Received December 24, 2020, accepted February 14, 2021, date of publication February 26, 2021, date of current version March 9, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3062545

Magnetometer Robust Deep Human Pose Regression With Uncertainty Prediction Using Sparse Body Worn Magnetic Inertial Measurement Units

HAMMAD TANVEER BUTT[1,2,3], **BERTRAM TAETZ**[2,3], **MATHIAS MUSAHL**[3], **MARIA A. SANCHEZ**[3], **PRAMOD MURTHY**[2,3], **AND DIDIER STRICKER**[2,3]

[1]Department of Computer and Software Engineering, College of Electrical and Mechanical Engineering, National University of Sciences and Technology, Islamabad 44000, Pakistan
[2]Department of Computer Science, University of Kaiserslautern, 67663 Kaiserslautern, Germany
[3]Augmented Vision Group, German Research Center for Artificial Intelligence (DFKI), 67663 Kaiserslautern, Germany

Corresponding author: Hammad Tanveer Butt (hammad.butt@dfki.de)

This work was funded by the Body Information on an Intelligent Chip (BIONIC) Project at DFKI Germany in collaboration with 10 other partners, under the European Commission Horizon 2020 Programme Grant Agreement number 826304.

ABSTRACT We propose a deep learning based framework that learns data-driven temporal priors to perform 3D human pose estimation from six body worn Magnetic Inertial Measurement units sensors. Our work estimates 3D human pose with associated uncertainty from sparse body worn sensors. We derive and implement a 3D angle representation that eliminates yaw angle (or magnetometer dependence) and show that 3D human pose is still obtained from this reduced representation, but with enhanced uncertainty. We do not use kinematic acceleration as input and show that it improves the generalization to real sensor data from different subjects as well as accuracy. Our framework is based on Bi-directional recurrent autoencoder. A sliding window is used at inference time, instead of full sequence (offline mode). The major contribution of our research is that 3D human pose is predicted from sparse sensors with a well calibrated uncertainty which is correlated with ambiguity and actual errors. We have demonstrated our results on two real sensor datasets; DIP-IMU and Total capture and have come up with state-of-art accuracy. Our work confirms that the main limitation of sparse sensor based 3D human pose prediction is the lack of temporal priors. Therefore fine-tuning on a small synthetic training set of target domain, improves the accuracy.

INDEX TERMS 3D human pose, deep learning, uncertainty estimation, inertial motion capture, sparse sensing, magnetic inertial measurement unit (MIMU), 9-axis IMU, wearable sensors.

I. INTRODUCTION

Estimation of 3D human pose is an important goal in computer vision, augmented and virtual reality, robotics and human motion capture. Either extrinsic sensors like cameras or body fixed sensors are utilized for this pose estimation. In later category, inertial and magnetic sensors have become quite common with advent of low cost Microelectromechanical systems (MEMS) in recent years. This technology is called inertial motion capture (*i*-Mocap). Compared with camera based 3D pose estimation, body worn inertial motion capture is robust to occlusion and also suitable for pose estimation in the wild. However the number of sensors or special clothing makes *i*-Mocap more obtrusive. Commercially available *i*-Mocap systems like Xsens use upto 13–17 sensors; one per each body segment for full kinematic capture [1]. The setting up and calibration of so many wearable sensors take time. Each sensor node comprises of magnetic-inertial measurement unit (MIMU), also often called 9-axis IMU. It employs sensor fusion of rate gyro, accelerometer and magnetometer to obtain an orientation estimate and linear acceleration in a global frame. The human body has constrained degree of freedom and temporal coherence and smoothness is an important feature of human motion. Many existing kinematic or inverse kinematic based *i*-Mocap frameworks, therefore uses

The associate editor coordinating the review of this manuscript and approving it for publication was Masood Ur-Rehman.

predefined constraints to reduce measurement errors or drifts [2]–[6]. In past research [7]–[9], a small set of inertial sensors is shown to estimate 3D pose to a reasonable accuracy. The data-driven approaches using reduced sensors ($\leq$6 instead of 13–17) [10]–[13] are more suitable for ambulatory data capture than full kinematic approach [14]–[19]. The scalability of data driven approach to 3D human pose estimation using reduced sensor set has been demonstrated using deep learning [20] and a large synthetic dataset. Reducing the number of sensors and their flexible placement on body makes the 3D pose estimation less obtrusive and thus this modality can be used for daily activity monitoring, ergonomics and wearable health more easily. However a learned model estimating the 3D pose with reduced sensors, depends greatly on the correlation in the data. Thus the predictions with inputs outside the training data are often inaccurate. In such a case, an estimate of uncertainty of predicted 3D pose becomes important.

Our presented work focuses on deep learning based uncertainty aware framework that learns data-driven temporal and spatial priors in a latent manifold to perform constrained 3D human pose estimation from sparsely worn Magnetic Inertial Measurement units (MIMU) sensors as input. The main contribution of this work are,

(1) Data-driven uncertainty estimation of 3D human pose from reduced sensors.

(2) A robust deep learning model which leads to a straightforward generalization to real sensor data, by training on 'synthetic' data.

(3) Though our work use full orientation estimation using magnetometer, we also show that a reduced orientation estimation (comprising only pitch/roll) from a sparse set of body worn MIMUs is 'sufficient' to estimate 3D human pose but increases the uncertainty. Thus dependence on magnetometer can be eliminated, which is desired in magnetically disturbed indoor environment or when IMU without magnetometers are used.

(4) As compared to [20], [21], our work shows that *linear* acceleration if used as input reduces the generalization to real sensors, due to different skeletal lengths and variable placement of sensors on real subjects. Our model achieves better performance in terms of generalization and accuracy than state-of-art [20] by *not* using the *linear* acceleration from sensors.

Also compared to existing state-of-art data-driven approaches [20], [21], we not only predict 3D human pose from a reduced number of sensors, but also provide a well calibrated estimate of uncertainty. This has an advantage of uncertainty driven information fusion with other sensor modalities [22] or with the output of other uncertainty based algorithms like Kalman Filter [3], [5], [23]. Our model also works in the inhomogeneous magnetic field, by ignoring the heading or yaw information but it shows more uncertainty in this case. The later problem is not addressed by [20], [21], but is a well-known limitation of *i*-Mocap. To the best of our knowledge, our work demonstrates a magnetometer robust 3D human pose estimation using *reduced* or *sparse* sensors for the first time. Previous work [2], [3], [5], [23], [24] has addressed this problem in the context of full body worn sensors (typical 13–17 sensors).

II. RELATED WORK

A. INERTIAL HUMAN MOTION SENSING

The accelerometers have been used as wearable sensors for human gait analysis [25] and one degree of freedom knee joint angle estimation [26]. The main limitation of accelerometers is that these only measure the pitch and roll angles (heading angle is missing) with reference to gravity vector. Also during the movement, the accelerometer not only registers the gravity but also a *linear* acceleration, depending on the displacement from joint axis. Few works in the past [8], [27] demonstrated that a reduced number of body worn accelerometers can obtain complete 3D human pose. The shortcoming of accelerometer based approaches arises from the fact that real subjects have different lengths of body segments and placement of sensors also varies, which lead to variations in sensor readings for same pose. This coupled with the fact that accelerometers do not provide absolute heading angles has been a major hurdle in accelerometer based human motion sensing. However when MEMS accelerometers are integrated with rate gyros and magnetometers in a magnetic-inertial measurement units (MIMU) or 9-axis IMU, sensor fusion [28] can be used to obtain full 3D orientation (pitch, roll and heading). More recently 9-axis or 6-axis IMUs (less magnetometers) have been used [2], [3], [5], [23], [24] for full body human motion capture. The human body is assumed as comprising of rigid segments articulated at joints and one sensor per segment is sufficient to compute 3D joint angle if adjacent segment orientations as rigid body are known. The main limitation and challenge to *i*-Mocap frameworks however is long term drift of sensors. Thus accurate calibration and robust sensor fusion [29, 30] is essential prerequisite to inertial human motion sensing. The kinematic and inverse kinematic based *i*-Mocap frameworks employ optimization [31] or stochastic filtering [23].

B. SPARSE SENSING OF HUMAN POSE

Past work has shown that owing to kinematic and temporal constraints of 3D human pose, it is possible to use only a reduced set of sensors (as opposed to one sensor per segment) and still obtain 3D human pose; though in general ambiguity of this ill-posed problem can lead to high uncertainty. Liu *et al.* [32] and Chai and Hodgins [33] demonstrated this with statistical human body model fit to reduced marker set. A small set of inertial sensors [7]–[9] is shown to estimate 3D pose accurately. Andrews *et al.* [34] have used an inverse dynamic solver for joint torques and internal/contact forces which satisfies motion priors and sparse sensor measurements, and thus generates physically plausible human motion. The data-driven approaches using reduced sensors ($\leq$6 instead of 13–17) [10]–[13] are well suited for online implementation than optimization or constrained stochastic filtering [15]–[19]. But good scalability of data driven approaches typically require a large dataset. Huang *et al.* [20] used deep learning and a large SMPL synthetic dataset for learning 3D human pose from sparse (six) 'synthetic' sensors and demonstrated state-of-art results after fine-tuning on real MIMU data. Their work is an extension of earlier optimization based approach by von Marcard *et al.* [14]. Wouda *et al.* [21] show that shallow temporal convolution (TC) and multilayer perceptron (MLP) yield similar results but 'jerkiness' error is less using deep learning approach. In their later work, Wouda *et al.* [35] employed a Kalman Filter to address this jerkiness from shallow network. Recently, Ekhoff *et al.* [36] demonstrated the condition for observability in a kinematic chain comprising of double hinge joints using sparse magnetometer free inertial tracking. Their work highlights that a sliding window (moving horizon) filter can estimate kinematics of two connected single hinge joints, using measurements from magnetometer-free IMUs only at the end links.

C. LEARNING OF HUMAN MOTION MANIFOLD

The data-driven learning of human motion has come to fore with large human motion capture datasets and deep learning. The human motion can be represented in a latent manifold space. Both convolutional and recurrent neural networks are used to learn human motion manifolds [37]–[39]. This section only identifies the representative work to emphasize the relatedness to our research. Li *et al.* [40] show that the latent representation of human motion learnt by a bidirectional recurrent autoencoder is robust to both input noise and missing information. A related issue in human motion manifold learning is minimal representation of 3D joint angles (SO3) so that data-driven learning is not complicated by singularity, duality and discontinuity problems occurring in input or output data. Earlier work by Murthy *et al.* [41] has compared Euler angles, quaternions and more intuitive swing-twist representation. Also Zhou *et al.* [42] show that 5-parameter and 6-parameter representations (instead of full 9-parameter rotation matrices) are continuous and best for deep learning.

D. UNCERTAINTY IN DEEP LEARNING

Estimating uncertainty of deep regression is relatively new research direction. Both the uncertainty in the data (aleatoric) and model uncertainty (epistemic) affect the final error in the output of deep model. The aleatoric SO3 uncertainty is dealt with by negative log likelihood (NLL) cost function and quaternions in the context of 3D rotation by Peretroukhin *et al.* [43] who demonstrated it on KITTI visual odometery dataset. Russel and Reale [44] extend uncertainty estimation in deep learning to multivariate outputs and used a Kalman filter for training and evaluation. Salinas *et al.* [45] and Zhu and Laptev [46] deal with uncertainty estimation in time series forecasting using LSTM. Most frameworks assume independent Gaussian distributions for outputs and estimate both mean and standard deviation. Wen *et al.* [47] and Kivaranovic *et al.* [48] have proposed a distribution free approach and predicted quantiles or prediction intervals of outputs. The main challenge in learning uncertainty from data in deep learning lies in calibration of regression uncertainty [49]. Laves *et al.* [50] propose a framework for calibration of test data uncertainty by scaling of standard deviation (σ) with a scalar value. The robust uncertainty from deep regression allows detection of unreliable predictions.

