Online Research @ Cardiff

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: https://orca.cardiff.ac.uk/id/eprint/152710/

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Chen, Xiaowei, Jiang, Xiao, Zhan, Lishuang, Guo, Shihui, Ruan, Qunsheng, Luo, Guoliang, Liao, Minghong and Qin, Yipeng ORCID: https://orcid.org/0000-0002-1551-9126 2022. Full-body human motion reconstruction with sparse joint tracking using flexible sensors. ACM Transactions on Multimedia Computing, Communications, and Applications 10.1145/3564700 file

Publishers page: https://doi.org/10.1145/3564700
<https://doi.org/10.1145/3564700>

Please note:
Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See
http://orca.cf.ac.uk/policies.html for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.

information services
gwasanaethau gwybodaeth

Full-body Human Motion Reconstruction with Sparse Joint Tracking Using Flexible Sensors

XIAOWEI CHEN, School of Informatics, Xiamen University, China
XIAO JIANG, School of Informatics, Xiamen University, China
LISHUANG ZHAN, School of Informatics, Xiamen University, China
SHIHUI GUO*, School of Informatics & Jiujiang Research Institute, Xiamen University, China
QUNSHENG RUAN, School of Informatics, Xiamen University, China
GUOLIANG LUO, East China Jiao Tong University, China
MINGHONG LIAO, School of Informatics, Xiamen University, China
YIPENG QIN, School of Computer Science and Informatics, Cardiff University, UK

Human motion tracking is a fundamental building block for various applications including computer animation, human-computer interaction, healthcare, etc. To reduce the burden of wearing multiple sensors, human motion prediction from sparse sensor inputs has become a hot topic in human motion tracking. However, such predictions are non-trivial as i) the widely adopted data-driven approaches can easily collapse to average poses. ii) the predicted motions contain unnatural jitters. In this work, we address the aforementioned issues by proposing a novel framework which can accurately predict the human joint moving angles from the signals of only four flexible sensors, thereby achieving the tracking of human joints in multi-degrees of freedom. Specifically, we mitigate the collapse to average poses by implementing the model with a Bi-LSTM neural network that makes full use of short-time sequence information; we reduce jitters by adding a median pooling layer to the network, which smooths consecutive motions. Although being bio-compatible and ideal for improving the wearing experience, the flexible sensors are prone to aging which increases prediction errors. Observing that the aging of flexible sensors usually results in drifts of their resistance ranges, we further propose a novel dynamic calibration technique to rescale sensor ranges, which further improves the prediction accuracy. Experimental results show that our method achieves a low and stable tracking error of 4.51 degrees across different motion types with only four sensors.

CCS Concepts: • **Computing methodologies → Motion capture**.

Additional Key Words and Phrases: flexible sensors, sparse signal processing, temporal convolutional network, median pooling

1 Introduction

Human motion tracking is widely used in the animation industry [19, 43], computer games [30, 47], human-computer interaction [3] and medical rehabilitation applications [21, 24, 28, 41]. Currently, optical solutions and inertial measurement units (IMUs) are the most popular approaches to tracking human motion that have mature applications [18, 31]. Between them, optical solutions suffer from bad environmental conditions (*e.g.* occlusions or poor lighting [5]) and are restricted by the clothes worn for the placement and visibility of optical markers [22]; inertial measurement unit (IMU) bypasses the above limitations and stands out as a better alternative [48]. For example, Yasuo and Hirotaka proposed a method that can perform 3D motion tracking with 32 IMUs [14]. Although

*Corresponding author: guoshihui@xmu.edu.cn.

Authors' addresses: Xiaowei Chen, School of Informatics, Xiamen University, Xiamen, Fujian, China, 361005; Xiao Jiang, School of Informatics, Xiamen University, Xiamen, Fujian, China, 361005; Lishuang Zhan, School of Informatics, Xiamen University, Xiamen, Fujian, China, 361005; Shihui Guo, School of Informatics Xiamen University, Fujian, China, 361005 and Jiujiang Research Institute, Xiamen University, Jiujiang, Jiangxi, China, 332105; Qunsheng Ruan, School of Informatics, Xiamen University, Xiamen, Fujian, China, 361005; Guoliang Luo, East China Jiao Tong University, Nanchang, Jiangxi, China, 330013; Minghong Liao, School of Informatics, Xiamen University, Xiamen,Fujian, China, 361005; Yipeng Qin, School of Computer Science and Informatics, Cardiff University, Cardiff, UK.

effective, their method based on a dense arrangement of IMUs is intrusive. The commercial system also employs even more than 10 IMUs [32]. Nonetheless, measuring by multiple flexible sensors is time-consuming and laborious work, because of the long-wearing time and the expensive cost for the facilities, and it is also inconvenient for users to wear multiple flexible sensors [12]. Recently, researchers have validated the feasibility of using a small set of sensors to track human motion [40]. Even though this method adopts 6 IMUs, their method requires a heavy computational cost, needing offline optimization of a non-convex problem over the entire sequence; The following work, DIP [12], achieves real-time performance with higher tracking accuracy with a bidirectional RNN. Their approach also uses 6 IMUs. However, their frame rate is 30 fps, which is not sufficient to capture fast movements. TransPose [46] leverages 6 IMUs to realize fast and realistic movements. With the same amount of IMUs, TIP [13] further improved the tracking result with the structure of the transformer. While researchers have adopted 6 IMUs to realize good tracking results, it is meaningful to explore tracking human motion with less number of sensors for better usability. Compared to works based on 6 sensors, motion tracking based on 4 sensors set more constraints for tracking motions currently. Schwarz et al. [35] proposed Gaussian Process regression to reconstruct a full-body human pose. However, their method is limited to 6 types of motion. Ha et al. [9] utilized 2 pressure sensing platforms and a hand tracking device to track the user's locomotion. But their method requires the users' feet to be bound to a small area. Thus, we propose to leverage 4 sensors to avoid the inconvenience of wearing multiple sensors, decrease cost and facility requirements, and achieve human motion tracking with good usability [46]. Considering that IMUs are easily affected by electromagnetic interference [11] and less flexible as they require a drift-avoidance configuration that prevents position deviation and inaccurate direction measurement [2, 33], we employ four flexible sensors to track human motion.

Albeit flexible sensors have the characteristics of good malleability, robustness in indoor and outdoor environments, bio-compatibility, and supporting long-time monitoring, they detect less information than IMUs without the fusion of gyroscopes, accelerometers and magnetometers and ordinarily show high nonlinearity and hysteresis in response [26]. Besides, since flexible sensors are attached to the clothes, it inevitably deforms as the body moves, causing sensor aging. To realize restoration from the low-dimensional inputs of flexible sensors to high-dimensional human

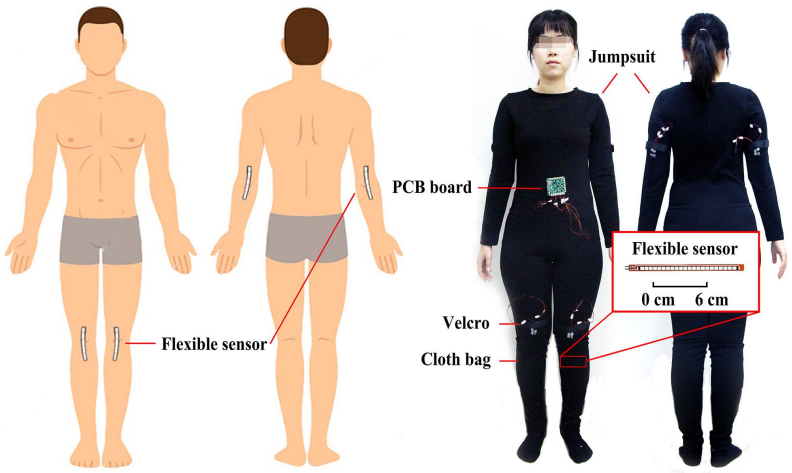


Fig. 1. Left: sensor locations. Right: jumpsuit configuration.

motion information, we employed four flexible sensors (placed in a jumpsuit lastly (Figure 1)) and adopted a data-driven method which is a Bi-LSTM neural network to regress human motion. Apart from that, considering the aging sensors, we utilized average pooling layers and a min pooling layer to estimate the changing baseline (the readings when the sensor is static without any bending caused by human motion) and calibrate them. To cope with some unnatural shaking movements, we leveraged a median pooling layer to smooth the movements. Our method improves the wearing experience of the motion-tracking flexible sensors' suit while keeping the stable tracking of human daily activities through only four sensors. Besides, our work supports the aging and replacement of flexible sensors. Our core contribution lies in three aspects:

- A data-driven method based on flexible sensors to track human daily activities. With several pooling layers, our method improves the robustness against sensor aging and sensor replacement.
- An approach to ease the jitters of the predicted output and deal with the sparse input. The Bi-LSTM model and the median pooling not only help to reduce the jitters of the output but also generalize well on the testing set. The tracking error is 4.51 degrees with only four sensors on the testing set.
- A system prototype based on four flexible sensors to track human daily activities. This system combines the data collecting, recording and visualizing modules, establishing the technical route from the flexible sensor to visual presentation.

