3D Canonical Pose Estimation and Abnormal Gait Recognition with a Single RGB-D Camera

Yao Guo, *Member*, *IEEE*, Fani Deligianni, Xiao Gu, Guang-Zhong Yang, *Fellow*, *IEEE*

***Abstract*— Assistive robots play an important role in improving the quality of life of patients at home. Among all the monitoring tasks, gait disorders are prevalent in elderly and people with neurological conditions, which increases the risk of fall. Therefore, the development of mobile systems for gait monitoring at home in normal living conditions is important. Here we present a mobile system that is able to track humans and analyze their gait in canonical coordinates based on a single RGB-D camera. Firstly, view-invariant 3D lower limb pose estimation is achieved by fusing information from depth images along with 2D joints derived in RGB images. Next, both the 6D camera pose and the 3D lower limb skeleton are real-time tracked in a canonical coordinate system based on Simultaneously Localization and Mapping (SLAM). A mask-based strategy is exploited to improve the re-localization of the SLAM in dynamic environments. Abnormal gait is detected by using the Support Vector Machine (SVM) and the Bidirectional Long-Short Term Memory (BiLSTM) network with respect to a set of extracted gait features. To evaluate the robustness of the system, we collected multi-camera, ground truth data from sixteen healthy volunteers performing six gait patterns that mimic common gait abnormalities. The experiment results demonstrate that our proposed system can achieve good lower limb pose estimation and superior recognition accuracy compared to previous abnormal gait detection methods.**

I. INTRODUCTION

Gait disorders usually result from neurological or musculoskeletal conditions and they are common in the elderly [1], [2]. In neurological diseases, such as Parkinson's, they mark the disease progression and severity. In elderly, they are associated with a high risk of falls, poor quality of life and increased risk for depression. Therefore, there is a pressing need for real-time gait analysis in patients' homes [3].

In the past decades, various gait analysis systems have been developed for pathological gait detection [4]. These include multi-camera motion capture system, multiple Inertial Measurement Units (IMU) system, force plates, and pressure insoles. Although these systems can monitor the kinematics and dynamics of the lower limbs' movements with promising precision, they require participants to wear specialized markers/sensors and involved complicated setup, which limits their application in specialized hospitals or rehabilitation centers.

Recent advances in computer vision have demonstrated good performance in offline markerless gait analysis [5]–[7]. Furthermore, real-time 2D full body pose estimation from RGB images has been achieved [8]. However, most of the pathological gait involves atypical joint kinematics and dynamics. The evaluation of the gait parameters in 2D space impedes objective clinical evaluation, since the relation to standard clinical indices is ill-defined [9]. On the other hand, depth sensors technology, such as Kinect, enables real-time 3D human skeleton tracking [10], [11]. Supported by recent advances in robot vision and artificial intelligence, it is possible to monitor a patient's health while carrying out daily activities [12]–[14].

To this end, a number of challenges should be addressed in order for robots to be aware of their surroundings, follow humans and track their gait accurately. Firstly, gait analysis based on depth images ignores the abundant texture features in the color space and sometimes results in the unsatisfactory estimation of the lower limb joints. Secondly, the lower-limb movement detection accuracy may vary with respect to the distance between the camera and the person while the human is observed from a fixed perspective. Thirdly, the 3D lower limb pose is represented in the moving camera frame, which means the joint trajectories and many significant gait parameters (e.g., the gait speed and the step length) cannot be estimated without prior knowledge of the moving camera position.

To address the aforementioned challenges, this paper presents an RGB-D based mobile 3D gait analysis system for tracking both the 6D camera and the 3D lower limb pose in a canonical coordinate system. The 3D map of the environment is pre-built and stored offline. The 3D lower limb pose is estimated in the camera frame coordinate system by fusing state-of-the-art RGB-based 2D pose estimation [8] with the depth inputs. Subsequently, 3D human pose estimation is mapped to a canonical world representation via information fusion of the 6D pose of the camera estimated via ORB-SLAM [15] and RGB-D images. Kalman Filter (KF) is used to predict the 3D joints that lose track as well as smooth the joint trajectories. Camera re-localization in the SLAM is sensitive to the high dynamic environment induced by the moving human target and a lack of salient world features [16]. Inspired by [17], we introduce a mask-based strategy to enhance the robustness of the re-localization accuracy. Fig. 1 demonstrates an overview of the proposed system for canonical pose estimation and gait analysis. For validating the robustness of the system, we compare the lower limb pose and joint angle estimation results with respect to the ground-truth data from sixteen healthy volunteers. The ground-truth data were recorded with the Vicon motion capture

This work was supported by Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/R026092/1.

Y. Guo, F. Deligianni, X. Gu, and G.-Z. Yang are with the Hamlyn Centre, South Kensington Campus, Imperial College London, London, SW7 2AZ, United Kingdom. {yao.guo, fani.deligianni, xiao.gu17, g.z.yang}@imperial.ac.uk). G.-Z. Yang is also with the Institute of Medical Robotics, Shanghai Jiao Tong University, China.

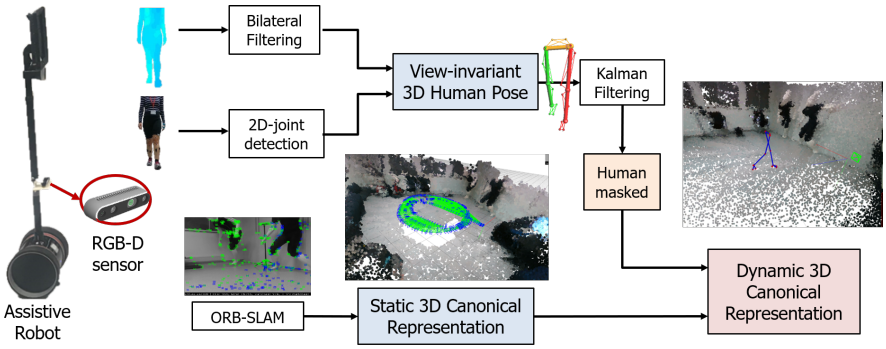


Fig. 1. Overview of the proposed mobile 3D canonical gait analysis system.

system, while subjects performed six gait patterns (normal, in-toeing, out-toeing, drop foot, supination, and pronation). Next, we investigate the potential of the proposed system to the abnormal gait recognition tasks. In the recognition stage, the Support Vector Machine (SVM) classifier and the Bidirectional Long-Short Term Memory (BiLSTM) [18] network were used to classify gait patterns.