To the best of our knowledge, our work is the first to implement and discuss uncertainty of 3D human pose estimation obtained from sparse MIMUs using deep regression. It is motivated by the fact that human motion has spatial and temporal constraints which may be learnt in latent space and such a latent space representation is then robust to missing or noisy information in the input space [40]. Our work differs from Huang *et al.* [20] and Wouda *et al.* [21] in that it also reports the data-driven uncertainty estimation of 3D human pose from reduced sensors. It also does not use acceleration as input like [20], [21], which is shown to improve generalization to real MIMU datasets. Also our work finds out that the 'jerkiness' which is reported by shallow networks approach [29] results from fixed finite temporal context. We train a bi-directional recurrent autoencoder and at inference time use a temporally skewed time window for real-time prediction with minimum 'jerkiness'.

A significant improvement that we make over [20], [21] is the use of 6D parametrization for input 3D orientation and exponential map for 3D joint angles at the output. The later allows us to predict the uncertainty in interpretable terms (radians) directly at the output of network without any post processing. Both rotation matrix and quaternion require an orthogonalization step [20], [21] and incorporation of upper limit of uncertainty is not straightforward especially when a parameter is near unity.

Apart from uncertainty estimation, most important aspect of our work is that we develop a robust model and show that even a reduced orientation estimation (comprising only pitch/roll) from a sparse set of body worn MIMUs is 'sufficient' to estimate 3D human pose with enhanced uncertainty. Thus dependence on magnetometer can be eliminated, which is desired in magnetically disturbed indoor environments. The estimation of kinematic uncertainty of 3D human pose obtained from sparse sensors may be used for uncertainty driven information fusion with other sensor modalities [22], or with the output of other uncertainty based algorithms like Kalman Filter [3], [5], [23].

III. PROPOSED METHODOLOGY

In our work we train a deep bidirectional recurrent autoencoder to learn a rich set of temporal priors for human pose in latent space using SMPL datasets [51]. The model is driven using an input of five body segment orientations (left wrist, right wrist, left leg, right leg, head), normalized with respect orientation of the root of human skeleton. These are 'synthesized' using forward kinematics. The model outputs a full human pose (joint angles) in SMPL space, less rotation

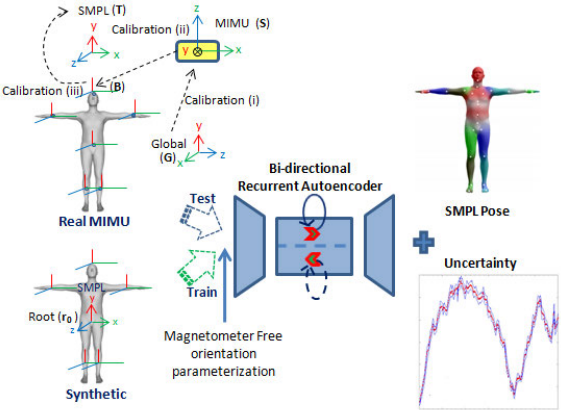

FIGURE 1. Overview of our framework with key contributions.

and translation of root. Only major[1]joint angles in SMPL pose are estimated, assuming no 3D rotation at other joints. The overview of our framework is shown in Fig. 1 (graphical abstract). The trained model has then been tested on two real MIMU datasets in Section V-B. The results of our framework are also compared with existing frameworks (SIP [14] and DIP [20]) on same datasets. A robust input parametrization is further suggested to eliminate the yaw drift/errors due to magnetometer in Section V-C. Ablation studies are performed to understand the limitations and to come up with a more optimal model in Sections V-D to V-I.

A. SYNTHETIC DATASET

The deep learning for human motion requires an abundant dataset comprising of many subjects, varying movements and activities, at different temporal speeds and range of motion of human joints. Many motion capture datasets are available, but these do not use standardized 3D skeleton and to the best of our knowledge only few [52], [53] include data from inertial sensors with associated calibration. Thus [20] have used a large synthetic dataset developed based on SMPL [51] for training their model. We have employed the synthetic dataset made publically available by [20]. But we also added data augmentation to achieve a more robust training of our models. This SMPL dataset includes Human3.6 [54], CMU human activities [55], Human Eva [56], Joint Limits [57] and eight other datasets all transformed to SMPL skeleton using AMASS framework [51]. The frame rate of standardized SMPL dataset is 60Hz. The body segment orientations are then 'synthesized' using forward kinematics from root sensor. The orientation of root sensor is simply obtained from SMPL pose. Assuming that inertial sensors body frames are aligned with human body segment (i.e. sensor-to-segment alignment is identity), the 'synthesized' body segment orientations then represent sensor orientations as well. For augmenting this ideal senor data, we have introduced both zero mean Gaussian white noise as well as random impulse noise to ideal sensor orientations. For each sequence in the dataset, we also introduced a drift in yaw angle of 3D orientation (based on random small bias value integrated over time). Our model was initially trained on raw ideal synthetic data and then fine-tuned on noise/ drift augmented data.

B. REAL MIMU SENSOR DATASETS

Two real MIMU sensor datasets are used for testing our model trained on 'synthetic' augmented dataset. DIP-IMU dataset has been provided by Huang *et al.* [20] as open source. Total capture MIMU dataset [52] is a publically available dataset with MIMU orientations and calibration. Both these datasets have an advantage that Huang *et al.* [20] have performed testing on these and our results are thus directly comparable. Also in case of total capture dataset, SMPL ground truth pose are obtained using AMASS framework [51]. In comparison Wouda *et al.* [21] has also used own real MIMU dataset, but their ground truth 3D poses are not in SMPL and instead are based on biomechanical model of XSens MVN. All these datasets are obtained using Xsen motion tracking MIMUs, hence apart from experimental conditions or calibration accuracy, the test results demonstrated on DIP-IMU and total capture, would be applicable to [21] as well.

C. SENSOR PLACEMENT

Six MIMU sensors are placed near left wrist, right wrist, left lower leg, right lower leg, lower spine and head. The sensor at lower spine is treated as reference or root sensor. The five sensors located at lower/upper limbs and head provide orientation measurements w.r.t. root sensor which is used to predict full 3D human pose. An alternate sensor configuration with lower legs sensors moved to feet and wrist sensors moved to hands, i.e. the end effectors of kinematic chain, is tested to be worse in performance.

Another interesting configuration is with sensors located on left/ right upper arms and left/right upper legs. Theoretically, this predicts shoulder joint and hip joint angles with high accuracy. However it is completely unable to constrain and predict the elbow and knee joints during arbitrary movements.

The acceleration readings from these sensors depend on their exact location w.r.t joint axis and center of rotation. Although Huang *et al.* [20] has used the sensor accelerations like earlier work by von Marcard *et al.* [14], we notice that the accelerations vary based on where the sensors are mounted on skeleton and hence are not a reliable input feature. von Marcard *et al.* [14] got better results with accelerations (SIP versus SOP), because they obtained SMPL model shape with laser scans. The training by Huang *et al.* [20] uses synthetic data from a standard SMPL skeleton and authors are able to generalize to real subjects with real sensors only after fine-tuning. Even on synthetic data, we show that the error obtained by Huang *et al.* [20] is more than our work, which only uses sensor orientations.

[1] Major SMPL joints include Left Hip, Right Hip, Spine0, Left Knee, Right Knee, Spine1, Spine2, Neck, Left Clavicle, Right Clavicle, Head, Left Shoulder, Right shoulder, Left Elbow and Right Elbow.

D. CALIBRATION OF INPUT DATA

Both the 'synthetic' and real MIMU 3D orientation data needs proper calibration before it can be used as input to the model. The orientation of the root of the human skeleton varies as subjects perform movement. But the overall 3D human pose is invariant to the rotation and translation of the root. Thus the orientation of five end effectors (left wrist, right wrist, left leg, right leg, head) is normalized w.r.t. orientation of the root (base of the spine near hip) in SMPL according to,

$$R_t^{R_0 B_i} = \left(R_t^{T R_0}\right)^{-1} R_t^{T B_i} \tag{1}$$

where T and B_i represents the SMPL frame of reference and reference frame attached to bone segment respectively as shown in Fig. 1 and R_0 is frame of reference fixed to root. $R_t^{R_0 B_i}$: $B_i \rightarrow R_0$ is the rotation matrix from root frame to respective bone frame of reference at time instant t and same convention holds for other rotation matrices representing orientations. As evident we need 3D orientations from root to SMPL and bone to SMPL to compute our 'invariant' orientation input to the model using (1). This is trivial for 'synthetic' data, since both the orientations can be obtained from SMPL ground truth pose using forward kinematics.

But when real MIMU sensors are employed, the frame of reference of sensor S_i is not always aligned to reference frame of body segment B_i. We need to compute a sensor-to-segment calibration matrix $R^{B_i S_i} : S_i \rightarrow B_i$. This is done at startup time using a 'static' calibration pose. We have implemented static I-pose for this calibration due to ease of implementation for elderly and functionally impaired subjects. More details are available in [58]. Also due to residual intra-sensor startup and calibration errors, each MIMU sensor after sensor fusion provides an orientation, $R^{I_i S_i}$, where the 'perceived' inertial frame I_i is slightly offset from actual global inertial frame of reference G. Hence we also need to obtain $R^{G I_i} : I_i \rightarrow G$ for each sensor at startup. The full calibration of real MIMU data then proceeds in following steps,

$$R_t^{G S_i} = R^{G I_i} R_t^{I_i S_i} \tag{2a}$$

$$R_t^{G B_i} = R_t^{G S_i} \left(R^{B_i S_i}\right)^{-1} \tag{2b}$$

$$R_t^{R_0 B_i} = \left(R_t^{G R_0}\right)^{-1} R_t^{G B_i} \tag{2c}$$

$$\overline{R}_t^{R_0 B_i} = R^{T G} R_t^{R_0 B_i} \tag{2d}$$

We obtain R^{TG} from root sensor as opposed to head sensor (unlike [20]) at initial body model calibration. This is found to be more robust to inter and intra subject variations.

As opposed to [20], [21], we have not used acceleration for reasons discussed before and later in results section, we demonstrate the advantage as well. Hence the calibration of acceleration is not discussed.

E. INPUT & OUTPUT PARAMETERIZATION

Two different 3D angle parametrization have been used in our work for input body segment orientations and output human pose respectively and the motivation for using both is explained in this section.

Earlier work by Huang *et al.* [20] employed full 9-parameter rotation matrix for both input and outputs, whereas Wouda *et al.* [21] have used quaternions. Both rotation matrix and quaternion require norm constraints since underlying degree of freedom (DoF) is only three in case of 3D rotation. This is accomplished as post-processing of output in [20] and [21] and thus cause additional jittering error. Moreover the input is over-parameterized in case of either quaternions (4 parameters) or rotation matrix (9 parameters). This redundancy both of input and output also increases the number of model parameters, which increase the training and test time by order of O(n), where n is number of model parameters. On the other hand, if a minimal 3DoF parametrization is used, it introduces gimbal lock, singularity and discontinuity issues in input space as highlighted in [41].

We have employed for input orientations, a 6-parameter representation (motivated by Zhou *et al.* [42]). A complete 3×3 rotation matrix is over-parametrized. But cross product of any two rows and columns in a right handed order leads to third row/column. Zhou *et al.* [42] chose first two columns for its 6D representation, since it uniquely determines the remaining column. Our 6-parameter representation (derived in Appendix. A) is a simple extension of the fact that yaw, pitch and roll angles (which completely define 3DoF) can be conveniently derived from rotation matrix as follows (ZXY order),

$$\gamma = tan^{-1}\left(r_{21}/r_{11}\right) \tag{3a}$$

$$\theta = tan^{-1}\left(-r_{31}/\sqrt{r_{32}^2 + r_{33}^2}\right) \tag{3b}$$

$$\varphi = tan^{-1}\left(r_{32}/r_{33}\right) \tag{3c}$$

Here γ, θ, φ are yaw, pitch and roll angles respectively and $r_{11}, r_{21}, r_{31}, r_{32}, r_{33}$ are components of 3×3 rotation matrix. As evident only 5 components of 3×3 rotation matrix are sufficient to obtain 3DoF. However to avoid an indeterminate case (Appendix. A) we also include r_{22} to make a 6-parameter representation. Compared to Zhou *et al.* [42], our 6-parameter representation can also be used to derivea reduced 3-parameter representation derived from (3) comprising of r_{31}, r_{32}, r_{33} which only account for pitch and roll angle information. Though the reduced 3-parameter representation increases the ambiguity and uncertainty, the motivation of it is 'justified' for reasons related to magnetometer error in inhomogeneous magnetic field and discussed later in Section V-I.