2 Related Work

Full-body motion tracking has a long history. In this work, we focus on sparse motion tracking and will review its two mainstream solutions, *i.e.* multi-type sensors and single-type sensors, as follows.

2.1 Sparse Motion Tracking based on Multi-type Sensors

The multi-type sensor approach was proposed to make the most of the advantages of different types of sensors, which minimizes the number of sensors placed on the human body and is thus less intrusive [46]. For example, Zhang et al. [49] explored the tracking of human motion with 3 optical cameras attached to the rear wheel of a bicycle, 2 IMUs on the bicycle and 4 small size tri-axial gyroscopes on the human body under specific motion constraints. Andrews et al. [1] proposed another human motion tracking method using a combination of 6 IMUs and 5 optical marker sensors. Guzov et al. introduced a joint optimization [8] which integrates camera localization, 8 IMUs-based tracking and scene constraints, resulting in smooth and accurate human motion estimation. To further relax experimental constraints, and improve accuracy and user experience, some works resort even more to easily-accessible optical data and make less use of IMUs. For example, [39] combined 4-8 cameras and 5 IMUs to track human motion. [50] adopted a single depth camera and 8 IMUs to capture human motion. [38] formulated the tracking as a novel graph-based optimization problem that associated the 2D pose detection to only the persons equipped with IMUs in each frame.

In summary, most multi-type sensor methods use optical sensors for higher accuracy and better user experiences, together with IMUs to get accurate limb orientations that can be challenging for pure optical systems under fast motion or occlusion scenarios. Even though researchers have made efforts to reduce the hardware setup, the facility requirements (eg. the equipment maintenance cost and user familiarity with the cost) of such an integrated technical route are higher than a single-type sensor system. The method combined with the optical system is susceptible to occlusions and lightning. Therefore, we follow the single-type sensor approach that is much easier to deploy in real-world scenarios.

Table 1. Overview of existing single-type sensor solutions for human motion tracking.

Passage	Year	Types and quantity of sensors	Limitation
Tautges et al.	2011	5 accelerometers	i) The acceleration data are noisy. ii) The space of possible postures was huge.
Schwarz et al.	2009	4 IMUs	Limited to only 6 types of motion.
Ha et al.	2011	2 pressure sensing platforms and a hand tracking device	Users' feet are required to be bounded to a small area.
Mousas	2017	1 IMU	Can only reconstruct simple periodic motion.
von Marcard et al.	2017	6 IMUs	Their method does not support real-time tracking.
Yang et al.	2021	4 IMUs on upper limbs	Do not support lower body tracking.
Huang et al.	2018	6 IMU sensor on limbs and pelvis	Relatively low accuracy (17.54 degrees).
Yi et al.	2021	6 IMU sensor on limbs and pelvis	The sensor number and tracking error can be further reduced.
Jiang et al.	2022	6 IMU sensor on limbs and pelvis	The sensor number and tracking error can be further reduced.

2.2 Sparse Motion Tracking based on Single-type Sensors

To extend the applications of human motion tracking beyond laboratory settings, massive efforts were made to reduce the number of sensors used that not only reduce costs but also increase wearing comfort. Table 1 shows an overview of previous single-type sensor solutions. Specifically, Tautges et al. [36] used 4 accelerometers to track human motion. They leveraged a cross-domain retrieval procedure to build up a lazy neighborhood graph in an online fashion. However, their method is relatively less accurate as the acceleration data obtained by their sensors are noisy and the huge space of possible postures makes it difficult to train a machine learning model reliably. Improving the type of sensors, Schwarz et al. [35] proposed using Gaussian Process regression to reconstruct full-body human pose using only 4 IMUs. Since the models are trained on specific movements of individual users for each activity of interest, which greatly limits its applicability, their method is limited to 6 types of motion. Ha et al. [9] utilized 2 pressure sensing platforms and a hand tracking device to track the user's locomotion. The adopted method based on ground reaction forces and cascade ANN makes the process effective. However, the users' feet are bound to a small area. Mousas et al. [25] employed Hidden Markov Model (HMM) to reconstruct human motion with only one IMU. However, their method can only be used to capture periodic motions. Marcard et al. [40] exploited a statistical body model and jointly optimized pose over multiple frames to fit both orientation and acceleration data. Since they adopted an iterative optimization-based method, their method is offline that does not support real-time tracking. Yang et al. [44] introduced a deep neural network (DNN) based method for real-time prediction of the lower body pose only from the tracking signals of the upper-body joints with an average error of 8.53 degrees. As they mainly placed sensors on the upper limb, their tracking result performs not well on the lower body. Even though those methods deal with human motion reconstruction with sparse sensors, they are not applicable to realistic scenarios since they set limits on tracking only several kinds of motions or the tracking accuracy should be further improved. Besides, the sensor number and placement

should be seriously judged. There are some researchers adopting 6 IMUs with sensor placement that is enough to produce good tracking results. Huang et al. [12] placed the sensors on limbs and pelvis. They proposed a deep neural network using 6 IMUs that was trained with a novel loss function based on normal distributions. Even though they produced results that were smooth and generally without penetrating the human model, the average joint angle error can be further improved. With the same sensor placement, Yi et al. [46] proposed a supporting-foot-based method and an RNN-based method to robustly solve human motion tracking as well as global translations. Their confidence-based fusion technique achieved a better result than [12]. The state-of-the-art method [13] leveraged an attention-based deep learning neural network together with a physics-based learning objective to predict "stationary body points" to track human motion with 6 IMUs (The placement is the same as [46] and [12]) and their neural network is easy to implement and supports fast running.

As above-mentioned, it can be observed that researchers mainly adopt IMUs to track human motion due to their small size, low cost, and that they can be easily configured outside a laboratory environment [7, 16, 20, 29, 49]. Nonetheless, IMUs have some inherent problems: i) although the number of sensors used is reduced, they are inflexible and thus still intrusive for human motion; ii) they are prone to be affected by the electromagnetic environment; iii) the data captured by the sensors can be quite noisy: the motions are occasionally exaggerated too much because of the errors in the measured acceleration. Addressing these limitations, in this work, we propose to use flexible sensors instead of IMUs for human motion tracking, which yields better user experiences. Existing solutions either adopt dense flexible sensors [15] that are not user-friendly or focus on tracking local motion [4, 15, 21, 23]. In contrast, our method uses only 4 flexible sensors but can track 48 types of full-body human motion covering most of those in our daily lives. Thanks to the sparse flexible sensors, our method enables a better user experience without sacrificing motion tracking accuracy.

