IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 23, 2019, accepted January 30, 2020, date of publication February 13, 2020, date of current version February 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2973898

Multi-Set Canonical Correlation Analysis for 3D Abnormal Gait Behaviour Recognition Based on Virtual Sample Generation

JIAN LUO[1] **AND TARDI TJAHJADI**[2]

[1]Hunan Provincial Key Laboratory of Intelligent Computing and Language Information Processing, Hunan Normal University, Changsha 410000, China
[2]School of Engineering, University of Warwick, Coventry CV4 7AL, U.K.

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

This work was supported in part by the National Natural Science Foundation of China under Grant 61701179 and Grant 41604117, 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 Small sample dataset and two-dimensional (2D) approach are challenges to vision-based abnormal gait behaviour recognition (AGBR). The lack of three-dimensional (3D) structure of the human body causes 2D based methods to be limited in abnormal gait virtual sample generation (VSG). In this paper, 3D AGBR based on VSG and multi-set canonical correlation analysis (3D-AGRBMCCA) is proposed. First, the unstructured point cloud data of gait are obtained by using a structured light sensor. A 3D parametric body model is then deformed to fit the point cloud data, both in shape and posture. The features of point cloud data are then converted to a high-level structured representation of the body. The parametric body model is used for VSG based on the estimated body pose and shape data. Symmetry virtual samples, pose-perturbation virtual samples and various body-shape virtual samples with multi-views are generated to extend the training samples. The spatial-temporal features of the abnormal gait behaviour from different views, body pose and shape parameters are then extracted by convolutional neural network based Long Short-Term Memory model network. These are projected onto a uniform pattern space using deep learning based multi-set canonical correlation analysis. Experiments on four publicly available datasets show the proposed system performs well under various conditions.

INDEX TERMS 3D body modelling, abnormal gait behaviour recognition, long short-term memory model, multi-set canonical correlation analysis.

I. INTRODUCTION

Gait refers to the periodic movements of the human feet and legs, with different body shapes when walking. It can be used to identify the human subject, analyze pedestrian behaviour, and diagnose gait-related health problems [1]–[3]. Abnormal gait behaviour recognition (AGBR) has important research values, e.g., for real time prevention of accidental fall among elderly subjects, predicting signs of gait-related illness for unhealthy subjects, medical diagnosis, and rehabilitative evaluation. Approaches to AGBR can either be based on vision data (i.e., videos or static images) or signal from a wearable physiological sensor. The latter requires the subjects' cooperation and is not suitable for continuous long-term monitoring of subjects with gait-related illnesses or in an elderly home. Furthermore, accurate 3-dimensional (3D) wearable motion analysis systems are usually expensive and more commonly installed in a laboratory environment equipped with numerous motion sensors and plantar pressure sensors. The paper [4] compared a passive vision-based system and a wearable inertial-based system for estimating temporal gait parameters and concludes that low-cost vision-based gait analysis is possible for real-world applications.

The use of 2-dimensional (2D) images for AGBR is intuitive and easy to implement. However, since gait is performed in 3D space, the robustness of a 2D detection model is limited when the camera view changes significantly due to the lack of depth information in 2D gait images. Furthermore, regardless of 2D or 3D system, another important problem with real-world applications is the lack of large abnormal

The associate editor coordinating the review of this manuscript and approving it for publication was Siddhartha Bhattacharyya.

gait dataset. In particular, when compared with the data of thousands of subjects used in biometrics analysis (i.e., face datasets, normal gait datasets, and fingerprint databases) 3D abnormal gait datasets are small and rare. This is partly because 3D abnormal gait behaviour data are difficult to create and rely on professional actors to act out. There are numerous approaches to AGBR, i.e., machine learning, statistical and pattern recognition. Most approaches perform well with sufficient training samples, i.e., it is important to address the small sample set (SSS) issues. In this paper, a different approach based on virtual sample generation (VSG) is proposed to achieve 3D AGBR against the SSS problem. The system could be used in continuous monitoring and classification of different abnormal gaits. In order to deal with SSS issues, virtual samples are generated using 3D parametric model to extend the training dataset.

AGBR can be influenced by various conditions, i.e. variations in viewing angle and body shape, and individual differences. The same abnormal gait features under different conditions are very different in low-level feature space, e.g., the appearances of the same subject at views 0° and 90°. However, they are related in the high-level pattern space. Like canonical correlation analysis (CCA), a pair of projection vectors can be determined to transform the two sets of abnormal gait features to a new subspace by correlation analysis. Unlike CCA, it is necessary for our recognition method to analyze the relationships for more than two sets, i.e., multi-set canonical correlation analysis (MCCA). However, MCCA determines the linear relationships among different vector sets and cannot flexibly focus on temporal feature representation. The feature representation process and the correlation analysis are independent from each other. To address these shortcomings, an end to end uniform spatial-temporal deep network based on MCCA is proposed to gain better performances under different variations.

The novel contributions of this paper are as follows. First, we propose a 3D human similarity measuring function based on body contour and key body joints, with penalty items based on a priori knowledge on human body shape and pose. The function helps morph the standard parametric 3D body model to fit the unstructured point cloud body data as reasonably and efficiently. Second, an asymmetric pose-perturbation virtual body sample generation method is proposed which is based on generative adversarial nets and the information-expanded function via triangular membership (TMIE). Third, a novel MCCA based Long Short-Term Memory model (LSTM) Deep Network (MCCA-DNet) is investigated for better performance of AGBR. We transform the traditional numerical solving method of MCCA to the deep learning problem by minimizing the corresponding loss function, using GPUs to accelerate the learning. MCCA-DNet projects the multi-set abnormal gait features into a uniform pattern space against various abnormal behaviour conditions, i.e., viewing angles and body shape of individuals.

The rest of this paper is organized as follows. Section II presents the related work. Section III presents the method for estimating 3D abnormal gait model from point cloud data. Section IV presents the virtual abnormal gait sample generation. Section V presents the spatial-temporal deep network based on MCCA. Section VI presents and discusses the experimental results. Finally, Section X concludes the paper.

II. RELATED WORK

2D colour surveillance cameras are commonly used to analyze abnormal gait behaviour. Gridded binary 2D gait contour and support vector machine (SVM) are used in [5] to classify abnormal gait, i.e., swaying and falling gait, of seven normal subjects. In [6] abnormal gait features are represented and classified by using optical flow graphs of gait contour. Six gait behaviours of one subject, and five styles of pathological gait performed by a professional actor are collected. A 2D video-based gait assessment system for clinical use is investigated in [7]. The system comprises a walkway grid mat, flat paper bull's eye markers, four photo switches, a light-indicator, a video camera, and a computer. Twelve young and healthy subjects are used as subjects. A 2D vision-based normal and abnormal gait behaviour classification system that analyses foot movements is proposed in [8], where 2D body silhouette segmentation locates the position of toe and heel points of both feet. The abnormal gait classification experiment was performed on a dataset comprising 30 subjects with half normal gait and half abnormal gait. Only lateral 90° abnormal gait data are captured. A low-cost 2D gait analysis using a webcam is evaluated in [9], using intraclass correlation coefficients and minimal detectable change values. The dataset used consists of twenty-one healthy subjects walking on a treadmill. A vision-based gait impairment analysis system for aided diagnosis in [10] uses the INIT gait database which includes sequences of binary silhouettes of ten healthy subjects performing seven abnormal gait styles. A markerless 2D video system based on segmented gait silhouettes to estimate the temporal gait feature is proposed in [11].

Most 2D abnormal gait analysis systems analyze 2D binary silhouettes. Thus, bad segmented silhouettes significantly hamper the system performance. Also, most of the experiments were conducted at lateral view only, i.e., not under real-world conditions. Therefore, a system which uses depth camera for human behaviour recognition and a method of extracting human motion energy features via supernormal vectors is proposed in [12]. The use of a filter to obtain local spatial-temporal features in video, and a measure of similarity features of deep cuboids, are applied to behaviour detection in [13]. The Kinect sensor is used in [14] to provide the joints angle of human body, and the location of ankles and centre of feet, thus giving more accurate gait parameters whilst walking on a treadmill. However, the sensor cannot provide the locations of heels and toes that are required in gait analysis. Furthermore, the output data is noisy.

Most abnormal gait analysis methods can perform well if sufficient training samples are given, especially those involving data-driven deep learning methods, i.e., convolutional neural network (CNN). More data usually mean

more information and intrinsic characteristics, which aid the classifier to achieve higher learning rate [15]. However, large-size datasets are sometimes not available due to some reasons. Thus, some research has been undertaken to address the SSS problem, especially in manufacturing, i.e., building forecast model from scarce samples [16], [17]. In face recognition, the works in [18]–[20] tackle the SSS issues by using face expression recognition theory, regularization approach or virtual faces generation method. Methods to address the SSS problem can be grouped into three categories: gray forecasting model, VSG and feature extraction. Gray forecasting model focuses on population estimation [17], energy consumption [21], and quality control of manufacturing system [16]. The dimension of the data is usually limited and not suitable for high dimensional estimation, e.g., image estimation. VSG is most popular in face recognition. By extending the dataset using virtual samples, the problem of insufficient training samples is overcome to a certain degree. VSG can be achieved using a priori knowledge of the task to extend the samples and adding noise to the current data. In biometrics research, the structure of human body or face is usually used as a priori knowledge, i.e., use mirror face and face symmetry to enlarge the normal face data. Feature extraction uses dimension deduction theory or subset feature selection to choose the intrinsic features of the subjects (i.e., view-invariant face features), and face expression recognition theory [18]. Different methods have their advantages and disadvantages, and in this paper, we combine the advantages of existing methods for abnormal gait behaviour recognition, i.e., VSG and feature-level MCCA.

CCA is a well-known algorithm for determining the correlation between two sets of variables, and it is usually used in feature extraction and fusion [22]. The traditional CCA is usually introduced to extract the related features from only two sets of variables for biometric recognition. In [23], a sparse tensor canonical correlation analysis (ST-CCA) is proposed for colour face recognition. In the multi-view gait recognition in [24], two sets of high-dimensional vectors under different views are analyzed by complete CCA (C3A). However, it is still a challenge to find the intrinsic correlation of multiple biometric features, i.e., more than two sets, due to the computational complexity of high-dimensional correlation matrix. Thus, MCCA was introduced for more than two sets of multitemporal remote sensing data [25]. A linear discriminant multi-set canonical correlations analysis (LDMCCA) is used for fusion of finger biometrics: finger vein, fingerprint and finger shape.

The traditional biometric features are usually static representations, and are thus not appropriate for spatial-temporal feature representation. The feature extraction from original data and the canonical correlation analysis are separated, i.e., not in a uniform model. As a result, they and the classification cannot combine well to address the MCCA problem. In order to do so, an end to end MCCA based spatial-temporal deep network (MCCA-DNet) is proposed in this paper to achieve better recognition against different variations. Compared with traditional MCCA, the multiple high-dimensional correlation projected matrix is not directly computed using general singular generalized eigensystem. Instead, a fully connected network is used for its implicit representation and is trained with our MCCA-DNet which exploits both deep learning network and MCCA theory.

III. ESTIMATING 3D PARAMETRIC ABNORMAL GAIT MODEL FROM POINT CLOUD DATA

Most of the publicly available databases used for gait research are captured by 2D cameras. However, the accuracy in using 2D data to construct 3D model is usually low. Thus, more and more researches use point cloud data with depth information to realize 3D object modelling [27]. The depth information enables the original motion data to be obtained, which aids the extraction of features that can distinguish similar actions and behaviours, and addresses the problem of view change.

