clothing variations. It comprises 124 subjects captured at eleven views from front view 0° to back view 180°. Every viewing data is composed of six normal walks (nm), two bag-carrying walks (bg), and two clothing-varied walks (cl).

Three types of experiments emphasising different variations are introduced, i.e., cross-view, variation in clothing and object carrying. First, the evaluation for cross-view identification was conducted. Like the experiment setting in [23], we chose 100 normal gait subjects, i.e., ID from 025-124 in CASIA-B for evaluation. The remaining twenty-four subjects are used to construct the mixed gait dataset, as shown in Table 4, for training the 3D-SGPE-LSTM. For each of the 100 subjects, two normal sequences out of six are selected for each view. In Fig. 5, the probe angles data compared with other methods are 54°, 72°, 108° and 126°. In the view variation task, the virtual samples with different probe views are generated based on gallery gait view gait data and added to train our network to make its model more robust to cross-view interferences.

Fig. 9 and Table 7 show a comparison of our approach with the other methods, i.e., Stacked Progressive Auto-Encoders with Nearest Neighbor (SPAE-NN), PoseMapGait, Averaged Gait Image with Sparse Reconstruction-based Metric Learning (AGRI-SRML), Latent Conditional Random Field (LCRF), and Novel Deep Neural Network (NDNN). They show that our 3DGait method performs well in cross-view walking conditions. In most cases, when the gallery and probe views are similar, all approaches have a reasonable recognition rate, i.e., higher than 95%. This is because the silhouettes are like each other, and thus, so are their features. However, when the variation in view is significant, the recognition rates are significantly influenced. Despite this, our method still managed to achieve good performance.

To strengthen the reliability of cross-view gait recognition in most cases, the focus is to automatically learn and extract the common features under various views using multi-view datasets. As discussed in Section 3, our 3D descriptors (i.e., 3D-*Shape* and 3D-*Pose*) are robust to view changes. The physical body shape descriptor is learned directly using 3D parametric models. Unlike 2D gait recognition methods, 3D-*Shape* is a high-level semantic descriptor with physical meanings. Unlike most 2D view-invariant feature selection algorithms, 3D-*Shape* extracts the intrinsic body shape features. Regarding the 3D-*Pose*, our parametric 3D body template integrates the CMU skeleton structure. All joints are encoded relative to the root coordinates. To represent the skeleton similarly, the root coordinate is transformed to the same lateral view, i.e., 90°. Except for the set of 3D*Gait* descriptors, our feature fusion MCCA-DNet using the given a priori knowledge of gait automatically refines the features before classification. The refined features make the result more accurate. Furthermore, the

RNN-based 3D-SGPE-LSTM network trained by the mixed gait dataset with multi-view gait data enables our model to estimate the 3D*Gait* descriptors directly from the 2D gait images under different views and walking conditions.