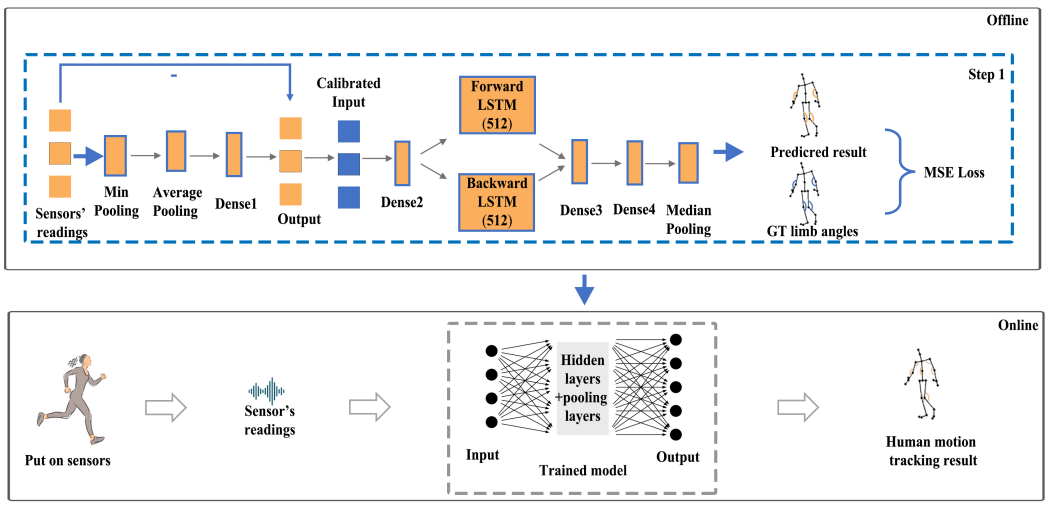


Fig. 2. Method overview. Our framework includes an offline training stage and an online inference stage: in the offline training stage, we trained a concatenation of a bidirectional LSTM neural network with pooling layers to learn the relationship between sensor signals and human motion data; in the online inference stage, we applied the trained model to predict motion from sensor readings in real-time.

3 Method

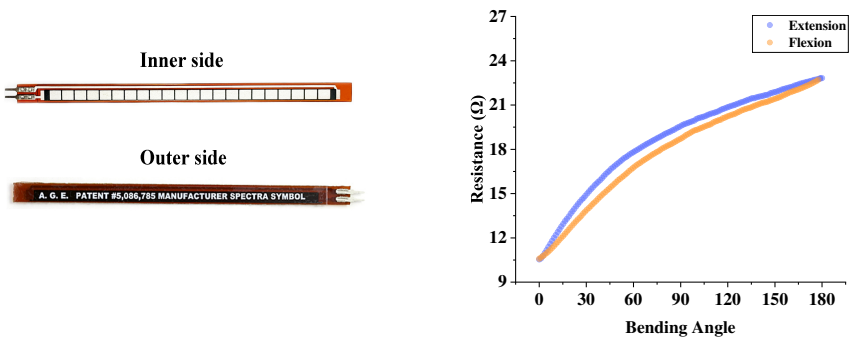
3.1 Method Overview

As Fig. 2 shows, our solution consists of an offline training stage and an online inference stage. In the offline training stage, we first guided each user to wear our jumpsuit and collected paired sensor data (via our PCB board) and motion data (via a Qualisys motion-capture system at a rate of 100 fps). During training, we fitted the training data obtained above with a bidirectional LSTM neural network (Bi-LSTM) [12] that learns the mapping between sensor signals and full-body human joint angles. To mitigate the sensor aging problem, we added a sequence of min and average pooling layers to calibrate aged flexible sensors. We also adopted a median pooling layer [17] to smooth the prediction, thereby decreasing the jitters. In the online inference stage, the predicted motion can be obtained in real-time by feeding the sensor readings through the Bi-LSTM model.

3.2 Prototype

Our work designs and develops a prototype containing a jumpsuit, flexible sensors and PCB board. We will introduce each of them below.

Jumpsuit Design Before we design the suit, we conducted a pilot study to decide the sensor placement given that the quantity of sensors is 4. Based on [46] and [13] that places sensors on the pelvis and limbs, we collected joint motion with an optical system and adopted a Bi-LSTM neural network to validate the optimal placement. The result shows placing sensors on limbs (Y-axis of Forearm and X-axis of Legs) will achieve the best tracking result. Thus, as a hardware prototype of our sparse-joint motion tracking solution, we placed four flexible sensors on the four joint positions (X-axis of two knees and Y-axis of two elbows) of a jumpsuit that can tightly fit different user bodies (Figure 1). Specifically, we reduced harmful sensor displacement by placing the sensors in four long cloth bags which are sewn on the jumpsuit. These sensors were secured on the cloth bags using velcros and hot-glue balls placed on wires. Note that we intentionally made the connection between a sensor and a wire pluggable as it facilitates: i) washing; ii) sensor replacement upon damage; iii) unnecessary damage caused by excessive bending and collisions of sensors when wearing. The conductor is a single strand of tinned copper conductor with an insulating layer, which was welded to the sensor at a temperature of 350 degrees. We placed the circuit board in front of users to facilitate interactions during data collection. The conductor was also designed to be pluggable



(a) (b)

Fig. 3. (a) The inner and the outer sides of our flexible sensor. (b) Sensor resistance vs. bending angles during extension and flexion.

against the circuit board with its end wrapped in plastic. Without sacrificing performance, we increased the aesthetic of our prototype by concealing most of the wires in either the garment's seam lines or the double-layer cloth bags attached to the seam lines.

Flexible Sensor Characteristics The sensors we adopted are resistive, bought from Spectra Symbol[1] (Fig. 3a). As Fig. 3b shows, although the sensor resistance monotonically rises when the side on which the grid is located (the inner side) is extended outward, the relationship between the resistance and the bending angle is non-linear, which makes it difficult to analytically derive their relationships. On the other hand, the resistance-bending relationships are different in the extension and flexion processes, which makes it challenging to learn a regression model. Since only the inner side of the flex sensor is designed to change dramatically with the bending, the inner side must be put outwards when put on.

PCB Board We designed a PCB board (the outlook and design are illustrated in Fig. 4a and Fig. 4b respectively) for sensor data collection. Specifically, our PCB board saves the collected sensor data to an SD card every 0.05 seconds. During data collection, we employed a multi-channel voltage divider to select the channels in order. The reference voltage is obtained via a pair of uniform resistance units that are depicted in the left bottom of Fig. 4b. We applied the Wheatstone bridge structure to compute the difference between the voltage of each sensor and reference voltage ($V_{ref} = V_{CC}/2$), where V_{CC} (Voltage Common Collector) denotes the access voltage of the circuit. Ignoring the effect of the low-pass filter, the input voltage to the digital-to-analog conversion can be defined as follows:

$$Vadc_{in} = \left(\frac{V_{CC} * Rsensor_i}{R_i + Rsensor_i} - V_{CC}/2\right) * Gain, \tag{1}$$

where R_i represents the divider resistor, $Rsensor_i$ denotes the resistance of the i^{th} flexible sensor, and $Gain$ indicates the magnification factor of the amplifier unit. The amplified voltage measurements are handled by a low-pass filter with a bandwidth of 300Hz. Eventually, the output signal is converted to a digital form within [0, 4096].

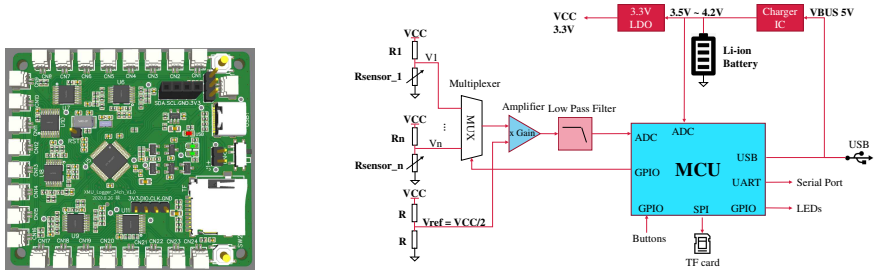


Fig. 4. (a) Outlook and (b) design of our PCB board.

[1] https://www.spectrasymbol.com/product/flex-sensors/

3.3 Data Collection and Preparation

Data Collection We invited ten subjects, including four males and six females, whose ages ranged from 19 to 41. With their consent, we asked them to put on the jumpsuits, install the sensor and the PCB board to the corresponding positions, and turn on the PCB board switch. Then, we simultaneously collected sensor data via the PCB board and human motion data using our optical measuring system (i.e. Qualisys Motion Capturing System). For human motion data, we only collected those of arm joints, forearm joints, upper leg joints, leg joints and spine joints (including all three spine joints). We did not collect the motion data of the head joint, thumb joints, finger joints, foot joints, hand joints, neck joints and the rotation of hips as the movements of these joints are independent of that of other joints. For example, a person's head can turn to the left or right independently while he is walking. The collected data were aligned into (sensor data and motion) pairs according to the system time and relative coefficient of the two collected sequences. Every subject was asked to repeat a set of pre-defined motions twice (listed in the appendix). To compare our solution with that using IMUs, we further asked two of our subjects to wear clothes with IMUs on top of our jumpsuit. The data collection of IMUs was synchronized with our sensors and optical measuring system.