For output human pose (joint angles), we have directly used exponential map (3-parmeters) representation of SMPL. By SMPL definition of human skeleton, the joint angles in exponential map representation are always continuous owing to joint constraints (explained in Appendix. B). It is not only a minimal DoF representation for human pose, but also allows us to learn uncertainty directly in quantitative terms in output space. If either rotation matrices or quaternions are used for output, then post-processing would be needed to obtain uncertainty measure in radians or degrees. The calibration of uncertainty in later case would also be complicated.

F. APPROXIMATE BAYESIAN MODEL

Given a training dataset, $\mathcal{D} = \left\{(x_t, y_t)^i\right\}_{i=1}^{N}$, where each training example (x_t, y_t) is a sequence or time series, we learn an approximate Bayesian model to infer probability distribution $\mathcal{P}\,(y_t|x_t)$ of 3D human pose from a sequence of sensor orientations x_t. If we assume a Normal distribution with diagonal covariance for the full 3D human pose $\mathcal{P}\,(y_t|x_t) = \mathcal{N}_{\mathrm{y}}(\mu_t, \sigma_t^2 I)$, we can write our model as,

$$\mathcal{N}_{\mathrm{y}}\left(\mu_t, \sigma_t^2 I\right) = \mathfrak{F}_{\mu_t,\sigma_t}\left(\{x_t\}_{t-n}^{t+m}\,;\, h_t, \{\theta\}_{1:p}\right) \tag{4}$$

where our model $\mathfrak{F}$ performs an approximate Bayesian inference using p perturbations of its parameters θ to obtain mean pose μ_t and its diagonal covariance vector σ_t^2, given a sequence of inputs x_t over a sliding time window $[t - n : t + m]$. Here, n are input samples from past and m are future samples; h_t represents the latent state of the model for a given sequence of inputs. In deep learning framework, this model has been implemented as a shallow (but wide) MLP neural network or a 1D CNN by Wouda *et al.* [21] and a recurrent temporal network by Huang et al[20]. In case of MLP neural network or a 1D CNN, the input time window remains fixed and cannot be changed at inference. To retain flexibility at inference time and learn a compact model, we have implemented a recurrent model which propagates the latent state h_t forward-backward recurrently and compose $(h_t^F \circ h_t^B)$ at time t to get the output. It is given as follows,

$$\mathcal{N}_{\mathrm{y}}\left(\mu_t, \sigma_t^2 I\right) = \mathfrak{F}\left(\{h_t|h_{t-1}, x_t\}_{t-n}^{t+m} \circ \{h_t|h_{t+1}, x_t\}_{t+m}^{t-n}\,;\, \{\theta\}_{1:p}\right) \tag{5}$$

Consistent with previous literature [45], [46], the aleatoric uncertainty σ_t^2 is learnt directly as an output alongside the mean pose μ_t, as an attentive regularization term in negative log likelihood (NLL) cost function (to be described in Section IV-D). The aleatoric uncertainty adapts to variance of data in the domain of training set. The model (epistemic) uncertainty is obtained using approximate Bayesian inference as explained above and for that Monte-Carlo dropout (MC- Dropout) [34] is used at inference time. Ensemble approach [59] is also possible for epistemic uncertainty but is computationally more expensive. The composition of two types of uncertainty (aleatoric and epistemic) is described in Section IV-F.

IV. IMPLEMENTATION

We have implemented our deep learning framework in *Tensorflow* 1.15.2 and *Python* 3.6 on a Desktop computer with GPU-NVidia GTX 1060 and CUDA 10.1. The data pre-processing, preparation and results evaluation is performed using *Matlab*2019a.

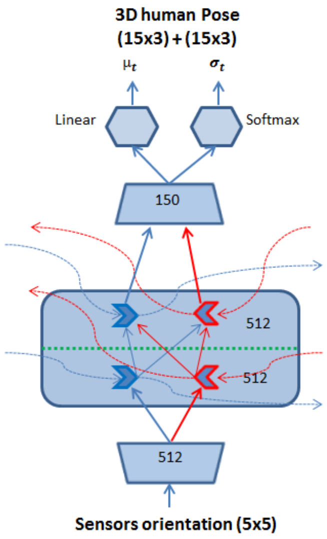

FIGURE 2. Our Bidirectional Recurrent Autoencoder (BiRAE) model.

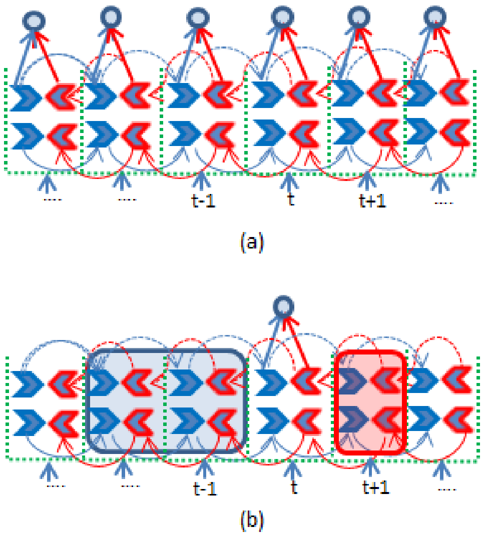

FIGURE 3. (a) Forward pass in BiRAE (offline/ training). Blue are the forward cells. Red are backward cells. (b) Sliding time window inference in BiRAE. The past (blue window) and future (red window) subsequence is used for real-time prediction of current time step. The size of future window plus computation time defines latency.

A. MODEL ARCHITECTURE

For the model given in (5), we have implemented Bidirectional LSTM auto-encoder with two stacked hidden Bi-LSTM layers each of size 512 (the size of our latent state). The input to Bi-LSTM is fully connected layer of size 512 with a drop-out of 0.2 and ReLU function. A fully connected output layer after Bi-LSTM stacked layers is chosen with size 150 and linear output to obtain mean pose μ_t. The same layer with independent weights predicts the diagonal covariance vector σ_t. The overall architecture is shown in Fig. 2. A forward pass on this bi-layer Bi-LSTM comprise of four sub-passes as shown in Fig. 3.

B. PREPROCESSING

Huang *et al.* [20] employed and tested different normalization schemes on the input to their model. We have also tested such normalization of input, but no significant difference is noticed with or without normalization. Since we are only using orientations (and no accelerations), both the 6-parameter input representation and output 3D pose in exponential map (in radians) vary in small range around zero, and normalization has no advantage.

C. POSTPROCESSING

Of 24 joints in SMPL model, only 15 major joints are predicted by our model which includes: Left Hip, Right Hip, Spine0, Left Knee, Right Knee, Spine1, Spine2, Neck, Left Clavicle, Right Clavicle, Head, Left Shoulder, Right shoulder, Left Elbow and Right Elbow. Since the rest of joints are located forward of sensors positions at limbs, these are not predicted and substituted by unit rotation as a post-processing step to get the full 3D human skeleton.

We predict the 3D pose directly in exponential map parameterization of SMPL, and obtain the uncertainty directly in radian for each component of exponential map representation at the output. Unlike quaternions [21] or rotation matrices [20], the uncertainty prediction using our proposed representation is smooth and does not need Unscented transform in post-processing.

D. COST FUNCTION FOR ALEOTORIC UNCERTAINTY

The aleatoric uncertainty deals with covariate uncertainty found in the training data for a given model structure. If the test data also lies within the domain of training data, it is sufficient to use aleatoric uncertainty. We learn this uncertainty, using a 'regularization' term in negative log likelihood (NLL) cost function that we use for our training as follows,

$$\mathcal{L}_{\text{NLL}}(\theta) = \sum_{i=1}^{m} \sum_{t=t-n}^{t+m} \left(\left(\sigma_t^{(i)}(\theta)\right)^{-2} \left\| \mathrm{y}_t^{(i)} - \mu_t^{(i)}(\theta) \right\|^2 + \ln\left(\sigma_t^{(i)}(\theta)\right)^2 \right) \tag{6}$$

Here the inner sum is over the temporal subsequence $[t-n : t+m]$ and m is the number of subsequences in a batch. Since we assume only diagonal covariance $\sigma_t I$ for the full 3D human pose, our cost function is simplified compared with multivariate case of [44]. Also we assume that Euclidian distance approximate SO(3) error in exponential map space. This assumption is true for small errors upon convergence of training.

Huang *et al.* [20] includes sensor accelerations as additional input. To force the network learning on this predictor, they also introduced an auxiliary task during training and their model was also forced to reconstruct the accelerations. Due to inherent problems with acceleration that we noted before, we have not used it as predictor. But we have used this 'auxiliary' loss as motivation to reconstruct 3D angular rate readings when we opted for reduced 3-parameter representation, which only account for pitch and roll angle information. Discarding yaw information, we used3D angular rate readings as additional input and also used auxiliary reconstruction loss on the later input; by this we got slightly better results (Table 1) than using 3-paremeter pitch/roll representation alone.

E. TRAINING

The sequences in our synthetic dataset used for training have variable length. Due to limitation of GPU memory size, we use truncated Backpropagation through time (BPTT), and divide our sequences in synthetic data into subsequences of length 300, discarding those with length ≤ 200, to avoid too much zero padding. We have randomly divided our synthetic data into training, validation and test set in 70/20/10 ratio. We also later perform testing on validation data, to understand better the poor learning of some pose subsequences.

Our model was trained using Adam algorithm with an initial learning rate of 0.001, exponentially decayed at rate of 0.9 with decay step 5000. Batch size for training was set at 16. Gradient clipping with a norm 1 was applied to Bi-LSTM training. The loss on validation set was used as early stopping criteria while training was set to max epoch size of 2000. The model with best error on validation set was saved during the training run.

F. INFERENCE WITH EPISTEMIC UNCERTAINTY

We carry out the evaluation using the trained model in two ways. In offline end-to-end inference mode, we use a maximum batch size for which the test dataset is divisible by an integer, for fast evaluation. During inference, four end-to-end sequential sub-passes are performed on the trained Bidirectional LSTM model as depicted in Fig. 3a (two passes per Bi-LSTM layer). At each time t, we obtain two hidden vectors after four sub-passes and compose them $(h_t^F \circ h_t^B)$ to obtain the output. This gives us mean pose μ_t and its diagonal covariance vector σ_t^2, given a sequence of inputs x_t. The later represent aleatoric uncertainty and henceforth we assign a superscript a in its symbol σ_t^a. Our model is trained using a dropout of 0.2 for regularization, and we use the same for Monte Carlo dropout (MC-Dropout) at the time of inference to obtain epistemic (model-based) uncertainty. This is essential to deal with out-of-domain data which was not seen in training. If $\mu_t^{(i)}(\theta_i)$ is the mean pose prediction for ith Monte Carlo iteration of model dropout, the we write for epistemic uncertainty,

$$\sigma_t^{e^2} = \frac{1}{M} \sum_{i=1}^{M} \left(\mu_t^{(i)}(\theta_i) - \frac{1}{M} \sum_{i=1}^{M} \mu_t^{(i)}(\theta_i) \right)^2 \tag{7}$$

The epistemic uncertainty is then combined with aleatoric part to get overall uncertainty as follows,

$$\sigma_t^2 = \sigma_t^{e^2} + \frac{1}{M} \sum_{i=1}^{M} \sigma_t^{a^2} \tag{8}$$

The problem with end-to-end bidirectional estimation as shown in Fig. 3a is that it can only be carried out *offline*. For a real-time application, we define a sliding time window with past frames and future frames and only predict the output $\left(\mu_t, \sigma_t^2 I\right)$ at time t, after four sequential sub-passes are performed within that time window, as shown in Fig. 3b. The advantage of our *online* approach is clearly evident over shallow temporal convolution (TC) and multilayer perceptron (MLP). Our BiRAE model can be trained end to end on sequences and then desired time window or unroll can be selected at the run-time (see also Fig. 6 for results). The former models take only fixed time window and hence their scope and performance is limited.

