Markerless Gait Analysis Based on a Single RGB Camera[1]

X. Gu, F. Deligianni, B. Lo, W. Chen and G.Z. Yang

Abstract— **Gait analysis is an important tool for monitoring and preventing injuries as well as to quantify functional decline in neurological diseases and elderly people. In most cases, it is more meaningful to monitor patients in natural living environments with low-end equipment such as cameras and wearable sensors. However, inertial sensors cannot provide enough details on angular dynamics. This paper presents a method that uses a single RGB camera to track the 2D joint coordinates with state-of-the-art vision algorithms. Reconstruction of the 3D trajectories uses sparse representation of an active shape model. Subsequently, we extract gait features and validate our results in comparison with a state-of-the-art commercial multi-camera tracking system. Our results are comparable to those from the current literature based on depth cameras and optical markers to extract gait characteristics.**

I. Introduction

Clinical gait analysis aims to describe human locomotion based on quantitative parameters such as step length, stride length, speed, and joint angles. Features extracted from gait analysis are used to characterize normal and abnormal gait in several clinical scenarios that range from orthopedics and posture control to functional decline in neurological conditions and elderly people [1]. In a normal gait cycle, the ankle drives the foot during the toe-off phase of the gait and it is the first joint to absorb the impact with the floor at heel strike. Therefore, the ankle's inversion/eversion, dorsiflexion/ plantarflexion and foot progression angle are of significance importance for detecting abnormal gait patterns [2].

Current methods adopted in gait analysis are based on either sensor systems or video/optical tracking. Wearable sensor system for gait analysis could use one or multiple sensors attached to the body [3]–[5] to measure acceleration, angular rate or pressure. Although inertial sensor systems can be used to monitor subjects 24/7, they are limited in that they do not measure joints angle directly [2].

To this end, multi-camera systems represent the current state-of-the-art in measuring joint displacements and angles based on markers [6]. However, these systems are expensive, difficult to set up and they cannot be used outside the clinic/laboratory. Therefore, the use of a single RGB or depth camera to extract gait parameters is desirable for monitoring patients/subjects in more natural settings [2]. Gait analysis based on a single RGB camera is challenging because it requires tracking the person while it is walking, to accurately locate the body joints as well as to be able to extract angular information invariant to the camera's perspective projection. Furthermore, gait indices extracted based on 2D joint angles only are difficult to interpret and compare with standard clinical systems. Another option is to directly reconstruct 3D information of the joint locations from the video recordings.

In the field of computer vision, human pose estimation from monocular videos has been extensively researched in recent years [7], [8]. For example, OpenPose is an open-source library for real-time multi-person joints detection. It is a bottom-up approach that consists of two-branch convolutional neural networks trained based on annotated 2D keypoints. Although it provides relatively accurate and stable estimation of human skeleton points, it does not reconstruct the 3D coordinates of the joints. Several challenges arise in the reconstruction of 3D human pose from 2D skeleton points that include the non-linear motion, occlusions and ambiguity between 2D and 3D poses. A promising method of estimating 3D human pose via an expectation maximization algorithm reduces uncertainty over the estimated 3D joints location along the entire video [8]. However, these approaches are focused on tracking the upper limbs and they are not tailored to gait analysis. Furthermore, the training datasets are based on specific activities that do not attempt to model gait variability in pathological scenarios.

In this article, we present a novel method for gait analysis based on a single RGB camera system on a cell phone. Compared to previous related work, we do not set any strict standards to camera position and background settings, whereas 3D reconstruction of the joint coordinates takes place [9]. The 2D-3D reconstruction methods we adopted has been proved to be able to efficiently achieve global optimality. Furthermore, here we focus only on the lower limbs with six key points that represent the knees, ankles, and toes. Therefore, it is not possible to impose anthropometric constraints that take into account the whole body [7], [10]. We validate our methods with a state-of-the-art multi-camera system. Our results compare well with the depth camera systems based on markers, which achieve up to 10 degrees accuracy [11].