Dataset Preparation Our dataset combined with optical measuring data and resistance data contains about 128,000 sequences of motion data (50Hz). For each action of each subject, we separated its data in a 0.1:0.1:0.8 split as the testing set $\mathcal{D}_{TE}$, the validation set $\mathcal{D}_{VA}$ and the training set $\mathcal{D}_{TR}$ respectively.

3.4 Model Implementation and Training

3.4.1 Network Architecture and Training Details We borrowed the Bi-LSTM network architecture comprising one hidden layer from [12] and implemented it using PyTorch [27]. Specifically, we used linear input and output layers together with one bidirectional LSTM layer containing 512 hidden units. We stopped the training when the decreasing of the average tracking error fell below 0.1 degrees over 10 consecutive epochs. We adopted an Adam optimizer with an initial learning rate of 0.001, decaying by a factor of 0.1 after every 10 epochs. To avoid the exploding gradient problem that damaged the training, we applied gradient clipping. Besides, we set the batch size to 64.

3.4.2 Dynamic Calibration for Sensor Aging Mitigation During data collection, we observed irreversible deformation of flexible sensors with bending (*i.e.* aging). As a result, the resistance values of our flexible sensors tend to increase with the collection, that is, the sensor baseline is increasing. This phenomenon can confuse the neural network during training.

To mitigate the influence of such sensor aging, we propose a novel dynamic calibration module:

$$Y_{out} = Y_{in} - (\mathrm{LN}(\mathrm{ave}(\mathrm{min}(Y_{in})))) \tag{2}$$

where Y_{in} denotes the normalized sensor readings with time sequence (the sensor readings are normalized with min-max normalization). Specifically, we adopted a sliding window to deal with the raw sensor readings. The sequence length is 100. The sliding window moved by two steps every time. Y_{out} denotes the calibrated sensor readings, LN is a linear layer, min is a min pooling layer and ave is an average pooling layer. The kernel sizes of min- and average-pooling layers are set to 25 and 10, respectively; the padding sizes of min- and the average-pooling layer are set to 0 and 8, respectively; *LN* recovers the number of features that are reduced by the min- and average-pooling layers and further learn the linear mapping between sensor baseline and $\mathrm{ave}(\mathrm{min}(Y_{in}))$; Y_{out} is passed through another two linear layers before being fed into the Bi-LSTM layer. Namely, the $\mathrm{ave}(\mathrm{min}(Y_{in})$ represents for the sensor baseline in every sample with sequence length of 100. Note

that only adopting either a min pooling layer or an average pooling layer will introduce a negative effect on tracking results. Thus, we assume that the sensor baseline computed by only the min pooling layer is not accurate enough because people sometimes may still be moving their joints to a lesser extent. Thus, the local-minimum sensor reading cannot represent for sensor baseline enough. However, averaging the result passing through the min pooling layer can reduce this effect. Similarly, sensor reading passing through the average pooling layer also cannot represent the baseline because the average value is susceptible to human changing moving ranges. Therefore, combining both of them is helpful for sensor calibration.

3.4.3 Model Training Before training, we normalized the sensor readings through min-max normalization. Then, we only fetched Y_{out} of five sequence length from Y_{out} (denoted as X=(x_0,x_1,. . . ,x_t,. . . ,x_T) and fed both X and its corresponding ground-truth motion data Y=(y_0,y_1,. . . ,y_t,. . . ,y_T) into Bi-LSTM (To distinguish the input sequence length of data from the sequence length of sensor readings before calibration, we used input sequence length and calibrated sequence length to represent them below). Finally, we optimized the parameters of Bi-LSTM and saved the one that performs the best on $\mathcal{D}_{VA}$. The loss function used during training is:

$$\mathcal{L}_S = \|\boldsymbol{X}(t) - \boldsymbol{Y}(t)\|_2^2 \tag{3}$$

where $\boldsymbol{X}(t)$ represents the sensor reading; $\boldsymbol{Y}(t)$ denotes the measured full-body joint angles collected by the optical system. Even though using future frames (the last data in the X series precedes the corresponding Y) is helpful for the tracking result, different from [12], we did not utilize the future frames because there is almost no difference whether to learn the future frames or not given that the sequence length of our data is 5. Moreover, we would not suffer from the decay caused by the learning of future frames.

3.4.4 Jitter Mitigation We observed jitters in our predicted motions that indicate a lack of smoothness and naturalness [6]. To mitigate such jitters, we embedded a median pooling layer to smooth the predicted value and get natural-looking results. The kernel size is set to 3 and the padding number is set to 1.

4 Results

4.1 Experimental Setup

We ran our experiments in a server configured with a three-core CPU with GPU support (NVIDIA GTX Titan Xp, 12G). The operating system was 64-bit Ubuntu 16.04. We measured jitters with the average jerk of all the joints we have predicted, which is the third derivative of the position (i.e. the degree changes in our task). The structures and parameters of models and techniques used are included in the appendix.

4.2 Quantitative Evaluation

4.2.1 Overall and Detailed Performance Overall, our model achieves a low average tracking error of 4.51°. To get more insights, we further break down the evaluation of our model into sub-evaluation tasks according to motion types (Table 2) and joint positions (Table 3):

Table 2. Average tracking errors of different motion types.

Motion	Error (Degree)	Motion	Error (Degree)
walk forward	3.67	swing	2.40
bend over	5.92	spin motion	2.26
two handed dribble	5.07	stride	4.09
climb	6.96	tai chi	5.76
walk backward	3.89	veer left	2.43
throw a baseball	4.15	punch	5.34
kick	3.80	sweep floor	4.97

Table 3. Average tracking errors of different joint positions.

Joint	Error (Degree)	Joint	Error (Degree)
Right-Shoulder	1.83	Spine1	1.07
Left-Shoulder	1.65	Spine2	1.07
Right-UpLeg	6.68	RightArm	8.88
RightLeg	5.28	Right-ForeArm	6.58
Left UpLeg	5.69	LeftArm	8.22
LeftLeg	4.78	Left-ForeArm	5.91
Spine	0.79		

- As Table 2 shows, it can be concluded that i) for the full-body movements with larger moving range (e.g. climbing), and motions of low limb utilization (*e.g.* bending over), the average tracking errors of our model are relatively high; ii) For motion involving only limb joints (e.g. throw a baseball), or full-body movements with smaller moving range (*e.g.* swing), the average tracking errors of our model are relatively low. iii) It can be observed that: apart from *bending over*, all the other common human motions have low tracking errors that are less than 7°. This indicates the superiority of our method in tracking human daily motions.
- As Table 3 shows, it can be observed that the average tracking errors of the arm joints and leg joints are larger than those of the shoulder and spine joints. We ascribe this to the greater motion range of arm and leg joints.

Table 4. Comparison with IMU solution.

	IMUs	Our Method
Error (Degree)	6.04	4.51
Jitter (10^4degree/s^3)	4.451	2.450

4.2.2 Comparison with IMU solution To demonstrate the superiority of our method against its IMU alternative, we evaluated their tracking errors and jitters using the same motion data collected from two subjects respectively. Specifically, we used an IMU jumpsuit produced by Beijing Noitom Technology Ltd. named Perception Legacy 2. This jumpsuit contains 18 IMU sensors in total. As Table 4 shows, it can be observed that the average tracking error of our method is lower than that of the IMU alternative by 1.53 degrees and the jitter of ours is lower than that of the IMU alternative by 2.090×10^4m/s^3. Hence, it can be concluded that our solution is more accurate and more stable than its IMU alternative.

4.2.3 Ablation Study To justify the effectiveness of the different components in our method, we conducted an ablation study as shown in Table 5:

Table 5. Ablation study.

