Human Motion Capture Based On Sparse Inertial Sensor

Huailiang Xia *[2],Xiaoyan Zhao †[1,2], Yan Chen *[1], and Zhaohui Zhang *[1,2,3]

*[1]School of Automation and Electrical Engineering, University of Science and Technology Beijing 30# Xueyuan Road, Haidian District, Beijing 100083, China
E-mail: zhaoxiaoyan@ustb.edu.cn
*[2]Shunde Graduate School, University of Science and Technology Beijing Fo Shan 528399, China
E-mail: 18489970952@163.com
*[3]Beijing Engineering Research Center of Industrial Spectrum Imaging, University of Science and Technology Beijing 30# Xueyuan Road, Haidian District, Beijing 100083, China
E-mail: zhangzhaohui@ustb.edu.cn

Abstract. In recent years, human motion capture has a broad market prospect, widely used in film and television animation production, human-computer interaction, virtual reality, sports training, medical rehabilitation and other interdisciplinary disciplines. Unlike the regular motion of mechanical objects, the movement of the human body is complex, changeable and unpredictable, and has a flexibility and variety that no mechanical object has. Therefore, at present, the capture, recognition and reproduction of human posture have become a research hotspot.
Body motion capture usually requires placing sensor markers above and below the joint to calculate joint angles, but this will lead to a large number of body sensors, affecting the performer's normal movements. Using the method of deep learning can reduce the number of sensors, but it requires a large number of training datasets and sufficient computing resources. In this paper, we propose a new deep learning-based approach, which learns data-driven time priors, merges past and future inertial sensor information into superimposed input vectors, which are then fed back to shallow neural networks to estimate human movements. Six inertial measurement unit (IMU) sensors were used to estimate human motion.

Keywords: parameterization, LSTM, deep learning, motion capture

1. INTRODUCTION

Human movement posture measurement and retrieval, is refers to the use by or other types of optical and inertial sensor accurately obtain the body after each moving part attitude data, to collect the posture and movement data calculation and processing, eventually converted to data and visualization can read stance action process of repetition, the main process is measured and the retrieval of two steps.

At present, the mainstream attitude measurement technology usually uses optical sensors such as inertial sensor or camera based on micro-electro-mechanical system. The system based on optical measurement usually sticks a number of cursor points on the designated position of the measured body, and then uses a high-speed camera to capture the cursor point displacement, and then transmits the captured data to the computer, according to the captured movement data for post-processing, and finally completes the whole process of attitude measurement. However, optical attitude measurement has its own shortcomings, for example, the whole equipment of optical attitude measurement is easily affected by occlusion or light interference. For some serious occlusion actions, such as sitting at a table and users walking around, are blocked by objects, which brings a lot of trouble to the post-processing work.

At present, the mainstream inertial sensor motion capture on the market generally adopts 13/17 sensors to be arranged on the whole body [1], which may cause interference and time-consuming due to the large number of sensors. Therefore, the data-driven method based on sparse sensors is more suitable for dynamic data acquisition than the full kinematics method. Using deep learning and a large synthetic data set, the scalability and usability of data-driven approaches to human pose estimation using sparse sensor combinations are improved. At the same time, reducing the number of sensors and their flexible placement on the body could reduce interference with motion estimation and make it easier to use for daily activity monitoring, ergonomics and wearable health. Deep learning models that use sparse sensors to estimate 3d pose depend largely on the correlation of data and the coherence of motion time series. The past and future inertial sensor information is combined into the superimposed input vector using time priori, which is fed back to the shallow neural network to estimate full-body postures [2].

2. SENSOR NODE SETTING

2.1. Inertial sensor

In this paper, mpu9250 nine-axis attitude angle module is used to output angle after filtering and fusion (commonly used XYZ angle), original data of nine-axis sensor and data after Kalman filtering (three-axis acceleration, three-axis gyroscope, three-axis magneto-meter) and quaternion output. The resolution is the degree of 0.01, static attitude measurement accuracy is 0.05, dynamic attitude measurement accuracy is 0.1, high stability. The return rate is 10-200Hz, the transmission

rate is fast, can be real-time transmission. Mpu9250 I2C mode circuit diagram is shown in **Fig.1**.