II. Methods

The framework of our method is displayed in Fig.1. We use OpenPose to track the 2D coordinates of human joints [12]. To accurately detect the foot orientation, the GrabCut algorithm [13] is applied. Subsequently, we use active shape modelling combined with sparse dictionary learning to reduce the parameters in 3D space [14]. Finally, a weak perspective projection is used to map the 2D extracted coordinates to 3D. Our methods are validated based on a multi-camera acquisition system (SMART-DX) as well as 2D manually annotated points of joints.

A. 2D Joint Points Detection

Firstly, we track human pose with OpenPose, which is an open-source state-of-the-art approach based on Part Affinity Fields [12], [15]. This allows multi-person tracking, which is important in accounting human interactions in natural settings.

[1] Xiao Gu was a visiting student at the Hamlyn Centre, Imperial College London. Fani Deligianni, Benny Lo and Guang-Zhong Yang are with the Hamlyn Centre, Imperial College London, UK ({fani.deligianni, benny.lo, g.z.yang}@imperial.ac.uk). Xiao Gu and Wei Chen are with the Center for Intelligent Medical Electronics, School of Information Science and Technology, Fudan University, China ({gux14,w_chen}@fudan.edu.cn).

In addition, it can accurately locate key points when occlusion occurs. OpenPose does not detect toe positions and it may fail to track the lower limbs when the camera view does not include the upper limbs. To circumvent this problem and distinguish foreground pixels of the foot from the background, we use GrabCut [13], which is a mixture-models image-segmentation method. This is refined based on morphological image processing that involved dilation and erosion operations. Subsequently, ellipse fitting is applied to the foot contour points and the foot orientation is defined based on the centre of the ellipse and the ankle joint. GrabCut is initialised based on a region of interest. We utilise the time consistency across different frames to detect and avoid failures. In this section, we denote as $\boldsymbol{W_j} \in \mathbb{R}^{2\times p}$ **,** 2D locations of lower limb points (p=6, left and right knees, ankles and toes) of the j^{th} frame (the frame number is *n*).

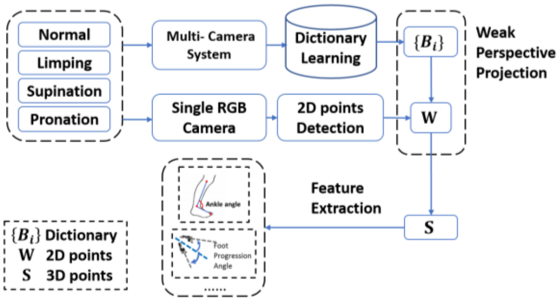


Fig.1 Framework of Our Method

B. 3D Sparse Reconstruction

It is well known that reconstructing 3D human pose from the perspective projection of the joints to the image plane is an ill-posed problem. However, the ambiguity of the 3D reconstruction can be reduced significantly if we express the 3D pose as a linear combination of basis postures based on the motion of the joints across time. Similarly, to active shape models, this will significantly reduce the reconstruction parameters. Subsequently, the 2D image coordinates of the joints are mapped to 3D based on a weak perspective camera projection. Below we describe this in details.

Instead of using a conventional active shape model, a sparse dictionary is used to further reduce dimensionality and to represent the sparsity in motion patterns [7], [9]. We have also used principal component analysis to verify the sparsity of the extracted 3D points. In fact, the first five principal components can well represent signals over 95%.

We denote the 3D lower limbs posture of the j^{th} frame as $\boldsymbol{S_j} \in \mathbb{R}^{3\times p}$ with the 3D coordinates of p points. Its sparse representation of our model takes the form:

$$\boldsymbol{S_j} = \sum_{i=1}^{k} \omega_{ij} \boldsymbol{B_i} \tag{1}$$

where $\{\boldsymbol{B_1}, \boldsymbol{B_2}, \dots, \boldsymbol{B_k}\}$ is an overcomplete dictionary of k atoms, $\boldsymbol{B_i} \in \mathbb{R}^{3\times \mathrm{p}}$is the i^{th} atom of the dictionary, and ω_{ij} is the coefficient of $\boldsymbol{B_i}$. $\{\omega_{\mathrm{ij}}\}$ is assumed to be sparse.