Method	Tracking Error (Degree)	Training Time (Min)	Run Time (Millisecond)	Jitter (10^4degree/s^3)
BiLSTM	4.75	3.80	0.002	3.168
Transformer	5.11	4.12	0.001	0.713
FCN	5.23	**2.30**	**0.0001**	1.51
TCN	5.02	5.11	0.001	**0.691**
BiLSTM-C	4.58	4.98	0.001	3.157
BiLSTM-M	4.69	4.50	0.001	2.443
BiLSTM-CM	**4.51**	5.18	0.001	2.452

- Row 1-4 justify the effectiveness of our Bi-LSTM network against other neural network architectures used in previous works [12, 13, 42, 45]. Specifically, we compared three alternatives: LSTM [34], Transformer [37] and fully-connected neural network. We trained each network solely for the motion prediction task. It can be observed that our Bi-LSTM network performs the best in all metrics.
- Row 5-7 justify the effectiveness of our median pooling layer (denoted as M) and dynamic calibration (denoted as C). It can be observed that adding our median pooling layer and dynamic calibration improves the accuracy and mitigates the jitter.

To further investigate the impact of our median pooling layer, we draw the distribution of the jitters in Fig. 5a. It can be observed that the jitters of predicted result with median pooling layer is mainly distributed between 0 and 0.5. Besides, the jitter percentage is lower with the median pooling layer, which indicates that our median pooling layer reduces jitters effectively. Fig. 5b depicts the change of the left forearm over 5 seconds. It can be observed that with a median pooling layer, the curve becomes smoother and closer to the ground truth, which justifies the necessity of the median pooling layer.

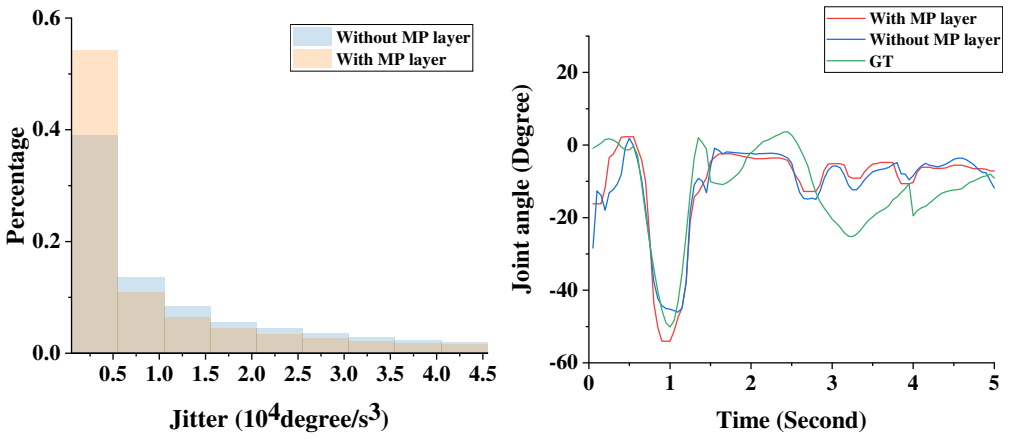


Fig. 5. (a) The distributions of jitters and (b) the bending angles of *left up leg* with and without the median pooling layer.

4.2.4 Justification of the Choice of Calibrated Sequence Length As Table 6 shows, we justified the hyper-parameters of our method by conducting the human motion prediction experiments with different calibrated sequence lengths and comparing the result. We fixed the kernel size of the min pooling layer and average pooling layer to 15 and 10; then, we changed the input length of data at one time. The result is shown in Table 6. It can be observed that when the calibrated sequence length is 100, the error on the $\mathcal{D}_{TE}$ is the smallest. Besides, when the calibrated sequence length is equal to 120 or 80, the error becomes slightly larger. When the calibrated sequence length is 60 or 40, the average tracking error goes up dramatically. In conclusion, the most suitable calibrated sequence length is equal to 100. Thus, we input data that sequence length is 100. After the sensor reading passes through two pooling layers, we only fetch the last 5 samples as input since the Bi-LSTM network performs best when the sequence length is equal to 5. There is no obvious difference between different kernel sizes of the min pooling layer and average pooling layer overall. However, when the kernel size of the min pooling layer is equal to 15 and the kernel size of the average pooling layer is equal to 10, the error is slightly lower than others. As a result, we adopted the kernel size of 15 and 10 for the min pooling layer and average pooling layer, respectively.

Table 6. Relationship between the calibrated sequence length and the average tracking error on testing dataset.

Calibrated Sequence Length	120	100	80	60	40
Error on $\mathcal{D}_{TE}$ (Degree)	4.58	**4.51**	4.58	4.56	5.6

4.3 Qualitative Evaluation

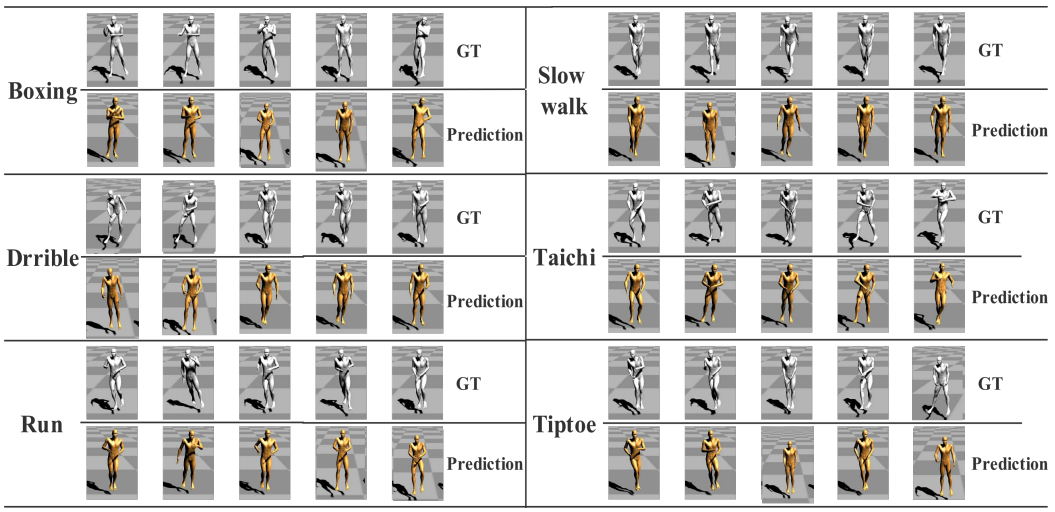


Fig. 6. Visual evaluation of our method against different motion types.

We also evaluated our method qualitatively by visualizing its output motions on virtual human bodies[2]. We show the overall performance of our method against different motion types in Fig. 6. Please see the supplementary materials for the accompanying videos.

[2]Note that we cloned the hip rotation from the ground truth to facilitate understanding.

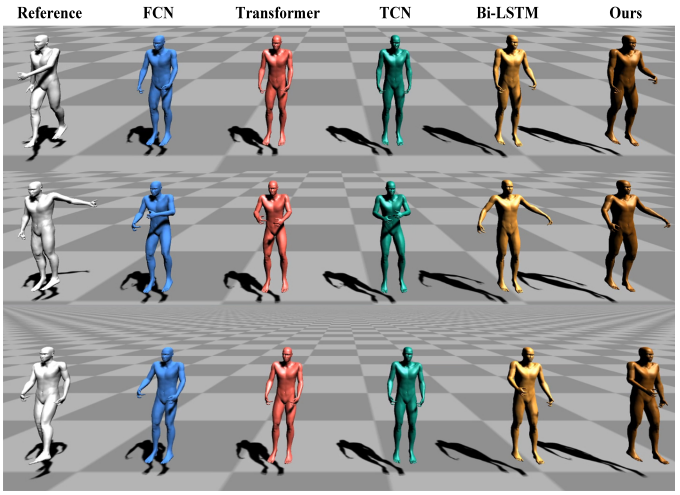


Fig. 7. Qualitative comparisons between different methods and ours.