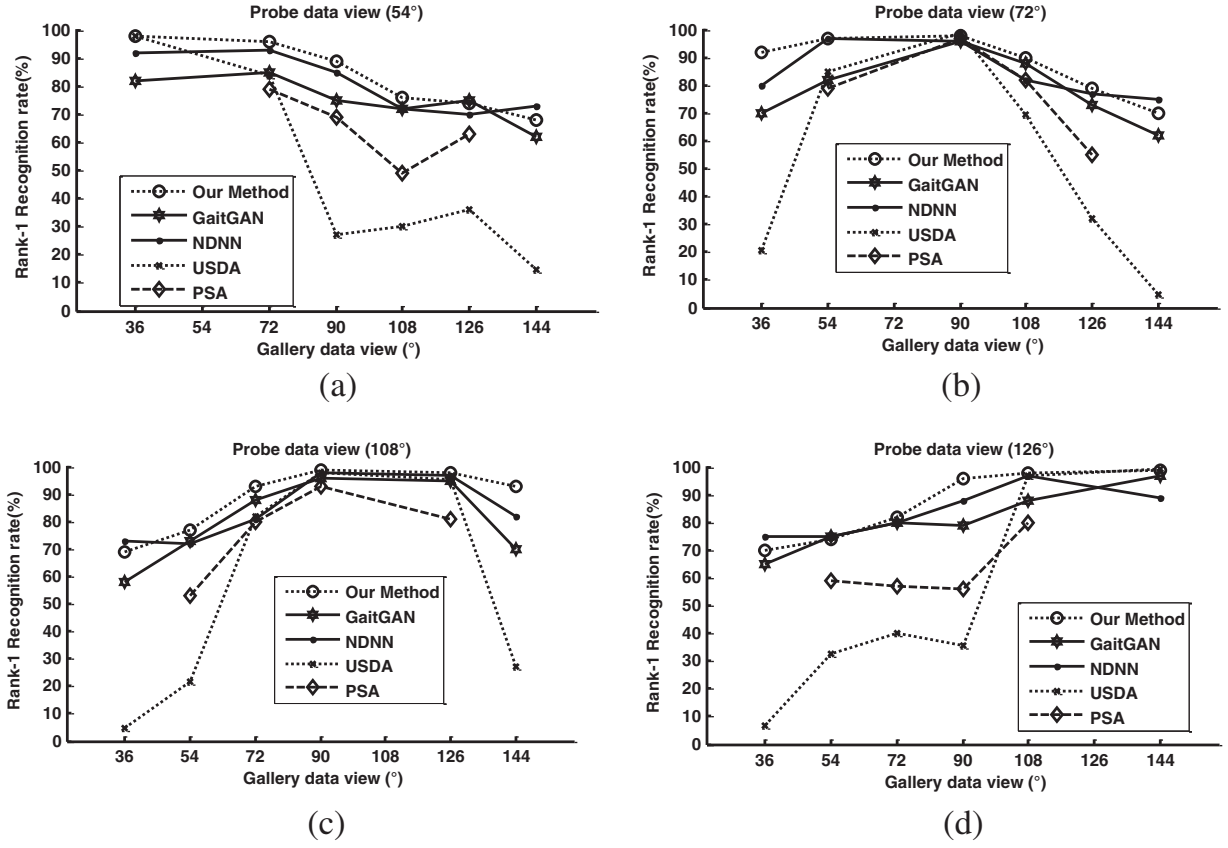


Figure 9: Comparisons of accuracy by varying the probe angles from 36° to 144°: (a) probe view is 54°; (b) probe view is 72°; (c) probe view is 108°; (d) probe view is 126°

Table 7: Rank-1 recognition rates (%) of cross-view identification under 90° gallery view

Method	Probe view						
	36°	54°	72°	108°	126°	144°	AVG
SPAE-NN [23]	52	70	95	95	81	56	74.8
PoseMapGait [35]	85	–	98	95	–	79	89.3
AGI-SRML [36]	–	68	94	96	70	–	82.0
LCRF [37]	68	93	98	99	93	67	86.3
NDNN [24]	78	89	97	99	92	76	88.5
3D*Gait* (our)	80	91	98	98	93	79	89.8

Experiments focused on clothing or object carrying with the same probe and gallery views were also conducted. Like the cross-view experiment, two normal gait sequences (i.e., nm01 and nm02) of

the 100 training subjects were chosen for training. Two gait sequences with different clothing (i.e., cl01 and cl02) and two with bag carrying (i.e., bg01 and bg02) were chosen for tests. An additional experiment was designed to test the mixed clothing and carrying conditions, i.e., nm-bg+cl, where walking with a bag and dressing with a coat condition were integrated. Tables 8 and 9 show the presentation of our method and other novel methods, i.e., Pose-Based Temporal-Spatial Networks (PTSN), Multi-task GANs (MGAN), Gait Convolutional Energy Maps (GCEM), Global and Local Feature Extractor (GLFE), and Gait recognition based on Global-Local network (GaitGL).

Table 8: Recognition rates (%) under different conditions in 90° gallery view

Gallery-probe (%)	PTSN [19]	MGAN [22]	GCEM [21]	GLFE [1]	OpenGait [5]	GaitGL [8]	NDNN [24]	3DGait
nm-cl	61.3	55	71.5	79.0	77.4	81.3	88.5	94
nm-bg	79.3	88	82.3	89.3	94.0	91.0	94.5	95
bg-cl	–	–	–	–	–	–	–	68
nm-bg+cl	–	–	–	–	–	–	–	89
Avg	70.3	71.5	76.9	84.15	85.7	86.2	91.5	86.5