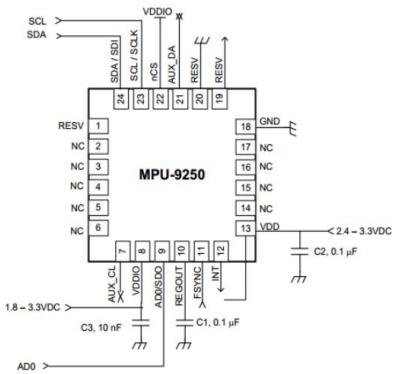


Fig.1 mpu9250 I2C mode circuit diagram

2.2. Wireless transmission module

As the data needs to be transmitted quickly and in real time with a high data rate, WIFI is selected as the data communication mode based on the following comparison. The four kinds of wireless communication technical is as follows shown in the **table.1**.

Table.1 Four kinds of wireless communication technical

Indicator	Bluetooth	WIFI	ZigBee	UWB
Operating frequency band	2.4GHz	2.4GHz	2.4GHz	-
Transmission distance	10-50m	3-100m	10-30m	10m
Peak current	<12mA	>50mA	20-30mA	10mA
Data rate	1Mbps	10Mbps	250kbps	100Mbps
Network supports	Star、P2P	Star、P2P	Star、P2P、Mesh	P2P

USR-wifi232-b2 module is an integrated 802.11b/g/N module, through which traditional serial port devices or MCU controlled devices can easily access WIFI wireless network, so as to achieve network control and management. Protocol conversion is completed inside the module, and two-way transparent transmission of data between serial port and WIFI can be realized through simple Settings.

In this paper, STA-based wireless network is selected as the networking mode. Usr-wifi232-b2 as STA is the most common networking mode, which consists of one route and multiple STAs to form a wireless network, as shown in the following **Fig.2**:

All STAs use the AP as the center of the wireless network, and the mutual communication between the STAs is completed through the AP forwarding. If the AP is already connected to external network, the device data can be transmitted to the server on the external network.

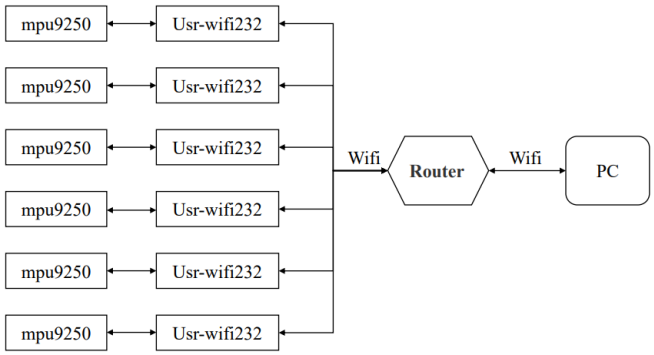


Fig.2 STA-based wireless network

The router sends out WIFI, and the PC is connected to the 232-B2 module wirelessly, so as to realize bi-directional data transmission between the PC and the 232-B2 module as in **Fig.3**.The hardware figure is shown in **Fig.4**.



Fig.3 USR-WIFI232 module



Fig.4 Hardware figure

2.3. Node Layout

Real-time estimation of human posture motion capture using 6 IMUs [3]. The six IMU sensors are located near the left wrist, right wrist, left calf, right calf, lower spine and head. Five sensors located in the lower/upper limbs and head, called leaf sensors, provide orientation measurements, while sensors located in the lower spine are considered root sensors, which are used to predict human posture. Schematic of sensor wear is shown in **Fig.5**.



Fig.5 Schematic of sensor wear

3. SYSTEM IMPLEMENTATION

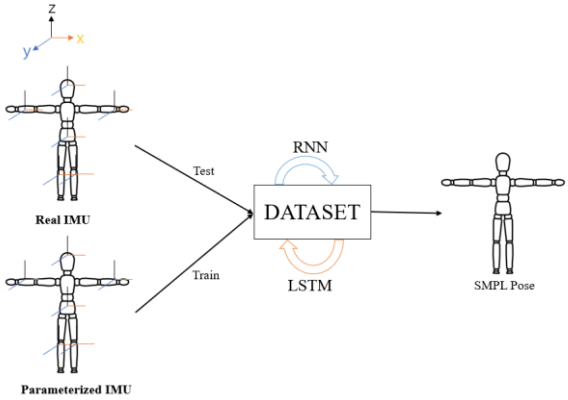
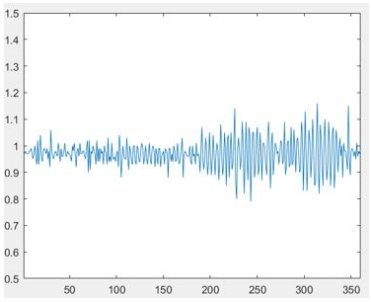