This paper is organized as follows. In Section II, two steps of information fusion are described, namely information fusion for 3D pose estimation and subsequently information fusion for human pose representation in the canonical coordinate system. Section III explains the gait feature selection and abnormal gait recognition. Section IV describes several experiments to evaluate the accuracy and robustness of the proposed framework.

II. RGB-D BASED 3D CANONICAL GAIT ANALYSIS

To construct a home-based assistive system for gait analysis, we used a light-weight telepresence robot equipped with a single RGB-D camera and without any additional sensing feedback. To facilitate the applications on light-weight mobile platforms with limited computing resources, the captured RGBD images were live-streamed to a remote workstation based on Real-Time Streaming Protocol (RTSP). The robot exploited the 3D lower limb pose estimation to follow the human and collect data with simple control commands (forward, backward, turn left, turn right, or the combination). This control policy is also beneficial for extending the proposed system across platforms.

A. Information Fusion for 3D Lower Limb Pose Estimation

Real-time, view-invariant, human pose estimation from RGB-D images has attracted much attention in recent years [7], [11], [19], in which they used the Kinect sensor to estimate 3D skeleton both for gait analysis and action recognition. The Kinect sensor was one of the first systems that allowed real-time 3D pose estimation and it is popular to the vision research community. However, its application in kinematic gait analysis provides unsatisfactory results and it is not suitable for clinical use [20]. Moreover, it only supports pose extraction from the live-streamed data. Zimmermann *et al.* also demonstrated that human pose estimation by leveraging both rgb and depth images performs better than using depth data alone [21].

To fuse information from depth images along with the abundance of texture features in RGB images, we firstly use a Part Affinity Fields approach that utilizes deep convolutional neural networks to detect human body parts and link them to a 2D skeleton [8]. The 2D skeleton consists of several key joints, and each joint j is represented as $\mathbf{p}_j = [x_j, y_j, \epsilon]^{\mathrm{T}}$, where $[x_j, y_j]$ is the 2D pixel coordinates and ϵ indicates the prediction probability of this joint.

To acquire reliable depth value z_j of joint j, the holes in the raw depth images are first filled. Subsequently, the bilateral filter in the spatial domain and the moving average filter in the temporal domain are adopted, respectively. To extract the 3D pose of the lower limb in real-time, we directly back-project the 2D points onto the 3D space as $\mathbf{P}_j = [X_j, Y_j, Z_j]$ based on the pin-hole camera model. Accordingly, the raw 3D lower limb skeleton ${}^{\{\mathrm{C}\}}\mathcal{S} = \{\mathbf{P}_1, \cdots, \mathbf{P}_j, \cdots\}$ expressed in the camera frame $\{\mathrm{C}\}$ can be determined. Furthermore, the Kalman filter is adopted to predict the possible 3D joint position while its 2D position loses tracking, which can also smooth the joint trajectories in 3D space. Specifically, the Kalman filter is applied for each joint respectively, in which the state vector $\mathbf{x}_{KF}$ consists of the 3D position $\mathbf{P}_j$ and its velocity vector $\mathbf{v}_j$.

B. Information Fusion for Human Pose Representation in the Canonical Coordinate System

The 3D joints over time ${}^{\{\mathrm{C}\}}\mathcal{S}(t)$ are represented in the moving camera frame $\{\mathrm{C}\}$. To extract joint trajectories as well as gait parameters such as joints velocities, gait speed and step/stride length, the 3D joint trajectory $\mathcal{S}(t)$ should be mapped to a canonical coordinate system $\{\mathrm{E}\}$. This requires to simultaneously extract information of the 6D trajectory $\mathbf{X}_C(t)$ of the moving camera in relation to the surrounding environment.

Although the camera pose can be estimated by using additional sensors, such as IMU and odometer, this paper is focused on a single RGB-D camera setup scenario. Vision-based SLAM algorithm is first used to localize both the 6D camera pose $\mathbf{X}_C(t)$ and the lower limb pose $\mathcal{S}(t)$ sequences in a canonical coordinate system $\{\mathrm{E}\}$, where $\{\mathrm{E}\}$ is determined by the initial pose of the camera in the

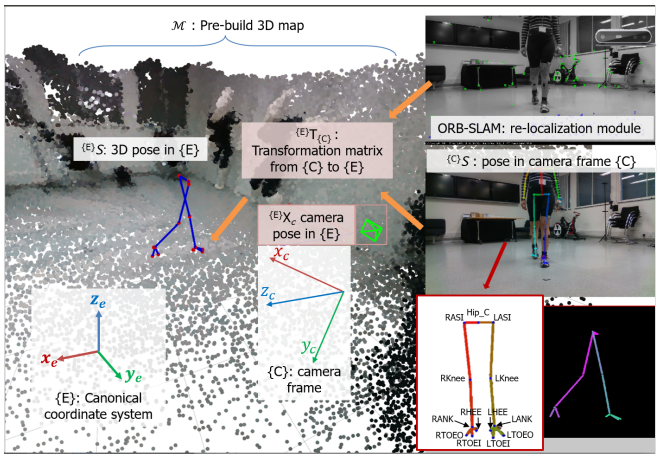


Fig. 2. Information fusion for human lower limb pose representation and 3D gait analysis in the canonical coordinate system. The 3D map $\mathcal{M}$ of the environment is pre-built and stored offline. During the mobile gait analysis, the 3D lower limb skeleton $\mathcal{S}$ is firstly estimated in the camera frame coordinate system $\{\mathrm{C}\}$. We utilize ORB-SLAM to track the real-time localization of the camera as well as the transformation matrix ${}^{\{\mathrm{E}\}}\mathbf{T}_{\{\mathrm{C}\}}$ indicating the transformation from camera space to a canonical coordinate system. Accordingly, the 3D human movement and the 6D camera motion can be represented in a canonical coordinate system $\{\mathrm{E}\}$.