Table 9: Recognition rates (%) comparison using different gallery views

Methods	Normal-gallery/condition-probe (%)								
	Conditions	18°/18°	36°/36°	54°/54°	72°/72°	108°/108°	126°/126°	144°/144°	162°/162°
SPAE-NN [23]	Bag	82	70	67	74	62	76	73	68
	Coat	49	47	47	43	47	44	40	41
NDNN [24]	Bag	–	–	93	–	–	92	–	–
	Coat	–	–	85	–	–	83	–	–
GLFE [1]	Bag	81	88	85	76	76	85	87	84
	Coat	59	67	65	58	60	63	61	57
PTSN [19]	Bag	95	93	88	84	85	83	83	90
	Coat	84	88	72.6	61	75	67	71.0	70.2
3D*Gait* (our)	Bag	84	85	**95**	**96**	**93**	**92**	84	82
	Coat	78	82	**88**	**90**	**91**	**86**	80	**77**

Table 8 shows that our method performs well, especially in bg-cl and nm-bg+cl experiments. Walking while burdened with a bag constituted the training phase for the bg-cl setting, with the coat-wearing scenarios reserved for testing. Due to occlusions and pose distortion, the gait silhouettes of the bag carrying are very different from the normal walking or coat-wearing. On one side of the body, where the hand stays on the bag, being non-oscillatory significantly affects the recognition rate. However, our 3D-SGPE-LSTM network trained by the mixed gait dataset still estimates the intrinsic body parameters under various walking conditions, including with bags. In feature fusion, the hand joint features related to bag variation are eliminated by paying more attention to unrelated joint parts.

The results of further experiments under other views, i.e., 18° to 162°, are shown in Table 9. The comparison shows that our 3D*Gait* is good at side views (near 90 degrees). This is mainly because gait in the side view carries more dynamic pose, shape, clothes and carrying information than other

views, especially when compared with the front view gait (near 0 degrees), as illustrated in Fig. 10. The performance of 3D*Gait* is furtherly improved by combining the gait data of other views to estimate 3D*Gait* descriptors better.

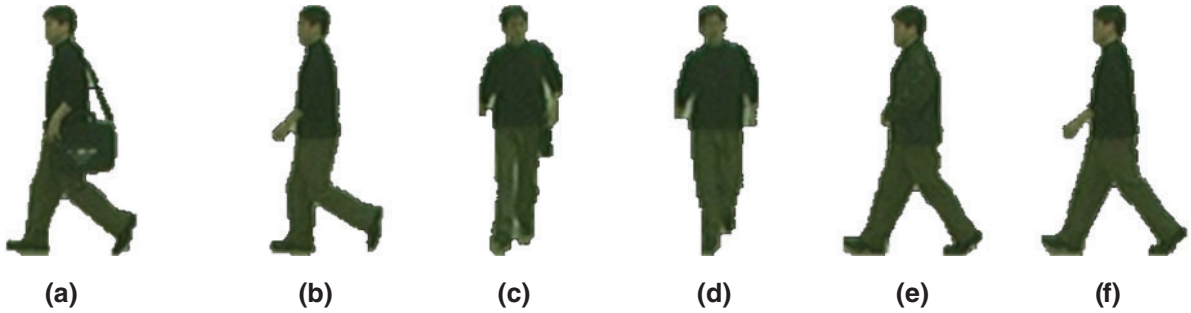


Figure 10: (a) 90° view with a bag; (b) 90° view without a bag; (c) 0° view with a bag; (d) 0° view without a bag; (e) 90° view wearing a coat; and (f) 90° view with normal clothes

Most multi-view gait feature extraction methods convert the gallery and probe data to a uniform representation or extract the unchanged traits. However, our framework finds the intrinsic body descriptors by the 3D-SGPE-LSTM network. As a knowledge-based and data-driven gait recognition method, our 3D*Gait* exploits both advantages of these approaches. It avoids the mismatched features in view transformation, especially for significant view changes.

4.3 Experiments Based on the TUM GAID Dataset

TUM GAID is a multi-model database. It comprises two separate recording parts. The prior was captured in January. The subsequent was recorded in April of the same year. During the elapsed time, various conditions occurred, i.e., hairstyle changes, clothing differences, bag carrying and changes in lighting. The dataset consists of 305 subjects in outdoor scenes and with only a lateral view. Each subject has six different normal walking videos (N1–N6). Two with a backpack carrying (B1–B2). The other two are wearing shoe covers (S1–S2). Like the settings in [7], four different walking variations were considered in our experiments, i.e., normal walk, taking a backpack, using shoe covers and time variation.

The experimental set is the same as in [7], which includes only thirty-two subjects to test the robustness against time variation. Object carrying, virtual dressing and pose-perturbation samples (i.e., backpack, coat, shoe covers, different hairstyles and subtly varied pose) were generated to extend the gait dataset for training 3D-SGPE-LSTM network, feature fusion network and recognition network to achieve better performance against time variation. The performance consequences are illustrated in Table 10.