An online dictionary learning method is applied to update the dictionary during an iterative learning process mentioned in [16]. It can solve the problem by alternately updating $\{\omega_{ij}\}$ and $\{\boldsymbol{B_i}\}$ in the following formulation (λ_1 is a regularization coefficient to constrain the sparsity):

$$\min_{\{\boldsymbol{B_i}\},\{\omega_{ij}\}} \sum_{j=1}^{n} \frac{1}{2} \left\| \boldsymbol{S_j} - \sum_{i=1}^{k} \omega_{ij} \boldsymbol{B_i} \right\|_F^2 + \lambda_1 \sum_{i,j} \omega_{ij} \tag{2}$$

$$s.t.\ \omega_{ij} \geq 0,\ \|\boldsymbol{B_i}\|_F \leq 1,\ \forall i \in [i,k],\ j \in [1,n]$$

Finally, a weak perspective projection is used to map 2D points to 3D coordinates [17]. As in our model, all postures are centred and normalized, we do not take the translation matrix into consideration. $\boldsymbol{W_j}$ is represented by the formula below,

$$\boldsymbol{W_j} = \boldsymbol{\Pi R S_j} = \boldsymbol{\Pi R} \sum_{i=1}^{k} \omega_{ij} \boldsymbol{B_i} = \sum_{i=1}^{k} \boldsymbol{M_{ij} B_i} \tag{3}$$

where $\boldsymbol{\Pi} = [\alpha, 0, 0; 0, \alpha, 0] \in \mathbb{R}^{2\times 3}$ is the scaling matrix (α is the scaling parameter), and $\boldsymbol{R} \in SO(3)$ is the rotation matrix. Equation (3) projects 3D into a 2D plane after rotation and scaling.

For 3D Reconstruction, we have to infer the parameters $\{\boldsymbol{M_i}\}$ in (3). The objective function, combination of (2) and (3), shows an optimization problem:

$$\min_{\boldsymbol{M}_1,\dots,\boldsymbol{M}_k} \frac{1}{2} \left\| \boldsymbol{W} - \sum_{i=1}^{k} \boldsymbol{M_i B_i} \right\|_F^2 + \lambda_2 \sum_{i}^{k} \|\boldsymbol{M_i}\|_2 \tag{4}$$

This is a least-squares problem of estimating the coefficient of each basis shape. We adopt the algorithm used in [9] to solve it globally, which is based on Alternating Direction Method of Multipliers [18] and the proximal operator of the spectral norm.

C. Gait Features Extraction

Inversion and eversion, as well as dorsiflexion and plantarflexion angles of the ankle, are important indices of abnormal gait [19]. In previous work, these measurements have been approximated based on 2D joint data alone [2]. In this scenario, dorsiflexion/plantarflexion was estimated as the angle of the foot with the vertical image plane axis, whereas inversion/eversion was estimated as the angle of the foot with the horizontal image plane axis [2]. Here, we also estimate key gait parameters based on the 3D reconstructed postures. These include foot progression angle and 3D foot-leg (ankle) angle, which are used in clinical settings and they are related to dorsiflexion/plantarflexion and inversion/eversion angles.

III. Results

A. Data Acquisition

We obtained data from 4 healthy volunteers (3 males and 1 female). These subjects were instructed to imitate different walking conditions in a straight trace, such as normal walking, limping, supination and pronation. A single RGB camera (30 Hz) was used to record the video. The minimal distance to the camera was set around 2m to render full lower body parts visible. We used manual annotation to mark the toe of the foot in each frame (2D ground truth data) and also obtained 3D ground truth data of lower limb points based on a multi-camera motion capture system (200Hz, Smart DX, BTS Bioengineering). For the acquisition of 3D ground truth data, we placed reflective markers on the knees, ankles and toes. The recorded points were labelled and tracked semiautomatically based on the Smart DX system to extract the 3D trajectories of the joints.

To synchronize the data between the single camera and the multi-camera system, we asked each subject to jump before and after walking. The peak of knee point locations was marked for synchronization.