V. RESULTS AND DISCUSSION

We analyze the baseline performance of our trained model first on the synthetic dataset to validate the extent of learning on *ideal* data, in order to choose the best performing architecture. Then we compare results of our best performing model on real MIMU data, mainly with state-of-art, i.e. DIP by Huang *et al.* [20] who have also tested on the same real MIMU datasets, i.e. DIP-IMU and Total Capture [52]. The results of *offline* mode of model inference are presented first. Since code and data by Wouda *et al.* [21] has not been publically made available and it does not use SMPL, it is left out of comparison. We also perform ablation studies for self-comparison between different variants of our model on real MIMU data. In particular we discuss and compare the results of 6-parameter and reduced 3-parameter representation and the significance of each. Next, the effect of real-time window length/configuration is also discussed based on results of *online* mode versus *offline* mode. Lastly, we report the uncertainty estimation obtained using our framework and discuss it. We also discuss the effect of covariate and domains shift, and show that sensor noise in real MIMU data of DIP-IMU and Total capture is not significant to cause covariate shift. But the trained model performs poorly for those data sequences for which similar ones are absent in training data. This is identified as the main limitation of pose tracking based on sparse sensors. We also show how much fine tuning can help address this problem. Since the real MIMU data used in evaluation is not much perturbed, we create simulated magnetic perturbation in yaw part of 3D orientation and demonstrate the performance advantage obtained using 3-parameter model in this scenario.

The metrics we have used throughout is *mean per joint angle error (MPJAE)* or *per joint angle error* of individual joints. Positional error is not used for two reasons; first we are interested in 3D human pose which is agnostic to scale of the skeleton and thus can be used for biomechanical or activity ergonomics across subjects, 2nd we observe that the position of 3D joints when used alone, loses the information of '*twist*' along a body segment [41] and hence is not useful for the above mentioned target applications focused by us.

A. PERFORMANCE ON SYNTHETIC DATA

After training our model, most important aspect to investigate was how well it performed on *ideal* sensors of 'synthetic' dataset. This sets a baseline on which we can then evaluate the performance of real MIMU datasets. We performed this evaluation on 10% test set drawn at random from 'synthetic test' and not used in training. The comparison was made between a model trained on synthetic data by Huang *et al.* (DIP) [20] and variants of our method. The results in Table 1 demonstrate the effectiveness of 3D angle parameterization chosen by us for our model both for the input and output, which is different from [20], who chose 9-parameter rotation matrix and also included sensors acceleration. We also build a model that uses 6D parameterization of Zhou *et al.* [42] for both input and output. In summary, results of our model with 6-parameter input and exponential map 3D pose are slightly better on synthetic data (our trained model performs much better on real MIMU data, as shown in next section) than DIP [20] and at par with using 6-parameters of [42]. It may be noted that our model parameters are also 20% less than DIP [20]. The performance comparison is depicted in Table 1

TABLE 1. Performance on synthetic data.

Model	MPJAE* (deg)		
	Distal §	**Tracking** §	**Other** §
DIP†[20](with Accel)	12.30	7.17	9.18
Our‡(No Accel, 6-param [42])	11.97	6.55	8.88
Our‡(No Accel, 6-param proposed in this paper)	**11.54**¶	**6.07**¶	**8.23**¶
Our‡(No Accel, 3-param proposed in this paper)	14.74	10.36	12.18
Our‡(No Accel, 3-param + angular rate)	14.37	10.13	11.91

† DIP [20] model uses 9-parameter sensor orientation input and 9-parameter 3D human pose output, with acceleration.
‡All Our models use exponential map 3D human pose. The input parameterization used is different as discussed in text. Last model also include rate of orientation change of a tracking sensor w.r.t root sensor. The model using 6-parameters of [42], uses this representation for both input/output.

*MPJAE is Mean Per Joint Absolute error in degrees.

§Distal Joints are shoulder and hip joints. Tracking joints are proximal to where sensors are located (Knee, Elbow, Head). Other include all others.
¶ **Highlighted numbers represent best performance**

Once we evaluated the results on individual sequences in test data, we found that *mean per joint angle error (MPJAE)* is particularly high for certain sequences as shown in Fig. 4 and it is uncorrelated with the length of sequences.

We investigated the *failure* cases further for 3D angle estimation of a joint as shown in Fig. 5. It was seen that the model converges to correct 3D angle at the start of sequence and also predicts the accurate values for periodic movement of joint, however intermittently it shows large error. We infer that main limitation of 3D pose prediction from sparse sensors comes from the *imbalance* or *absence* of certain subsequences in the training dataset. Detailed reasons with experimental evaluation would be discussed in section V-H on fine tuning.

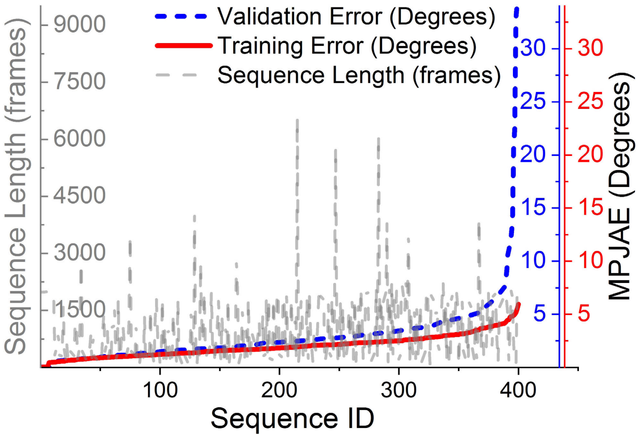

FIGURE 4. Performance of trained model on synthetic test data (Blue dotted). For comparison error on a random subset of training data is also shown (in red). The sequences are shown in order of ascending error. The x-axis is sequence number.

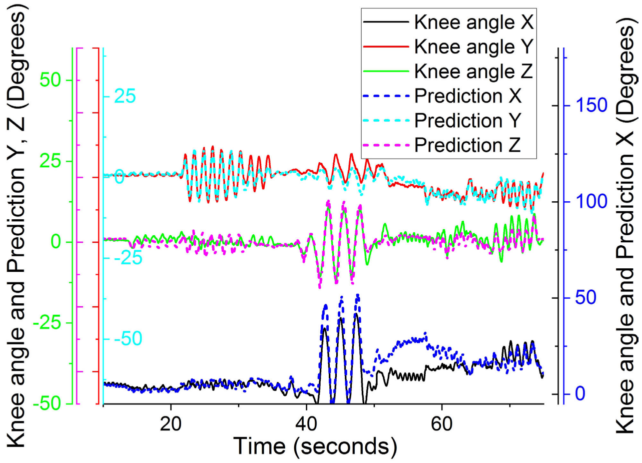

FIGURE 5. Trained model performance on 3D angles of a joint in synthetic test data. Intermittent failure on 'unseen' subsequence.

TABLE 2. Performance on real MIMu data.

Model	Dataset	MPJAE* (deg)		
		Distal §	Tracking §	Other §
DIP [20]	DIP-IMU (No Fine)	34.27	24.06	25.82
	DIP-IMU (Fine)	19.13	14.77	16.76
Our 6-param Model	DIP-IMU (No Fine)	**29.15**	**4.55**	**10.92**
	DIP-IMU (Fine)	**18.47**	**4.81**	**9.28**
Our 3-param Model	DIP-IMU (No Fine)	34.41	16.62	21.08
	DIP-IMU (Fine)	23.23	14.20	17.26
Our 6-param Model	Total capture (No Fine)	**16.06**	**12.83**	**14.83**
	Total capture (Fine)	**15.21**	**4.16**	**10.17**
Our 3-param Model	Total capture (No Fine)	18.49	15.33	17.04
	Total capture (Fine)	17.23	10.11	14.23
SOP [14]	DIP-IMU	27.78		
	Total capture	22.18		
SIP [14]	DIP-IMU	24.00		
	Total capture	16.98		

*MPJAE is Mean Per Joint Absolute error in degrees.
§Distal Joints are Left/right shoulder and hip joints. Tracking joints are proximal to where sensors are located (Knee, Elbow, Head). Other include all other major joints predicted with the model.

B. PERFORMANCE ON REAL MIMU DATA

The real MIMU dataset used in our evaluation are DIP-IMU and Total capture, we have used these datasets because these are publically available and two prior works [20] and [14] have reported results on the same datasets. For comparison of our trained model with [20] and [14] in this paper, we therefore use these datasets. The quality of sensor orientations obtained in these datasets has been evaluated by using forward kinematics on SMPL ground truth in either case. This point will be discussed in a section V-G on covariate and domain shift. We report better accuracy with and without fine-tuning on real datasets compared with DIP [20] and SIP/SOP [14], and results are depicted in Table 2.

The model errors in case of real MIMU data are large compared to 'synthetic' training data. As would be shown in section V-G and V-H, it is attributable less to the sensor errors than to the absence of similar sub-sequences in the training data. We have earlier seen that 'synthetic' test data also has such subsequences on which error is large, since those are under-represented in training data. We have trained our model on 'synthetic' data augmented with noise and disturbances, which might explain its better performance without fine-tuning on real MIMUs. But disturbances such as magnetic conditions and bias integration are non-stationary in nature. It is shown by Butt *et al* [28] that sensor fusion might be affected greatly by magnetic disturbances, while the orientation errors due to body acceleration are bounded. We therefore also tested a variant of our model with a reduced 3-parameter input from sensors which only represent pitch/roll, i.e. magnetometer or heading information is ignored. It will be discussed in next section. We however point out that DIP model [20] performs very poorly on DIP-IMU data without fine tuning, and we attribute it to linear acceleration used by [20]. For training data, they used a standard body model and fixed sensor positions, which explain why the results on synthetic data in Table 1 obtained by DIP [20] are comparable to our model. But in real scenario, linear accelerations vary based on where the sensors are mounted on skeleton and hence are not a reliable input feature. von Marcard *et al* [14] got better results with accelerations (SIP versus SOP), because they obtained SMPL model shape with laser scans and measured sensor position accurately for the subjects.

C. EFFECT OF REDUCED 3-PARAMETER INPUT

Our model with 3-parameter sensor input takes into account only the pitch/roll information and drops the other components with yaw angle information from 6-parameter representation. This provides intrinsic robustness against magnetic disturbances. Although magnetometers are still used for sensor fusion and proper calibration to fit in with the full framework, the yaw information is finally not available to the model for training or at inference time. The comparison of 3-parameter versus 6-parameter input on real MIMU datasets (Table 2), shows that the use of 3-parameter representation degrades the performance. However the periodic activities like walking etc. are still predicted good using 3-parameter model. The advantage of using 3-parameter input model over 6-parameter is not obvious on real MIMU datasets used in this study, since these datasets are not highly perturbed by magnetic interference or a yaw angle drift (Fig. 11). More challenging datasets (with SMPL ground truth and magnetically perturbed MIMU data) are not available at the time of this study. Later in this paper, we show a comparison on *simulated* strong magnetic interference on DIP-IMU to demonstrate the value of 3-parameter model in such situations.

We have further augmented the 3-parameter sensor pitch/roll input with 3D angular rate. It is derived as additional 3-vector from the difference of two consecutive 3D orientations in case of each of five sensors readings using quaternion formalism given in [28]. As explained in section IV, we use an auxiliary reconstruction loss on this later input; and by this we got better results than using 3-paremeter pitch/roll representation alone. Instead of raw sensor acceleration or rate gyro readings, we have used 3-parameter pitch/roll and 3-parameter angular rate *normalized* w.r.t root frame of reference, as described above. The advantage of this approach is that we compute all input parameters in root sensor frame of reference and our body calibration procedure as implemented in section III, works without any change.