In order to overcome the limitation of 2D-vision based abnormal gait recognition in VSG and to exploit a priori knowledge (i.e., structure of human body), the parametric 3D body model with embedded skeleton is introduced in this paper for pose and shape morphing. The body point cloud data are used as deformed observations that guide the standard parametric 3D to morph correctly to the real data by minimizing a similarity matching function. Point cloud data can be easily obtained by 3D depth sensors, e.g., Microsoft Kinect camera. Let $I_{depth} = \left\{\left(x, y, d_{(x,y)}\right), x \in [1..M], y \in [1..N]\right\}$ denote the $M \times N$ depth image from the camera.$d_{(x,y)}$ is the depth value of pixel (x, y). The world coordinates of 3D point cloud data are calculated using the Kinect geometrical model [28]

$$[XYZ]^T = \frac{1}{c_1 d + c_0} dis^{-1}\left(K^{-1}\left[x + u_0\ y + v_0\ 1\right]^T, k\right), \quad (1)$$

where d is depth value, c_1 and c_0 are parameters of the model, u_0 and v_0 are respectively the shifted parameters of IR structural image and depth images, dis is distortion function, k is distortion parameter of the Kinect camera, and K is the IR camera calibration matrix.

The point cloud data of the body are its 3D voxel representation. The data is unstructured and greatly influences the representation of the 3D shape and posture. In order to exploit the body structure as a priori knowledge, we introduce a 3D parametric body learned from the 3D body dataset using a statistical method. The parameterized body model refers to the description and construction of the corresponding body mesh via abstract high-order semantic features (e.g., height, weight, age, gender, skeleton joints, etc.). The parameters involved are based on statistical learning methods. The skeleton of the body is embedded in the model and the 3D parametric model can be deformed both in shape and pose.

As shown in Fig. 1, a standard parametric body model with CMU mocap motion skeleton embedded is introduced for shape and pose morphing, and all the semantic related deformation are conducted directly on this template. The initial template model can be learned from the 3D body dataset

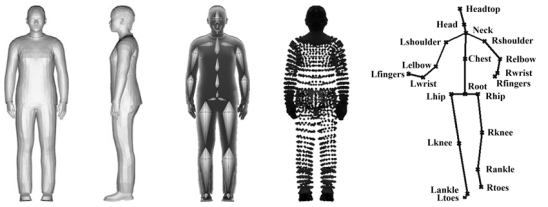

FIGURE 1. Standard parametric body model with I pose: (a) & (b) 3D gait model at front and lateral view (c) display of embedded implicit skeleton; (d) corresponding structured point cloud data; and (e) CMU mocap motion skeleton.

and constructed by principal component analysis, data-driven method or other statistical algorithms. In this paper, no more attention is paid to the construction and training of 3D parametric gait model as numerous works have been done and published in our papers [29], [30].

For simplification we define the deformation function as $F_{de}(\cdot)$, and the 3D model $\hat{Y}$ with pose parameter r and shape parameter β represented by

$$\hat{Y} = F_{de}(r, \beta) = P(r) \cdot S(\beta) \cdot X_{std}, \tag{2}$$

where X_{std} is the standard 3D body model. $P(r)$ denotes the pose deformation with joint input parameters $r \in \mathbb{R}^{N_p}$, where N_p defines the number of joints parameters, and $S(\beta)$ defines the shape deformation with shape parameters $\beta \in \mathbb{R}^{N_s}$, where N_s denotes the number of shape parameters. These two relatively independent deformation parameters, i.e., pose and shape deformation, are trained separately. After the training with 3D standard parametric model via 3D body dataset, the corresponding deformable function is obtained by setting the semantic values of body shape and pose. The body shape semantic parameters used in this paper is shown in Table 1 based on our previous work in [29]. The joints parameters with three degrees of freedom for pose representation are based on the CMU mocap motion skeleton as shown in Fig. 1(g).

TABLE 1. Semantic parameters of the human body shape.

Category	Parameters	Category	Parameters
Global	Gender	Legs	Leg length
	Height		Leg thickness
	Weight	Torso	Torso deep
Head	Head size		Breast size
Arms	Arm length		Stomach size
	Arm thickness		Hip size

Unlike the traditional methods that directly model the 3D body from point cloud data using point cloud reduction algorithm and triangular mesh grid method, the 3D parameterized body model is morphed to fit the point cloud data both in shape and posture. An observation function that measures the similarity of the deformed 3D model and the point cloud data of the human body is proposed. It helps to correctly deform the 3D body using an iterative process that minimizes the observation function. Based on the 3D model estimation, the features of point cloud data of the body are finally converted to high-level structured representation. This process not only abstracts the unstructured data to high-order semantic description, but also effectively completes the dimensionality reduction of the original data.

Let I_p^{α} be the point cloud data captured by the depth sensor at view α. After background subtraction and normalization, they are projected from world coordinate onto the x-y plane with depth information and denoted by P_{α}. As illustrated in Fig. 2, different grey shades indicate the various depth values. The standard 3D body model X_{std} with pose r_{η} and body shape β_{std} is denoted by $\hat{Y}_{\eta} = P(r_{\eta}) \cdot S(\beta_{std}) \cdot X_{std}$. Its corresponding projected depth image at view α is $\Upsilon_{\alpha}(r_{\eta}, \beta_{std})$.

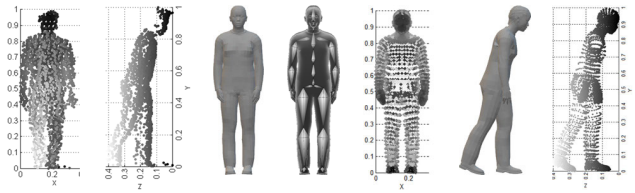

FIGURE 2. (a) & (b) respectively show the 0° and 90° projected depth images of point cloud data captured by Kinect; (c), (d) & (e) respectively show the standard I pose parametric body model, its skeleton structure and point cloud projected depth image; (f) morphed parametric 3D model according to point cloud data (b); and (g) point cloud projected depth image of (f).

To morph the standard parametric 3D body model to fit the body in the point cloud data, the similarity measuring function based on contour and joints matching,

$$\mathcal{L}_{\alpha} = \sum_{i=1}^{M \times N} \left\| \mathcal{D}_i(\Upsilon_{\alpha}(r, \beta)) - \mathcal{D}_i(P_{\alpha}) \right\|_2^2 + \sum_{k=1}^{K} \left\| J_{oint}^{k}(\Upsilon_{\alpha}(r, \beta)) - J_{oint}^{k}(P_{\alpha}) \right\|_2^2, \tag{3}$$

is used, where $\mathcal{D}_i(\cdot)$ is the depth value extraction function from the *ith* pixel in the given depth image. The M and N are respectively the height and width of the normalized image. P_{α} denotes projected depth images of point cloud data at view α, and Υ_{α} is the 3D model projected depth image at the same view after 3D rotation transformation. $J_{oint}^{k}(\cdot)$ denotes the kth key joints data of the human body estimated from the corresponding point cloud depth image. We use the joints estimated algorithm based on single depth image [31], and several key joints are chosen in Eqn. (3), i.e., head, neck, shoulder, elbow, wrist, chest, hip, knee and ankle. Therefore, guided by the extracted body contour of point cloud data and the locations of key joints, the optimization problem is solved by minimizing the contour and key joints similarities as shown in Fig. 3, i.e., arg $min\mathcal{L}_{\alpha}$ for optimal (r, β). The optimal pose and shape semantics parameters are respectively represented by r_{opt} and β_{opt}, and the corresponding morphed 3D model is $Y_{opt} = P(r_{opt}) \cdot S(\beta_{opt}) \cdot X_{std}$.

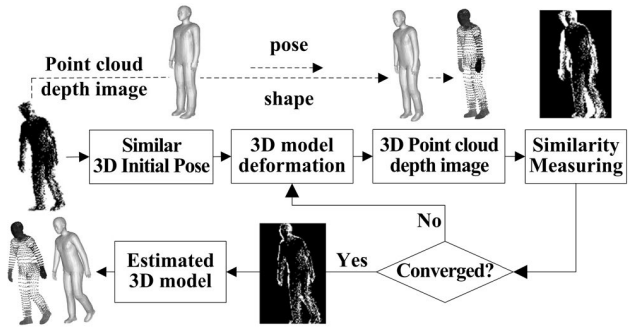

FIGURE 3. Illustration of 3D parametric abnormal gait model estimation.

Morphing a standard 3D parametric body model to fit a 2D body image or unstructured 3D point cloud data is a difficult and time-consuming task. Especially if the initial standard I pose is dissimilar from the target real posture, the result may not be the global optimum or not even resembling the structure of the body. To tackle this problem, a group of 3D gait and common action models with different postures are constructed based on our knowledge, i.e., normal walk, jog, bend, sitting down, wave, clap, kick, throw, etc. Some of the postures are illustrated in Fig. 4, and the base posture dataset of 3D gait or action related models are extensible according to the real application tasks.

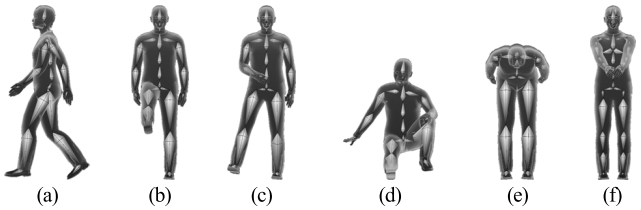

FIGURE 4. Some body postures: (a) walk (b) kick forward (c) kick sideway (d) sit down (e) bend and (f) clap.

Before morphing the 3D parametric model according to Eq. 3, the most similar 3D model is chosen based on the silhouettes from the base posture dataset using the method in our previous work [29]. It is set as an initial pose with standard shape parameters, which greatly helps speed up the morphing process. To make the pose estimation results more reasonable, an extra penalty item is added, i.e.,

$$\mathcal{L}_r = \mathcal{L}_\alpha + \sum\nolimits_{j\in[1J]} r_{ulej}(r) + \sum\nolimits_{l\in[1L]} \hat{r}_{ule_l(\beta)} \qquad (4)$$

where $\mathcal{R}_{pose} = \{r_{ule_j} | j \in [1\ J]\}$ denotes a set of rules about joints with J items, and $\mathcal{R}_{shape} = \{\hat{r}_{ule_l} | l \in [1\ L]\}$ denotes a set of rules about body shape with L items. The rule item function $r_{ule}(\cdot)$ inputs the current joints data r or shape data β to check for any violation of the current rule. It returns a large positive value when the joints data violate the rule, and zero otherwise.

Since the physical variables of body shape and pose are related to each other, i.e., length of head is about 1/8 of height [32], and weight is highly related to height and can be estimated using Body Mass Index (BMI) [33]. These are useful knowledge for controlling the body deformation and reducing the required computation. As for pose data, the constraints for the maximum or delta ranges of joints are added with an appropriate initial pose. The conditions for normal walking movement also aid to speed up the deformation.

IV. VIRTUAL ABNORMAL GAIT SAMPLE GENERATION

normal gait data can easily be obtained using monitoring cameras that are fixed in viewing angles. However, abnormal gait data are usually acted out by healthy subjects. Therefore, the size and the ground truth of the abnormal database are limited. If there are insufficient samples for training, then both the recognition accuracy and generalization ability of an agbr system are greatly curtailed, especially in unknown environments. In order to address this problem, we generated three types of samples, i.e., symmetrical virtual samples, various body-shape virtual samples and pose-perturbation virtual samples. The virtual samples aid in providing missing information among the available samples in order to improve the recognition performance.

A. GENERATING SYMMETRICAL VIRTUAL SAMPLES

The symmetry of the human body structure is widely used in synthesizing 2D or 3D mirror objects, i.e., 2D mirror faces, and 3D face reconstruction with mirror features. The symmetry is used as a priori knowledge to perform 3D reconstruction with missing data. However, there is little research that uses this property for VSG of 3D gait. This is partly because most of the 3D body data captured by 3D cameras or body scanning system are not normalized or unstructured. Noise, missing data and redundant information also make it difficult to use symmetry to synthesize gait-related body parts. In this paper, we introduce parametric 3D body model to alleviate these problems due to the structured body data being controlled by semantic parameters.