B. *Sparse Dictionary Learning*

Fig.2 shows the reconstruction error during training on the datasets of normal walking, limping, supination and pronation, separately, as well as the concatenated set of data across conditions. Here, reconstruction error means a normalized distance. As for training, the original units of the coordinates of point position is in millimeters, we normalized each motion with the standard deviation and the error reflects the Euclidean distance between reconstructed motion and ground truth. The error is less when the system is trained with each condition separately. However, in real life, it is more likely to have mixed training scenarios of several types of abnormal walking. In our work, we trained and validated the system based on the concatenated set of data across conditions. As the sparsity of the dictionary, namely the number of atoms owning non-zero coefficients, becomes larger, the error decreases. Therefore, a trade-off should be achieved between the sparsity of the learned dictionary and the reconstruction accuracy. In our work, we set the sparsity to eight in training mixed conditions as the error does not decrease significantly over eight.

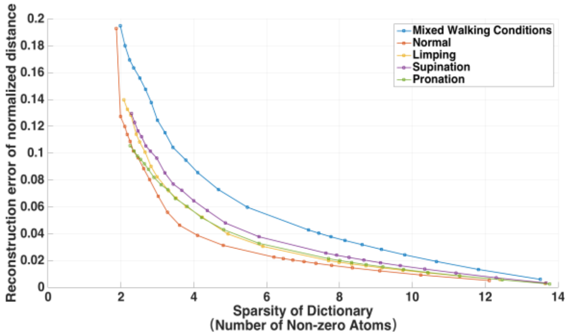


Fig.2 Reconstruction Error of Dictionary Learning

C. *2D Validation of Gait Angular Features*

We validated the accuracy of detecting the dorsiflexion/plantarflexion and inversion/eversion angle based on manually annotated data of the toes across all video frames. This reflects mainly the error in toes estimation and foot orientation based on the GrabCut algorithm. We estimated dorsiflexion/plantarflexion as the angle of the foot with the vertical image plane axis and inversion/eversion as the angle of the foot with the horizontal image plane axis similarly to [2]. Fig.3 summarizes the results of the angular error in degrees with histograms. Fig.3-a) shows two histograms of the total angular error of the inversion/eversion and dorsiflexion/plantarflexion angles, respectively, across all walking conditions. Fig.3-b) refers to the inversion/eversion angle and it displays the histograms of angular errors in each walking condition separately. Fig.3-c) refers to the dorsiflexion/plantarflexion angle and it also displays the histograms of angular errors in each walking condition separately.

The results show that in most of the frames the error is less than five degrees and compares well with previously published work [2]. The estimation of dorsiflexion/plantarflexion angle is more accurate than the estimation of inversion/eversion angle, especially in the pronation condition. This is because the fitted ellipse of the foot often has a larger horizontal component, thus having a larger bias from the true toe position in the horizontal direction, affecting inversion/eversion angle.