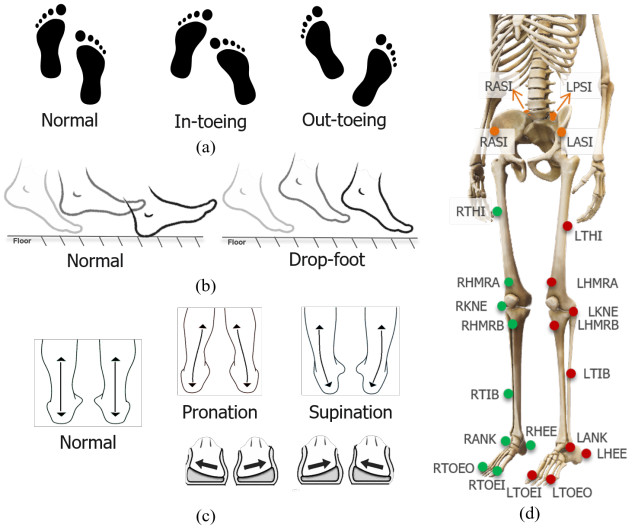


Fig. 3. Visualization of five abnormal gait patterns. a) In-toeing and out-toeing gait patterns; b) drop-foot gait; c) For the supination and pronation, participants were wearing correction insoles, which helps for simulating these two gait abnormalities in a natural manner. d) To capture the ground truth data by the Vicon motion capture system, reflective markers were attached to the human body and the camera.

mapping stage of SLAM. In this paper, the 3D map $\mathcal{M}$ of the environment is pre-built and stored offline by using the ORB-SLAM [15] with RGB-D input. It allows the robot to relocate itself in an offline world representation of its environment.

However, the recovery of camera pose in dynamic scenes is challenging [16]. In the scenario of the robot following a human, the human body typically occupies most of the image frame and it moves in a relatively fast pace, which would result in unsatisfactory re-localization in the pre-built 3D map due to the lack of salient features. To alleviate this problem and improve the camera re-localization performance, a mask-based strategy is used. A square human mask is generated in each frame according to the 2D joint estimation. The mask not only covers the whole human body but also involves the neighbor area of the human body. This is because the neighbor area is also influenced by the human movement and it is affected by motion blur and illumination vibration. In the re-localization module of the ORB-SLAM algorithm, features are only extracted from the image outside the human mask instead of the extraction from the whole image [15]. This simple method can improve the stability in matching features of the current dynamic view with the pre-built maps.

Given ${}^{\{\mathrm{E}\}}\mathbf{X}_C$ and the pre-built 3D map $\mathcal{M}$, ${}^{\{\mathrm{E}\}}\mathbf{T}_{\{\mathrm{C}\}}$ indicating the transformation from camera space to canonical coordinate system can be determined. Finally, the canonical human pose representation is derived by

$$ {}^{\{\mathrm{E}\}}\mathcal{S} = {}^{\{\mathrm{E}\}}\mathbf{T}_{\{\mathrm{C}\}}{}^{\{\mathrm{C}\}}\mathcal{S} \tag{1} $$

In this paper, the transformation matrix ${}^{\{\mathrm{C}\}}\mathbf{T}_{\{\mathrm{R}\}}$ between camera space $\{\mathrm{C}\}$ and robot space $\{\mathrm{R}\}$ is assumed as the identical matrix $\mathbf{I}$. This canonical representation allows gait analysis to be performed in a view-independent canonical space as demonstrated in Fig. 2, thus providing the feasible estimation of the 3D joint trajectories and the corresponding gait indices (heel and toe strikes, gait speed, and step length). Moreover, the spatial relationship among the human target, moving camera/robot, and the 3D structure of the surrounding environment can be captured.

III. 3D GAIT ANALYSIS FOR ABNORMAL GAIT RECOGNITION

In this section, we investigate the potential of the proposed system for abnormal gait recognition. To this end, we first collected a gait database from sixteen subjects using a Vicon motion capture system and our proposed system. Next, the joint angle features were extracted from 3D lower limb skeleton. Finally, two recognition approaches were given by using the joint angles as input for abnormal gait recognition.

A. Ground Truth Data Acquisition

Sixteen healthy volunteers (mean age = 27.5, fourteen males and two females) were recruited for this study. Each volunteer was asked to walk with normal style and imitate the other five abnormal gait patterns (in-toeing, out-toeing, drop foot, supination, and pronation) as shown in Fig. 3. This allows us to examine the accuracy of our proposed method under conditions that resemble abnormal gait patterns in a range of gait abnormalities. It has been proven that foot pronation/supination angles and inward/outward rotation angles are the critical indices for reducing mechanical stress and avoid sports injuries and osteoarthritis [22]. In-toeing and out-toeing gait indicate the foot forward direction point inward and outward instead of straight ahead during walking. Pronation refers to the inward rotation of the ankle joint, and supination indicates the outward roll. Especially, the correction insoles were provided to the participants to naturally

simulate the pronation and supination gait patterns without exaggeration.

For the acquisition of ground truth gait data, 22 reflective markers were attached to the lower limb as demonstrated in Figs. 3(d). The Vicon motion capture system tracked the 3D positions of these markers with high precision and frequency (120Hz). The ground truth values of the camera pose were also recorded by the Vicon system. The timestamps of the images and the Vicon system were recorded for synchronization. The participants initiated the gait from different directions and then walked along the diagonal line of a sensing area of size $2m \times 3m$. A subject repeated each gait pattern for eight times, and the total number of samples for each subject is 48.