D. EFFECT OF TIME WINDOW AT INFERENCE

As pointed out by Wouda *et al* [21], the main difference between shallow approaches using limited temporal context and recurrent neural networks is the 'jerkiness' that appears in the motion. While Huang *et al.* [20] have demonstrated that Bidirectional LSTM can be used for real-time 3D pose regression in online mode with limited context of past and future frames, they have not discussed the issue of 'jerkiness'. We also carry out this evaluation for our model and compare the online mode with offline mode for different configurations of time window in Fig. 6.

We carried out a grid search in variable increments over a range of [0,500] of past/future frames used in pose prediction. A window size of [130,30] for past/future frames gives an accuracy comparable to offline mode. The latency in this case is only 0.3 sec (at frame rate of 60Hz) plus computation time. The 'jerkiness' depicted in Fig. 6 for this window configuration, is also acceptable. The end-to-end offline mode is however recommended for all non-realtime applications.

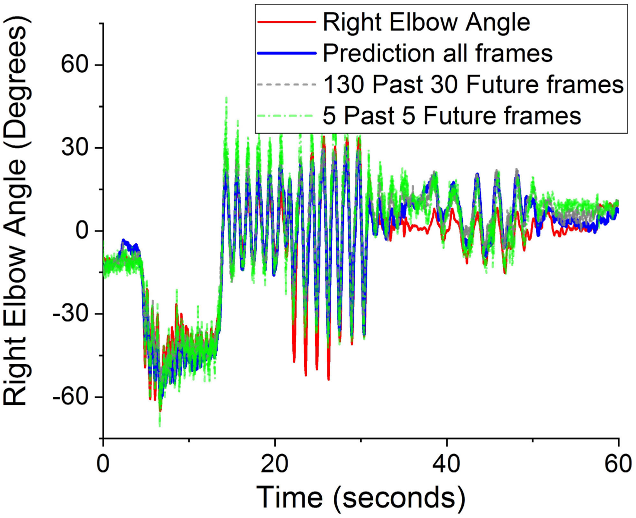

FIGURE 6. Comparison of 'Jerkiness' for different time window sizes (online vs. offline mode).

TABLE 3. Frame rate ablation study.

Frame rate	MPJAE[*] (deg)		
	Distal [§]	Tracking [§]	Other [§]
Total capture –Original (60Hz)	16.06	12.83	14.83
Total capture –Down Sampled (30Hz)	16.67	13.08	15.11
Total capture - Down Sampled (15Hz)	18.60	14.59	16.72
Total capture - Down Sampled (6Hz)	20.67	17.64	19.62

[*] MPJAE is Mean Per Joint Absolute error in degrees.
[§] Distal Joints are Left/right shoulder and hip joints. Tracking joints are proximal to where sensors are located (Knee, Elbow, Head). Other include all other major joints predicted with the model.

E. FLEXIBLE FRAME RATE AT INFERENCE

It is interesting to point out that at the time of inference our recurrent model is flexible to frame rate of input. We trained our 6-parameter model on 'synthetic' dataset obtained at 60Hz. But at time of inference, we performed under-sampling of recorded data from 60Hz to 30Hz and even 15Hz. We show that the degradation in performance is only gradual as reported in Table 3. We infer that the learned recurrent model is only acting by propagating latent state in time and is not affected by its actual rate. The gradual increase in error occurs from loss of high rate component of the motion. In comparison, the shallow MLP model proposed by Wouda *et al.* [21] not only depends on exact window configuration but also the frame rate on which it is trained. There is no flexibility to change either at the inference time.

F. UNCERTAINTY ESTIMATION

The main feature of our work is estimation of 3D human pose uncertainty, while regressing from a sequence of sparse

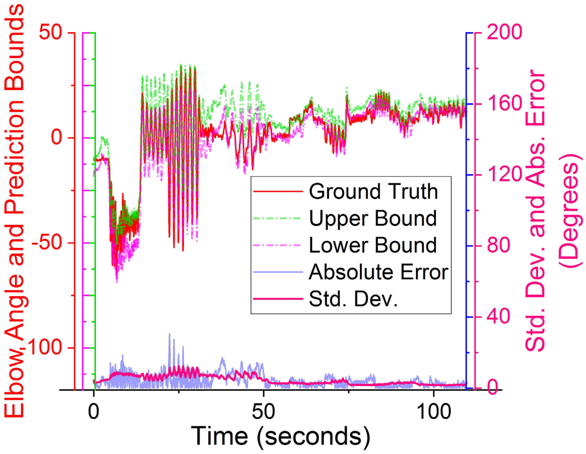

FIGURE 7. Uncertainty estimation and actual error on synthetic test data sequence, using 6-parameter trained model.

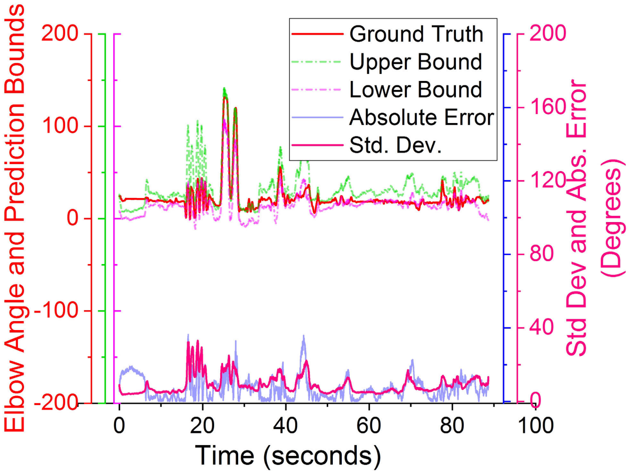

FIGURE 8. Uncertainty estimation and actual error on total capture data sequence, using 6-parameter trained model.

MIMU sensors. We estimate both aleatoric and epistemic uncertainty and then combine these in overall estimate using (8). Fig. 7 depicts the output 3D angle with uncertainty ($\pm 1\delta$) bounds for a test sequence from synthetic data. It is clear that uncertainty on 'synthetic' data from test set is predicted very well. In Fig. 8, we further show uncertainty ($\pm 1\delta$) bounds for a sequence from real Total capture dataset. Again the uncertainty of our model learned using synthetic data alone scales well to real MIMU data.

The investigation of uncertainty reveals that it correctly scales with the actual error. Since we assume a parametric model of uncertainty and predict Normal distribution with diagonal covariance for the full 3D human pose $\mathcal{N}_{\mathrm{y}}(\mu_t, \sigma_t^2 I)$, the actual error must lie within ($\pm 1\delta$) bounds, 68% of the time. We also note that both the synthetic sequence in Fig. 7 and Total capture sequence in Fig. 8, do not start at zero initial pose of SMPL (T-pose), yet model converges to low error after first few frames (at inference time, model hidden state is always initialized as zero at start of sequence).

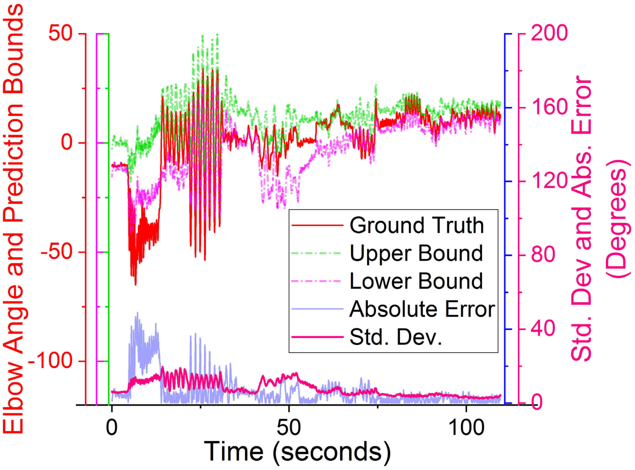

FIGURE 9. Uncertainty estimation and actual error on synthetic test data sequence, using 3-parameter trained model.

The uncertainty prediction during initial frames is not always accurate, but then it gets better. The uncertainty and absolute error is more when the amplitude of movement is more. The absolute error is stochastic (*max*: ±20 deg), but the mean absolute error is low (<5 deg). The model attempts to replicates the movement patterns.

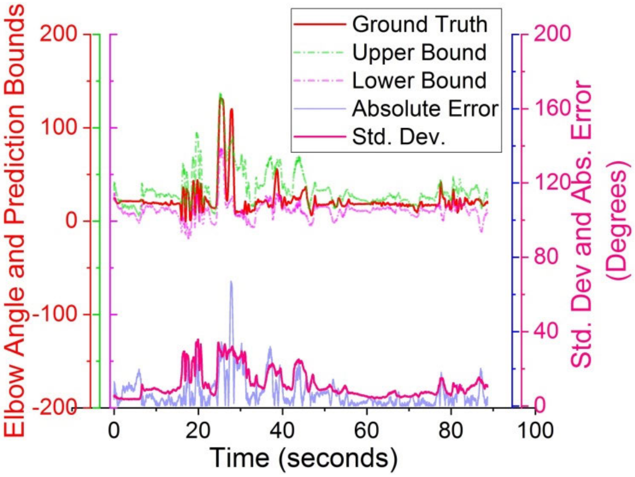

FIGURE 10. Uncertainty estimation and actual error on total capture data sequence, using 3-parameter trained model.

For a comparison Fig 9 and 10 show the results obtained with 3-parameter model for the same sequences in 'synthetic' and Total capture data respectively. We noted in Table 1 and Table 2 that 3-parameter model performance is lower than 6-parameter model. We can also identify from Fig. 9 and 10 that the uncertainty and error of prediction from 3-parmeter model is slightly higher, but it still predicts the movement and changes reasonable well. We demonstrate later in Section V-I that in case the sensor data is highly perturbed by magnetic interference (yaw angle), the performance of 3-parameter model is more robust than 5-parameter model.

G. COVARIATE AND DOMAIN SHIFT OF REAL DATA

Real MIMU data is never the same as *ideal* sensor data we created in synthetic dataset. Although we did data augmentation of synthetic data by adding noise and disturbances, but in fact the real artifacts are non-stationary. So it is necessary to investigate covariate and domain shift that exist in real data. Since we found that DIP-IMU dataset shows highest error when our model is not fine-tuned (Table 1), we focused on this dataset for analysis.

We synthesized an *ideal* MIMU orientation, from ground truth pose available for DIP-IMU using forward kinematics. In order to check how the real sensor data differed from 'synthetic' data on which we trained our model, we obtain the angular difference between the real and 'synthetic' orientations of respective sensors in SO(3) space, using axis-angle metric.

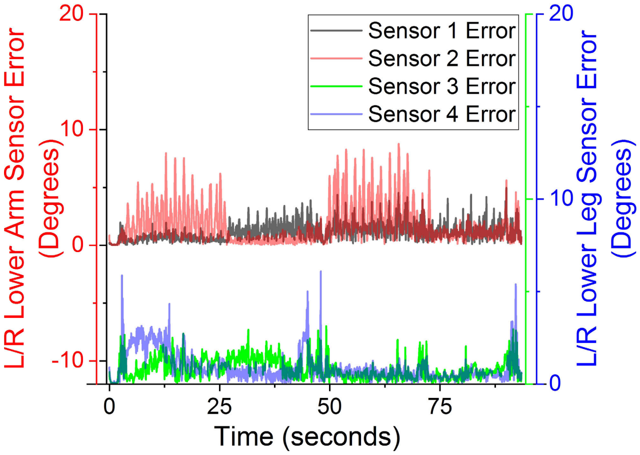

FIGURE 11. The error between real and ideal sensor orientation for a sequence in DIP-IMU dataset.

In Fig. 11, we depict the angular difference between *ideal* sensor data that we should have for a given 3D pose and the real sensor orientations for a typical sequence in DIP-IMU dataset. It is important to note that despite real sensors in DIP-IMU are perturbed, the nature and severity of perturbation is only 5–10 degrees and there is no drift. This alone does not explain the high error that we get on DIP-IMU, when we estimate 3D pose from this dataset using a synthetic data trained model. Although not shown here, but sensor errors on Total capture are also of same order, yet we obtain much better 3D pose estimation on Total capture, even without fine tuning.