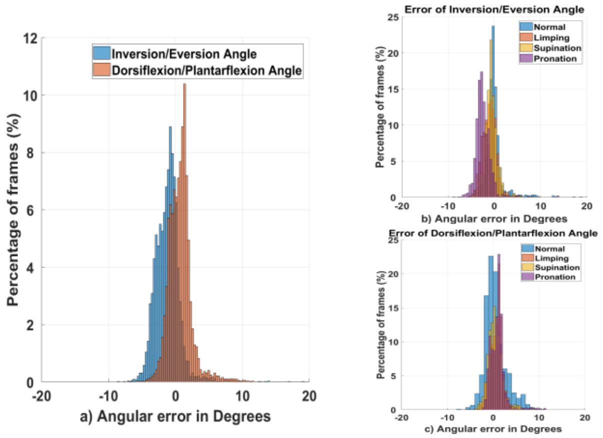


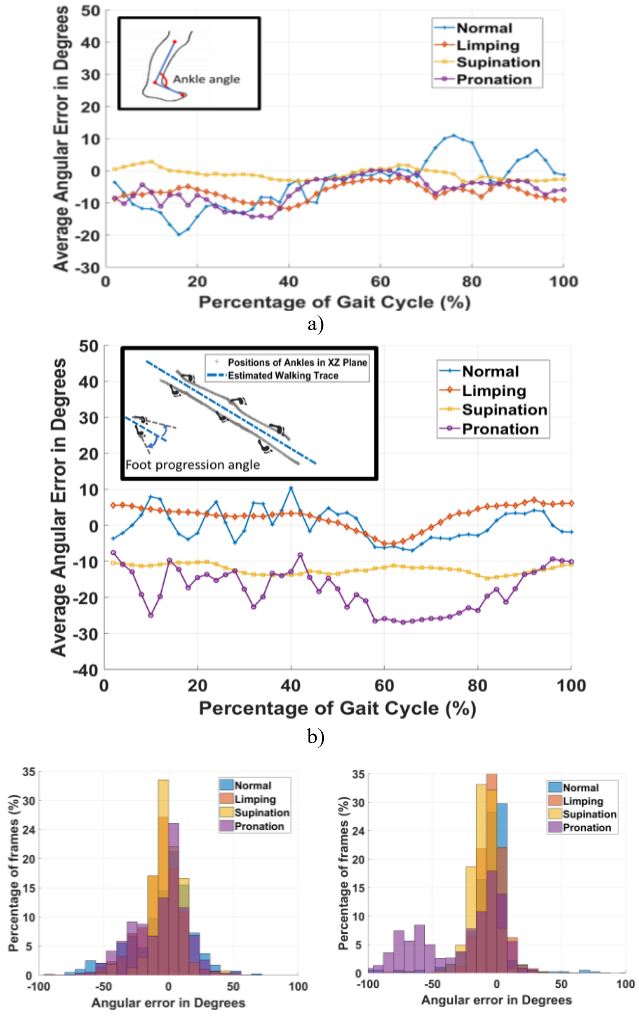
Fig.3 Angular error in degrees based on 2D manually annotated data, a) Total error across all walking conditions, b) Inversion/eversion angular error estimated in each walking condition separately, c) Dorsiflexion/plantarflexion angular error estimated in each walking condition separately.

D. *3D Validation of Gait Angular Features*

We used leave-one-out cross-validation to characterize the out-of-sample error of the 3D gait angular features. The 3D ground truth data were extracted from the multi-camera systems to train the sparse dictionary and extract the bases of motion kinematics.

Fig.4-a-b) shows the 3D angle between foot and shank as well as foot progression angle, respectively. To demonstrate the error across the gait cycle, we segmented the time-series based on Singular Spectrum Analysis (SSA) [2], [4]. Subsequently, we averaged the error of angular characteristics across gait cycles, Fig.4-a). We also plotted the histograms of the errors in each walking condition, normal walking, limping, supination and pronation, respectively, Fig.4-c). The error is smaller than 10 degrees in most of the frames and across all conditions.

Furthermore, Fig.4-b-d) shows the error in the foot progression angle estimation. As we asked that the subjects walked in a straight way to the lens, we used the least squares method to estimate the walking path and then extract the foot progression angle as the angle between the walking path and the ankle-toe line, Fig.4-b). For the estimated motions, the line connecting the camera and the subject would be Z axis and it can be approximated as walking path orientation. Fig.4-b) shows the error across the gait cycle, whereas Fig.4-d) shows the histogram of error across frames for normal walking, limping, supination and pronation, respectively. Although, the error is smaller than 10 degrees in most of the frames in normal walking, limping and supination, we observe error up to 30 degrees in pronation. This is because pronation differs from the other three conditions and this biases the training.



a)

b)

c)

d)

Fig.4 Angular error in degrees based on 3D ground truth data, a) Ankle angular error of different walking conditions in a gait cycle, b) Foot progression angular error of different walking conditions in a gait cycle, c) and d) Ankle and Foot progression angular error histogram

IV. Discussion and Conclusion

Gait analysis based on a single RGB camera is a challenging problem and to our knowledge, there is no work that estimates 3D gait parameters based on a markerless RGB scenario. Here we developed a novel framework to estimate 3D gait angular features of the lower limbs. To this end we have used joint detection algorithms and subsequently reconstructed the 3D coordinates of the joints based on a sparse dictionary representation. We have used a state-of-the-art 3D multi-camera system to validate our methods. We demonstrated that our system achieves accuracy that compares well with methods that are based on markers and depth information [11]. Further work should aim to validate our methods in clinical settings and larger datasets.