Most methods assume the face or body structures are symmetric. Some methods add noise in the symmetrical VSG in order to address the asymmetry of the human body. It is useful to introduce some noise to simulate the difference (caused by body asymmetry) between the left and the right regions of a body. However, it is not reasonable because the difference may be large for subjects with leg problems. In order to alleviate the problems due to asymmetry, the Extremely Learning Machine (ELM) model [34] is used to predict the symmetric data of abnormal gait based on semantic parameters derived from real samples, i.e., body shape and pose parameters.

ELM is capable of randomly assigning its input weights and biases, with its output weights determined by the least-squares method. These characteristics enable ELM to learn faster and better at generalization than single-hidden layer feedforward neural networks and SVM that are based on gradient descent. The ELM model must be trained using some real abnormal gait data. First, the training semantic data, i.e., body shape and pose parameters, and its ground-truth-predicted-symmetric data should be set. In this paper,

the similar abnormal gaits derived from different left or right part of the body are divided into two groups according to their region parts, i.e., the left leg limping and the right leg limping. By estimating their 3D parametric body model from point cloud data, they are denoted by the semantic parameters, i.e., shape and pose parameters. Let $X^L = \{x_k^L | k \in [1 \dots K]\}$ denote the left-body-part-related abnormal gait data where $x_k^L = \left[r_k^L, \beta_k^L\right]^T = \left[x_{k1}^L, x_{k2}^L, \dots x_{kn}^L\right]^T \in \mathbb{R}^n$. r_k^L are the pose parameters of the *kth* 3D model sample in left-related dataset and β_k^L are its body shape parameters. Let $X^R = \{x_k^R | k \in [1 \dots K]\}$ denote the corresponding right-body-part-related abnormal gait data where $x_k^R = \left[r_k^R, \beta_k^R\right]^T = \left[x_{k1}^R, x_{k2}^R, \dots x_{kn}^R\right]^T \in \mathbb{R}^n$. Let the input of the ELM model be x_k^L and the output with M hidden nodes is

$$x_k^R = \sum_{m=1}^{M} \mu_m g\left(w_m \cdot x_k^L + b_m\right), \tag{5}$$

where $g(x)$ denotes the activating function and $\cdot$ is the inner product operator. Rewriting Eq. (5) in matrix form gives

$$X^R = \mathrm{H}\mu = \begin{bmatrix} g\left(w_1 \cdot x_1^L + b_1\right) & \cdots & g\left(w_M \cdot x_1^L + b_M\right) \\ \vdots & \ddots & \vdots \\ g\left(w_1 \cdot x_K^L + b_1\right) & \cdots & g\left(w_M \cdot x_K^L + b_M\right) \end{bmatrix} \times \begin{bmatrix} \mu_1 \\ \dots \\ \mu_M \end{bmatrix}, \tag{6}$$

where H is the hidden layer matrix, $w_m = [w_{m1}, \dots, w_{mn}]^T$, and w_{mn} respectively denote the coefficients that connect and weight the input neurons and the *mth* hidden neuron. $\mu_m = [\mu_{m1}, \dots, w_{mn}]^T$ denotes the vector that connects the *mth* hidden neuron and the output neurons. $tanhx = \left(e^x - e^{-x}\right) / \left(e^x + e^{-x}\right)$ is chosen as the activating function for $g(x)$. For ELM, the input weights w_m and the biases b_m are fixed before training. The model parameters μ that need to be learned are determined using the least squares optimization

$$\hat{\mu} = \underset{\mu}{argmin} \left\| H\mu - X^R \right\|, \tag{7}$$

where $\|\cdot\|$ is the L_2 norm. The optimal solution for the single hidden layer neural network is transformed to solve the problem of the linear system [34]

$$\hat{\mu} = H^{\dagger} X^R, \tag{8}$$

where $H^{\dagger}$ is the Moore-Penrose generalized inverse of H, and the parameters $\hat{\mu}$ are uniquely determined by the given training samples. In real-world applications, the left-part abnormal gait data are used to predict its symmetric right-part data. Also, if the right-part-related data are given, the corresponding left symmetric data can be predicted. This only needs to train the ELM model with two different datasets whose inputs and outputs are symmetrically opposite, i.e., (X^L, X^R) and (X^L, X^R). The ELM model learns both symmetric and asymmetric information of abnormal gait, and performs better in VSG than noise-added methods. Fig. 5 illustrates the generation of the virtual symmetric pose from real samples.

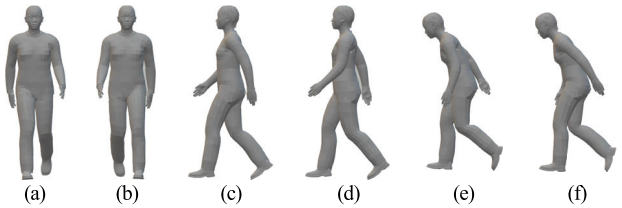

FIGURE 5. (a), (c) and (e) are original pose; and (b), (d) and (f) are respectively he generated symmetrical poses of (a), (c) and (e).

B. GENERATING ASYMMETRIC POSE-PERTURBATION VIRTUAL BODY SAMPLES

To recap, the ELM model is used to predict the symmetric virtual model, and the multi-dimension Gaussian distribution function is fitted for generating various body-shape virtual samples. Unlike the body shape parameters that are more likely to obey the Gaussian distribution in parametric 3D body models, the estimated parameters of 3D pose corresponding to different abnormal gaits may be bias distributed. The pose data derived from a small sample size may differ from the Gaussian probability distribution. Thus, variations in real-world abnormal gait conditions present a challenge to ABGR. To learn the data tendency in SSS tasks is important for VSG. The Gaussian based method can be used to generate the virtual pose data, but how to determine the mean and the standard error of Gaussian distribution is difficult in small dataset.

In this paper, TMIE [17] is introduced to determine the acceptable range of asymmetric perturbation for abnormal pose data. The conditional generative adversarial nets (CGAN) is used to train the conditional generative model for generating the data according to both the classification label and given perturbation. TMIE is derived from mega-trend diffusion (MTD) [35] and aims at estimating data tendency. Unlike single virtual surface data generation, ABGR is based on sequence data. The generated virtual gait data must also be sequence data that can determine a classification label.

Let $X = \{x_k^c | k \in [1 \dots K], c \in [1 \dots C]\}$ denote the K sequential observation samples of 3D abnormal gait, c is the classification label and $x_k^c = \left[r_k^1, \dots, r_k^l, \dots, r_k^L\right]^T \in \mathbb{R}^{L \times N_p}$, where $r_k^l = [r_{k,1}^l, \dots, r_{k,n}^l, \dots r_{k,N_p}^l]^T \in \mathbb{R}^{N_p}$ is the joints parameter vector with N_p data of the *lth* frame belonging to *kth* sequential observation sample. L is the maximum number of frames in a gait cycle for feature extraction and classification. In pose VSG, the body shape parameters are kept fixed and denoted by $\beta_k \in \mathbb{R}^{N_s}$. Let $\mathcal{D}_{c,n}^l = \{r_{k,n}^l, k \in \mathcal{R}_c\}$ be the pose perturbation dataset related to the *nth* joint parameter data belonging to the *lth* frames of samples with class label c. $\mathcal{R}_c$ defines a set of sample index of the same class c.The centre point of $\mathrm{D}_{c,n}^l$ is calculated as $U_{CP} = \frac{1}{N_c} \sum_{k \in \mathcal{R}_c} r_{k,n}^l$ where N_c denotes the total number of samples in class c. The lower bound (L_B) and the upper bound (U_B) for the asymmetric domain range of $\mathrm{D}_{c,n}^l$ are given by [17]

$$\begin{aligned} L_B &= U_{CP} - \frac{1}{S_U} \times (U_{CP} - U_{min}) \\ U_B &= U_{CP} + \frac{1}{S_L} \times (U_{max} - U_{CP}), \end{aligned} \tag{9}$$

where $S_U = N_U/(N_L+N_U+s_p)$, $S_L = N_L/(N_L+N_U+s_p)$, and $U_{min}(U_{max})$ is the minimum (maximum) of the observation value in $\mathrm{D}^l_{c,n}$. N_U denotes the number of observations larger than U_{CP}, N_L is the smaller number and s_p denotes the adjusting coefficient. The fuzzy triangular probability function is illustrated in Fig. 6 and mathematically expressed as

$$F(x) = \begin{cases} (x - L_B)/(U_{CP} - L_B), & L_B \le x \le U_{CP} \\ (U_B - x)/(U_B - U_{CP}), & U_{CP} < x \le U_B \\ 0, & otherwise. \end{cases} \quad (10)$$

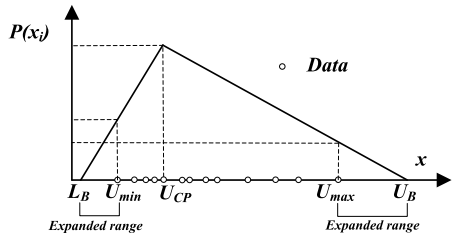

FIGURE 6. Possibility distribution of X by a probability function.

By using TMIE method to all pose perturbation data $\mathcal{D}^l_{c,n}$, the acceptable domain ranges of asymmetric perturbation for the detailed joints parameters of abnormal gaits are expanded based on their associated class and frame position. The next step is to train CGAN and obtain the generative model. CGAN is based on generative adversarial nets (GAN) [36], which is composed of two adversarial models: a generative model G and a discriminative model D. The model G concentrates on learning the data distribution, and the model D tries to distinguish a sample from the training data rather than G. CGAN tries to add the a priori knowledge feeding y_{pr} to both models G and D as an additional information. The CGAN model is

$$\min\max V(D, G) = E_{x\sim p_{data}(x)}\left[\log D\left(x|y_{pr}\right)\right] + E_{z\sim p_z(z)}[log(1 - D(G(z|y_{pr})))], \quad (11)$$

where x is training data, z is noise data that obeys the distribution $p_z(z)$, and y is auxiliary information. The abnormal gait class labels and the perturbations between the observation data and their mean are chosen as the additional information for training the CGAN model.

A multi-layer perceptron is used to construct the generative model G and the discriminative model D as shown in Fig. 7. The dimension of input x is determined by the maximum frame number L used for classification and the pose parameter vector r. L is fixed to 20 and the pose joints data are based on CMU mocap motion skeleton, i.e., 15 joints with 45 parameters are chosen for VSG. The input x dimension is 900 and the noise data z (derived from a uniform distribution) is set as 100 dimensions. The abnormal gait class labels are encoded as one-hot vectors, i.e., 10 dimensions, and denoted by O^c. Let $\Delta X = \{\Delta x^c_k | k \in [1 \dots K]\}$ define a set of corresponding perturbation vector

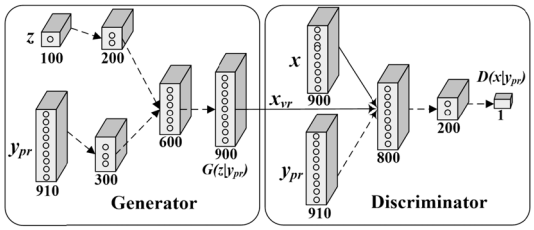

FIGURE 7. The a priori knowledge-based generative adversarial nets.

based on the sequential observation sample set X, $\Delta x^c_k = \left[I^c_1(x^c_k), \dots, I^c_l(x^c_k), \dots, I^c_L(x^c_k)\right]^T \in \mathbb{R}^{(L\cdot N_p)\times 1}$, and $I^c_l(x^c_k) = [I^c_{l,1}\left(r^l_{k,1}\right), I^c_{l,2}\left(r^l_{k,2}\right), \dots, I^c_{l,N_p}\left(r^l_{k,N_p}\right)] \in \mathbb{R}^{1\times N_p}$. $I^c_{l,n}(\cdot)$ is the indication function which divides the perturbations of the *nth* joint parameter in the *lth* frame with label c into the following levels:

$$I(x) = \begin{cases} -3, & L_B \le x < U_{min} \\ -2, & U_{min} \le x \le (U_{min} + U_{CP})/2 \\ -1, & (U_{min} + U_{CP})/2 < x < U_{CP} \\ 1, & U_{CP} < x \le (U_{max} + U_{CP})/2 \\ 2, & (U_{max} + U_{CP})/2 < x \le U_{max} \\ 3, & U_{CP} < x \le U_B \end{cases} \quad (12)$$

The conditional input is denoted by $y_{pr} = [\Delta x^c_k\ O^c]$. After training the CGAN model, the virtual sequential samples data $x_{vr} = G(z|y) = \{x^1_{vr}, x^2_{vr}, \dots, x^L_{vr}\}$ are generated by setting the noise data z and the a priori knowledge data y_{pr}. The new virtual abnormal gait are then generated using the parametric model $Y = F_{de}\left(x^l_{vr}, \beta_k\right) = P\left(r^l_{vr}\right) \cdot S\left(\beta_{fix}\right) \cdot X_{std}$, where $l \in [1\ L]$ and β_k denotes the body shape parameters of *kth* sequential observation sample.