Without the influence of elapsed time, most methods have a reasonable recognition rate, i.e., close to 100%. However, gait patterns changed rapidly over time and degraded the recognition rate significantly when the various walking conditions changed together. Even the result of the experiment TN without object carrying and clothing changes is below 90% due to time variation. In Exp. TB, the recognition rate is no higher than 80% when backpack-carrying and time variation are combined. By comparing the gait images without the bag carrying, we observed some subtle pose changes of the same subject between the two sessions. Take the subject's head as an example; in the data captured in January, the head faces forward, while in April, it faces the ground, as shown in Fig. 11. This slight pose variation directly changed both the silhouette and the skeleton joints.

Table 10: Comparations on TUM GAID dataset for robustness test to time variation

Exp.	Gallery	Probe	Methods						
			GEI	JITN [18]	DCS [16]	Fusion [7]	Siamese DAE [38]	AT-GCN [3]	3D*Gait*
N	N1–N4	N5–N6	99	100	100	100	99	99	100
B	N1–N4	B1–B2	27	98	99	94	94	96	100
S	N1–N4	S1–S2	53	99	99	96	98	99	98
TN	N1–N4	TN5–TN6	44	63	78	88	81	81	93
TB	N1–N4	TB1–TB2	6	63	62	80	76	79	90
TS	N1–N4	TS1–TS2	9	66	55	83	78	80	88
AVG	–	–	39.7	81.5	82.2	90.2	87.7	89.0	94.8

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

Figure 11: Synthesised pose-perturbation virtual samples: (a)–(c) 2D gait frames of the same subject at different time sessions; (d) 3D pose and shape estimates from (a); (e) texture mapping effect of (d); (f) right hand subtly pose-perturbated 3D*Gait* model based on (d); (g) neck pose-perturbated 3D*Gait* model based on (d); (h) virtual backpack carrying based on (g); and (i) with medium-length coat and new hairstyle dressing based on (d)

To overcome this problem, we generated pose-perturbation virtual samples and other variations, i.e., clothing, hairstyles and backpack carrying, as illustrated in Fig. 11, to extend the gallery set. The delta Gaussian-based method generates the virtual pose directly from the given 3D pose. Let the 3D-*Pose* of a subject be denoted by $P_{dcr} = \left(P^1_{key}, P^2_{key}, \ldots, P^{N_k}_{key}\right)$, where $P^n_{key} = [P^{n,1}, P^{n,2}, \ldots, P^{n,K_j}]^T$ denotes the K_j joint values including three degrees of freedom (DOF) based on body skeleton.

The virtual 3D-*Pose*, i.e., $P^{vr}_{dcr} = (P^1_{key} + \Delta P^n_{joint}, P^2_{key} + \Delta P^n_{joint}, \ldots, P^{N_k}_{key} + \Delta P^n_{joint})$, is generated. The delta changes of joints values are defined as $\Delta P^n_{joint} = [P^{n,1} + \Delta p^{n,1}, P^{n,2} + \Delta p^{n,2}, \ldots, P^{n,K_j} + \Delta p^{n,K_j}]^T$ where $\overline{p} = \left\{(\Delta p^{n,1}, \Delta p^{n,2}, \ldots, \Delta p^{n,Z}), n \in N\right\}$ is N samples generated by Gaussian distribution function such that

$$N\left(\overline{p}|p_{std}, \sigma\right) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\sigma|^{1/2}} exp\left[-\frac{1}{2}\left(\overline{p} - p_{std}\right)^T \Sigma^{-1}\left(\overline{p} - p_{std}\right)\right], \tag{9}$$

where σ denotes the covariance and p_{std} denotes the mean value with D dimensions. Using the generated pose-perturbation virtual samples to extend the gallery set, the feature-coupled projection network and

SoftMax recognition network are trained to be more robust to subtle pose variation, which is one of the main reasons for the good performances of our method in time-elapsed experiments.

4.4 Experiments on the GPJATK Dataset

GPJATK [39] is a 3D multi-view video and motion capture dataset. The dataset consists of 166 data sequences with thirty-two subjects (ten women and twenty-two men) and three different walking factors, i.e., clothes variation, carrying a backpack and four views. Each sequence consists of videos with RGB images of size 960 × 540. In 128 video sequences, thirty-two subjects are wearing clothes only. In twenty-four data sequences, six out of thirty-two subjects changed clothes (subjects 26–31). Finally, seven subjects have a backpack on their back in fourteen data sequences. The dataset helps evaluate methods for cross-variation gait recognition.