B. Gait Parameter Extraction

Among different type of gait parameters, joint angles of the 3D lower limb correlate well to the joint kinematics in the gait periodic movement [2], [6]. They are also sensitive to gait abnormalities and constitute typical measures in clinical evaluation. Recalling that slight differences exist between the marker positions of the ground truth data and the joint positions, thus joint angles are more appropriate to evaluate the accuracy of the proposed method compared to the absolute difference between reflective markers and detected joints.

We first calculate joint angular features in relation to a human-based local coordinate system $\{\mathrm{H}\}$. Based on this human coordinate system, Sagittal, Coronal, and Transverse planes are determined as illustrated in Figs. 4(a). Let denote the 3D positions of the RASI and LASI joints at time t as $\mathbf{P}_{RASI}(t)$ and $\mathbf{P}_{LASI}(t)$, respectively. The origin of the coordinate system $\{\mathrm{H}\}$ at time t is denoted as $\mathbf{o}(t) = 1/2(\mathbf{P}_{RASI}(t) + \mathbf{P}_{LASI}(t))$. Two unit vectors at time index t can be determined by $\mathbf{V}_1 = \mathbf{P}_{RASI}(t) - \mathbf{P}_{LASI}(t)/||\mathbf{P}_{RASI}(t) - \mathbf{P}_{LASI}(t)||$ and $\mathbf{V}_2 = (\mathbf{o}(t) - \mathbf{o}(t-1))/||\mathbf{o}(t) - \mathbf{o}(t-1)||$. Accordingly, $\{\mathrm{H}\} = [\hat{\mathbf{x}}_h, \hat{\mathbf{y}}_h, \hat{\mathbf{z}}_h]$ can be calculated by:

$$\hat{\mathbf{x}}_h = \mathbf{V}_1, \quad \hat{\mathbf{y}}_h = \frac{\mathbf{V}_1 \times \mathbf{V}_2}{||\mathbf{V}_1 \times \mathbf{V}_2||}, \quad \hat{\mathbf{z}}_h = \frac{\hat{\mathbf{x}}_h \times \hat{\mathbf{y}}_h}{||\hat{\mathbf{x}}_h \times \hat{\mathbf{y}}_h||} \tag{2}$$

Commonly, the 3D lower limb skeleton can be represented by six link segments: *Thigh(L)*, *Shank(L)*, *Foot(L)*, *Thigh(R)*, *Shank(R)*, and *Foot(R)*, where L and R indicate *left* and *right*. As demonstrated in Fig. 4(a), we first calculate the joint angles $\{\phi_x, \phi_y, \phi_z\}$ between each link segment l with respect to the normal vectors of the Sagittal, Coronal, Transverse planes, respectively.

The angle between link l and floor normal vector $\mathbf{n}_f$ is referred to ϕ_f. Note that the floor plane can be determined in the mapping process of the ORB-SLAM. Next, the joint angles between two segments are also considered in this paper. As illustrated in Fig. 4(b), the knee angle θ_{knee} is calculated by the thigh and shank segments, and the ankle angle θ_{ankle} indicates the angle between shank and foot segments. Finally, another critical gait parameter, which indicates the angle between the foot direction and walking path, is the foot

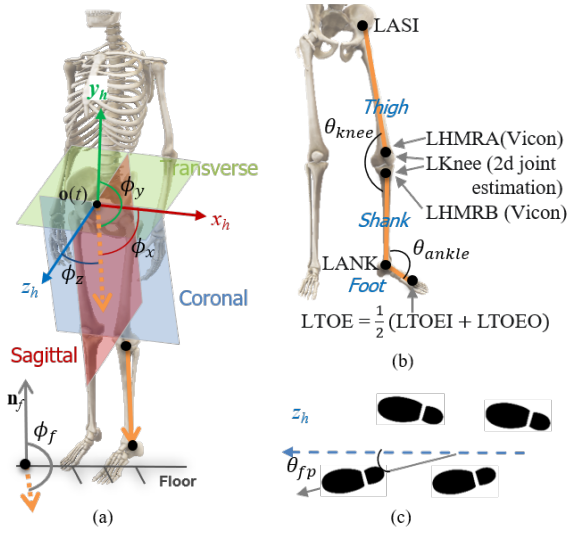


Fig. 4. Illustration of the human body coordinate system and various extracted joint angle features for abnormal gait recognition. a) A human body coordinate system $\{\mathrm{H}\}$ can be defined as Eq. (2). From this local coordinate system, we can define three planes as Sagittal, Coronal, and Transverse, respectively. The joint angles between link segments and the normal vectors of the plane are calculated; b) The knee angle and the ankle angle denote the angle between two link segments; c) foot progression angle is the angle between foot orientation vector and the body forward direction.

progression angle θ_{fp}. The walking path orientation at each frame can be regarded as the z-axis (forward body direction) of the human body coordinate system as shown in Fig. 4(c).

It should be noted that we only consider the abnormal gait recognition from forward walking styles in this study, which means that those regarding turn events are not taken into consideration. The features within each gait cycle are estimated, and the gait cycle is defined as the time interval between two consecutive heel-strike events. A heel-strike is the moment when the heel touches the ground. In this paper, the meta-feature of the joint angle for representing the gait pattern at each timestamp $\mathbf{\Phi}(t) \in \mathbb{R}^K (K = 30)$ is the combination of the aforementioned features, which are summarized in Table I.

TABLE I
EXTRACTED JOINT ANGLE FEATURES

Joint angle	Feature quantity	Description
ϕ_x	6: Left, Right	angle between link segments {Thigh, Shank, Foot} and $\hat{\mathbf{x}}_h$
ϕ_y	6: Left, Right	angle between link segments {Thigh, Shank, Foot} and $\hat{\mathbf{y}}_h$
ϕ_z	6: Left, Right	angle between link segments {Thigh, Shank, Foot} and $\hat{\mathbf{z}}_h$
ϕ_f	6: Left, Right	angle between link segments {Thigh, Shank, Foot} and $\hat{\mathbf{n}}_f$
θ_{knee}	2: Left, Right	angle between thigh and shank
θ_{ankle}	2: Left, Right	angle between shank and foot
θ_{fp}	2: Left, Right	angle between foot orientation and forward direction $\hat{\mathbf{z}}_h$

C. Abnormal Gait Recognition

In most of the previous works for skeleton-based gait analysis approaches [11], [23], only binary classification of the normal and abnormal gait is achieved. In this paper, we focus on more challenging abnormal gait recognition tasks for six gait patterns (normal, in-toeing, out-toeing, drop foot, supination, and pronation) as shown in Fig. 3. Two recognition methods are utilized for achieving abnormal gait recognition.