Fig. 8 illustrates some virtual abnormal gait samples of pose perturbation. Each pose-perturbation level according to Eq. 12 has its corresponding meaning, i.e., ±3-medium perturbation, ±2-small perturbation, and ±1-very small perturbation. The pose pose-perturbations level is specially for VSG samples.

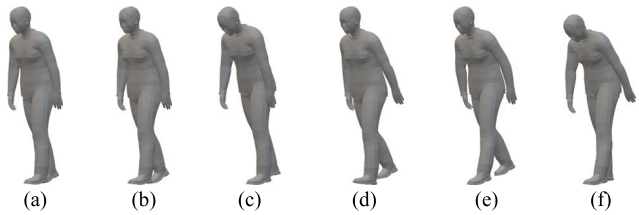

FIGURE 8. Synthetized pose-perturbation virtual samples: (a) original 3D body model; and (b)-(f) pose-perturbation virtual samples generated from (a).

C. GENERATING VARIOUS BODY-SHAPE VIRTUAL SAMPLES WITH MULTI-VIEWS

The human body shape varies, i.e., fat or slim, tall or short. However, in real-world applications, it is impractical to collect all the 3D abnormal gait models with various shapes,

especially when facing the SSS problem. 3D voxel body model with unstructured point cloud data is limited in body shape deformation. But the parametric 3D body model can be morphed to synthesize different shapes, e.g., from adult body to infant body. This advantage is exploited in 3D VSG with a priori knowledge. To synthesize various body shape samples, the observation pose parameter vector r_{opt} of 3D models that are estimated from point cloud data must be given. By varying the shape parameters, different body shape samples are generated. Let the pose vector be r_{opt}, and the model with standard shape parameters be given by $\hat{Y} = P\left(r_{opt}\right) \cdot S\left(\beta_{std}\right) \cdot X_{std}$. The multi-dimensional Gaussian normal distribution function is introduced to generate various body shape parameters, i.e.,

$$N\left(\bar{\beta}|\beta_{std},\sigma\right) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\sigma|^{1/2}} exp \times\left[-\frac{1}{2}\left(\bar{\beta}-\beta_{std}\right)^{T}\sigma^{-1}\left(\bar{\beta}-\beta_{std}\right)\right], \quad (13)$$

where σ denotes covariance, and β_{std} is a D-dimensional standard shape parameter vector.$\Phi = \{\beta_1, \beta_2, \ldots, \beta_T\}$ defines the generated virtual data with different shape parameter vectors that obey the Gaussian normal distribution. The generated 3D gait samples are represented by $\Omega_s = \left\{\hat{Y}\left(r_{opt}, \bar{\beta}\right)\right\}$. Define the multi-views set as $\Theta = \{\alpha_0, \alpha_1, \ldots, \alpha_N\}$, the pose set as $\emptyset = \left\{r_{opt_0}, r_{opt_1}, \ldots, r_{opt\,M}\right\}$ and their corresponding projected depth image from point cloud data denoted as

$$\Psi = \left\{\Upsilon_\alpha\left(r, \bar{\beta}\right)|\alpha \in \Theta, r \in \emptyset, \bar{\beta} \in \Phi\right\}. \quad (14)$$

Fig. 9 illustrates virtual abnormal samples of a body shape from a point cloud frame with multi-views.

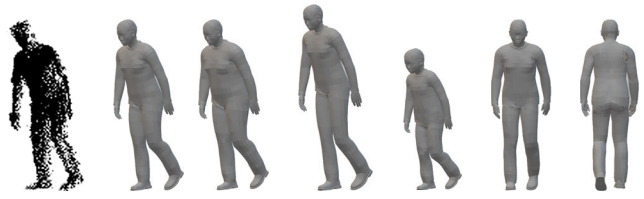

FIGURE 9. Synthetized virtual gait data: (a) point cloud data; (b) estimated parameterized body model; (c)-(e) are respectively the virtual shape transformation of weight, height and age; (f) front view projection of (b); and (g) back view projection of (b).

V. SPATIAL-TEMPORAL DEEP NETWORK BASED ON MULTI-SET CANONICAL CORRELATION ANALYSIS

The structured point cloud depth images (SPCDIs) $\Upsilon_\alpha\left(r_{opt}, \beta_{opt}\right)$ from morphed 3D parametric gait models at view α are used to extract the spatial-temporal features of abnormal gait. Unlike face and fingerprint, gait has periodic and dynamic properties that are spatial-temporal features. Thus, the combination of CNN based Resnet-50 network and RNN based LSTM [37] is introduced to extract abnormal gait features from gait sequence data.

Every gait sequence data is segmented into fixed L frames with abnormal gait class labels and additional flags including viewing angles $v_i \in [0360]$, and two typical normalized shape parameters s_i including body height and weight. Fig. 10 illustrates the exploited uniform MCCA-DNet, which comprises two phases, i.e., spatial-temporal abnormal gait feature extraction and invariant pattern feature projection based on MCCA. Let $X = \{x_i, i = 1, \ldots, I\}$ where $x_i = \left[x_i^1, x_i^2\right]$ be the input of MCCA-DNet, which consists of two sub inputs. One is $x_i^1 = \left[\Upsilon_1^\alpha, \Upsilon_2^\alpha, \ldots, \Upsilon_L^\alpha\right]$ which defines the L consecutive frames of SPCDIs. $\Upsilon_L^\alpha \in \mathbb{R}^{m\times n}$ is the Lth $m \times n$ resolution SPCDI of the corresponding parametric gait model derived from a 3D video at view α. The other is the conditional parameter by concatenating viewing angle, body shape data, i.e., $x_i^2 = \left[v_i^T s_i^T\right]^T \in \mathbb{R}^{N_v+N_s}$.

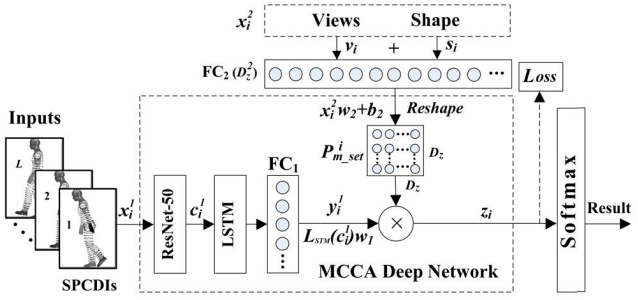

FIGURE 10. Uniform LSTM Deep Network based on MCCA.

The spatial-temporal abnormal gait feature extraction scheme consists of CNN layers, a temporal LSTM layer and a fully connected layer (FC_1). The ResNet-50 network without output layer is introduced for feature extraction in CNN layers. Its L frames output $c_i^1 = F_{resnet}\left(x_i^1\right) = [c_1^1 c_2^1 \ldots c_L^1] \in \mathbb{R}^{D_c\times L}$ are fed into the LSTM network for temporal feature extraction where D_c is the ResNet-50 output dimension. The FC_1 maps the LSTM output to a given dimension D_z using the projecting matrix $\omega_z \in \mathbb{R}^{D_L\times D_z}$, where D_L is the output dimension of LSTM network, i.e., $y_i = \left[L_{stm}(c_1^1, c_2^1 \ldots, c_L^1)^T \cdot \omega_z\right]^T \in \mathbb{R}^{D_z}$. $L_{stm}(\cdot)$ denotes the spatial-temporal information encoding using LSTM network. After the spatial-temporal feature extraction process to all samples in X based on CNN and RNN, the output feature set $Y = [y_1 \ldots y_i \ldots y_I] \in \mathbb{R}^{D_z\times I}$ is further projected by a multi-set canonical matrix $P_{m\_set}^i$ based on different conditions. The new feature of the ith sample in X after projecting to the new pattern space is

$$\begin{aligned} z_i{}^T &= y_i^T P_{m_{set}}^i \\ &= y_i{}^T \cdot Reshape\langle Relu\left[(x_i)^{2T}\omega_2 + b_2\right]\rangle, \end{aligned} \quad (15)$$

where $w_2 \in \mathbb{R}^{(N_v+N_s)\times D_z}$, $P_{m\_set}^i = [p_i^1 \ldots p_i^i \ldots p_i^{D_z}] \in \mathbb{R}^{D_z\times D_z}$, and b_2 is the bias. $Reshape\langle\cdot\rangle$ transforms the vector with D_z^2 dimension to a matrix with $D_z \times D_z$ dimension. The new pattern space is constructed and described by a new feature set $Z = \left\{z_i \in \mathbb{R}^{D_z}\right\}$ where $i \in [1\ I]$. The idea of

MCCA-DNet as given by Eq. (15) is similar to the attention model (AM) which is widely used in machine translation and speech recognition. However, AM uses the context information by mapping a query and a set of key-value pairs to an output. Our MCCA-DNet uses the pattern projecting matrix $P^i_{m\_set}$ to achieve feature transformation based the final LSTM output rather than based on each frame. After feature transformation based on their correlations, the subjects have different conditions of gait behaviour, i.e., view changes and typical body shape variation, in the new unified space. These enable to determine the optimal parameters in MCCA-DNet that maximize the correlations of the similar abnormal gait behaviours but under different conditions. The mechanism pays more attentions to the information of interest, and ignores the disturbed data, i.e., views and body shape, in the high dimensional space.

In order to describe the correlations among gait behaviours under different conditions, we classify the samples according to variations, i.e., different viewing angles, body height and weight. Let $X_n \in X, n = [1\ N]$ be the subset of X in which all samples belong to the same conditions of gait behaviour, i.e., the same view or similar body shape, and N is the total number of conditions. $Y_n \in Y$ are the corresponding sets after feature extraction. Let $p_j, p_k \in \mathbb{R}^{D_z}$, and the correlation coefficient of two sets X_j and X_k is determined by their spatial-temporal features Y_j and Y_k, i.e.,

$$\rho_{j,k} = \frac{p_j^T Y_j Y_k^T p_k}{\sqrt{(p_j^T Y_j Y_j^T p_j)(p_k^T Y_k Y_k^T p_k)}}. \tag{16}$$

The criterion used by our MCCA-DNet is described by the following optimization problem:

$$\begin{aligned} &\max_{P_{m_{set}}} \sum_{j=1}^{N} \sum_{k \in U_+^j} \rho_{j,k}, \\ &s.t. \quad p_j^T Y_j Y_j^T p_j = p_k^T Y_k Y_k^T p_k = 1. \end{aligned} \tag{17}$$

U_+^j defines the indexes of all positive sets of X_j which means they have the similar abnormal gait or action samples but under different walking conditions, i.e., different views or body shape features. It can be reformulated as the following minimization problem:

$$\begin{aligned} &\arg\min_{P_{m_{set}}} \sum_{j=1}^{N} \sum_{k \in U_+^j} \left\| Y_j^T P^j_{m_{set}} - Y_k^T P^k_{m_{set}} \right\|_2^2, \\ &s.t. \quad p_j^T Y_j Y_j^T p_j = p_k^T Y_k Y_k^T p_k = 1, \end{aligned} \tag{18}$$

where P_{m_set} is a tensor with size of $D_z \times D_z \times N$. According to Eq. (15), $P^j_{m\_set}$ is indirectly controlled by the conditional vector x_i^2 through FC_2 network. The calculation of P_{m_set} is transformed to learning the parameters of FC_2 network. The loss function of our MCCA-DNet is set as

$$\mathcal{L}_{oss} = \sum_{j=1}^{N} \sum_{k \in U_+^j} \sum_{i=1}^{M} \left\| y_j^i P^j_{m\_set} - y_k^i P^k_{m\_set} \right\|_2^2, \tag{19}$$

where M is the sample number of sets Y_j and Y_k. Compared with the traditional numerical method of solving the problem, the deep learning is achieved by minimizing the loss function and using GPU to accelerate the training of MCCA-DNet. The MCCA-DNet transforms the multi-sets data under different conditions into a uniform pattern space. Thus, it is trained prior to the SoftMax classifier. Its training data can be the same gallery set for the SoftMax classifier. The uniformed pattern features z_i are then used to train the SoftMax classifier with class labels in the gallery set, and test the probe data after training.