H. EFFECT OF FINE TUNING

Since no significant covariate shift was found out by comparing the real and ideal sensor data, we investigated the domain shift of input sequences. Earlier, we observed for synthetic test data, that certain sequences displayed unusually high error (Fig. 1). Similarly, we hypothesize that the temporal patterns in the DIP-IMU dataset are different than the sequences on which our model is trained; therefore the error is more on DIP-IMU. In order to validate this *hypothesis*, we first tested the performance of a synthetic data trained model (without fine tuning) on real DIP-IMU data and then using the synthesized *ideal* DIP-IMU data. The results are depicted in Fig. 12. It is evident that both the real and ideal MIMU data result in almost same error for different sequences in DIP-IMU test data. Therefore we conclude that the noise and data augmentation that we used during training, make the model robust against real sensor errors.

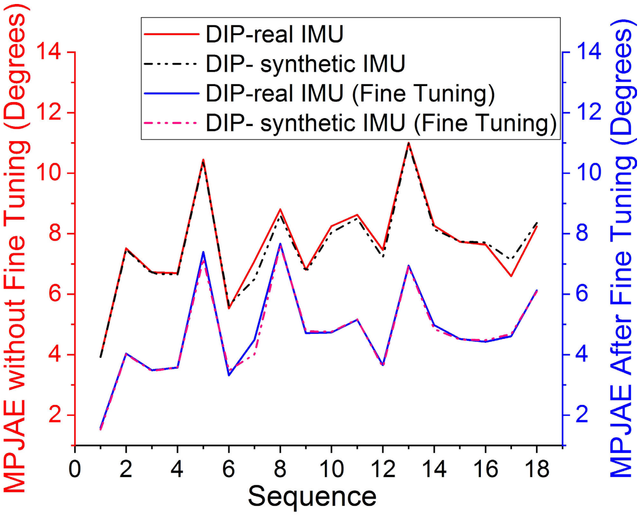

FIGURE 12. Performance of trained model before and after fine tuning on DIP-IMU sequences.

Next, we fine-tuned the model on a subset of 'synthetic' DIP-IMU data (not the same on which test results are depicted in Fig. 12). Using this fine-trained model, we again tested for the error on DIP-IMU test sequences, using both *ideal* sensors and real sensor readings. The results again shown in Fig. 12, clearly depict a decrease in error. Therefore we conclude that fine-tuning on a class of activities in a dataset, obtains better error performance on test data from the same set, especially if the temporal patterns in such a dataset are under-represented in previous training.

It is also interesting to note the effect of fine-tuning on uncertainty estimation in case of DIP-IMU. Fig. 13 shows the output 3D joint angle with uncertainty ($\pm 1\delta$) bounds for a test sequence from DIP-IMU, estimated using a trained model without fine-tuning. Unlike Total capture dataset, where the estimation error was lower and uncertainty also scaled well with the estimation error (Fig. 8), we observe that in case of DIP-IMU not only error is higher but also the uncertainty is underestimated.

We then depict the results in Fig. 14 for same sequence of DIP-IMU, obtained using a model fine-tuned on training data of DIP-IMU. Clearly not only the error has much reduced in this case, but uncertainty is also better calibrated now, after fine-tuning.

Therefore we conclude that only a small training set even with 'synthetic' sensor orientations is sufficient to achieve good accuracy and well-calibrated uncertainty on unseen temporal patterns. For instance, the ground truth for

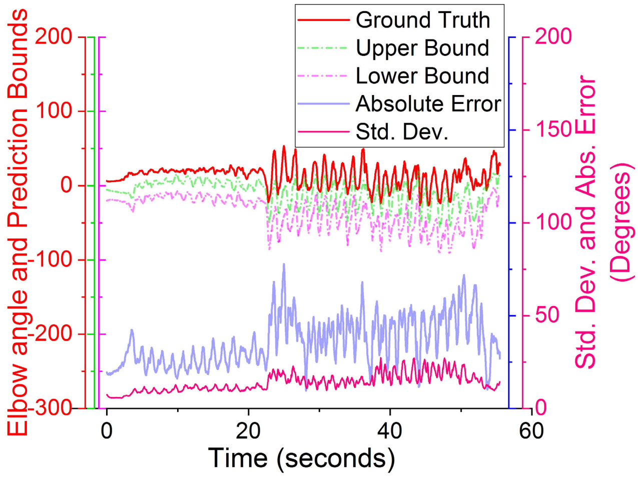

FIGURE 13. Uncertainty estimation and actual error on DIP-IMU data sequence, using trained 6-param model (without fine-tuning).

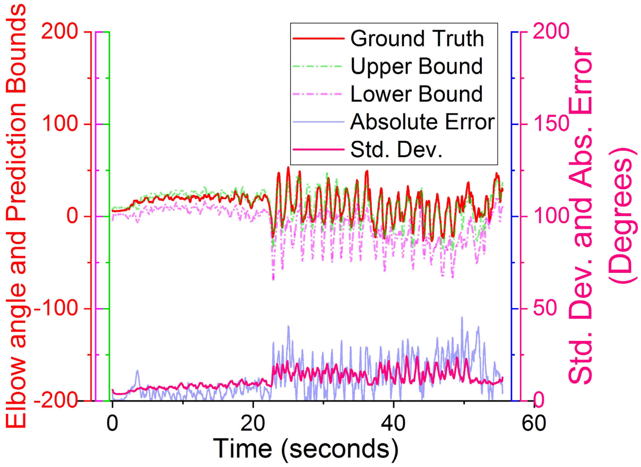

FIGURE 14. Uncertainty estimation and actual error on DIP-IMU data sequence, using trained 6-param model (after fine-tuning).

a set of training exercises can be obtained in an elaborate optical or inertial capture set-up. The 'synthetic' sensor data can then prepared from ground truth poses. Once the pre-trained model is fine-tuned using this small dataset, it can then be deployed in arbitrary setting with real sensors (only six) for inference.

I. SIMULATED MAGNETC INTERFERENCE

We see from Fig. 11 that DIP-IMU data is not strongly perturbed by magnetic interference (error between 5–10 degrees only). The covariate shift for Total capture data has been found out to be of same order. But in general the MIMU orientation is strongly affected by magnetic interference especially indoors and it can be quite significant [28]. Also when only IMU (rate gyro/accelerometer) is used without magnetometer, it leads to constant drift in yaw part of 3D angle. It is with these considerations in mind, that we trained a model on 3-parameter (pitch/roll) representation of sensor input. Since the two real MIMU datasets are not affected much by either yaw drift or strong interference, we tested the 3-parameter versus 6-parameter model on DIP-synthetic IMU data (as obtained in previous section), after it is corrupted by *simulated* yaw drift and strong magnetic interference [58]. The results of performance comparison are shown in Table 4. The models used are first fine-tuned for DIP-synthetic IMU data without magnetic perturbations and then tested on DIP-IMU magnetically perturbed data.

TABLE 4. Performance under magnetic interference.

Model	MPJAE[*] (deg)		
	Distal [§]	Tracking [§]	Other [§]
Our 6-param Model(Fine-tuned on DIP-IMU)	18.27	4.04	9.12
Our 3-param Model(Fine-tuned on DIP-IMU)	22.54	13.10	16.24
Our 6-param Model(Tested on DIP-IMU perturbed)	**45.70**	**24.02**	**35.13**
Our 3-param Model(Tested on DIP-IMU perturbed)	23.11	13.90	17.01

[*] MPJAE is Mean Per Joint Absolute error in degrees.
[§] Distal Joints are Left/right shoulder and hip joints. Tracking joints are proximal to where sensors are located (Knee, Elbow, Head). Other include all other major joints predicted with the model.

As evident from Table 4, when DIP-IMU data is perturbed by yaw drift and strong magnetic interference, the performance of 6-parameter model is much worse than 3-parameter model. There is almost no significant degradation in performance of 3-parameter model for perturbed (yaw) and unperturbed DIP-IMU data. However when perturbation is negligible (or none), 6-parameter model indeed performs better than 3-parameter model, as shown earlier.

VI. LIMITATIONS

As evident from discussion in Section V-H, the major limitation of our work is the dependence on temporal priors due to sparse sensors information. Although we used a synthetic dataset rich in activities and movements with data augmentation, it is not exhaustive for data-driven learning. We found that fine-tuning on a small subset of activities or movements on which prediction is required, addresses this limitation quite well. Another limitation of our work is that we did not use an inertial motion capture dataset with sequences collected in both homogenous magnetic field as well as in highly perturbed magnetic environment. The DIP-IMU and Total Capture datasets have been found out to be very 'clean' (both employ full 13-sensor MVN algorithm and then isolate 6-sensors data for evaluation). However if there would be only 6 sensors on the body then the individual sensor fusion (orientation estimation) would not close the gap between ideal synthetic and real MIMU data so well. In order to do more realistic testing, we had to simulate the magnetic interference to test the robustness of our proposed 3-parameter (pitch/roll) sensor input model. Although we demonstrated its robustness vis a vis full 6-parameter sensor orientation in magnetically perturbed scenario, this still needs to be evaluated more with real perturbations. A dataset collection with SMPL ground truth in magnetically challenging environments is our next goal for this testing.

VII. CONCLUSION

Our work proposes an uncertainty aware bi-directional deep recurrent model to estimate 3D human pose from only six magnetic-inertial measurement units. To the best of our knowledge, our model is the first to provide and test uncertainty estimation for this problem. Our model output the 3D pose directly in exponential map representation of SMPL. This avoids the renormalization of output as required in case of quaternions or rotation matrices. Also the estimation of uncertainty and its interpretation is straightforward. The definition of SMPL avoids the discontinuities in exponential map, owing to joint constraints. For sensors input, we propose a new 6-parameter representation for 3D orientation, which avoids the singularity and ambiguity in input space. In order to deal with magnetic perturbations, we further introduce a 'reduced' 3-parameter representation for input sensor orientation. This ignores the yaw part in 3D orientation. Our results show that even this reduced 3-parameter (pitch/roll) representation accomplishes 3D pose estimation albeit with higher uncertainty. The uncertainty estimated as a part of our model output, is found to be well correlated with ambiguity and actual error. We test our model on two real MIMU datasets and show that the major limitation in sparse sensor based 3D human pose estimation is the need to train on representative motion sequences, on which prediction is required. Our model can be used both in offline mode for end-to-end bi-directional inference or in online-mode using a moving window over inputs at run-time.

APPENDIX A

In this appendix, we derive our 6-parameter representation from 3 × 3 rotation matrix. We also demonstrate that theoretically, it is at par with 6-parameter representation of Zhou *et al.* [42], but gives us an advantage that we can easily obtain a reduced 3-parameter representation from our 6-parameters, by *masking* three parameters in it.

The 3 × 3 rotation matrix is an over-complete representation of a 3D angle. There are six constraints on its 9 parameters, which reduce it to 3DoF. These constarints arise from vector cross-product of its three rows and columns. Zhou *et al.* [42] have used this fact to define a 6-parameter representation. They show that the remaining 3-parameters can be uniquely determined from the first two columns of a rotation matrix by a righ handed cross-product.

Motivated by [42], we also define a 6-parameter representation using cross-product constraints. We first identify 5 parameters comprising first column and last row of a rotatio matrix. Using a cross product of column 1 and 2 and 3 and 1 respectively, we obtain following

$$(c_{21}c_{32} - c_{31}r_{22}) = r_{13} \tag{9a}$$

$$(c_{11}c_{32} - c_{31}r_{12}) = -r_{23} \tag{9b}$$

$$(c_{11}r_{22} - c_{21}r_{12}) = c_{33} \tag{9c}$$

$$(c_{31}r_{23} - c_{21}c_{33}) = c_{32} \tag{9d}$$

where $c_{11}, c_{21}, c_{31}, c_{32}, c_{33}$ are the components of our 6-parameters representation which are same as the corresponding components of 3 × 3 matrix. The unknown components $r_{12}, r_{13}, r_{22}, r_{23}$ of rotation matrix can be unambiguously obtained from (9). The only exception is when $c_{11}, c_{21}, c_{31}, c_{32}, c_{33} := [0, 0, 1, 0, 0]$, where above equations converge to a singleton solution, whereas infinite number of solution exist. Therefore in order to resolve this ambiguity, we also include $c_{22} = r_{22}$ in our representation.