Fig.6 System implementation

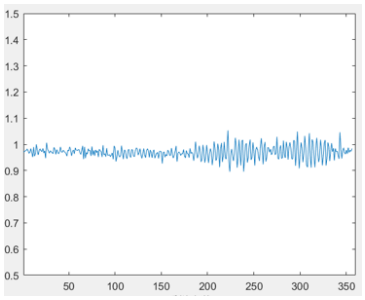
As shown in the **Fig.6**, virtual sensors placed on the SMPL grid synthesize IMU signals. In the middle: IMU signal of recursive neural network was used as input, and the coherence of time before and after movement was used to predict SMPL attitude parameters in combination with RNN and LSTM. Right: Full body motion capture with sparse inertial sensors.

3.1. Data preprocessing

The sampling frequency is set as 50Hz in the program. In order to reduce the impact of noise, the moving average filter is adopted in the measurement process, and the window size is 5. Since each filtered data is calculated according to the mean value of the sampling point and the preceding four points, weighted average is adopted to ensure that the filtered data is as close to the true value as possible. The time weight of the filtered data starts from the first data and increases as it gets closer to the sampling point. The whole test scheme for filtering is to place the sensor horizontally and do jitter. The moving average filter can effectively suppress the intensity of acceleration change and make the acceleration change smoother. The abscissa is the number of sample points and the ordinate is the acceleration as in **Fig.7.**



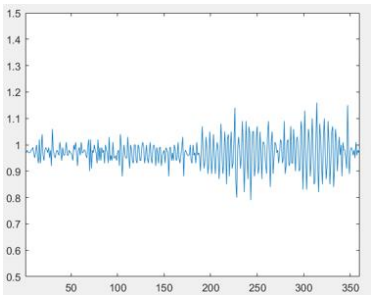
(a)before filtering



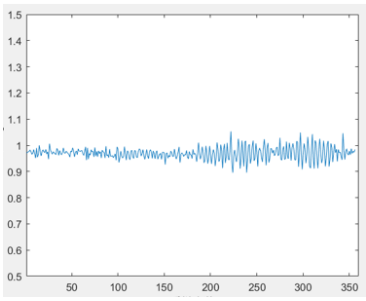
(b) after filtering

Fig.7 Accelerometer filtering

The moving average filtering is also used in the preprocessing of gyroscope data. (c) is the graph before filtering, and (d) is the graph after filtering. The abscissa is the number of sample points and the ordinate is the angular acceleration as in **Fig.8**.



(c)before filtering



(d) after filtering

Fig.8 Gyroscope filtering

3.2. IMU Dataset

In deeping study of human movement requires extensive data sets, including many topics, different movements and activities of the human body, at different times and speeds, and the range of motion of human joints. While many motion capture datasets are available, these datasets do not use standardized 3D skeletons, and only a few contain data from inertial sensors along with associated calibration. In this paper, the DIP-IMU[4], a large synthetic data set developed based on SMPL, is used to train the model. At the same time, the forward kinematics of the root sensor is used to "synthesize" the body segmentation direction, where the direction of the root sensor is simply obtained from the SMPL attitude. On the assumption that the inertial sensor body frame is aligned with the segments of the body (that is the sensor-to-segment alignmentis consistent), then the "composite" segment orientation also represents the sensor orientation. The three kinds of IMU dataset are as follows shown in the **table 2**.

Table.2 IMU Dataset

Name	AMASS	Total-Capture	DIP-IMU
Type	Synth	Real	Real
Mode	M	M,I,R	I
Frames	9,730,526	179,176	330,178
Minutes	2,730	50	92
Subjects	603	5	10
Motions	11,234	46	64

("M" is for MoCap, "I" is for IMU, and "R" is for RGB images)

Real IMU sensor datasets are used to test models trained on "synthetic" enhanced datasets. The DIP-IMU dataset provided as open source by Huang et al[5], the Totalcapture dataset is a public dataset with IMU orientation and calibration[6].

3.3. Recalibration of input data

To reduce error, the data obtained need to be properly calibrated before model input. The orientation of the root of the human bone changes with the subject's movement, but the rotation and translation of the root of the whole body posture remains the same. The orientation of the five sensors (left wrist, right wrist, left leg, right leg, and head) is standardized W.R.T by[7]:

T and B_i respectively represent the SMPL reference frame and the reference frame connected to the bone segment, while R_0 is the reference frame fixed to the root joint. $R_t^{R_0B_i}$: $B_i \rightarrow R_0$is the rotation matrix from the root frame to the corresponding skeleton reference frame at time T. The same convention applies to other rotation matrices representing direction. Obviously, we need 3D orientation from root to SMPL and from bone to SMPL to use the "constant" orientation input of the computational model.