The criteria setting in [39] was used for our experiments. Let s1 and s2 be straight walk with clothing 1. s1 denotes walking from right to left and s2 from left to right. s3 and s4 are diagonal walks with clothes1, where s3 is walking from right to left and s4 is walking in reversed direction. s5 and s6 are straight walks like s1 and s2 but with different clothing 2. s7 and s8 are diagonal walks with clothing 2. s9 and s10 walk with backpacks where s9 walks from right to left and s10 walks in the reverse direction. The rank-1 recognition accuracy under clothes1 *vs*. clothes1, clothes1 *vs*. clothes2, and clothes1 *vs*. backpack were determined. In the clothes1 *vs*. clothes1 experiment, sequences s1 and s2 were used as gallery, and sequences s3 and s4 as test samples. In the clothes1 *vs*. clothes2 experiments, sequences s1, s2, s3 and s4 containing subjects p26–p31 with clothes1 were used as a gallery, and the same subjects with clothes2 were used as test samples. In the clothes1 *vs*. backpack experiment, sequences s1, s2, s3 and s4 containing subjects p26–p32 with clothes1 were used as gallery and sequences s9 and s10 containing subjects with backpacks as test samples. We compared our 3D*Gait* with Naïve Bayes (NB), Support Vector Machine (SVM), Multilayer Perceptron (MLP) and CNN methods using marker-less motion data according to the results reported in [39]. Unlike in [39], where only 3D motion data are used for features, both 3D-*Pose* and 3D-*Shape* features are used in our experiment.

Table 11 compares our method with other methods, i.e., NB, SVM, MLP, CNN, and Regularized Discriminant Analysis and Whale Optimization Algorithm (RDA-WOA) [40]. It shows that our 3D*Gait* method achieves the best performance in terms of robustness. It is suggested in [39] that with marker-free gait data, the precision of motion estimation significantly impacts gait recognition performance. A small between-class distance exists when only the motion or pose is used. Furthermore, multiple feature information is usually required for accurate recognition [41]. However, our 3D*Gait* considers motion data and intrinsic body shape features, i.e., 3D-*Shape*. Taking subject 31 as a sample, the recognition rate achieved by the MLP classifier in [39] is approximately 60% for clothes1 *vs*. clothes2 and even lower, i.e., 50% for clothes1 *vs*. a backpack.

Table 11: Rank-1 recognition rates (%) on the GPJATK dataset against views, clothes and backpack variations

Experiments	Covariate	Methods					
		NB	SVM	MLP	CNN	RDA-WOA [40]	3D*Gait* (ours)
clothes1 *vs*. clothes1	views	56	68	80	–	87	91
clothes1 *vs*. clothes2	clothes	57	64	76	67	85	90
clothes1 *vs*. backpack	carryings	71	68	78	68	94	96

The details of subject 31 are shown in Fig. 12, which shows a significantly large body shape, which is different from the other subjects. Using our 3D*Gait* method, the characteristic 3D-*Shape* features (i.e., weight, leg thickness, torso scale, stomach size and hip size) are estimated and fully used for 3D*Gait* model reconstruction. Thus, the proposed gait recognition method is more robust against views, dressing and carrying variations.

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

Figure 12: 3D*Gait* model estimation for subject ID-31 in GPJATK dataset: (a) 2D gait silhouettes of ID-31 subject; (b) estimated 3DGait model from (a); (c) 3DGait model with 30° rotation of (b); (d) 2D gait silhouettes of ID-31 subject with backpack; (e) estimated 3DGait model from (d); and (f) 3DGait model with −90° rotation of (e)

5 Conclusions

This paper proposes a novel 3D*Gait* model based on 3D*Gait* descriptors and a 3D parametric body model with virtual dressing. It is the first attempt to blend 3D intrinsic shape descriptor, 3D*Gait* poses descriptor and 3D external gait factors in a uniform framework to construct the target 3D*Gait* model. Directly using the 3D*Gait* descriptors or making reasonable changes enables easy reconstruction of the corresponding 3D*Gait* model. Thus, the approach has great potential for generating 2D or 3D virtual gait data under various walking conditions, especially clothing, carrying and viewing changes. Our virtual sample generation strategy based on 3D*Gait* is evaluated on different challenging datasets, focusing on when different walking conditions are combined.

