ActivityPoser: Activity driven Full-Body Pose Estimation from Sparse IMU Configurations

Karan Ahuja
Carnegie Mellon University
Pittsburgh, PA, USA
kahuj@cs.cmu.edu

Eric Whitmire
Reality Labs Research, Meta
Redmond, WA, USA
ewhitmire@meta.com

Joseph D Greer
Reality Labs Research, Meta
Portland, OR, USA
jdgreer@meta.com

Wolf Kienzle
Reality Labs Research, Meta
Redmond, WA, USA
wkienzle@meta.com

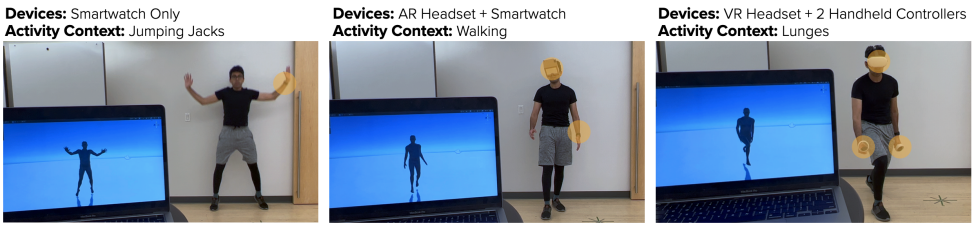

Figure 1: Using the activity context of jumping jacks, walking and lunges and inertial data from a smartwatch, smartwatch and AR headset and a consumer-grade VR headset and hand-held controllers respectively our system predicts the full-body pose.

ABSTRACT

On-body IMU-based pose tracking systems have gained prevalence over their external tracking counterparts due to their mobility, ease of installation and use. However, even in these systems, an IMU sensor placed on a particular joint can only estimate the pose of that particular limb. In contrast, activity recognition systems contain insights into the whole body's motion dynamics. In this work, we present ActivityPoser, which uses the activity context as a conditional input to estimate the pose of limbs for which we do not have any direct sensor data. ActivityPoser compensates for impoverished sensing paradigms by reducing the overall pose error by up to 17%, compared to a model bereft of activity context. This highlights a pathway to high-fidelity full-body digitization with minimal user instrumentation.

ACM Reference Format:
Karan Ahuja, Eric Whitmire, Joseph D Greer, and Wolf Kienzle. 2022. ActivityPoser: Activity driven Full-Body Pose Estimation from Sparse IMU Configurations. In *Symposium on Spatial User Interaction (SUI '22), December 1–2, 2022, Online, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3565970.3567687

1 INTRODUCTION

In recent years, on-body sensing systems have gained prevalence over their external tracking counterparts due to their mobility, ease of setup and use. In particular, Inertial Measurement Unit (IMU) based systems have gained a lot of traction due to their affordable cost, low power consumption and ubiquity in consumer devices such as smartphones and smartwatches. However, even in such systems, a sensor placed on a particular joint can generally only estimate the pose of that particular body keypoint. For example, a contemporary VR system such as the Oculus Quest can only track the user's instrumented head and hands.

In contrast, a user's activity contains information about the whole body's motion dynamics. Furthermore, Human Activity Recognition (HAR) techniques that make use of IMU data from a single sensor such as a smartphone or a smartwatch, or an array of on-body devices have become increasingly prevalent. In such cases, making use of the activity context inferred from the instrumented joint can help narrow down the pose of the non-instrumented joints. In response, we present ActivityPoser, a learning framework that takes the activity context as a conditional input for a higher fidelity full-body pose inference. Using activity as a prior, we create a custom neural network pose model that can compensate for the lack of sensors placed on the body and affords high-fidelity pose with minimal user instrumentation.

2 METHODS

There are many plausible configurations for placing the IMU sensors around the body, each unique in terms of their device placement, the number of sensors employed and the fidelity of captured pose. We consider 6 placement scenarios showcased in Figure 2.

We make use of a priori activity context and its corresponding inertial data as the system input. To encode our full-body pose output, we make use of the SMPL [2] body framework (72 pose angles representing 24 joints). For our learning framework, make use of a conditional bi-directional Recurrent Neural Network (RNN) with 128 long short-term memory (LSTM) cells. We initialize the states of our LSTM with a learned representation of the conditional input, in this case, the activity prior.

While it is intuitive that there exist multiple possible body poses for a given input - especially for cases where only a handful of joints are instrumented - our method aims to recover the best fitting full-body pose in a least-square sense, weighting joints inversely to

SUI '22, December 1–2, 2022, Online, CA, USA

ACM ISBN 978-1-4503-9948-7/22/12.
https://doi.org/10.1145/3565970.3567687

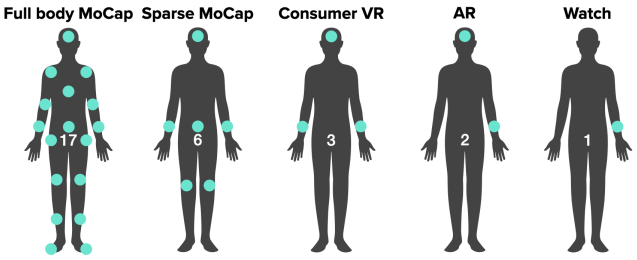

Figure 2: On-body inertial sensor placements.

their Euclidean distance from the point of instrumentation. We train a separate model for each 5 form factors. For model training, we use a batch size of 64 and update the weights using RMSProp solver with a learning rate of 0.001 for training. We train our model for 1000 epochs with TensorFlow on an NVIDIA Titan X GPU.