Dynamic acceleration and magnetic interference should also be considered in sensor fusion. Azimuth estimation should be robust to these disturbances, otherwise they need to be estimated and compensated. At the same time the reference frame S_i of the sensor is not always aligned with the reference frame B_i of the body segment. Therefore, a sensor is introduced to segment the calibration matrix $R^{B_iS_i}$: $B_i \rightarrow S_i$. The subject's sensor is put on, powered on, the sensor works normally, and the subject's Tshape calibration posture is completed.

Due to residual startup and calibration errors in the sensor, each sensor after sensor fusion provides a direction $R^{I_iS_i}$, and the inertial frame I_i of the sensor is offset from the actual body inertial frame G. Therefore, when starting each sensor, we also need to obtain GR^{GI_i}: $I_i \rightarrow \mathrm{G}$. R^{TG} is obtained from the root sensor during initial calibration of the human body model, and then complete calibrationof the actual sensor data according to the following steps[8]:

$$\begin{aligned} R_t^{R_0B_i} &= \left(R_t^{TR_0}\right)^{-1} R_t^{TB_i} \\ R_t^{GS_i} &= R^{GI_i} R_t^{I_iS_i} \\ R_t^{GB_i} &= R_t^{GS_i} (R^{B_iS_i})^{-1} \\ R_t^{R_0B_i} &= \left(R_t^{GR_0}\right)^{-1} R_t^{GB_i} \\ \bar{R}_t^{R_0B_i} &= R^{TG} R_t^{R_0B_i} \end{aligned} \tag{1}$$

3.4. Input and output parameterization

The most commonly used method to express attitude is yaw Angle, pitch Angle, roll Angle these three attitude angles. The three-axis accelerometer can obtain the acceleration of the measuring node in three directions and calculate the attitude Angle of the measuring node based on the influence of gravity, but the result is not accurate. Similarly, the gyroscope can also measure the real-time angular acceleration of the measuring node, obtain the Angle through integration, and estimate the attitude roughly, but the results are still not accurate. Considering that the accelerometer has too much high-frequency noise in the measurement process, but its dynamic response is poor, while the measurement effect is ok in the static state. Gyroscopes, on the other hand, have static response but accumulated errors over time, but have excellent dynamic response. Therefore, an algorithm is needed to combine the advantages of both to solve the attitude.

Quaternion method is usually used to solve the attitude for the convenience of calculation, and then the quaternion is converted into the attitude Angle to output. AHRS attitude calculation usually includes the following seven steps:

1)Quaternion initialization: the accelerometer calculates the current attitude Angle according to the gravity components in each direction, and then converts the attitude Angle into the initial quaternion through the formula.

2)Acceleration normalization:

$$\begin{cases} \text{norm} = \sqrt{ax^2 + ay^2 + az^2} \\ ax = \dfrac{ax}{\text{norm}} \\ ay = \dfrac{ay}{\text{norm}} \\ az = \dfrac{az}{\text{norm}} \end{cases} \quad (2)$$

3)The quaternion equivalent cosine matrix is listed and the gravity component is extracted by multiplying with the gravity matrix.

4)Quaternion equivalent matrix is as follows:

$$M = \begin{bmatrix} q_0^2 + q_1^2 - q_2^2 - q_3^2 & 2(q_1q_2 - q_0q_3) & 2(q_1q_3 + q_0q_2) \\ 2(q_1q_2 + q_0q_3) & q_0^2 - q_1^2 + q_2^2 - q_3^2 & 2(q_2q_3 - q_0q_1) \\ 2(q_1q_3 - q_0q_2) & 2(q_2q_3 + q_0q_1) & q_0^2 - q_1^2 - q_2^2 + q_3^2 \end{bmatrix} \quad (3)$$

5)Gravity matrix [0 0 1] can be multiplied by m-matrix to get gravity component:

$$\begin{cases} vx = 2(q_1q_3 - q_0q_2) \\ vy = 2(q_2q_3 + q_0q_1) \\ vz = q_0^2 + q_1^2 + q_2^2 + q_3^2 \end{cases} \quad (4)$$

6)Using the vector cross product to find the attitude error:

$$\begin{cases} ex = vy \times az - vy \times ay \\ ey = vz \times ax - vx \times az \\ ez = vx \times ay - vy \times ax \end{cases} \quad (5)$$

According to the error of the proportional integral operation, update the gyroscope.