1) Statistical features with SVM classifier: We first adopt the nonlinear SVM as the classifier for multiple classes recognition. As joint angles are represented as temporal sequences of different lengths, the statistical feature will be calculated to normalize these sequences into the same dimension. Given a joint angle sequence, the corresponding histogram $\mathbf{F}_i$ with N_b bins representing the distribution of the angles within $[-\pi, \pi]$ can be extracted. For the K extracted joint angle sequences, the histograms $\mathbf{F}_i$ will be concatenated as the final feature representation $\mathcal{F} = [\mathbf{F}_1, \cdots, \mathbf{F}_i, \cdots, \mathbf{F}_K] \in \mathbb{R}^{1\times 30N_b}$. The SVM classifier is trained using the RBF kernel $K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma||\mathbf{x}-\mathbf{x}'||_2^2)$. Finally, classification results will be given based on a one-against-one rule.

2) Angle sequences with Bi-LSTM classifier: We also adopt the single layer BiLSTM [18] for abnormal gait recognition. Recurrent Neural Networks (RNN) is a popular neural network architecture to recognize patterns in sequential data. Specifically, bidirectional RNN (BRNN) connects two hidden layers of opposite directions to the same output, and the output layer can get information from the past and future simultaneously. The joint angle sequences $\mathbf{\Phi}(t) \in \mathbb{R}^K$ is first segmented into segments according to the heel strike detection. Then all the segmented sequences in the training set are divided into mini-batches of size N_m. We also pad the sequences within each mini-batch to the same length as the longest sequence. The size of the input layer is $K = 30$, which equals the dimension of the joint angle features. In the BiLSTM layer, we use N_h LSTM units and output the last element, in which each cell consists of an input gate, an output gate and a forget gate. Finally, a fully connected layer of size six is followed by a softmax layer and a classification layer. Considering a test sample contains multiple gait cycles as well as joint angle segments, we vote the predicted label of the sample as the one with the maximum occurrence frequency among the predicted labels of segments.

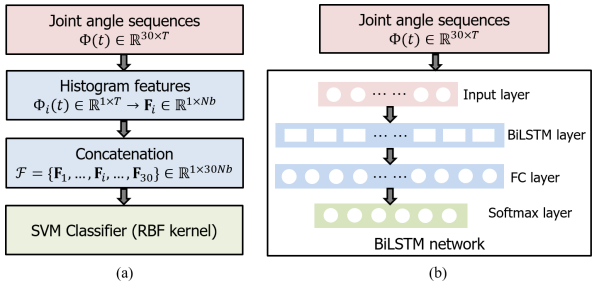


Fig. 5. Two approaches for abnormal gait recognition with respect to joint angle features.

IV. EXPERIMENTS AND RESULTS

A. Experiment Settings

We used a remote workstation with the following technical characteristics: Intel(R) Core(TM) i7-7700K CPU @4.2 GHz with an NVIDIA Titan 1080Ti GPU. For the mobile platform, a laptop was deployed on the robot to real-time stream the RGB-D images. It should be pointed out that the streaming protocols only support 8-bit images, whereas depth information is usually represented in 16-bit grayscale images. To this end, we transformed 16-bit depth images into 8-bit 3-channel images. Finally, the RGB image and the encoded depth image were concatenated as the streaming resource, and the RTSP was applied for low-latency streaming. The workstation decoded the received depth images into the real distance values. The resolution of the RGB-D sensor was 640×480 and the frame rate was 30. We summarize the parameters used in the 2D pose estimation, ORB-SLAM, and recognition methods in Table II.

TABLE II
PARAMETERS IN DIFFERENT METHODS

Algorithms	Parameter	Value
pose estimation	Net input resolution	640×384
	Tracked joint threshold	ϵ >0.2
ORB-SLAM	Close/Far threshold	50
	No. of ORB features per image	2000
	Scale factors in scale pyramid	1.2
	No. of levels in the scale pyramid	4
SVM	Histogram bins N_b	37
	RBF kernel γ	10
BiLSTM	Mini-batch size N_m	16
	No. of hidden units N_h	512
	Max epochs	100
	Initial learning rate	0.001
	Learning rate drop period (epochs)	20
	Learning rate drop factor	0.1

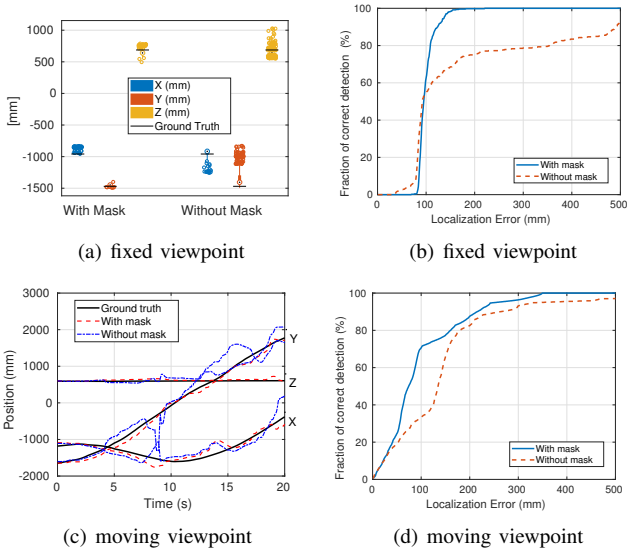
In the recognition experiments, we evaluated the performance of the proposed approach for detecting subtle gait changes and compared with the methods proposed in [7], [11], [19]. For our proposed approach, four setups (**SVM-GT**, **BiLSTM-GT**, **SVM-EST**, and **BiLSTM-GT**) were examined, where *GT* indicates the angular features calculated from ground truth data and *EST* refers to the estimated joint angles based on the proposed algorithm. For the comparison methods, **SVM- [7]** uses the quantitative gait features {step length, gait cycle time, and gait symmetric measure} proposed in [7] as the input to the SVM classifier. As the seven joint angles {left hip angle, right hip angle, left knee angle, right knee angle, left ankle angle, right ankle angle and two feet angle} [11] and the DSRF descriptors of six rigid bodies {left thigh, right thigh, left shank, right shank, left foot, and right foot} [19] are also temporal sequences, we validated their performance with the BiLSTM classifier, which are denoted as **BiLSTM- [11]** and **BiLSTM- [19]**, respectively. It should be pointed out that these features were extracted from the lower limb pose estimated by our system.