4.3.1 Ablation Study The visual result of the ablation study is shown in Fig. 7. It can be observed that the approach of Bi-LSTM and our method will have obvious actions in most cases, especially when the subject is moving arm, the tracking is closer to the ground truth. By contrast, even though the TCN, FCN and Transformer networks will also track the large range and fast arm motion, they tend to move slightly or move at wrong angles. This may be due to their limited fitting power for the small range motion, which makes it collapse to the average pose slightly. The performance of our method is closer to the effect of Bi-LSTM. However, when there are some large arm movements, our method performs better than Bi-LSTM.

4.3.2 Tracking Results with and Without Median Pooling Layer Fig. 8a shows the effect of adopting the median pooling layer. From the static figure, we can not find obvious differences between the method with a median pooling layer and the method without a median pooling layer. While in the video, the effect of easing the jitter is easier to be observed (i.e. decreasing the extra bending angle that comes from jitters), it still takes some careful observation to see where the improvements are. Hence, although the median pooling layer takes the effect of decreasing the jitters numerically as well as increasing the accuracy, the visual effect of easing the jitter is not obvious.

4.3.3 Subjective Results Compared to IMU Alternative From Fig. 8b, it can be concluded that our method is better than the IMUs, especially in arm and leg joints. It is noticeable that the gait of IMUs is different from the ground truth sometimes, which could be due to the magnetic environment underground and the noisy data captured by the sensors. So the motions are occasionally exaggerated too much by the errors in the acceleration term of IMUs. Compared to IMUs, our method is unaffected by magnetic fields. Besides, we adopted the dynamic calibration as well as median pooling layers to mitigate sensor aging and potential jitters and our tracking is much closer to the ground truth.

4.4 Results Compared with IMUs

Procedure In order to quantify the convenience of wearing the IMU jumpsuit and ours, we invited five participants to evaluate the wearing time of our jumpsuit and IMU jumpsuit. They were required to do motions mentioned in Table 9 after we recorded the wearing time. Since we

randomly decided the order they put on the IMUs jumpsuit or ours, they were assigned to wear a different type of jumpsuit in a different order and perform daily exercises for 20 minutes when wearing every jumpsuit. Then, they ranked the comfort level with a 5-point Likert Scale (1 = very uncomfortable, 5 = very comfortable).

Results The result is shown in Table 7. Lastly, we interviewed the 5 participants on four aspects: comfort, the convenience of wearing, whether it would affect sports and possible suggestions for improvement. Table 8 records the ranking scores of the comfort level of the IMU jumpsuit and ours. All our scores on the jumpsuit are higher than IMU's, which proves that our comfort level is higher than IMU's.

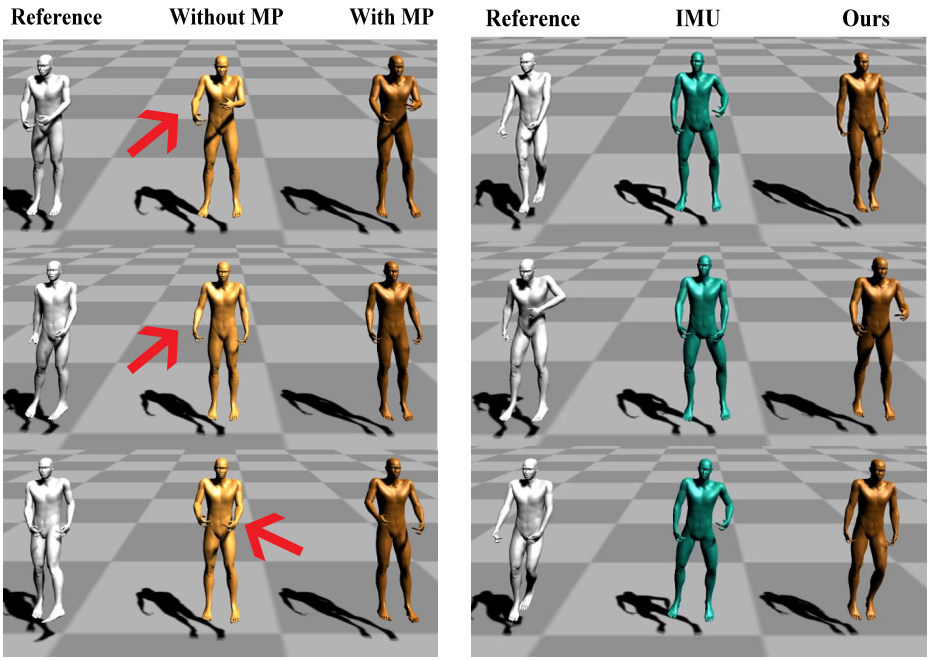


Fig. 8. (a) Qualitative comparisons between our method and method without median pooling layer. MP refers to the median pooling layer. (b) Qualitative comparisons between our method and IMUs.

During the interview, all five participants agreed that our jumpsuit was more comfortable than the IMU jumpsuit even though they had several concerns related to foreign body sensation. P1, P2 and P3 said there are no obvious differences between the jumpsuit and daily wearing clothes (P2:"it's like a daily wearing base, a bodysuit."). However, although some participants mentioned some foreign body sensations and some psychological concerns brought by sensors, visible wires and circuit boards, it was not affect their motion (P1:"There will still be a little bit of foreign body sensation, but I do not think it will affect motion." P2:"I was worried that the circuit board would be knocked by my movement."). As for the degree of tightness, all users said they did not feel tight (P3: "the tightness is equal to the tightness of ordinary undergarments.").

Table 7. The Likert point according to the wearing comfort of the IMU jumpsuits and ours.

Subject	IMU jumpsuit	Our jumpsuit
Subject1	1	**4**
Subject2	2	**5**
Subject3	2	**4**
Subject4	2	**4**
Subject5	2	**3**

Table 8. Comparison on wearing convenience with IMU solution.

Subject	IMU jumpsuit (Min)	Our jumpsuit (Min)
Subject1	21	**6**
Subject2	15	**4**
Subject3	10	**7**
Subject4	12	**6**
Subject5	11	**4**

By contrast, we received more negative feedback for the IMU jumpsuit, mainly on comfortability and tightness. Users generally reported that the IMU was too tight and they felt "weird" (P1: "The IMU jumpsuit is too tight, which makes me uncomfortable. Besides, I can feel the presence of many wires which is strange"). Participants also reported that the IMU jumpsuits impeded them from doing motion freely (P1: "there are so many velcro strips on the surface that they can easily stick to each other out of sight during movement. Hence, I had to stop and wait for other's help."). In addition, P1, P2 and P3 felt that the IMU suit was heavy (P3: "Perhaps because of the equipment embedded in the IMU suit, the suit was bulky, which made me not free to move."). When talking about the wearing process, our clothes also received a relatively positive response. P1, P2, P3 and P4 acknowledged that our jumpsuit was convenient to wear and they could wear it independently. However, they could not wear the IMU jumpsuit by themselves (P4: "We need to adjust the sensors that slip away when wearing the IMU jumpsuit, and we need help within that process. By contrast, the jumpsuit with only four sensors is much simpler than the IMU jumpsuit."). P5 mentioned that he could wear an IMU jumpsuit by himself but it would take a long time.

In conclusion, by using fewer, lighter and more flexible sensors, our clothes do not need fussy fixtures and the weight of our jumpsuit is small. Hence, our jumpsuit is superior to the IMU jumpsuit in the aspect of the wearing experience and wearing process.

5 Limitation

This work broadens the application of flexible sensors in human motion tracking. However, if we want to further decrease the tracking error, some factors like the interference from the external environment such as electromagnetic radiation, static electricity from the human body and so on should be taken into consideration. These factors limit the design of the circuit board, which makes it read sensor data with low precision, attributing to more one-to-more mappings of the sensor value to human joint angle. In future work, we will wrap the circuit board with insulation and revise the design of the circuit board to read sensor data with high precision. Additionally, although the CM-BiLSTM method is helpful to regress human motion accurately, the leg tracking result will collapse to the average pose when the range of leg movement is small. Besides, since our method is based on a time-sequence model, the movement of the predicted joints would have some interaction effects on each other. For example, the tracking is not accurate during some challenging motion types, such as climb, Tai Chi and so on, which can be seen in the video. It may be further improved by some context awareness (e.g., the subject is doing some exercise or some daily activities) from the long-term observation.