7)According to Runge-Kutta method update quaternion ("ω" _"x" "ω" _"y" "ω" _"z" for the angular velocity of the fused) :

$$\begin{pmatrix} q_0 \\ q_1 \\ q_2 \\ q_3 \end{pmatrix}_{t+T} = \begin{pmatrix} q_0 \\ q_1 \\ q_2 \\ q_3 \end{pmatrix}_t + \frac{T}{2} \begin{pmatrix} -\omega_x q_1 - \omega_y q_2 - \omega_z q_3 \\ \omega_x q_0 - \omega_y q_3 + \omega_z q_2 \\ \omega_x q_3 - \omega_y q_0 - \omega_z q_1 \\ -\omega_x q_2 - \omega_y q_1 - \omega_z q_0 \end{pmatrix} \quad (6)$$

Transform the quaternion to the attitude angle output.

Overall **Fig.9** of attitude solution:

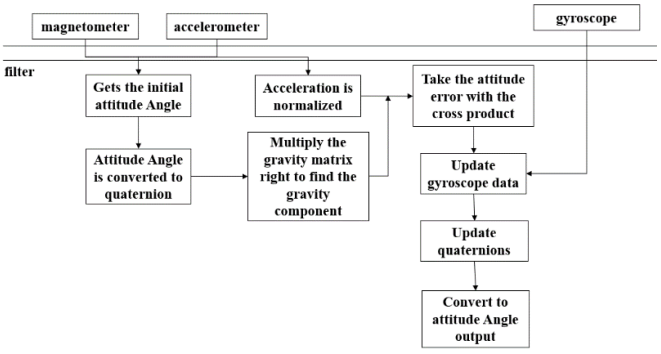


Fig.9 chart of attitude solution

Since a complete 3×3 rotation matrix is overparameterized, any right-handed cross product of two rows and columns results in a third row, column. The first two columns are chosen for the representation because it uniquely identifies the remaining columns. Then, γ, θ and ψ are yaw angle, pitch angle and roll angle respec-tively. r11、r21、r31、r32 and r33 are components of 3×3 rotation matrix. The whole rotation matrix can be obtained by five components of 3×3 rotation matrix.The six-parameter notation is a simple extension of the fact that yaw, pitch, and roll angles can be easily derived from the rotation matrix, as shown below [9]:

$$\begin{aligned} \gamma &= \tan^{-1}\left(\frac{r_{21}}{r_{11}}\right) \\ \theta &= \tan^{-1}\left(-\frac{r_{31}}{\sqrt{r_{32}^2 + r_{33}^2}}\right) \\ \varphi &= \tan^{-1}\left(\frac{r_{32}}{r_{33}}\right) \end{aligned} \quad (7)$$

3.5. Model architecture

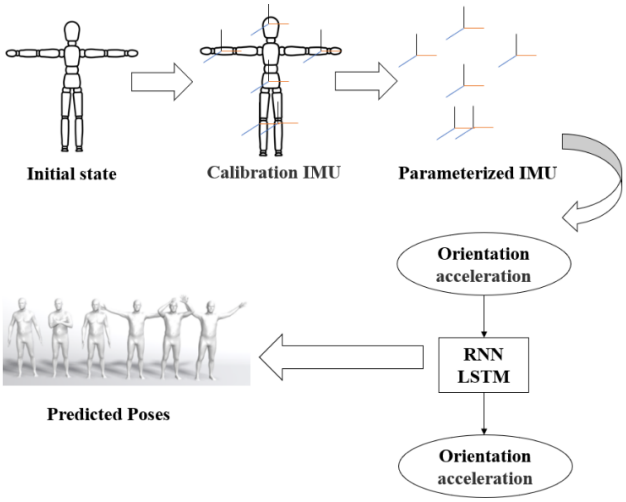


Fig.10 System implementation

As it is shown in the **Fig.10** above, the subject wears the inertial sensor and starts for several seconds to complete the calibration. The direction and acceleration were obtained by parameterization of IMU collected data, and the simulated attitude parameters were outputed by RNN and LSTM models to construct SMPL model to complete attitude prediction.