VI. EXPERIMENT

In order to evaluate our proposed 3D-AGRBMCCA, 3D abnormal gait datasets or 3D gait related action databases are required. Since the whole silhouettes are required for 3D abnormal gait modelling, 3D CSU abnormal gait dataset [38] and the 3D walking gait dataset [39] are appropriate for evaluations. The MSR-Action 3D Dataset [40] and UTD multimodal human action dataset (UTD-MHAD) [41] are also chosen due to the lack of related 3D abnormal gait datasets. In these datasets normal walk is defined as normal gait and the others, e.g., swaying, limping, kicking, bending, punching, hammering and falling, are defined as abnormal gait that are distinguished from the normal ones.

A. EXPERIMENTS ON CSU ABNORMAL DATASET

The CSU abnormal dataset (see Fig. 11) is in the form of 3D point cloud data, created using Microsoft Kinect and consists of 10 subjects. The actions include six states: normal walk, sitting down, falling forward, falling backward, limping of left foot and limping of right foot. It also includes three viewing angles: 0°, 45° and 90°. The original video is RGB and depth image with a resolution of 640 x 480. The abnormal 3D point cloud data of body are generated from depth images, and normalized after denoising and background subtraction.

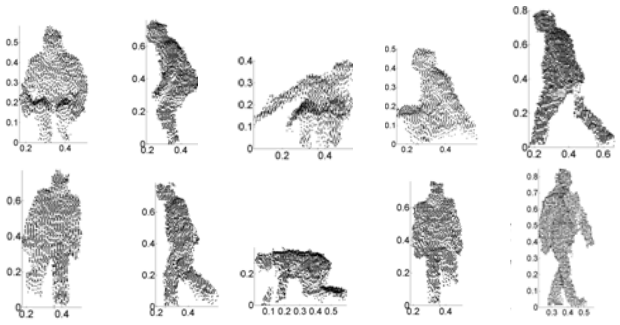

FIGURE 11. Samples of 3D abnormal gait point cloud data at different views from CSU abnormal dataset [38].

Our proposed method first extracts the abnormal gait features of real samples and generated virtual samples, and assigns them with different abnormal gait labels. The MCCA-DNet and SoftMax classifier are then trained based on the enlarged sample set as follows. First, the real abnormal gait samples are divided into two groups according to different subjects for cross subject test. The first group is

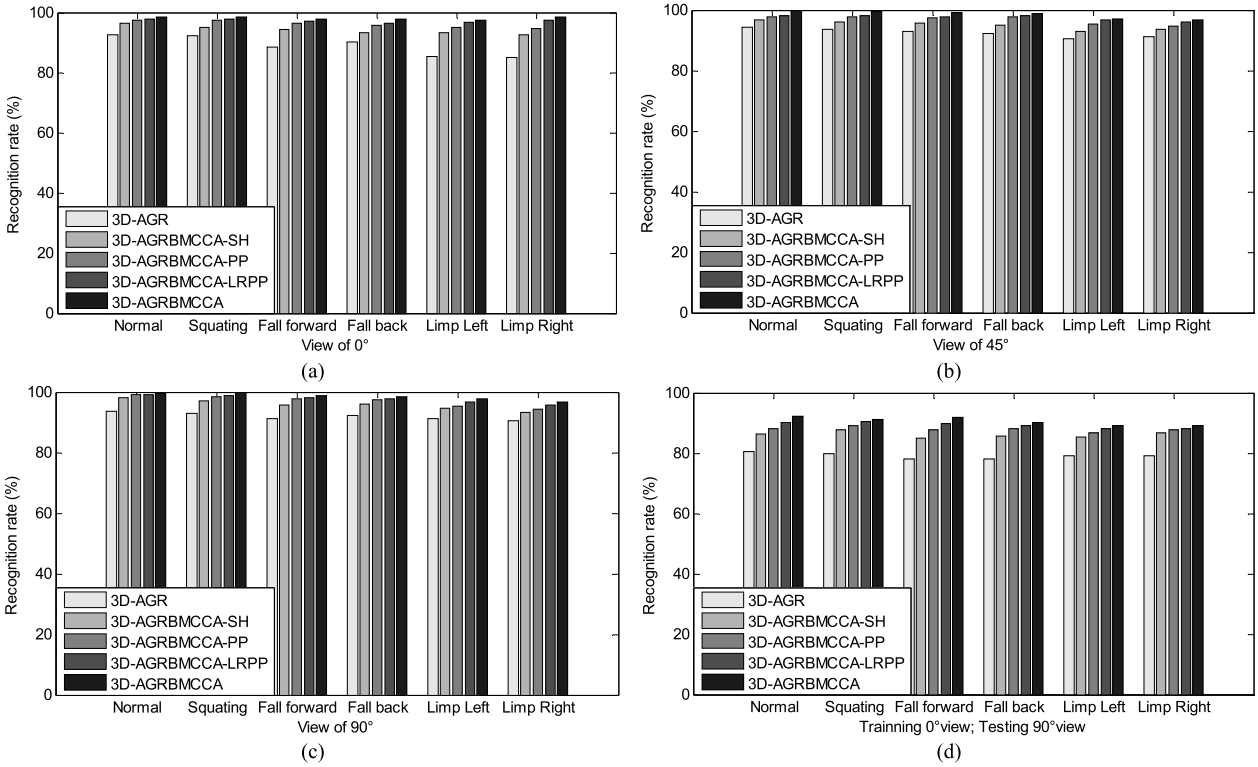

FIGURE 12. Abnormal gait recognition results at different views: (a) 0° view; (b) 45° view (c) 90° view; and (d) training at 0° and testing 90° view.

composed of samples from the first two subjects for constructing the training/gallery set. Six styles of each subject's gait at three different views are collected. The remaining probe samples belonging to the other eight subjects are assigned to the second group. The number of training samples is much smaller than the test samples used to evaluate the efficiency of our proposed 3D-AGRBMCCA. Second, all gait point cloud sequences are subsampled and segmented into subsets with $L = 20$ frames, which can have overlapping frames.

The 3D gait models corresponding to each frame of point cloud are then estimated by using the parameterized 3D gait estimation method based on point cloud contour. The training dataset is enlarged using the estimated 3D abnormal gait models to generate virtual samples. Using the VSG methods in Section IV, virtual samples with 60 styles of small and medium level pose-perturbation virtual data, 60 styles of virtual shape parameters, and left-to-right (right-to-left) symmetrical virtual samples are generated. These virtual samples are then projected onto different views $\Theta = \{0°, 45°, 90°\}$ with real samples. Before VSG, the basic number of real sequence samples for training is 36. The training sequence samples are then enlarged to about 21672 by three types of VSG.

The data in the first group and their virtual samples are used to train the MCCA-DNet and SoftMax classifier with their own loss functions. MCCA-DNet aims to extract the correlation among different sets, i.e., different views and shape features. The projected depth images at different views $\alpha \in \Theta$ are used as input to MCCA-DNet. The viewing angles, normalized body height and weight values are fed to MCCA-DNet as a priori knowledge to generate the correlation projecting matrix under different conditions for multi-sets data.

Fig. 12 compares the results of different methods, i.e., 3D-AGR, 3D-AGRBMCCA-SH, 3D-AGRBMCCA-PP, 3D-AGRBMCCA-LRPP, 3D-AGRBMCCA, at different views. 3D-AGR is our abnormal gait recognition method without VSG, 3D-AGRBMCCA-SH denotes the method which uses only various body-shape virtual samples, 3D-AGRBMCCA-PP denotes the method which uses only pose-perturbation virtual data, and 3D-AGRBMCCA-LRPP denotes the method which uses both left-to-right (right-to-left) symmetrical virtual samples and pose-perturbation virtual data. 3D-AGRBMCCA uses all the virtual samples. The figure shows that the robustness and generalization ability of the abnormal gait recognition model have been improved after training with virtual samples. The abnormal gait recognition results of left limping and right limping are slightly impaired at viewing angles 45° and 90° due to self-occlusion. The mean recognition rate at 0° is slightly lower than the other two views. This is mainly because the gait contour features are not so obvious at 0° and only slightly affect the 3D gait model estimation accuracy. As a result, the recognition results are only affected to a certain extent. Fig. 12(d) shows the recognition result for a classifier trained using 0° samples

and their related virtual samples. In testing, 90° samples are used to evaluate the robustness of our 3D gait recognition method when faced with large view changes. Both shape and pose-perturbation virtual samples greatly improve the recognition rate and the symmetrical virtual samples work well when the symmetrical abnormal gait data existed, i.e., limp left and limp right. To summarize, the recognition with VSG are less degraded than those without.

Table 2 compares the performance using different action features, i.e., GEI [42], D-DMHI [43], HP-DMM-CNN [44], and MVSM-CGCI [45], with our 3D-AGRBMCCA. GEI is gait energy image, D-DMHI is depth difference motion history image, HP-DMM-CNN is a descriptor based on hierarchical pyramid DMM deep convolutional neural network, and MVSM-CGCI is colour gait curvature image (CGCI) based on multi-view point cloud registration and synthesis. GEI, D-DMHI and HP-DMM-CNN are silhouette-based methods, and MVSM-CGCI is a 3D-based approach. 3D-AGRB-CNNLSTM is based on our 3D features, and is also introduced to evaluate the power of our MCCA-DNet with the same virtual sample setting as 3D-AGRBMCCA. 3D-AGRB-CNNLSTM takes the same CNN (ResNet-50) and LSTM structure as illustrated in Fig. 10 but directly uses the LSTM output y_i for classification without feature transformation by the MCCA mechanism. Due to recognition in varied views as in Table 2, only the training samples at 0° and testing 90° views were used. The results show that 3D-AGRBMCCA outperforms the other methods, and there are two reasons for this.

TABLE 2. Performance comparison using different depth descriptors.

Method	Cross-subject Accuracy (%)	View-varied Accuracy (%)	Avg.
GEI [42]	85.2	38.9	62.1
D-DMHI [43]	90.3	55.6	73.0
HP-DMM-CNN[44]	93.1	61.1	77.1
MVSM-CGCI [45]	91.7	68.9	80.3
3D-AGRB-CNNLSTM	93.6	78.0	85.8
3D-AGRBMCCA	97.2	90.6	93.9

First, 3D-AGRBMCCA is based on 3D parametric body model which overcomes view-invariant gait recognition problems easily with limited training samples. The 3D body structure is useful in gait feature extraction, enabling the feature extraction to be robust to noise and view changes. 3D-AGRBMCCA also exploits varied-view virtual samples by the proposed VSG method. MVSM-CGCI performs better than GEI and D-DMHI due to its multi-view synthesized virtual samples based on point cloud data. HP-DMM-CNN also performs better than GEI and D-DMHI due to its three feature representations DMM_f, DMM_s, and DMM_t, i.e., respectively projected front, side and top views. Since it is not a 3D-based method, i.e., using only depth maps for feature extraction, it performs worse than MVSM-CGCI and 3D-AGRB-CNNLSTM when faced with large view changes.