Compared to [42], the advantage of our 6-parameter representation is that we can uniquely identify, the components which correspond to yaw, pitch and roll, as given by (3) and hence drop the components c_{11}, c_{21} and c_{22}, in order to obtain a yaw-free 3-parameter representation that we propose to be used for magnetically perturbed environments. The results given in Table 1 show that our 6-parameter representation is at par with 6-parameter representation proposed by Zhou *et al.* [42].

APPENDIX B

In this appendix, we justify the use exponential map representation for the output 3D human pose of our model, instead of quaternion [21], rotation matrix [20] or 6-paramater [42]. We show that SMPL skeleton, owing to joint constraints of human body does not present any discontinuity in exponential map representation of 3D joint angles. Since demonstrating this rigorously for human joints is non-trivial, we chose the data-driven approach [41] and check if by performing extreme range of human motion, any of the 3D joint angle reach their limits $[\pi, -\pi]$ radians in exponential map representation of SMPL, where such discontinuity might arise. We did this analysis for the complete SMPL dataset and found out that none of the joints ever reach the limits of $[\pi, -\pi]$ radians and that discontinuity does not occur. In Fig. 15 for clarity we show only the sequences in which extreme range of motion of a joint is performed. The SMPL data for this was obtained from both Joint Limit [54] and Total capture [52], in which extreme motions are present.

Clearly there is no discontinuity seen in SMPL data. The maximum range of motion occurs for x and y-component of Left/Right Knee and Elbow Joints in SMPL exponential map representation respectively but these are still within [150, −150] degrees. The shoulder and hip joints are well within [100, −100] degrees in SMPL exponential map representation. Thus the discontinuity of expmap representation i.e. at [180, −180] degrees is completely avoided in SMPL.

There is obvious advantage of regressing 3D human pose directly in exponential map representation of SMPL. No orthogonalization of model output is needed like quaternions or rotation matrix. The uncertainty which is predicted as a part of output can be also directly interpreted. If 3D angle output is estimated in quaternions or rotation matrix, it is

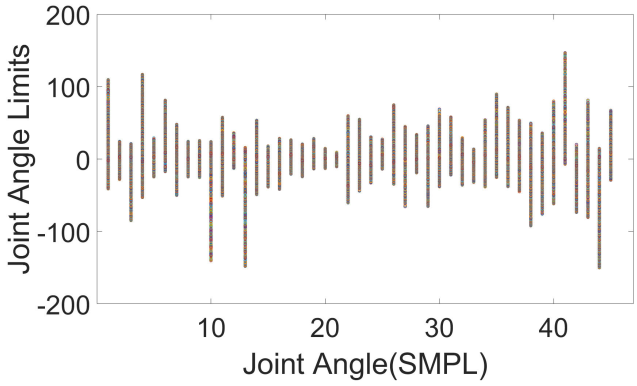

FIGURE 15. Joint limits constraint in exponential map (SMPL).

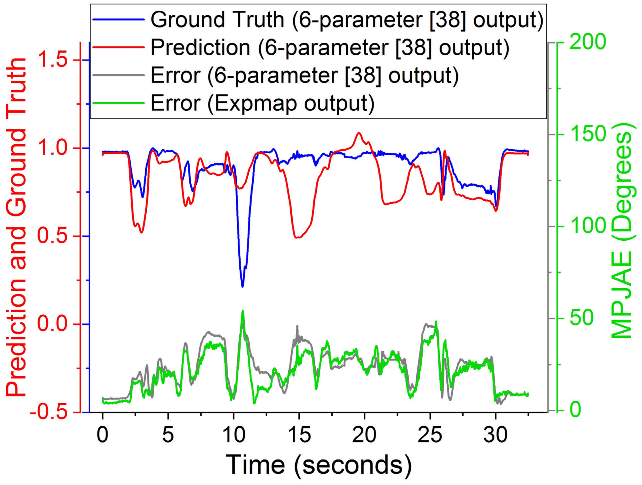

FIGURE 16. Exponential map output versus 6-parameters [38] output.

observed (Fig. 16) that few components violate the constraint of unit norm. The orthogonalization is therefore necessary as a post-processing or additional step.

This on average leads to slightly more error compared to exponential map as shown in Fig. 16. Also the computation of uncertainty in case of quaternion, 6-parameter or 9-parameter representations, require the use of unscented transform and we note that due to orthogonalization step, the probability distribution of output samples is distorted. This again introduces errors in the uncertainty prediction as well.

ACKNOWLEDGMENT

The research was sponsored as a part of BIONIC (Body Information on an Intelligent Chip) project at German Research Centre for Artificial Intelligence (DFKI) Germany, funded by European Commission under Horizon 2020 Programme Grant Agreement No. 826304 in collaboration with 10 other partners. The authors thank the National University of Sciences and Technology (NUST), Pakistan, and Technische Universtät (TU) Kaiserslautern, Germany for their support. They also thank Huang *et al.* [20] for making their datasets/code public which enabled us to report comparisons with their work.

REFERENCES

[1] *Xsens Homepage—Motion Capture Products-MVN Awinda, MVN Link*. Accessed: Sep. 1, 2020. [Online]. Available: https://www.xsens.com/products/

[2] I. Weygers, M. Kok, H. De Vroey, T. Verbeerst, M. Versteyhe, H. Hallez, and K. Claeys, "Drift-free inertial sensor-based joint kinematics for long-term arbitrary movements," *IEEE Sensors J.*, vol. 20, no. 14, pp. 7969–7979, Jul. 2020, doi: 10.1109/JSEN.2020.2982459.

[3] T. Maruyama, M. Tada, A. Sawatome, and Y. Endo, "Constraint-based real-time full-body motion-capture using inertial measurement units," in *Proc. IEEE-SMC*, Miyazaki, Japan, Oct. 2018, pp. 4298–4303.

[4] J. K. Lee and T. H. Jeon, "Magnetic condition-independent 3D joint angle estimation using inertial sensors and kinematic constraints," *Sensors*, vol. 19, no. 24, p. 5522, Dec. 2019, doi: 10.3390/s19245522.

[5] G. Tao, Z. Huang, Y. Sun, S. Yao, and J. Wu, "Biomechanical model-based multi-sensor motion estimation," in *Proc. IEEE-SAS*, Galveston, TX, USA, Feb. 2013, pp. 156–161, doi: 10.1109/SAS.2013.6493577.

[6] W. Teufl, M. Miezal, B. Taetz, M. Fröhlich, and G. Bleser, "Validity, test-retest reliability and long-term stability of magnetometer free inertial sensor based 3D joint kinematics," *Sensors*, vol. 18, no. 7, p. 1980, Jun. 2018, doi: 10.3390/s18071980.

[7] C. Mousas, "Full-body locomotion reconstruction of virtual characters using a single inertial measurement unit," *Sensors*, vol. 17, no. 11, p. 2589, Nov. 2017, doi: 10.3390/s17112589.

[8] J. Tautges, A. Zinke, B. Krüger, J. Baumann, A. Weber, T. Helten, M. Müller, H.-P. Seidel, and B. Eberhardt, "Motion reconstruction using sparse accelerometer data," *ACM Trans. Graph.*, vol. 30, no. 3, pp. 1–12, May 2011, doi: 10.1145/1966394.1966397.

[9] H. Eom, B. Choi, and J. Noh, "Data-driven reconstruction of human locomotion using a single smartphone," *Comput. Graph. Forum*, vol. 33, no. 7, pp. 11–19, Oct. 2014, doi: 10.1111/cgf.12469.

[10] L. A. Schwarz, D. Mateus, and N. Navab, "Discriminative human full-body pose estimation from wearable inertial sensor data," in *Proc. 3DPH*, Zermatt, Switzerland, 2009, pp. 159–172, doi: 10.1007/978-3-642-10470-1_14.

[11] L. A. Schwarz, D. Mateus, and N. Navab, "Multiple-activity human body tracking in unconstrained environments," in *Proc. AMDO*, Mallorca, Spain, 2010, pp. 192–202, doi: 10.1007/978-3-642-14061-7_19.

[12] F. Wouda, M. Giuberti, G. Bellusci, and P. Veltink, "Estimation of full-body poses using only five inertial sensors: An eager or lazy learning approach?" *Sensors*, vol. 16, no. 12, p. 2138, Dec. 2016, doi: 10.3390/s16122138.

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

[14] T. von Marcard, B. Rosenhahn, M. J. Black, and G. Pons-Moll, "Sparse inertial poser: Automatic 3D human pose estimation from sparse IMUs," in *Proc. Eurographics*, Lyon, France, 2017, pp. 349–360.

[15] X. Hu, C. Yao, and G. S. Soh, "Performance evaluation of lower limb ambulatory measurement using reduced inertial measurement units and 3R gait model," in *Proc. IEEE-ICORR*. Singapore: NTU, Aug. 2015, pp. 549–554, doi: 10.1109/ICORR.2015.7281257.

[16] L. Sy, M. Raitor, M. Del Rosario, H. Khamis, L. Kark, N. H. Lovell, and S. J. Redmond, "Estimating lower limb kinematics using a reduced wearable sensor count," 2019, *arXiv:1910.00910*. [Online]. Available: http://arxiv.org/abs/1910.00910

[17] M. El-Gohary, L. Holmstrom, J. Huisinga, E. King, J. McNames, and F. Horak, "Upper limb joint angle tracking with inertial sensors," in *Proc. IEEE-EMBC*, Boston, MA, USA, Aug. 2011, pp. 5629–5632, doi: 10.1109/IEMBS.2011.6091362.

[18] J. Kim, Y. Seol, and J. Lee, "Realtime performance animation using sparse 3D motion sensors," in *Proc. MIG*, Rennes, France, 2012, pp. 31–42, doi: 10.1007/978-3-642-34710-8_4.

[19] H. Liu, X. Wei, J. Chai, I. Ha, and T. Rhee, "Realtime human motion control with a small number of inertial sensors," in *Proc. ACM-I3D*, San Francisco, CA, USA, 2011, pp. 133–140, doi: 10.1145/1944745.1944768.

[20] 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, doi: 10.1145/3272127.3275108.

[21] F. J. Wouda, M. Giuberti, N. Rudigkeit, B.-J.-F. van Beijnum, M. Poel, and P. H. Veltink, ''Time coherent full-body poses estimated using only five inertial sensors: Deep versus shallow learning,'' *Sensors*, vol. 19, no. 17, p. 3716, Aug. 2019, doi: 10.3390/s19173716.

[22] A. Gilbert, M. Trumble, C. Malleson, A. Hilton, and J. Collomosse, ''Fusing visual and inertial sensors with semantics for 3D human pose estimation,'' *Int. J. Comput. Vis.*, vol. 127, no. 4, pp. 381–397, Apr. 2019, doi: 10.1007/s11263-018-1118-y.

[23] M. Miezal, B. Taetz, and G. Bleser, ''Real-time inertial lower body kinematics and ground contact estimation at anatomical foot points for agile human locomotion,'' in *Proc. ICRA*, Singapore, May 2017, pp. 3256–3263, doi: 10.1109/ICRA.2017.7989371.

[24] G. Ligorio, E. Bergamini, L. Truppa, M. Guaitolini, M. Raggi, A. Mannini, A. M. Sabatini, G. Vannozzi, and P. Garofalo, ''A wearable magnetometer-free motion capture system: Innovative solutions for real-world applications,'' *IEEE Sensors J.*, vol. 20, no. 15, pp. 8844–8857, Aug. 2020, doi: 10.1109/JSEN.2020.2983695.

[25] J.-A. Lee, S.-H. Cho, J.-W. Lee, K.-H. Lee, and H.-K. Yang, ''Wearable accelerometer system for measuring the temporal parameters of gait,'' in *IEEE-EMBC*, Lyon, France, Aug. 2007, pp. 483–486, doi: 10.1109/IEMBS.2007.4352328.