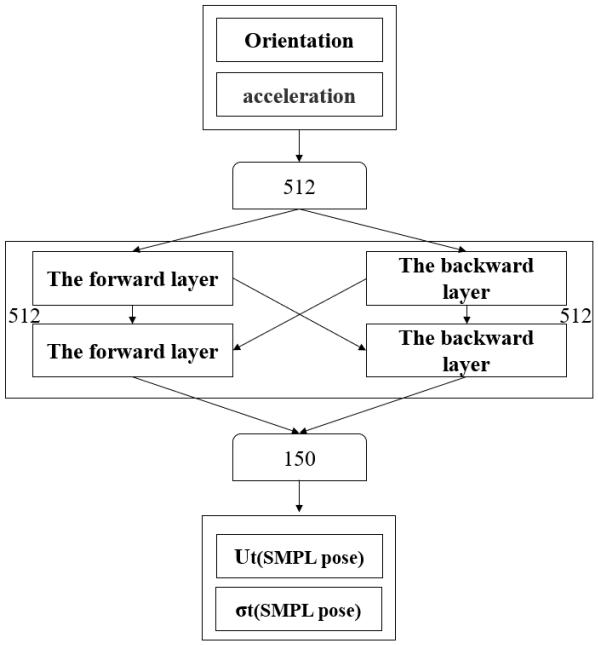


Fig.11 the details of the model architecture

The **Fig.11** above is the details of the model architecture. A bidirectional LSTM autoencoder is proposed. Two stacked bidirectional LSTM hidden layers (forward layer and back layer) are 512 in size. The attitude Ut and the standard deviation σt in SMPL are obtained from its linear output. The next work will be data-training around Bi-LSTM and Bi-RNN.

4. CONCLUSION

Our work proposes a recursive model of six inertial measurement units estimate three-dimensional human posture, and a pure inertial sensor capture system can avoid the problem of camera occlusion in vision. Testing and training will be conducted by using the Large Mocap database, Total-Capture, DIP-IMU database. Direction and acceleration data will be synthesized for IMU using SMPL body model. Using the information before and after the movement time, the Bi-RNN and Bi-LSTM model would be used to propagate the information forward and backward in time, and to reconstruct the human postures.

In the future, sparse sensor motion capture will reduce joint angle error, improve system rapidity, and solve the problem of capturing multi-person interaction and environmental interaction. A small number of wearable sensors for real-time attitude estimation will play a key role in emerging interactive technologies such as VR and AR.

Acknowledgements

This work was supported by the National Key Research and Development Project (2020YFF0304704) and Scientific and Technological Innovation Foundation of Shunde Graduate School (BK19BE020).

REFERENCES:

[1] Xsens Homepage—Motion Capture Products-MVN Awinda, MVN Link. Accessed: Sep.1, 2020.

[2] S. Vlutters, ‘‘Long short-term memory networks for body movement estimation,’’ M.S. thesis, Dept. EEMCS, UT, Enschede, The Netherlands,2016.

[3] Ijaz Akhter and Michael J Black. 2015. Pose-conditioned joint angle limits for 3D human pose reconstruction. In Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition. IEEE, 1446–1455.

[4] Y. Huang, M. Kaufmann, E. Aksan, M. J. Black, O. Hilliges, and G. Pons-Moll, ‘‘Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time,’’ ACM Trans. Graph., vol. 37, no.6, pp. 1–15, Jan. 2019.

[5] C. Yi, et al., "Estimating Three-Dimensional Body Orientation Based on an Improved Complementary Filter for Human Motion Tracking," Sensors, vol. 18, p. 3765, 2018.

[6] M. Trumble, A. Gilbert, C. Malleson, A. Hilton, and J. Collomosse, ‘‘Total-capture: 3D human pose estimation fusing video and inertial sensors,’’ in Proc. BMVC, London, U.K., 2017, pp. 1–13.

[7] R.J. Hyndman and Y. Khandakar, “Automatic time series forecast-ting: the forecast package for R,” Journal of Statistical Software, vol. 26,no. 3, pp. 1–22, 2008.

[8] H. T. Butt, M .Pancholi, M.Musahl, P . Murthy, M. A. Sanchez, and D. Stricker, ‘‘Inertial motion capture using adaptive sensor fusion and joint angle drift correction,’’ in Proc. FUSION, Ottawa, ON, Canada, 2019,pp.1–8.

[9] Y . Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, ‘‘On the continuity of rotation representations in neural networks,’’in Proc.CVPR,Long Beach,CA,USA,Jun.2019,pp.57455753.