The second reason is due to our proposed VSG method. By using the 3D human body as a priori knowledge, virtual samples are generated to improve the robustness of 3D-AGRBMCCA. Unlike the statistical features, i.e., GEI and D-DMHI, that use compressed energy image for feature representation, our abnormal gait features are extracted by CNN based LSTM, which exploits both CNN and RNN. For examples, take the limping left and limping right that are more like normal gait due to self-occlusion. If the 3D depth information and the temporal features before and after cannot be effectively used, it will inevitably affect the classification of small distinct parts. By using MCCA our method achieved better performance against various walking conditions, including gait views, by transforming the features into a uniform pattern space.

B. EXPERIMENTS ON 3D WALKING GAIT DATASET

The 3D walking gait dataset [39] created using Microsoft Kinect 2 enables abnormal gait detection. Nine subjects performed nine walking styles on a treadmill, i.e., one normal walk and eight simulated abnormal walks. Eight abnormal walks are symmetrical, including padding the sole of shoes with a thickness of 5 cm, 10cm and 15cm under left or right foot. The rest of the abnormal gait data involved attaching a weight (4 kilograms) to the subject's left or right ankle. The dataset contains 3D point cloud data of the human body with the background subtracted. Each walking gait contains 1200 consecutive image frames with several gait cycles. We separated sequences of ten gait cycles from each video to give a total of 810 gait sequences. Fig. 13 illustrates some data from the dataset.

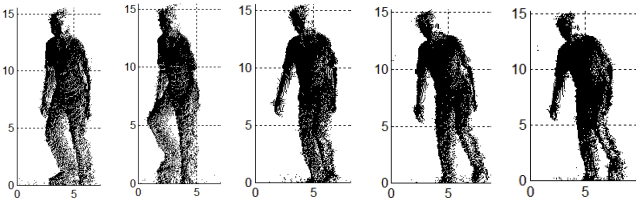

FIGURE 13. Point cloud gait data of 3D walking gait dataset at 45° from the 3D walking gait dataset [39].

In order to demonstrate the robustness of our method under SSS condition, the gait samples are split into two groups according to different subjects, i.e., three training/gallery subjects and six testing subjects. The basic number of training sequence samples before VSG is smaller than the testing sequence samples, i.e., 270 vs. 540, which is contrary to most cases in traditional methods of gait recognition. All the samples from 3D walking gait dataset are at the same view, i.e., 0°. As the gait features from front 0° are less obvious than in lateral view, all samples are rotated to the lateral 45° for training and testing in our experiments, as shown in Fig. 13. All gait point cloud sequences are segmented into subsets with $L = 20$ frames. The processes are the same as with the experiments on CSU abnormal gait dataset. After VSG,

the training sequences are extended from 270 to 7740 using the same settings in CSU dataset experiment, but with only one viewing angle.

TABLE 3. Rank-1 recognition rates (%) of nine gait styles.

Category	Style	3D-AGR	3D-AGRBMCCA	ABR. rate
Normal	NW.	96.7	100	-
Abnormal	L|5cm	85.0	95.0	98.3
	L|10cm	86.7	93.3	100
	L|15cm	93.3	98.3	100
	L|4kg	91.7	96.7	100
	R|5cm	83.3	93.3	96.7
	R|10cm	81.7	91.7	100
	R|15cm	91.7	96.7	100
	R|4kg	90.0	95.0	98.3

Table 3 shows the rank-1 recognition rate of nine gait styles. L|n cm means padding a sole with a thickness of n cm under left foot and R|4kg denotes attaching a weight (4 kilograms) to right ankle. The Abnormal recognition rate (ABR. rate) of 3D-AGRBMCCA is the accuracy that the classifier can distinguish between the abnormal gait from the normal. The classification is binary with the class being either normal or abnormal. The multi-styles 3D-AGR and 3D-AGRBMCCA recognition rates mean that the classifier must recognize the nine gait styles, i.e., normal walk, L|5cm, L|10cm, etc. Table 3 shows that the L|15cm and R|15cm styles can be easily distinguished from the other gait styles. This is mainly because the lower leg is significantly influenced by the highest sole (due to more padding). The difference among normal walk, L|5cm and L|4kg are sometimes confused which led to lower recognition rates. 3D-AGRBMCCA achieves much higher performance rate due to VSG. The average binary abnormal recognition rate is much higher than the multi-styles recognition rate. The binary abnormal recognition could be very useful in certain scenarios, i.e., high accurate detection of fall for elderly healthcare.

In order to compare our 3D-AGRBMCCA with other gait recognition methods when faced with the SSS problem, additional experiments were conducted. In the 3D walking gait dataset, eight abnormal walking styles are symmetrical and such property is very useful in evaluating the effects of our symmetrical VSG. Thus, all the abnormal gait samples are divided into two parts, i.e., left-leg-related and right-leg-related abnormal gait. The left ones are used for training and the right ones for testing. Before the symmetrical sample generation, the ELM symmetric body data prediction model is first trained by only one subject with all the symmetrical nine styles data. Following that, only symmetrical virtual samples are included and labelled for training the classifiers in 3D-AGRBMCCA. However, other methods without VSG will obviously have difficulties in classifying the right leg abnormal gait samples. For a fair comparison, the former subject with all the nine symmetrical styles data that are used for ELM model learning is included in training the classifiers.

Fig. 14 shows the recognition rates for different methods, i.e., GEI [42], D-DMHI [43], HP-DMM-CNN [44], and MVSM-CGCI [45]. It is observed that with only one right-leg-related abnormal gait training subject, i.e., all other training subjects are left-leg-related, the silhouette-based method performs unsatisfactorily. This is because the 2D binary or

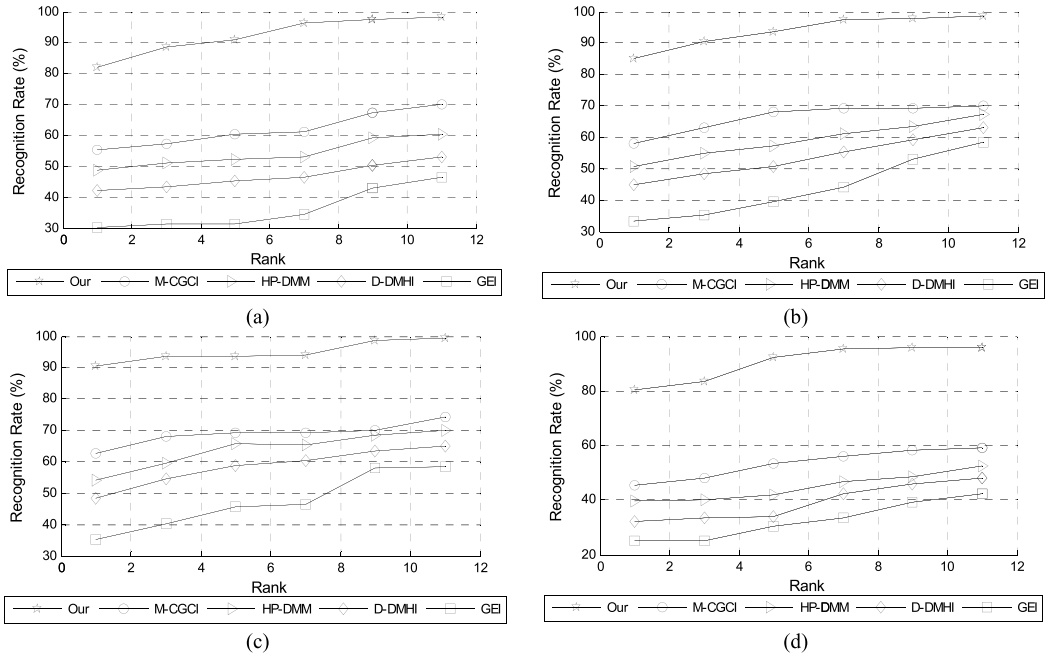

FIGURE 14. Recognition rates of different methods: (a) R|5cm (b) R|10cm (c) R|15cm and (d) R|4kg.

depth silhouettes between the left and right symmetrical abnormal gait look much the same. The left and right legs (or hands) are difficult to distinguish in 2D binary images if they are in symmetrical gait style as illustrated in Fig. 15.

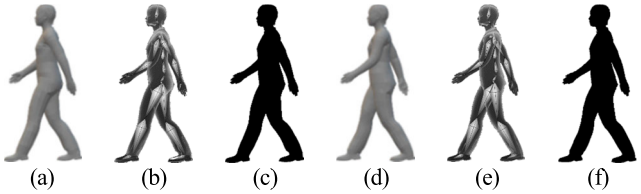

FIGURE 15. (a) and (d): symmetrical gait image; (b) and (e): thei skeleton images; and (b) and (d): their binary images.

C. EXPERIMENTS ON 3D WALKING GAIT DATASET

Fig. 16 shows the MSR-Action 3D dataset in the form of 3D point cloud data. The original data are sequences of depth maps. They are translated to 3D world coordinate according to Eq. (1). The dataset was created using Microsoft Kinect sensor and consists of twenty actions performed by ten subjects.

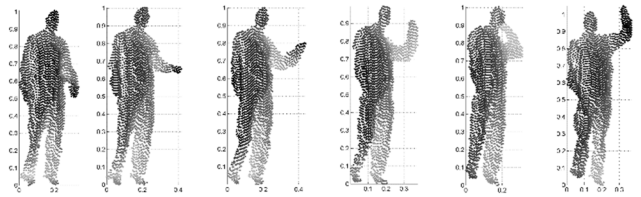

FIGURE 16. Examples of the sequences of point cloud action data.

The actions are high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two-hand wave, side-boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, and pickup & throw [40]. Each subject is captured two or three times, and the total action samples is about 4020. The dataset is collected at the front view (0°) and all samples in Fig. 16 are rotated to the lateral 15° to illustrate more information of the actions.

Experiments were conducted following the protocol of [40], and the dataset was grouped into three subsets, i.e., AS1, AS2 and AS3. Each subset includes eight actions. Actions in AS1 and AS2 comprise similar movements, while actions in AS3 are more complex that involve more joints. The recognition experiments were conducted on each subset separately, and the cross-subject test was conducted. In cross-subject test, half of the subjects were chosen for training, i.e., 1,3,5,7 and 9 subjects (if exist), and the other half were used for testing. But before training, the basic training samples are enlarged with our VSG process using the same setting in CSU dataset experiment, i.e., 60 styles of pose-perturbation virtual data, 60 styles of virtual shape parameters, and left-to-right (right-to-left) symmetrical virtual samples. Table 4 reports the comparations of our 3D-AGRBMCCA with other methods that used 3D joints features or depth map features for recognition. Table 4 clearly shows that AS2 gives the lowest recognition results compared with AS1 and AS3. This is mainly because the actions are much more similar than the other two groups. Also, our AGRBVSG method performs the best.

TABLE 4. Comparison: Rank-1 recognition rate (%) with the state-of-the-art methods on MSR-Action3D dataset.