6 Conclusion

This work proposes a route of human daily activities tracking with sparse flexible sensors, which can be used as a reference for context-awarding in terms of human motion. The current average

tracking error is 4.51 degrees by using readings of only four flexible sensors across daily motion. To achieve this, we employed a Bi-LSTM neural work. Then, we adopted the min pooling layer and average pooling layer to calibrate the sensor readings dynamically. In addition, we leverage a median pooling layer to ease the jitters of the predicted output. This work presents the significance of the Bi-LSTM neural network in extracting the sequence features of the motion data, and the ability to regress the motion data. Besides, this work presents a solution to deal with aging or broken flexible sensors. In terms of specific application, this work is meaningful to rehabilitation (It is not convenient for patients to wear many sensors) or training of the athletes (wearing too many sensors will impede them from doing accurate movements), etc. Furthermore, its auto-calibration based on pooling layers further increases the possibility of leveraging flexible sensors for long-time tracking, mitigating the limitations of the flexible sensors. Hence, this work can be a reference for the choosing of the neural network, filtering method and flexible sensor calibration when dealing with the predicting of human motion with sparse flexible sensors.

Future work should focus on tracking the movements of other body parts such as heads. Since the current sensor readings are sparse, it is promising to judge those movements from prior knowledge related to the law of human movement to judge joints. Besides, under the popularity of the meta-universe and big data, it will be interesting to address human emotion annotation through the sensor readings collected from sparse flexible sensors. Unlike the EMG signal, which is easy to introduce noise [10], the soft sensor readings may be a potential solution to predict human emotion.

7 Acknowledgments

This work is supported by National Natural Science Foundation of China (62072383, 61702433, 62077039, 61962021), the Fundamental Research Funds for the Central Universities (20720210044, 20720190006), the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (VRLAB2020B17), Science and Technology Guiding Project of Fujian Province, Key Research Program of Jiangxi Province, and Natural Science Foundation of Fujian Province of China (No.2021J011169, No.2020J01435), the Key Project of National Key R&D Project (No.2017YFC1703303), Industry-University-Research Cooperation Project of Fujian Science and Technology Planning (No:2022H6012), Industry-University-Research Cooperation Project of Ningde City and Xiamen University (No.2020C001). Also thanks the China Scholarship Council (CSC) for providing financial support (program no. 202006310161). This work is partially supported by Royal Society (IEC\NSFC\211022).

References

[1] Sheldon Andrews, Ivan Huerta, Taku Komura, Leonid Sigal, and Kenny Mitchell. 2016. Real-time physics-based motion capture with sparse sensors. In *Proceedings of the 13th European conference on visual media production (CVMP 2016)*. 1–10.

[2] Brice Bouvier, Sonia Duprey, Laurent Claudon, Raphaël Dumas, and Adriana Savescu. 2015. Upper limb kinematics using inertial and magnetic sensors: Comparison of sensor-to-segment calibrations. *Sensors* 15, 8 (2015), 18813–18833.

[3] Manuel Caeiro-Rodríguez, Iván Otero-González, Fernando A Mikic-Fonte, and Martín Llamas-Nistal. 2021. A Systematic Review of Commercial Smart Gloves: Current Status and Applications. *Sensors* 21, 8 (2021), 2667.

[4] Dorin Copaci, Enrique Cano, Luis Moreno, and Dolores Blanco. 2017. New design of a soft robotics wearable elbow exoskeleton based on shape memory alloy wire actuators. *Applied Bionics and Biomechanics* 2017 (2017).

[5] Rita Cucchiara and Matteo Fabbri. 2022. Fine-grained human analysis under occlusions and perspective constraints in multimedia surveillance. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 18, 1s (2022), 1–23.

[6] Tamar Flash and Neville Hogan. 1985. The coordination of arm movements: an experimentally confirmed mathematical model. *Journal of neuroscience* 5, 7 (1985), 1688–1703.

[7] Mattia Guidolin, Razvan Andrei Budau Petrea, Oboe Roberto, Monica Reggiani, Emanuele Menegatti, and Luca Tagliapietra. 2021. On the accuracy of IMUs for human motion tracking: a comparative evaluation. In *2021 IEEE*

International Conference on Mechatronics (ICM). IEEE, 1–6.

[8] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. 2021. Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4318–4329.

[9] Sehoon Ha, Yunfei Bai, and C Karen Liu. 2011. Human motion reconstruction from force sensors. In *Proceedings of the 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. 129–138.

[10] Mohammad Mehedi Hassan, Md Golam Rabiul Alam, Md Zia Uddin, Shamsul Huda, Ahmad Almogren, and Giancarlo Fortino. 2019. Human emotion recognition using deep belief network architecture. *Information Fusion* 51 (2019), 10–18.

[11] Amir Hooshiar, Amir Sayadi, Javad Dargahi, and Siamak Najarian. 2021. Integral-free spatial orientation estimation method and wearable rotation measurement device for robot-assisted catheter intervention. *IEEE/ASME Transactions on Mechatronics* (2021).

[12] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J Black, Otmar Hilliges, and Gerard Pons-Moll. 2018. Deep inertial poser: learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics (TOG)* 37, 6 (2018), 1–15.

[13] Yifeng Jiang, Yuting Ye, Deepak Gopinath, Jungdam Won, Alexander W Winkler, and C Karen Liu. 2022. Transformer Inertial Poser: Attention-based Real-time Human Motion Reconstruction from Sparse IMUs. *arXiv preprint arXiv:2203.15720* (2022).

[14] Yasuo Katsuhara and Hirotaka Kaji. 2019. Towards multi-person motion forecasting: Imu based motion capture approach. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 97–100.

[15] Dooyoung Kim, Junghan Kwon, Seunghyun Han, Yong-Lae Park, and Sungho Jo. 2018. Deep full-body motion network for a soft wearable motion sensing suit. *IEEE/ASME Transactions on Mechatronics* 24, 1 (2018), 56–66.

[16] Manon Kok, Jeroen D Hol, and Thomas B Schön. 2017. Using inertial sensors for position and orientation estimation. *arXiv preprint arXiv:1704.06053* (2017).

[17] Yong Lee and S Kassam. 1985. Generalized median filtering and related nonlinear filtering tech niques. *IEEE Transactions on Acoustics, Speech, and Signal Processi ng* 33, 3 (1985), 672–683.

[18] Miaopeng Li, Zimeng Zhou, and Xinguo Liu. 2020. Cross Refinement Techniques for Markerless Human<? brk?> Motion Capture. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16, 1 (2020), 1–18.

[19] Yanran Li. 2021. *Human motion prediction*. Ph. D. Dissertation. Bournemouth University.

[20] Gandolla Marta, Ferrante Simona, Costa Andrea, Bortolotti Dario, Sorti Stefano, Vitale Federico, Bocciolone Marco, Braghin Francesco, Masiero Stefano, and Pedrocchi Alessandra. 2019. Wearable biofeedback suit to promote and monitor aquatic exercises: A feasibility study. *IEEE Transactions on Instrumentation and Measurement* 69, 4 (2019), 1219–1231.

[21] Aizan Masdar, BSKK Ibrahim, Dirman Hanafi, M Mahadi Abdul Jamil, and KAA Rahman. 2013. Knee joint angle measurement system using gyroscope and flex-sensors for rehabilitation. In *The 6th 2013 Biomedical Engineering International Conference*. IEEE, 1–4.

[22] Takuya Matsumoto, Kodai Shimosato, Takahiro Maeda, Tatsuya Murakami, Koji Murakoso, Kazuhiko Mino, and Norimichi Ukita. 2020. Human Pose Annotation Using a Motion Capture System for Loose-Fitting Clothes. *IEICE Transactions on Information and Systems* 103, 6 (2020), 1257–1264.

[23] Yiğit Mengüç, Yong-Lae Park, Hao Pei, Daniel Vogt, Patrick M Aubin, Ethan Winchell, Lowell Fluke, Leia Stirling, Robert J Wood, and Conor J Walsh. 2014. Wearable soft sensing suit for human gait measurement. *The International Journal of Robotics Research* 33, 14 (2014), 1748–1764.