The results show that our 3D*Gait* methods are very robust against multi-cross variations that 2D methods cannot handle well, i.e., normal *vs*. inclining walking with speed change and ball carrying, subtle pose changes together with different hairstyle and bag carrying caused by time variation, viewing changes together with clothes and carrying variations, etc. In addition, we proposed a semantic gait parameter estimation LSTM network, 3D-SGPE-LSTM, which aids in estimating 3D*Gait* descriptors directly from 2D gait sequences.

Regarding potential real-world applications, the outcomes of this study can be used in several domains. In surveillance and security, 3D*Gait* enhances existing 2D gait surveillance systems by taking advantage of 3D*Gait* using our 3D-SGPE-LSTM network, even in situations where their appearance or attire may have changed. The study provides a virtual sample synthesis method for 2D or 3D*Gait* recognition research against variant scenarios. Based on our 3D*Gait* model, virtual samples can be easily generated by its 3D descriptors. The model significantly and logically extends the 2D or 3D*Gait* data under various walking conditions for research.

In the future, we will try to improve our gait sensory experiment and develop an automatic approach to labelling more semantic 3D*Gait* descriptors. This will enable our 3D-SGPE-LSTM network to be trained with more targeted data to improve its precision and performance against various walking conditions.

Acknowledgement: This work was supported by the Research Foundation of the Education Bureau of Hunan Province and the National Natural Science Foundation of China.

Funding Statement: This work was partly funded by the Research Foundation of Education Bureau of Hunan Province, China, under Grant Number 21B0060 and the National Natural Science Foundation of China, under Grant Number 61701179.

Author Contributions: All authors participated in designing and implementing the study. All authors discussed the basic structure of the manuscript. Jian Luo drafted the main parts of the manuscript. Tardi Tjahjadi reviewed and edited the draft. Bo Xu and Jian Yi participated in the experiments. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The CASIA B Gait database is available at http://www.cbsr.ia.ac.cn/english/Gait%20Databases.asp, accessed on 23/05/2024. The CMU MoBo Gait database is available at https://www.ri.cmu.edu/publications/the-cmu-motion-of-body-mobo-database, accessed on 23/05/2024. The TUM GAID Gait database is available at https://www.ce.cit.tum.de/en/mmk/misc/tum-gaid-database, accessed on 23/05/2024. The GPJATK Gait database is available by Email at bkw@agh.edu.pl.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] B. Lin, S. Zhang, and X. Yu, "Gait recognition via effective global-local feature representation and local temporal aggregation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Montreal, QC, Canada, Feb. 28, 2022, pp. 1–9.

[2] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan, "A comprehensive study on cross-view gait based human identification with deep CNNs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 209–226, Feb. 2017. doi: 10.1109/TPAMI.2016.2545669.

[3] W. Sheng and X. Li, "Multi-task learning for gait-based identity recognition and emotion recognition using attention enhanced temporal graph convolutional network," *Pattern Recognit.*, vol. 114, no. 1, pp. 107868, Jun. 2021. doi: 10.1016/j.patcog.2021.107868.

[4] S. Liu, S. Wang, X. Liu, C. Lin, and Z. Lv, "Fuzzy detection aided real-time and robust visual tracking under complex environments," *IEEE Trans. Fuzzy Syst.*, vol. 29, no. 1, pp. 90–102, 2021. doi: 10.1109/TFUZZ.2020.3006520.

[5] C. Fan, J. Liang, C. Shen, S. Hou, Y. Huang and S. Yu, "OpenGait revisiting gait recognition towards better practicality," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, Canada, Jun. 18–22, pp. 9707–9716.

[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. Comput. Vis. Appl.*, vol. 10, no. 4, pp. 1–14, Dec. 2018. doi: 10.1186/s41074-018-0039-6.