B. Camera Localization in Dynamic Environment

In contrast to the RGB-D images used in the construction of 3D map points, the human target in the gait analysis can be regarded as a dynamic object, which will increase the difficulty in tracking the camera pose via the re-localization module of the ORB-SLAM. To evaluate this, we first conducted two experiments to validate the proposed mask-based strategy for improving the camera localization performance in the dynamic environment. In the results shown below, "*With Mask*" and "*Without Mask*" denote the proposed mask-based method and the standard ORB-SLAM, respectively.

In the first experiment, the camera was located at a fixed location, and a human subject was asked to randomly move within the field-of-view of the camera. The Vicon motion capture system recorded the ground truth data of the human lower limb as well as the 6D camera pose. In Fig. 6(a), the camera localization errors in x, y, and z axes by *With Mask* and *Without Mask* are compared. The circle dots indicate the estimated position, and the black lines are the ground truth values. It can be seen that the localization result with mask-based strategy is more stable than without it. Fig. 6(b) shows the cumulative error distributions of the camera localization w/o the mask, respectively. With mask-based strategy, the fraction of the correction detection increases quickly and reaches 100% while the localization error threshold is about $160mm$. However, the fraction of correct detection without mask-based strategy is still 90% while the error threshold is $500mm$.

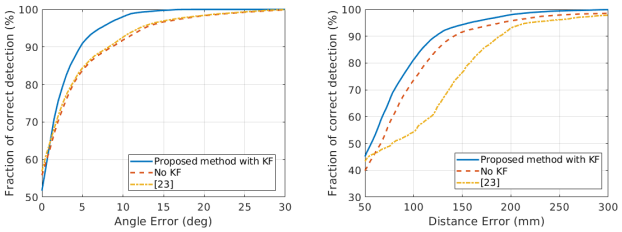
Next, we validated the stability of the camera re-localization while both the camera and the human target were moving. In Fig. 6(c), the solid black line is the ground truth value, the blue dot line indicates the result using mask-based strategy during the re-localization, and the red slash line shows the result with the standard ORB-SLAM without the mask. The precision and the stability of the camera localization using *With Mask* are superior to those by *Without Mask*. The cumulative error distribution with respect to different distance error thresholds using two methods are compared in Fig. 6(d). The fraction of the correction detection of the mask-based method increases to 100% while the localization error threshold is about $320mm$. Due to the low texture existed in some images and the image motion blur, the localization error of the moving camera is higher than that measured from a fixed viewpoint.



(a) fixed viewpoint (b) fixed viewpoint

(c) moving viewpoint (d) moving viewpoint

Fig. 6. Comparison results of the camera localization in terms of dynamic human movement w/o the mask-based strategy for ORB-SLAM, respectively. a) The error of the x, y, and z axes between ground truth data and the estimated positions from a fixed viewpoint. b) The cumulative error distribution of correct detection on the fixed camera position with respect to different acceptance threshold in the distance; c) The comparison results of the camera localization on three axes from a moving camera. d) The cumulative error distribution of correct detection on the moving camera position with different acceptance threshold under the mask-based and without mask methods, respectively.



(a) (b)

Fig. 7. Comparison results by the proposed method w/o KF (blue/red) and the VoxelPoseNet (yellow) [21] in 3D lower limb estimation. a) The cumulative error distribution of correct detection on the joint angle with respect to different acceptance thresholds; b) the cumulative distribution plots of the pose estimation in the distance.

In summary, the camera localization via the ORB-SLAM with the proposed mask-based strategy is more stable and accurate than the standard algorithm. The proposed mask-based strategy can substantially improve the re-localization performance of SLAM algorithm in terms of the dynamic environment, thus providing the reliable pose estimation in the canonical coordinate system.

C. Lower Limb Pose and Joint Angle Estimation

In this study, we validated the joint positions and joint angles estimated by the proposed system and a state-of-the-art method [21] based on our collected gait database. Zimmermann *et al.* estimated 2D pose from color images via human keypoint detectors, and then incorporated depth information to predict 3D joints via a deep network, called VoxelPoseNet [21]. Figs. 7(a) and (b) show the cumulative distribution of correct angle and position estimation regarding different error thresholds. For our proposed method, the results w/o KF for predicting and smoothing 3D joints are given. As Kalman filter can not only smooth the 3D joints but also predict the 3D positions of the missing joints, the proposed method with KF has better detection rates in both joint positions and angles estimation. Especially, the 100% correct detection rate is achieved when the angle error threshold is 16° and the distance error threshold is about $250mm$. One of the main reason for the measurement error is that the depth values of key joints are extracted from the

surface of the human body, and these values may vary a little from different viewpoints. Moreover, the inherent distance error in joint position measurements results from different displacements between estimated joints and ground truth markers. The yellow slash line indicates the results estimated by VoxelPoseNet [21]. The large distance error occurs due to the tracking failure when only the lower limbs of the target human are present in the rgb and depth images.