[24] Maria F Montoya, John E Muñoz, and Oscar A Henao. 2020. Enhancing Virtual Rehabilitation in Upper Limbs With Biocybernetic Adaptation: The Effects of Virtual Reality on Perceived Muscle Fatigue, Game Performance and User Experience. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 28, 3 (2020), 740–747.

[25] Christos Mousas. 2017. Full-body locomotion reconstruction of virtual characters using a single inertial measurement unit. *Sensors* 17, 11 (2017), 2589.

[26] Yong-Lae Park, Daniel Tepayotl-Ramirez, Robert J Wood, and Carmel Majidi. 2012. Influence of cross-sectional geometry on the sensitivity and hysteresis of liquid-phase electronic pressure sensors. *Applied physics letters* 101, 19 (2012), 191904.

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019), 8026–8037.

[28] Sen Qiu, Hongkai Zhao, Nan Jiang, Zhelong Wang, Long Liu, Yi An, Hongyu Zhao, Xin Miao, Ruichen Liu, and Giancarlo Fortino. 2022. Multi-sensor information fusion based on machine learning for real applications in human

activity recognition: State-of-the-art and research challenges. *Information Fusion* 80 (2022), 241–265.

[29] Sen Qiu, Hongkai Zhao, Nan Jiang, Donghui Wu, Guangcai Song, Hongyu Zhao, and Zhelong Wang. 2022. Sensor network oriented human motion capture via wearable intelligent system. *International Journal of Intelligent Systems* 37, 2 (2022), 1646–1673.

[30] Ioannis Rallis, Apostolos Langis, Ioannis Georgoulas, Athanasios Voulodimos, Nikolaos Doulamis, and Anastasios Doulamis. 2018. An embodied learning game using kinect and labanotation for analysis and visualization of dance kinesiology. In *2018 10th international conference on virtual worlds and games for serious applications (VS-Games)*. IEEE, 1–8.

[31] Michael Rietzler, Florian Geiselhart, Janek Thomas, and Enrico Rukzio. 2016. Fusionkit: a generic toolkit for skeleton, marker and rigid-body tracking. In *Proceedings of the 8th ACM SIGCHI Symposium on Engineering Interactive Computing Systems*. 73–84.

[32] Daniel Roetenberg, Henk Luinge, and Per Slycke. 2009. Xsens MVN: Full 6DOF human motion tracking using miniature inertial sensors. *Xsens Motion Technologies BV, Tech. Rep* 1 (2009), 1–7.

[33] Gaspare Santaera, Emanuele Luberto, Alessandro Serio, Marco Gabiccini, and Antonio Bicchi. 2015. Low-cost, fast and accurate reconstruction of robotic and human postures via IMU measurements. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2728–2735.

[34] Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing* 45, 11 (1997), 2673–2681.

[35] Loren Arthur Schwarz, Diana Mateus, and Nassir Navab. 2009. Discriminative human full-body pose estimation from wearable inertial sensor data. In *3D physiological human workshop*. Springer, 159–172.

[36] Jochen Tautges, Arno Zinke, Björn Krüger, Jan Baumann, Andreas Weber, Thomas Helten, Meinard Müller, Hans-Peter Seidel, and Bernd Eberhardt. 2011. Motion reconstruction using sparse accelerometer data. *ACM Transactions on Graphics (ToG)* 30, 3 (2011), 1–12.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[38] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. 2018. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 601–617.

[39] Timo Von Marcard, Gerard Pons-Moll, and Bodo Rosenhahn. 2016. Human pose estimation from video and imus. *IEEE transactions on pattern analysis and machine intelligence* 38, 8 (2016), 1533–1547.

[40] Timo von Marcard, Bodo Rosenhahn, Michael J Black, and Gerard Pons-Moll. 2017. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer Graphics Forum*, Vol. 36. Wiley Online Library, 349–360.

[41] W-W Wang and L-C Fu. 2011. Mirror therapy with an exoskeleton upper-limb robot based on IMU measurement system. In *2011 IEEE International Symposium on Medical Measurements and Applications*. IEEE, 370–375.

[42] Frank J Wouda, Matteo Giuberti, Giovanni Bellusci, Bert-Jan F van Beijnum, and Peter H Veltink. 2019. Improving Full-Body Pose Estimation from a Small Sensor Set Using Artificial Neural Networks and a Kalman Filter. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 10063–10064.

[43] Shihong Xia, Lin Gao, Yu-Kun Lai, Ming-Ze Yuan, and Jinxiang Chai. 2017. A survey on human performance capture and animation. *Journal of Computer Science and Technology* 32, 3 (2017), 536–554.

[44] Dongseok Yang, Doyeon Kim, and Sung-Hee Lee. 2021. Lobstr: Real-time lower-body pose prediction from sparse upper-body tracking signals. In *Computer Graphics Forum*, Vol. 40. Wiley Online Library, 265–275.

[45] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. 2022. Physical Inertial Poser (PIP): Physics-aware Real-time Human Motion Tracking from Sparse Inertial Sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13167–13178.

[46] Xinyu Yi, Yuxiao Zhou, and Feng Xu. 2021. TransPose: Real-time 3D Human Translation and Pose Estimation with Six Inertial Sensors. *ACM Transactions on Graphics* 40, 4, Article 86 (08 2021).

[47] Qing You, Wenjie Chen, and Ye Li. 2020. 3D Human Motion Capture Based on Neural Network and Triangular Gaussian Point Cloud. In *2020 39th Chinese Control Conference (CCC)*. IEEE, 7481–7486.

[48] Zikang Yuan, Dongfu Zhu, Cheng Chi, Jinhui Tang, Chunyuan Liao, and Xin Yang. 2019. Visual-inertial state estimation with pre-integration correction for robust mobile augmented reality. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1410–1418.

[49] Yizhai Zhang, Kuo Chen, Jingang Yi, Tao Liu, and Quan Pan. 2015. Whole-body pose estimation in human bicycle riding using a small set of wearable sensors. *IEEE/ASME Transactions on Mechatronics* 21, 1 (2015), 163–174.

[50] Zerong Zheng, Tao Yu, Hao Li, Kaiwen Guo, Qionghai Dai, Lu Fang, and Yebin Liu. 2018. Hybridfusion: Real-time performance capture using a single depth sensor and sparse imus. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 384–400.

A Motion List

Table 9 shows all of the motion types subjects have performed when collecting the dataset. Among them, some basic movements are subdivided to ensure a variety of movements, which are listed in Table 10

Table 9. Motion types collected from subjects.

climb	run	veer right	veer left	shake hands
90-degree left turn	shoot basketball	swordplay	dribble	lay-up shot
Punch	spin kick	straight kick	spin motion	jumping jacks
squats	bend over	putt	throw a frisbee	mop floor
skip rope	scoop up	turn around	pick up	freestyle
play the violin	get dressedt	hrow a baseball	underhand toss	hop on one foot
tennis	dance	swing golf	sweep floor	boxing
swing	drum	jump	tai chi	balance
jump up to grab	point	laugh	pull	shoot a gun
arm toss	wash windows	explain with hand gestures	dribble	

Table 10. The subdivided motion of some basic movements.

Basic motion	Subdivided motion
walk	walk stiffly, tiptoe, walk and 90-degree left turn, walk and 90-degree right turn walk forward, walk backward, brisk walk, slow walk, stride, hobble
jump	180 jump, 2 jump, high jump, long jumps, hop on one foot, forward jump backward jump, jump and 90-degree left turn
dribble	two handed dribble, forward dribble, backward dribble, sideways dribble dribble and go forward, crossover dribble forward dribble and turn left forward dribble and turn right, shoot basketball dribble and shoot, lay-up shot

B Implementation Details

- TCN: We implemented it with Pytorch. A linear layer was added before the TCN block to change in input size to 64. The kernel size is 7, the number of hidden units is 64 and the dilation factor is [1],[2]. The sequence length is 5.
- Bi-LSTM: The structure of this neural network can be referred to 2. The number of hidden units of Dense1, Dense2, Dense3, Dense4 is 100, 512, 1024,150, respectively.
- Transformer: We adopted the encoder structure and one linear layer (output layer) as the decoder. The number of the head is 4. There is one layer in the encoder. The sequence length is 5.
- FCN: There are two hidden layers in total. The number of hidden units is 200 and 100, respectively.