[7] M. Deng, C. Wang, F. Cheng, and W. Zeng, "Fusion of spatial-temporal and kinematic features for gait recognition with deterministic learning," *Pattern Recognit.*, vol. 67, pp. 186–200, Jul. 2017. doi: 10.1016/j.patcog.2017.02.014.
[8] B. Lin, S. Zhang, M. Wang, L. Li, and X. Yu, "GaitGL: Learning discriminative global-local feature representations for gait recognition," arXiv preprint arXiv:2208.01380, 2022, pp. 1–12.
[9] G. Ariyanto and N. Mark, "Model-based 3D*Gait* biometrics," in *2011 Int. Joint Conf. Biometr. (IJCB)*, Washington DC, USA, Dec. 29, 2011, pp. 1–7.
[10] R. Liao, S. Yu, W. An, and Y. Huang, "A model-based gait recognition method with body pose and human prior knowledge," *Pattern Recognit.*, vol. 98, no. 2, pp. 107069, Feb. 2020. doi: 10.1016/j.patcog.2019.107069.
[11] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, Feb. 2006. doi: 10.1109/TPAMI.2006.38.
[12] S. D. Choudhury, Y. Guan, and C. Li, "Gait recognition using low spatial and temporal resolution videos," in *2nd Int. Workshop Biometr. Forensics*, Valletta, Malta, Mar. 2014, pp. 27–28.
[13] W. Li, C. Kuo, and J. Peng, "Gait recognition via GEI subspace projections and collaborative representation classification," *Neurocomputing*, vol. 275, no. 1, pp. 1932–1945, Jan. 2018. doi: 10.1016/j.neucom.2017.10.049.
[14] V. Narayan, S. Awasthi, N. Fatima, M. Faiz, and S. Srivastava, "Deep learning approaches for human gait recognition: A review," in *2023 Int. Conf. Artif. Intell. Smart Commun. (AISC)*, Greater Noida, India, Jan. 2023, pp. 763–768.
[15] V. Rani and M. Kumar, "Human gait recognition: A systematic review," *Multimed. Tools Appl.*, vol. 82, no. 24, pp. 37003–37037, Oct. 2023. doi: 10.1007/s11042-023-15079-5.
[16] F. Castro, M. Marín-Jiménez, and N. Guil, "Multimodal features fusion for gait, gender and shoes recognition," *Mach. Vis. Appl.*, vol. 27, no. 8, pp. 1213–1228, Nov. 2016. doi: 10.1007/s00138-016-0767-5.
[17] M. Deng, C. Wang, and T. Zheng, "Individual identification using a gait dynamics graph," *Pattern Recognit.*, vol. 83, no. 2, pp. 287–298, Nov. 2018. doi: 10.1016/j.patcog.2018.06.002.
[18] X. Li, Y. Makihara, C. Xu, Y. Yagi, and M. Ren, "Joint intensity transformer network for gait recognition robust against clothing and carrying status," *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 12, pp. 3102–3115, Dec. 2019. doi: 10.1109/TIFS.2019.2912577.
[19] R. Liao, C. Cao, E. Garcia, S. Yu, and Y. Huang, "Pose-based temporal-spatial network (PTSN) for gait recognition with carrying and clothing variations," in *Proc. 12th Chin. Conf. Biometr. Recognit.*, Shenzhen, China: Springer, 2017, pp. 474–483.
[20] D. Muramatsu, A. Shiraishi, Y. Makihara, M. Uddin, and Y. Yagi, "Gait-based person recognition using arbitrary view transformation model," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 140–154, Jan. 2015. doi: 10.1109/TIP.2014.2371335.
[21] A. Sepas-Moghaddam and A. Etemad, "View-invariant gait recognition with attentive recurrent learning of partial representations," *IEEE Trans. Biometr., Behav., Identity Sci.*, vol. 3, no. 1, pp. 124–137, Jan. 2021. doi: 10.1109/TBIOM.2020.3031470.
[22] Y. He, J. Zhang, H. Shan, and L. Wang, "Multi-task GANs for view-specific feature learning in gait recognition," *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 1, pp. 102–113, Jan. 2019. doi: 10.1109/TIFS.2018.2844819.
[23] S. Yu, H. Chen, Q. Wang, L. Shen, and Y. Huang, "Invariant feature extraction for gait recognition using only one uniform model," *Neurocomputing*, vol. 239, no. 2, pp. 81–93, May 2017. doi: 10.1016/j.neucom.2017.02.006.
[24] M. Khan, M. Farid, and M. Grzegorzek, "A non-linear view transformations model for cross-view gait recognition," *Neurocomputing*, vol. 402, no. 12, pp. 100–111, Aug. 2020. doi: 10.1016/j.neucom.2020.03.101.
[25] R. Birdal and A. Sertbas, "3-D gait identification utilizing latent canonical covariates consisting of gait features," *Comput. Mater. Contin.*, vol. 76, no. 3, pp. 2727–2744, Oct. 2023. doi: 10.32604/cmc.2023.032069.