D. Abnormal Gait Recognition

Finally, to evaluate the performance of the proposed system and the previous methods for abnormal gait recognition, seven setups (**SVM-GT**, **BiLSTM-GT**, **SVM-EST**, and **BiLSTM-GT**, **SVM- [7]**, **BiLSTM- [11]** and **BiLSTM- [19]**) were examined. Considering the individual differences in the walking styles, we used the following three protocols to provide a reliable evaluation of the proposed recognition methods.

Intra-subject protocol. Considering the individual differences in the walking styles, we first evaluate different methods with intra-subject protocol, which means that the personalized classifier is trained for each subject. In addition to evaluating the estimation accuracy of the proposed system, we randomly divided eight samples into two groups with the same sizes. Hence, the experiments were repeated for 50 times and the average recognition rates were given.

Leave-one-subject-out (LOSO) protocol. We considered more challenging inter-subject validation protocols. For the LOSO protocol, the training set consisted of the samples from 15 subjects, and the samples from the remaining subject were for testing. In total, the experiments repeated 16 times and in each time one of the sixteen subjects was excluded. The LOSO is beneficial for testing the robustness of an algorithm when the number of samples is small. The average recognition rate over 16 runs was estimated.

Cross-subject protocol In the cross-subject protocol, we randomly selected 50 permutations for dividing 16 subjects into two groups, in which gait samples by 8 subjects were used for training, and the test data came from the remaining 8 subjects. The average recognition accuracy over 50 experiments was reported.

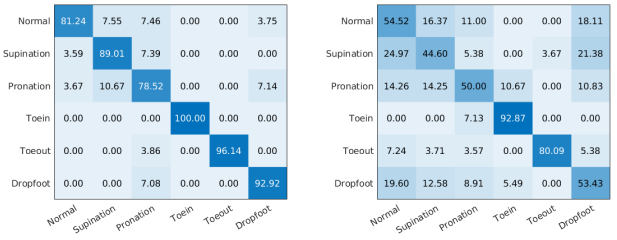
In Table III, we report the comparison results in recognizing six gait patterns with different methods and protocols. Bold denotes the highest recognition rate under each protocol, and underline indicates the second-best result. It can be seen that our proposed method outperforms the previous methods in terms of recognition accuracy under various protocols. The classification results achieved by SVM- [7] are inferior to other recognition methods validated in this work. This method is affected severely from differences in body sizes and body movements across subjects. BiLSTM- [11] and BiLSTM- [19] achieve similar recognition accuracies. However, they are still inferior to those by our proposed approach. In addition to the features proposed in [11], we also calculate the {thigh angle, shank angle, and foot angle} that reflect the relationship between the lower skeleton and the floor plane. Moreover, the joint angles with respect to the local human body coordinate system also help capturing subtle differences among abnormal gait patterns. On the other hand, the DSRF descriptor in [19] was devised to be invariant to rigid transformations by rotational normalization and trajectory length normalization. Apparently, these invariant properties reduce the ability to detect subtle changes among various gait patterns.

TABLE III
CLASSIFICATION RESULT ON SIX ABNORMAL GAIT RECOGNITION

Methods\Protocol	Intra-subject	LOSO	Cross-subject
SVM-GT	<u>89.86</u>%	<u>60.82</u>%	<u>54.36</u>%
SVM-EST	88.19%	60.21%	54.27%
SVM-[7]	38.98%	30.64%	23.83%
BiLSTM-GT	**90.75**%	**61.55**%	**54.43**%
BiLSTM-EST	87.79%	60.67%	54.08%
BiLSTM-[11]	68.54%	47.08%	36.86%
BiLSTM-[19]	63.41%	44.71%	39.51%

As can be observed, the recognition accuracies under the intra-subject protocol are much higher than those with inter-subject validation protocols (LOSO and cross-subject). This is because distinct people have different body sizes and additionally exhibit individual differences in both normal and abnormal walking styles. Although a personalized classifier is beneficial for discriminating the gait patterns, it is not applicable to collect all the candidate gait abnormalities of a person in the real-world applications. The BiLSTM methods achieve superior recognition results with respect to the GT data under the LOSO (61.55%) and cross-subject (54.43%) protocols, respectively. Due to the reduced size of the training data, the recognition rates under cross-subject protocol are slightly lower than those with the LOSO protocol. More important, the classification rates using the EST data are slightly inferior to those using the GT data, which emphasizes the effectiveness of the proposed system in terms of lower pose estimation.

However, the recognition results on the six gait pattern recognition task are unsatisfactory, in which the accuracies under LOSO and cross-subject protocols are below 62%. Figs. 8(a) and (b) plot the confusion matrix of the highest recognition rates under intra-subject (90.75%) and LOSO (61.55%) protocols. It can be seen that the supination and pronation are difficult to differentiate by the proposed recog-



(a) Intra-subject: 90.75% (b) LOSO protocol: 61.55%

Fig. 8. Confusion matrix of six classes abnormal gait recognition. Supination and pronation are easily confused with the normal gait.

nition approaches. This is because the correction insoles introduce only slight differences in pronation and supination compared to normal gait. We expect the sensitivity of the classification to increase with the addition of more subjects for training.

Moreover, we also assessed four gait patterns (normal, in-toeing, out-toeing, and drop-foot gait) as reported in Table IV for demonstrating the potential clinical value of the proposed technique. The highest recognition rates with different methods under the intra-subject protocol are 99.31% (SVM-GT) and 98.55% (BiLSTM-GT), which emphasizes that the personalized classifier can detect the gait abnormalities with high precision. As for the LOSO protocol, the BiLSTM method achieves the better recognition rates 88.90% and 87.44% with GT and EST data, respectively. Compared to results under the intra-subject protocol, the recognition rates using the LOSO protocol drop around 15% due to the individual differences in normal/abnormal gait styles. Without considering the challenging supination and pronation gait patterns, the results in recognizing four gait patterns are significantly increased.