[26] M. Ockendon and R. Gilbert, ''Validation of a novel smartphone accelerometer-based knee goniometer,'' *J. Knee Surg.*, vol. 25, no. 4, pp. 341–346, May 2012, doi: 10.1055/s-0031-1299669.

[27] R. Slyper and J. K. Hodgins, ''Action capture with accelerometers,'' in *Proc. ACM-SIGGRAPH*, Dublin, Ireland, 2008, pp. 193–199, doi: 10.5555/1632592.1632620.

[28] H. T. Butt, M. Pancholi, M. Musahl, M. A. Sanchez, and D. Stricker, ''Development of high rate wearable MIMU tracking system robust to magnetic disturbances and body acceleration,'' in *Proc. SAI-IntelliSys*, London, U.K., 2019, pp. 1178–1198.

[29] A. Cereatti, D. Trojaniello, and U. D. Croce, ''Accurately measuring human movement using magneto-inertial sensors: Techniques and challenges,'' in *Proc. IEEE-ISISS*, Hapuna Beach, HI, USA, Mar. 2015, pp. 1–4, doi: 10.1109/ISISS.2015.7102390.

[30] G. Bleser, B. Taetz, M. Miezal, C. A. Christmann, D. Steffen, and K. Regenspurger, ''Development of an inertial motion capture system for clinical application: Potentials and challenges from the technology and application perspectives,'' *i-com*, vol. 16, no. 2, pp. 113–129, Aug. 2017, doi: 10.1515/icom-2017-0010.

[31] M. Kok, J. D. Hol, and T. B. Schön, ''An optimization-based approach to human body motion capture using inertial sensors,'' in *Proc. IFAC*, Cape Town, South Africa, 2014, pp. 79–85, doi: 10.3182/20140824-6-ZA-1003.02252.

[32] G. Liu, J. Zhang, W. Wang, and L. McMillan, ''Human motion estimation from a reduced marker set,'' in *Proc. ACM-I3D*, 2006, pp. 35–42, doi: 10.1145/1111411.1111418.

[33] J. Chai and J. K. Hodgins, ''Performance animation from low-dimensional control signals,'' *ACM Trans. Graph.*, vol. 24, no. 3, pp. 686–696, Jul. 2005, doi: 10.1145/1073204.1073248.

[34] S. Andrews, I. Huerta, T. Komura, L. Sigal, and K. Mitchell, ''Real-time physics-based motion capture with sparse sensors,'' in *Proc. CVMP*, London, U.K., 2016, pp. 1–10, doi: 10.1145/2998559.2998564.

[35] F. J. Wouda, M. Giuberti, G. Bellusci, B.-J. F. van Beijnum, and P. H. Veltink, ''Improving full-body pose estimation from a small sensor set using artificial neural networks and a Kalman filter,'' in *Proc. AAAI*, Honolulu, HI, USA, 2019, pp. 10063–10064, doi: 10.1609/aaai.v33i01.330110063.

[36] K. Eckhoff, M. Kok, S. Lucia, and T. Seel, ''Sparse magnetometer-free inertial motion tracking—A condition for observability in double hinge joint systems,'' 2020, *arXiv:2002.00902*. [Online]. Available: http://arxiv.org/abs/2002.00902

[37] I. Katircioglu, B. Tekin, M. Salzmann, V. Lepetit, and P. Fua, ''Learning latent representations of 3D human pose with deep neural networks,'' *Int. J. Comput. Vis.*, vol. 126, no. 12, pp. 1326–1341, Dec. 2018, doi: 10.1007/s11263-018-1066-6.

[38] D. Holden, J. Saito, and T. Komura, ''A deep learning framework for character motion synthesis and editing,'' *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–11, Jul. 2016, doi: 10.1145/2897824.2925975.

[39] Y. Park, S. Moon, and I. H. Suh, ''Tracking human-like natural motion using deep recurrent neural networks,'' 2016, *arXiv:1604.04528*. [Online]. Available: http://arxiv.org/abs/1604.04528

[40] S. Li, Y. Zhou, H. Zhu, W. Xie, Y. Zhao, and X. Liu, ''Bidirectional recurrent autoencoder for 3D skeleton motion data refinement,'' *Comput. Graph.*, vol. 81, pp. 92–103, Jun. 2019, doi: 10.1016/j.cag.2019.03.010.

[41] P. Murthy, H. T. Butt, S. Hiremath, A. Khoshhal, and D. Stricker, ''Learning 3D joint constraints from vision-based motion capture datasets,'' *IPSJ-Trans. Comput. Vis. Appl.*, vol. 11, pp. 1–9, Jun. 2019, doi: 10.1186/s41074-019-0057-z.

[42] 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. 5745–5753, doi: 10.1109/CVPR.2019.00589.

[43] V. Peretroukhin, B. Wagstaff, and J. Kelly, ''Deep probabilistic regression of elements of $SO_{(3)}$ using quaternion averaging and uncertainty injection,'' in *Proc. CVPR Workshops*, Long Beach, CA, USA, 2019, pp. 83–86.

[44] R. L. Russell and C. Reale, ''Multivariate uncertainty in deep learning,'' 2019, *arXiv:1910.14215*. [Online]. Available: http://arxiv.org/abs/1910.14215

[45] D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski, ''DeepAR: Probabilistic forecasting with autoregressive recurrent networks,'' *Int. J. Forecasting*, vol. 36, no. 3, pp. 1181–1191, Jul. 2020, doi: 10.1016/j.ijforecast.2019.07.001.

[46] L. Zhu and N. Laptev, ''Deep and confident prediction for time series at Uber,'' in *Proc. IEEE-ICDMW*, New Orleans, LA, USA, Nov. 2017, pp. 103–110, doi: 10.1109/ICDMW.2017.19.

[47] R. Wen, K. Torkkola, B. Narayanaswamy, and D. Madeka, ''A multi-horizon quantile recurrent forecaster,'' 2017, *arXiv:1711.11053*. [Online]. Available: http://arxiv.org/abs/1711.11053

[48] D. Kivaranovic, K. D. Johnson, and H. Leeb, ''Adaptive, distribution-free prediction intervals for deep networks,'' 2019, *arXiv:1905.10634*. [Online]. Available: http://arxiv.org/abs/1905.10634

[49] V. Kuleshov, N. Fenner, and S. Ermon, ''Accurate uncertainties for deep learning using calibrated regression,'' 2018, *arXiv:1807.00263*. [Online]. Available: http://arxiv.org/abs/1807.00263

[50] M.-H. Laves, S. Ihler, J. F. Fast, L. A. Kahrs, and T. Ortmaier, ''Well-calibrated regression uncertainty in medical imaging with deep learning,'' in *Proc. PMLR*, Montreal, QC, Canada, 2020, pp. 393–412.

[51] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. Black, ''AMASS: Archive of motion capture as surface shapes,'' in *Proc. IEEE-ICCV*, Seoul, South Korea, Oct. 2019, pp. 5442–5451, doi: 10.1109/ICCV.2019.00554.

[52] 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, doi: 10.5244/C.31.14.

[53] T. V. Marcard, G. Pons-Moll, and B. Rosenhahn, ''Human pose estimation from video and IMUs,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1533–1547, Aug. 2016, doi: 10.1109/TPAMI.2016.2522398.

[54] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, ''Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1325–1339, Jul. 2014, doi: 10.1109/TPAMI.2013.248.

[55] F. De la Torre, J. Hodgins, A. Bargteil, X. Martin, J. Macey, A. Collado, and P. Beltran, ''Guide to the Carnegie Mellon University multimodal activity (CMU-MMAC) database,'' Robot. Inst., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-RI-TR-08-22, Apr. 2008. [Online]. Available: https://www.ri.cmu.edu/pub_files/pub4/de_la_torre_frade_fernando_2008_1/de_la_torre_frade_fernando_2008_1.pdf

[56] L. Sigal, A. O. Balan, and M. J. Black, ''HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion,'' *Int. J. Comput. Vis.*, vol. 87, nos. 1–2, pp. 4–27, Mar. 2010, doi: 10.1007/s11263-009-0273-6.

[57] I. Akhter and M. J. Black, ''Pose-conditioned joint angle limits for 3D human pose reconstruction,'' in *Proc. CVPR*, Boston, MA, USA, Jun. 2015, pp. 1446–1455, doi: 10.1109/CVPR.2015.7298751.

[58] 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.

[59] B. Lakshminarayanan, A. Pritzel, and C. Blundell, ''Simple and scalable predictive uncertainty estimation using deep ensembles,'' in *Proc. NIPS*, Long Beach, CA, USA, 2017, pp. 6405–6416, doi: 10.5555/3295222.3295387.

HAMMAD TANVEER BUTT was born in Pakistan, in July 1976. He received the B.E. degree in avionics engineering and the M.Sc. degree in electrical engineering from the National University of Sciences and Technology, Islamabad, Pakistan, in 1999 and 2013, respectively. He is currently pursuing the Ph.D. degree in computer science with the Technische Universität (TU), Kaiserslautern, Germany.

Since 2016, he has been a Research Assistant with the Augmented Vision Group, German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany, where he is a part of the Body Information on an Intelligent Chip (BIONIC) project. He has a prior research experience in Technology CAD of III-IV semiconductor devices and metal organic Solar cells. He has focused on the core area of uncertainty aware deep learning for sensors and biomedical data. His current research interests include machine learning, deep learning, MEMS signal processing, and biomechanical and biomedical wearable sensor technology.

BERTRAM TAETZ received the B.Sc. and M.Sc. degrees in applied mathematics, minor subject physics, and the Ph.D. degree from Ruhr University Bochum, Germany, in 2009 and 2012, respectively.

Since 2013, he has been working as a Senior Researcher with the German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany. He joined the Department of Computer Science, University of Kaiserslautern, in 2015. During his Ph.D., he focused on numerical methods for dynamical systems. His research interests include numerical and statistical methods for motion estimation and dynamical systems.

MATHIAS MUSAHL received the Diploma degree in electrical engineering from TU Kaiserslautern, Kaiserslautern, Germany, in 2012.

From 2012 to 2017, he was working as a Software Developer in the field of network-based Intercom systems. He is currently working as a Researcher with the Augmented Vision Group, German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany. His research interests include the development of intelligent sensor networks and the design and implementation of highly portable and modular hardware and software architectures for on body processing.

MARIA A. SANCHEZ received the B.S. degree in electronics engineering from Simon Bolivar University, Venezuela, in 2014, and the M.S. degree in embedded systems from the Technical University of Kaiserslautern, Germany, in 2017.

From 2015 to 2016, she was a Research Assistant with the Fraunhofer Institute of Industrial Mathematics ITWM, Kaiserslautern, Germany. She is currently working with the German Research Center for Artificial Intelligence (DFKI), Augmented Vision Department, as part of the Body Sensor Network Group. Her research interests include the development of robust, flexible embedded systems to do motion tracking and posture estimation, to be used in a wide variety of applications like wearable, sports, rehabilitation, and collaborative gaming systems.

PRAMOD MURTHY received the B.E. degree in computer science and engineering from S.R.T.M University, Nanded, India, in 2005, and the M.S. degree in computer science from TU Kaiserslautern, Germany, in 2016, where he is currently pursuing the Ph.D. degree.

From 2006 to 2011, he worked as different engineering roles on System software and Natural Language Processing for Indic languages. Since 2016, he has been working as a Research Assistant with the Department of Augmented Vision, German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany. His current research interests include modeling human motion dynamics from monocular images and uncertainty estimation using Bayesian deep learning.

DIDIER STRICKER led the Department of Virtual and Augmented Reality, Fraunhofer Institute for Computer Graphics, Darmstadt, Germany, from 2002 to 2008. He is currently a Professor with the Department of Computer Science, University of Kaiserslautern, Germany. He is also a Scientific Director with the German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany, where he leads the Augmented Vision Research Group. His research interests include 3D computer vision, autonomous driving, wearable health, augmented reality applications, and deep learning.

He received the Innovation Prize of the German Society of Computer Science, in 2006. He serves as a Reviewer for noteworthy journals in the area of VR/AR and computer vision.