[26] J. Luo, J. Tang, T. Tjahjadi, and X. Xiao, "Robust arbitrary view gait recognition based on parametric 3D human body reconstruction and virtual posture synthesis," *Pattern Recognit.*, vol. 60, no. 2, pp. 361–377, Dec. 2016. doi: 10.1016/j.patcog.2016.05.030.

[27] J. Tang, J. Luo, T. Tjahjadi, and F. Guo, "Robust arbitrary-view gait recognition based on 3D partial similarity matching," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 7–22, Jan. 2017. doi: 10.1109/TIP.2016.2612823.

[28] X. Liang, K. Gong, and X. Shen, "Look into person: Joint body parsing & pose estimation network and a new benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 871–885, Apr. 2019. doi: 10.1109/TPAMI.2018.2820063.

[29] S. Bates, J. Sienz, and V. Toropov, "Formulation of the optimal latin hypercube design of experiments using a permutation genetic algorithm," in *45th AIAA/ASME/ASCE/AHS/ASC Struct., Struct. Dyn. Mater. Conf.*, Palm Springs, CA, USA, Apr. 2004, pp. 1–7.

[30] R. Gross and J. Shi, "The CMU motion of body (MoBo) database," in *Technical Report CMU-RI-TR-01-18*, Pittsburgh, Pennsylvania: Carnegie Mellon University, Jun. 2001, pp. 1–11.

[31] A. Roy, S. Sural, and J. Mukherjee, "Gait recognition using pose kinematics and pose energy image," *Signal Process.*, vol. 92, no. 3, pp. 780–792, Mar. 2012. doi: 10.1016/j.sigpro.2011.09.022.

[32] W. Zeng, C. Wang, and F. Yang, "Silhouette-based gait recognition via deterministic learning," *Pattern Recognit.*, vol. 47, no. 11, pp. 3568–3584, Nov. 2014. doi: 10.1016/j.patcog.2014.04.014.

[33] F. Castro, M. Marín-Jiménez, N. Guil, and R. Muñoz-Salinas, "Fisher motion descriptor for multiview gait recognition," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 31, no. 1, pp. 1756002, Jan. 2016. doi: 10.1142/S021800141756002X.

[34] M. Deng, T. Fan, J. Cao, S. Fung, and J. Zhang, "Human gait recognition based on deterministic learning and knowledge fusion through multiple walking views," *J. Franklin Inst.*, vol. 357, no. 4, pp. 2471–2491, Mar. 2020. doi: 10.1016/j.jfranklin.2019.12.041.

[35] R. Liao, Z. Li, S. Bhattacharyya, and G. York, "PoseMapGait: A model-based gait recognition method with pose estimation maps and graph convolutional networks," *Neurocomputing*, vol. 501, no. 2, pp. 514–528, Aug. 2022. doi: 10.1016/j.neucom.2022.06.048.

[36] J. Lu, G. Wang, and P. Moulin, "Human identity and gender recognition from gait sequences with arbitrary walking directions," *IEEE Trans. Inf. Forensics Secur.*, vol. 9, no. 1, pp. 51–61, Jan. 2014. doi: 10.1109/TIFS.2013.2291969.

[37] X. Chen, J. Weng, W. Lu, and J. Xu, "Multi-Gait recognition based on attribute discovery," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 7, pp. 1697–1710, Jul. 2018. doi: 10.1109/TPAMI.2017.2726061.

[38] W. Sheng and X. Li, "Siamese denoising autoencoders for joints trajectories reconstruction and robust gait recognition," *Neurocomputing*, vol. 395, pp. 86–94, Jun. 2020. doi: 10.1016/j.neucom.2020.01.098.

[39] B. Kwolek, A. Michalczuk, T. Krzeszowski, A. Switonski, H. Josinski and K. Wojciechowski, "Calibrated and synchronized multi-view video and motion capture dataset for evaluation of gait recognition," *Multimed. Tools Appl.*, vol. 78, no. 22, pp. 32437–32465, Nov. 2019. doi: 10.1007/s11042-019-07945-y.

[40] T. Krzeszowski and K. Wiktorowicz, "Combined regularized discriminant analysis and swarm intelligence techniques for gait recognition," *Sensors*, vol. 20, no. 23, pp. 6794, 2020.

[41] S. Liu, S. Huang, S. Wang, K. Muhammad, P. Bellavista and J. Del Ser, "Visual tracking in complex scenes: A location fusion mechanism based on the combination of multiple visual cognition flows," *Inf. Fusion*, vol. 96, no. 22, pp. 281–296, 2023. doi: 10.1016/j.inffus.2023.02.005.