Method	AS1	AS2	AS3	Average
Bag of 3D Points [40]	72.9	71.9	79.2	74.7
Eigenjoints [46]	74.5	76.1	96.4	83.3
LARP [47]	95.29	83.87	98.22	92.46
3RB-tLDS [48]	96.81	89.14	98.83	94.85
Sparse Pose-based [49]	91.23	90.09	99.5	93.61
MM Information [50]	92.0	85.0	93.0	90.0
ST Patterns [51]	91.70	72.2	98.6	87.5
Depth Motion Maps [52]	96.2	83.2	92.0	90.47
Key Poses Model [53]	91.53	90.23	97.06	92.94
Our 3D-AGRBMCCA	98.11	97.34	99.1	98.18

There are several reasons for this. First, most of the skeleton-based methods, i.e., [47]–[49], [53], perform much better than the silhouette based or contour-feature-points based methods on AS1 and AS2. One reason is due to the unstructured point cloud data corrupted with much noise. This makes the body contour incomplete and influences the efficiency of the feature extraction. Although the body silhouettes usually contain rich shape information, it has redundancy due to the unstructured and noise point cloud data. The silhouettes are also very varied due to the body shape, i.e., shape differences of adults and kids. This makes the silhouettes-based method less robust, especially when there are insufficient training samples. Skeleton-based methods can also be influenced by these problems. Tracking and estimating body joints from depth images are also an unsolved problem. They can be easily influenced by occlusions and noise that commonly exist in depth maps [45]. They are more likely to be influenced if the key joints are not extracted especially in some simple actions, i.e., hand wave, throw, side and forward kick.

In summary, it is inefficient to extract abnormal gait features directly from the data, especially from insufficient training samples. The recognition model is easily influenced by noise and missing data. Small parts of the body data captured by the Kinect camera that absorb the projecting light, and thus reduces the corresponding reflection, can sometimes be lost, e.g., black hair and black shoes. Our AGRBVSG uses the parametric 3D body model to fit point cloud data of body and estimates its 3D model. By transforming the unstructured point cloud data to structured, our method overcomes the self-occlusion problem to a certain extent.

In the cross-subject test there are large intra-class variations due to variations of the same action performed by different subjects. Taking the pickup and throw action as an example, some subjects use two hands while others use only one hand. The body shapes of subjects are also different.

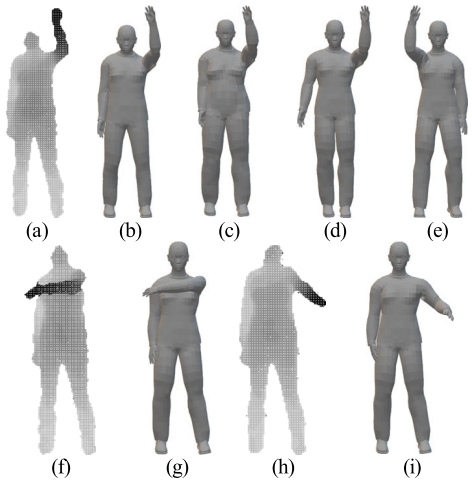

FIGURE 17. (a), (f) and (h) are three key depth maps from draw X sequence; (b), (g) and (i) are their estimated 3D parametric model; and (c), (d), (e) are respectively the shape varied, pose perturbed and symmetrical VSG models of (b).

These issues cannot be addressed if there are insufficient samples for training. However, our body shape and pose-perturbation VSG methods greatly help eliminate the variations under cross-subject testing conditions. In addition, only the right arm or leg is involved if the action is performed by a single arm or leg. However, our VSG methods synthesize symmetrical poses to cope with SSS problem which most of the other methods cannot achieve. Fig. 17 illustrates some 3D parametric models morphed to fit the noisy depth maps in MSR-Action 3D dataset with three types of VSG samples. The three key depth maps in Fig. 17 are derived from draw X action of the 6*th* subject.

Fig. 18 shows the confusion matrix of our 3D-AGRBMCCA for cross-subject test with different subsets. For most actions, the rank-1 recognition rate is 100%. Due to the 3D parametric model and VSG, our recognition method is robust in distinguishing similar actions in cross-subject test.

D. EXPERIMENTS ON UTD MULTIMODAL HUMAN ACTION DATASET

The UTD-MHAD [41] is created with a Kinect camera. The depth map is 16-bit 320×240 in size and every RGB image has a resolution of 640×480 pixels. The dataset is composed of 27 actions performed by 8 people, i.e., 4 males and 4 females. Each action is repeated 4 times with RGB images, depth maps and inertial data.

The actions are walk, wave, slap, jog, sit, swing, catch, push, throw, wipe, squat and some sports activities, i.e., boxing, bowling, tennis serve and basketball shoot. The dataset has 861 data sequences in total. In our experiments, cross-subject test was conducted, i.e., depth sequences of odd subjects were used for training after VSG using the same setting in the CSU dataset experiment, and the rest were for testing. The VSG are similar with the experiments on CSU abnormal gait dataset, but just the front view data is essential.

Table 5 shows the comparison of state-of-art on UTD-MHAD using depth maps. It shows that our proposed method performs well due to the 3D VSG process and MMC-DNet. The papers [41], [54] also discuss the fusion of RGB images to assess their performance. In our experiments and comparisons, despite only using depth maps for modelling and feature extraction, the performances of our method are better than most of the fusion schemes.

TABLE 5. Rank-1 recognition rates of seventeen action styles using depth map.

Method	Accuracy (%)
DMMs/ CRC [41]	66.1
DMMs& Inertial/ CRC [41]	79.1
HP-DMM-CNN/SVM [44]	82.8
3DHOT/MBC [54]	84.4
BoA/SVM [44]	85.4
JDMs/CSF [55]	88.1
SAC/SVM [56]	91.7
OUR 3D-AGRBMCCA	93.2

UTD-MHAD is challenging due to the similarity of many actions based only on arm motion. The confusion matrix of our 3D-AGRBMCCA method on UTD-MHAD is shown in Fig. 19 with above 90% accuracy on 19/27 actions with robust performance for most actions. In [56], it is shown that drawing circle clockwise and counterclockwise, which act in opposite direction, are usually confused. This is mainly

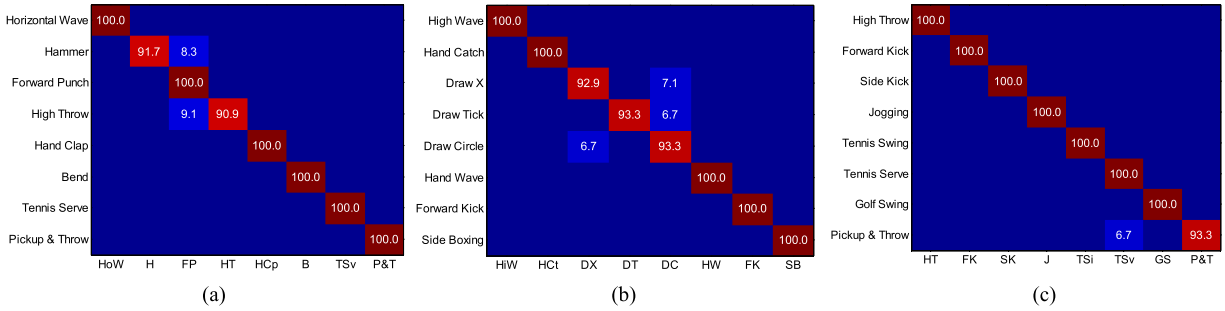

FIGURE 18. Confusion matrix of 3D-AGRBMCCA for cross-subject test: (a) AS1; (b) AS2; and (c) AS3.

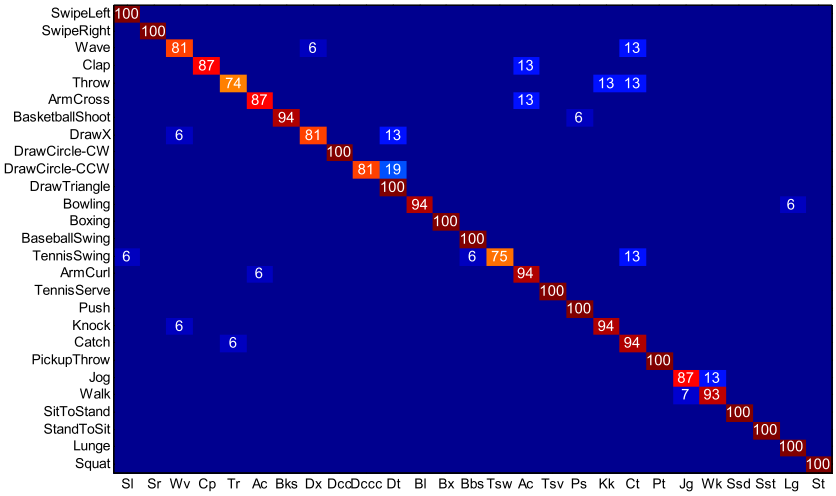

FIGURE 19. Confusion matrix of 3D-AGRBMCCA on UTD-MHAD dataset.

because the motion energy-based methods, i.e. Depth Motion Maps (DMMs), are poor in spatial-temporal action feature extraction. By compressing the sequence depth maps into a motion energy map in the video, the temporal feature is obliterated. Some skeleton-base methods also perform poorly when the joints extraction results are poor. Our 3D-AGRBMCCA is robust to noise by fully utilising a priori knowledge of body structure. The use of CNN based LSTM network which combines CNN and Recurrent Neural Network (RNN) also enables our method to perform well in spatial-temporal feature extraction.

VII. DISCUSSION AND CONCLUSION

A. DISCUSSION

The comparative results on the CSU 3D abnormal gait dataset, 3D walking gait dataset, MSR-Action 3D dataset and UTD-MHAD clearly show that our method with VSG best addresses the SSS issues. They also clearly show that 2D methods have difficulties in dealing with 3D VSG due to the lack of 3D body structure information.

Our 3D parametric body model is greatly aided with learned a priori knowledge for generating symmetrical, various body-shape, and pose-perturbation virtual samples. This makes our method robust and effective for classification tasks with small sample sets. Compared with the traditional VSG forecast tasks, the VSG for classification has its own characteristics that the synthesized samples may not influence the current subject recognition result, i.e., bad virtual samples with negative impact are excluded. In this paper CGAN is used to achieve better a priori knowledge-based generative model which generates data according to both classification label and the given perturbation. The CGAN model greatly helps in generating the good virtual samples based on the two-player min-max game with value function. To combine the feature extraction, multi-sets correlations analysis and classification in a uniform model for better recognition performance, a novel MCCA based deep learning network is proposed in our paper.

The results on the CSU dataset show that the rank-1 mean detection and recognition rate of abnormal gait is about 22% higher than GEI, D-DMHI and HP-DMM-CNN when faced with big view changes. The efficiency of symmetrical VSG has been clearly shown in the experiment on the 3D walking gait dataset. The a priori knowledge of the human body is fully used for a better performance which is demonstrated on the MSR-Action 3D Dataset and UTD-MHAD. One of the key processes of our proposed 3D-AGRBMCCA is the 3D body estimation by point cloud gait data which aids the extraction of the 3D parametric gait semantic data. A fast and more accurate 3D body estimated approach should be further studied by fully exploiting the spatial-temporal information of the human gait.

B. CONCLUSION

Based on the 3D parametric gait model, a VSG approach is proposed to address the problem of SSS in abnormal gait recognition. The abnormal gait point cloud data are abstracted to high-order semantic description, i.e., shape and pose, using the proposed 3D gait model estimation method. The pose and shape parameters are then fully used as a priori knowledge to generate virtual samples using three different VSG methods. Using MCCA-DNet, the spatial-temporal features of abnormal gait behaviours are extracted effectively. By exploiting VSG and MCCA, good classification and recognition of abnormal gait data at different views are achieved. Compared with the traditional 2D abnormal gait recognition methods, 3D based methods can deal with view-invariant problem and object occlusion more easily. The proposed method not only improves the recognition accuracy of abnormal gait recognition, but also provides a new idea for recognizing abnormal actions.

REFERENCES

[1] 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.

[2] J. Hariyono and K.-H. Jo, "Detection of pedestrian crossing road: A study on pedestrian pose recognition," *Neurocomputing*, vol. 234, pp. 144–153, Apr. 2017.

[3] B. Pogorelc, Z. Bosnić, and M. Gams, "Automatic recognition of gait-related health problems in the elderly using machine learning," *Multimedia Tools Appl.*, vol. 58, no. 2, pp. 333–354, May 2012.