Acknowledgment

We would like to thank Charence Wong for helping with data acquisition for this study.

References

[1] A. Leardini, C. Belvedere, F. Nardini, N. Sancisi, M. Conconi, and V. Parenti-Castelli, “Kinematic models of lower limb joints for musculo-skeletal modelling and optimization in gait analysis,” *J. Biomech.*, May 2017.

[2] F. Deligianni, C. Wong, B. Lo, and G.-Z. Yang, “A fusion framework to estimate plantar ground force distributions and ankle dynamics,” *Inf. Fusion*, vol. 41, no. Supplement C, pp. 255–263, May 2018.

[3] C. Wong, Z. Q. Zhang, B. Lo, and G. Z. Yang, “Wearable Sensing for Solid Biomechanics: A Review,” *IEEE Sens. J.*, vol. 15, no. 5, pp. 2747–2760, May 2015.

[4] D. Jarchi, C. Wong, R. M. Kwasnicki, B. Heller, G. A. Tew, and G. Z. Yang, “Gait Parameter Estimation From a Miniaturized Ear-Worn Sensor Using Singular Spectrum Analysis and Longest Common Subsequence,” *IEEE Trans. Biomed. Eng.*, vol. 61, no. 4, pp. 1261–1273, Apr. 2014.

[5] T. Han *et al.*, “Gait Analysis Based on Ear-worn Wireless Motion Sensor,” presented at the 11th IEEE-EMBS International Summer School and Symposium on Medical Devices and Biosensors (MDBS’ 2017), 2017.

[6] D. J. Kuhman, M. R. Paquette, S. A. Peel, and D. A. Melcher, “Comparison of ankle kinematics and ground reaction forces between prospectively injured and uninjured collegiate cross country runners,” *Hum. Mov. Sci.*, vol. 47, pp. 9–15, Jun. 2016.

[7] V. Ramakrishna, T. Kanade, and Y. Sheikh, “Reconstructing 3D Human Pose from 2D Image Landmarks,” in *Computer Vision – ECCV 2012*, 2012, pp. 573–586.

[8] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis, “Sparseness Meets Deepness: 3D Human Pose Estimation From Monocular Video,” presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4966–4975.

[9] X. Zhou, S. Leonardos, X. Hu, and K. Daniilidis, “3D Shape Estimation From 2D Landmarks: A Convex Relaxation Approach,” presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4447–4455.

[10] I. Akhter and M. J. Black, “Pose-Conditioned Joint Angle Limits for 3D Human Pose Reconstruction,” presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1446–1455.

[11] M. Ye, C. Yang, V. Stankovic, L. Stankovic, and A. Kerr, “A Depth Camera Motion Analysis Framework for Tele-rehabilitation: Motion Capture and Person-Centric Kinematics Analysis,” *IEEE J. Sel. Top. Signal Process.*, vol. 10, no. 5, pp. 877–887, Aug. 2016.

[12] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields,” *ArXiv161108050 Cs*, Nov. 2016.

[13] C. Rother, V. Kolmogorov, and A. Blake, “‘GrabCut’: Interactive Foreground Extraction Using Iterated Graph Cuts,” in *ACM SIGGRAPH 2004 Papers*, New York, NY, USA, 2004, pp. 309–314.

[14] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active Appearance Models,” *IEEE Trans Pattern Anal Mach Intell*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[15] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, “Convolutional Pose Machines,” presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4724–4732.

[16] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, “Online Learning for Matrix Factorization and Sparse Coding,” *J Mach Learn Res*, vol. 11, pp. 19–60, Mar. 2010.

[17] N. Sarafianos, B. Boteanu, B. Ionescu, and I. A. Kakadiaris, “3D Human pose estimation: A review of the literature and analysis of covariates,” *Comput. Vis. Image Underst.*, vol. 152, pp. 1–20, Nov. 2016.

[18] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers,” *Found Trends Mach Learn*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[19] H. Xia, J. Xu, J. Wang, M. A. Hunt, and P. B. Shull, “Validation of a smart shoe for estimating foot progression angle during walking gait,” *J. Biomech.*, vol. 61, pp. 193–198, Aug. 2017.