TABLE IV

CLASSIFICATION ACCURACY IN RECOGNIZING NORMAL, IN-TOEING, OUT-TOEING, AND DROP-FOOT GAIT PATTERNS

Methods\Protocol	Intra-subject	LOSO	Cross-subject
SVM-GT	**99.31**%	85.15%	**76.77**%
SVM-EST	96.27%	84.55%	73.79%
SVM-[7]	56.08%	39.01%	31.10%
BiLSTM-GT	98.55%	**88.90**%	76.32%
BiLSTM-EST	97.43%	87.44%	75.32%
BiLSTM-[11]	73.99%	67.02%	59.81%
BiLSTM-[19]	70.70%	60.08%	54.15%

V. CONCLUSIONS

In summary, we have presented in this paper a mobile 3D gait analysis system for monitoring patients at their home based on an RGB-D sensor and a light-weight mobile robot. To deal with the 3D pose estimation in the moving camera frame, visual-SLAM was used to localize both camera movement and the human gait in a canonical coordinate system, which enables the calculation of the joint kinematics and quantitative gait features. Our work does not merely combine two approaches but it integrates them in a way that the accuracy of each of the subsystems improves. We have demonstrated this with results that show the improvement of camera localization as well as several comparisons with state-of-the-art methods in abnormal gait detection. We use both SVM and BiLSTM to classify gait patterns of normal, in-toeing, out-toeing, drop-foot, supination and pronation gait. Our proposed method outperforms the previous approaches in terms of recognition accuracy. Pronation and supination are based on correction insoles and therefore only result in subtle changes in the foot position. Future work would aim to increase the number of subjects to avoid overfitting of the neural network and improve the feature selection for detecting subtle gait changes.

REFERENCES

[1] G.-Z. Yang, *Body Sensor Networks*. Springer, 2006.

[2] S. Chen, J. Lach, B. Lo, and G.-Z. Yang, "Toward pervasive gait analysis with wearable sensors: a systematic review," *IEEE J. Biomed. Health Inform.*, vol. 20, no. 6, pp. 1521–1537, 2016.

[3] L. Penteridis, G. D'Onofrio, D. Sancarlo *et al.*, "Robotic and sensor technologies for mobility in older people," *Rejuvenation Res.*, vol. 20, no. 5, pp. 401–410, 2017.

[4] T. Seel, J. Raisch, and T. Schauer, "Imu-based joint angle measurement for gait analysis," *Sensors*, vol. 14, no. 4, pp. 6891–6909, 2014.

[5] A. Pfister, A. M. West, S. Bronner, and J. A. Noah, "Comparative abilities of microsoft kinect and vicon 3d motion capture for gait analysis," *J. Med. Eng. Technol.*, vol. 38, no. 5, pp. 274–280, 2014.

[6] F. Deligianni, C. Wong, B. Lo, and G.-Z. Yang, "A fusion framework to estimate plantar ground force distributions and ankle dynamics," *Inform. Fusion*, vol. 41, pp. 255–263, 2018.

[7] S. Bei, Z. Zhen, Z. Xing *et al.*, "Movement disorder detection via adaptively fused gait analysis based on kinect sensors," *IEEE Sens. J.*, vol. 18, no. 17, pp. 7305–7314, 2018.

[8] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 7291–7299.

[9] X. Gu, F. Deligianni, B. Lo, W. Chen, and G.-Z. Yang, "Markerless gait analysis based on a single rgb camera," in *Proc. IEEE Int. Conf. Wearable Implant. Body Sens. Netw. (BSN)*. IEEE, 2018, pp. 42–45.

[10] J. Shotton, T. Sharp, A. Kipman *et al.*, "Real-time human pose recognition in parts from single depth images," *Commun. ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[11] T.-N. Nguyen, H.-H. Huynh, and J. Meunier, "Skeleton-based abnormal gait detection," *Sensors*, vol. 16, no. 11, p. 1792, 2016.

[12] W. Chi, J. Wang, and M. Q.-H. Meng, "A gait recognition method for human following in service robots," *IEEE Trans. Syst., Man, and Cybern., Syst.*, vol. 48, no. 9, pp. 1429–1440, 2017.

[13] G. Wilson, C. Pereyda, N. Raghunath *et al.*, "Robot-enabled support of daily activities in smart home environments," *Cogn. Syst. Res.*, vol. 54, pp. 258–272, 2019.

[14] E. W. McClain and S. Meek, "Determining optimal gait parameters for a statically stable walking human assistive quadruped robot," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*. IEEE, 2018, pp. 1751–1756.

[15] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE Trans. Robot.*, vol. 33, no. 5, pp. 1255–1262, 2017.

[16] T. Taketomi, H. Uchiyama, and S. Ikeda, "Visual slam algorithms: A survey from 2010 to 2016," *IPSJ Transactions on Computer Vision and Applications*, vol. 9, no. 16, pp. 1–11, 2017.

[17] B. Bescos, J. M. Fácil, J. Civera, and J. Neira, "Dynaslam: Tracking, mapping, and inpainting in dynamic scenes," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 4076–4083, 2018.

[18] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[19] Y. Guo, Y. Li, and Z. Shao, "Dsrf: A flexible trajectory descriptor for articulated human action recognition," *Pattern Recognit.*, vol. 76, pp. 137–148, 2018.

[20] R. A. Clark, B. F. Mentiplay, E. Hough, and Y.-H. Pua, "Three-dimensional cameras and skeleton pose tracking for physical function assessment: a review of uses, validity, current developments and kinect alternatives," *Gait & posture*, 2018.

[21] C. Zimmermann, T. Welschehold, C. Dornhege *et al.*, "3d human pose estimation in rgbd images for robotic task learning," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*. IEEE, 2018, pp. 1986–1992.

[22] M. Edo and S. Yamamoto, "The difference in kinematic chain behavior between pronation/supination of calcuneus and rotation of shank in standing position by individual, age, gender and right and left: Analysis of kinematic using optical three dimensional motion analysis system," *International Journal of Physiotherapy*, vol. 5, pp. 31–35, 2018.

[23] A. A. Chaaraoui, J. R. Padilla-López, and F. Flórez-Revuelta, "Abnormal gait detection with rgb-d devices using joint motion history features," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, vol. 7. IEEE, 2015, pp. 1–6.