[4] I. González, I. H. López-Nava, J. Fontecha, A. Muñoz-Meléndez, A. I. Pérez-SanPablo, and I. Quiñones-Urióstegui, "Comparison between passive vision-based system and a wearable inertial-based system for estimating temporal gait parameters related to the GAITRite electronic walkway," *J. Biomed. Informat.*, vol. 62, pp. 210–223, Aug. 2016.

[5] C. Bauckhage, J. K. Tsotsos, and F. E. Bunn, "Automatic detection of abnormal gait," *Image Vis. Comput.*, vol. 27, nos. 1–2, pp. 108–115, Jan. 2009.

[6] L. Wang, "Abnormal walking gait analysis using silhouette-masked flow histograms," in *Proc. 18th IEEE Int. Conf. Pattern Recognit.*, Hong Kong, Aug. 2006, pp. 1–4.

[7] U. C. Ugbolue, E. Papi, K. T. Kaliarntas, A. Kerr, L. Earl, V. M. Pomeroy, and P. J. Rowe, "The evaluation of an inexpensive, 2D, video based gait assessment system for clinical use," *Gait Posture*, vol. 38, no. 3, pp. 483–489, Jul. 2013.

[8] M. Nieto-Hidalgo, F. J. Ferrández-Pastor, R. J. Valdivieso-Sarabia, J. Mora-Pascual, and J. M. García-Chamizo, "A vision based proposal for classification of normal and abnormal gait using RGB camera," *J. Biomed. Informat.*, vol. 63, pp. 82–89, Oct. 2016.

[9] R. J. Saner, E. P. Washabaugh, and C. Krishnan, "Reliable sagittal plane kinematic gait assessments are feasible using low-cost webcam technology," *Gait Posture*, vol. 56, pp. 19–23, Jul. 2017.

[10] J. Ortells, M. T. Herrero-Ezquerro, and R. A. Mollineda, "Vision-based gait impairment analysis for aided diagnosis," *Med. Biol. Eng. Comput.*, vol. 56, no. 9, pp. 1553–1564, Sep. 2018.

[11] T. T. Verlekar, H. De Vroey, K. Claeys, H. Hallez, L. D. Soares, and P. L. Correia, "Estimation and validation of temporal gait features using a markerless 2D video system," *Comput. Methods Programs Biomed.*, vol. 175, pp. 45–51, Jul. 2019.

[12] X. Yang and Y. Tian, "Super normal vector for human activity recognition with depth cameras," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 1028–1039, May 2017.

[13] L. Xia and J. K. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 2834–2841.

[14] X. Xu, R. W. McGorry, L.-S. Chou, J.-H. Lin, and C.-C. Chang, "Accuracy of the Microsoft Kinect for measuring gait parameters during treadmill walking," *Gait Posture*, vol. 42, no. 2, pp. 145–151, Jul. 2015.

[15] J. Yang, X. Yu, Z.-Q. Xie, and J.-P. Zhang, "A novel virtual sample generation method based on Gaussian distribution," *Knowl.-Based Syst.*, vol. 24, no. 6, pp. 740–748, Aug. 2011.

[16] C.-J. Chang, D.-C. Li, Y.-H. Huang, and C.-C. Chen, "A novel gray forecasting model based on the box plot for small manufacturing data sets," *Appl. Math. Comput.*, vol. 265, pp. 400–408, Aug. 2015.

[17] Z.-S. Chen, B. Zhu, Y.-L. He, and L.-A. Yu, "A PSO based virtual sample generation method for small sample sets: Applications to regression datasets," *Eng. Appl. Artif. Intell.*, vol. 59, pp. 236–243, Mar. 2017.

[18] G. Guo and C. R. Dyer, "Learning from examples in the small sample case: Face expression recognition," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 35, no. 3, pp. 477–488, Jun. 2005.

[19] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition," *Pattern Recognit. Lett.*, vol. 26, no. 2, pp. 181–191, Jan. 2005.

[20] H.-M. Moon, M.-G. Kim, J.-H. Shin, and S. B. Pan, "Multiresolution face recognition through virtual faces generation using a single image for one person," *Wireless Commun. Mobile Comput.*, vol. 2018, pp. 1–8, Nov. 2018.

[21] Y.-L. He, P.-J. Wang, M.-Q. Zhang, Q.-X. Zhu, and Y. Xu, "A novel and effective nonlinear interpolation virtual sample generation method for enhancing energy prediction and analysis on small data problem: A case study of Ethylene industry," *Energy*, vol. 147, pp. 418–427, Mar. 2018.

[22] L. Sun, S. Ji, and J. Ye, "Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 194–200, Jan. 2011.

[23] S. Huang, J. Chen, and Z. Luo, "Sparse tensor CCA for color face recognition," *Neural Comput. Appl.*, vol. 24, nos. 7–8, pp. 1647–1658, Jun. 2014.

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

[25] A. A. Nielsen, "Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data," *IEEE Trans. Image Process.*, vol. 11, no. 3, pp. 293–305, Mar. 2002.

[26] J. Peng, Q. Li, A. A. Abd El-Latif, and X. Niu, "Linear discriminant multiset canonical correlations analysis (LDMCCA): An efficient approach for feature fusion of finger biometrics," *Multimedia Tools Appl.*, vol. 74, no. 13, pp. 4469–4486, Jul. 2015.

[27] F. O. Fabio, A. S. S. Anderson, A. C. F. Marcelo, B. G. Rafael, and M. G. G. Luiz, "Efficient 3D objects recognition using multifoveated point clouds," *Sensors*, vol. 18, no. 7, pp. 2302–2329, Jul. 2018.

[28] J. Smisek, M. Jancosek, and T. Pajdla, "3D with Kinect," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Barcelona, Spain, Nov. 2011, pp. 1154–1160.

[29] 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, pp. 361–377, Dec. 2016.

[30] 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.

[31] Q. Wu, G. Xu, M. Li, L. Chen, X. Zhang, and J. Xie, "Human pose estimation method based on single depth image," *IET Comput. Vis.*, vol. 12, no. 6, pp. 919–924, Sep. 2018.

[32] J. Lorenzo-Navarro, M. Castrillón-Santana, and D. Hernández-Sosa, "On the use of simple geometric descriptors provided by RGB-D sensors for re-identification," *Sensors*, vol. 13, no. 7, pp. 8222–8238, Jun. 2013.

[33] WHO Expert Consultation, "Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies," *Lancet*, vol. 363, no. 9403, pp. 157–163, Mar. 2004.

[34] G. B. Huang, Q. Y. Zhu, and C. K. Siew, "Extreme learning machine: A new learning scheme of feedforward neural networks," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Budapest, Hungary, Jul. 2004, pp. 439–501.

[35] D.-C. Li, C.-S. Wu, T.-I. Tsai, and Y.-S. Lina, "Using mega-trend-diffusion and artificial samples in small data set learning for early flexible manufacturing system scheduling knowledge," *Comput. Oper. Res.*, vol. 34, no. 4, pp. 966–982, Apr. 2007.

[36] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Proc. Int. Conf. Neural Inf. Process. Syst.* Cambridge, MA, USA: MIT Press, 2014, pp. 1–9.

[37] F. Karim, S. Majumdar, H. Darabi, and S. Chen, "LSTM fully convolutional networks for time series classification," *IEEE Access*, vol. 6, pp. 1662–1669, 2018.

[38] J. Luo, J. Tang, P. Zhao, F. Mao, and P. Wang, "Abnormal behavior detection for elderly based on 3D structure light sensor," (in Chinese), *Opt. Technique*, vol. 42, no. 2, pp. 145–151, Mar. 2016.

[39] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3D points," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.-Workshops*, San Francisco, CA, USA, Jun. 2010, pp. 9–14.

[40] T.-N. Nguyen, H.-H. Huynh, and J. Meunier, "3D reconstruction with time-of-flight depth camera and multiple mirrors," *IEEE Access*, vol. 6, pp. 38106–38114, 2018.

[41] C. Chen, R. Jafari, and N. Kehtarnavaz, "UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *Proc. IEEE Int. Conf. Image Process.*, Quebec City, QC, Canada, Sep. 2015, pp. 168–172.

[42] 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.

[43] Z. Gao, H. Zhang, G. P. Xu, and Y. B. Xue, "Multi-perspective and multi-modality joint representation and recognition model for 3D action recognition," *Neurocomputing*, vol. 151, pp. 554–564, Mar. 2015.

[44] N. E. D. Elmadany, Y. He, and L. Guan, "Information fusion for human action recognition via biset/multiset globality locality preserving canonical correlation analysis," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5275–5287, Nov. 2018.

[45] J. Tang, J. Luo, T. Tjahjadi, and Y. Gao, "2.5D multi-view gait recognition based on point cloud registration," *Sensors*, vol. 14, no. 4, pp. 6124–6143, Mar. 2014.
[46] X. Yang and Y. Tian, "Effective 3D action recognition using EigenJoints," *J. Vis. Commun. Image Represent.*, vol. 25, no. 1, pp. 2–11, 2014.
[47] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a lie group," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, Jun. 2014, pp. 588–595.
[48] W. Ding, K. Liu, E. Belyaev, and F. Cheng, "Tensor-based linear dynamical systems for action recognition from 3D skeletons," *Pattern Recognit.*, vol. 77, pp. 75–86, May 2018.
[49] I. Theodorakopoulos, D. Kastaniotis, G. Economou, and S. Fotopoulos, "Pose-based human action recognition via sparse representation in dissimilarity space," *J. Vis. Commun. Image Represent.*, vol. 25, no. 1, pp. 12–23, Jan. 2014.
[50] Z. Gao, J.-M. Song, H. Zhang, A.-A. Liu, Y.-B. Xue, and G.-P. Xu, "Human action recognition via multi-modality information," *J. Electr. Eng. Technol.*, vol. 9, no. 2, pp. 739–748, Mar. 2014.
[51] A. W. Vieira, E. R. Nascimento, G. L. Oliveira, Z. Liu, and M. F. Campos, "On the improvement of human action recognition from depth map sequences using space-time occupancy patterns," *Pattern Recognit. Lett.*, vol. 36, pp. 221–227, Jan. 2014.
[52] C. Chen, K. Liu, and N. Kehtarnavaz, "Real-time human action recognition based on depth motion maps," *J. Real-Time Image Process.*, vol. 12, no. 1, pp. 155–163, Jun. 2016.
[53] X. Li, Y. Zhang, and J. Zhang, "Improved key poses model for skeleton-based action recognition," in *Proc. 18th Pacific-Rim Conf. Multimedia*, Harbin, China, Sep. 2017, pp. 358–367.
[54] B. Zhang, Y. Yang, C. Chen, L. Yang, J. Han, and L. Shao, "Action recognition using 3D histograms of texture and a multi-class boosting classifier," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4648–4660, Oct. 2017.
[55] C. Li, Y. Hou, P. Wang, and W. Li, "Joint distance maps based action recognition with convolutional neural networks," *IEEE Signal Process. Lett.*, vol. 24, no. 5, pp. 624–628, May 2017.
[56] S. Zeng, G. Lu, and P. Yan, "Enhancing human action recognition via structural average curves analysis," *Signal, Image Video Process.*, vol. 12, pp. 1551–1558, Nov. 2018.

JIAN LUO received the B.Sc. degree in communication engineering from Hunan Normal University, China, in 2007, the M.Sc. degree in electronic science and technology from Hunan University, China, in 2010, and the Ph.D. degree in information science and engineering from Central South University, China, in 2016. He has been a Lecturer with Hunan Normal University, China, since 2017. His research interest is gait recognition.

TARDI TJAHJADI received the B.Sc. degree in mechanical engineering from University College London, in 1980, and the M.Sc. degree in management sciences and the Ph.D. degree in total technology from UMIST, U.K., in 1981 and 1984, respectively. He has been an Associate Professor with Warwick University, since 2000, and a Reader, since 2014. His research interests include image processing and computer vision.