Our model outputs the rotations as 72 SMPL axis-angle pose parameters. We can perform forward kinematics to estimate the full-body pose in cartesian space from these rotations. However, a problem with such rotation-based regressors is that the rotation of a joint for which we have direct orientation data may differ from the predicted value. To account for this we make use of an inverse kinematic solver that adjusts the joint rotation along the kinematic chain according to the known rotation provided directly from the IMU sensor. This further allows us to animate an avatar and correct for unnatural poses.

3 EVALUATION

In order to study the relationship between different sensor configurations and placements, we need data with activity context, precise acceleration and orientations of each joint, along with the corresponding full-body pose. For our first dataset, we make use of the AMASS Dataset [3] which provides a large collection of MoCap data across a multitude of human activities. Similar to prior work [1], we synthesize synthetic inertial data from the AMASS dataset by placing virtual IMUs on the corresponding vertices of the SMPL mesh to generate acceleration and orientation data.

For preliminary evaluation, we make use of Deep Inertial Poser's (DIP) [1] IMU-based MoCap, to test the performance of our model on real-world data. It consists of real-world IMU data spanning across 5 activity classes: upper-body, lower-body, interaction, freestyle

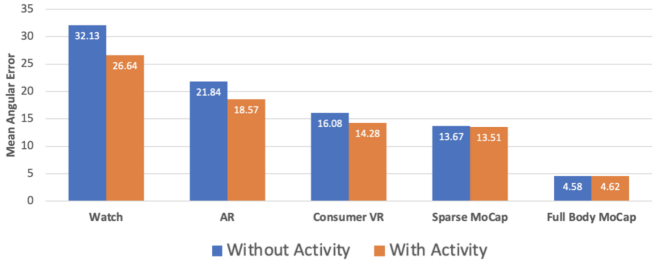

Figure 3: Mean Per Joint Rotation Error (in degrees) on DIP-IMU dataset across different sensor configurations.

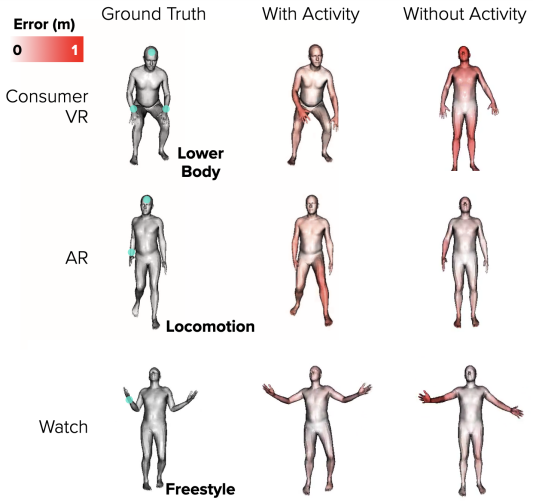

Figure 4: Predicted pose with activity and without.

and locomotion. Consistent with prior work [1] we fine-tune our model trained on the AMASS dataset on the DIP-IMU dataset making use of data from the first 8 participants as training, with the last 2 participants used for testing.

Figure 3 shows accuracy across different IMU configurations. The activity-driven model outperforms the pose-only model by 17.1% for the Watch scenario (1 sensor). For the AR (2 sensors) and Consumer VR (3 sensors) scenario, the accuracy increase is 14.9% and 11.2% respectively. This accuracy difference is negligible (1%) for the sparse MoCap (6 sensors) and full-body MoCap (17 sensors) scenario. Thus, as the sensor instrumentation of the whole body increases, we get a more holistic picture of the full-body pose. However, for impoverished sensing paradigms, the use of activity context helps compensate for the lack of sensors present.

This benefit of the activity context for pose prediction can further be seen in Figure 4. In the Consumer VR configuration, when we do not take the activity context into account, the motion of the lower body cannot be estimated when the user is squatting. The use of activity also produces more temporally coherent motions. This can be seen in the Locomotion example in the AR configuration, where the temporal gait of the hands corresponds to the walking motion and animates the corresponding legs. The same can be seen in the freestyle motion example when the person is skipping (Watch scenario). With activity context, the sensor placed on the right hand can glean insights about the position of the left hand despite having no direct sensor data for the same.

REFERENCES

[1] 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.

[2] Matthew Loper, Naureen Mahmood, and Michael J Black. 2014. MoSh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics (TOG)* 33, 6 (2014), 1–13.

[3] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. 2019. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5442–5451.