A Robust Low-cost Mocap System with Sparse Sensors

Seong Uk Kim
Department of Computer Science and Engineering
Kangwon National University

Hanyoung Jang
Motion AI Team
Game AI Lab.
NCSOFT

Jongmin Kim
Department of Computer Science and Engineering
Kangwon National University

ABSTRACT

In this paper, we propose a robust low-cost mocap system (mocap) with sparse sensors. Although the sensor with an accelerometer, magnetometer, and gyroscope is cost-effective and offers the measured positions and rotations from these devices, it potentially suffers from noise, drift, and lost issues over time. The resulting character obtained from a sensor-based low-cost mocap system is thus generally not satisfactory. We address these issues by using a novel deep learning framework that consists of two networks, a motion estimator and a sensor data generator. When the aforementioned issues occur, the motion estimator feeds the newly synthesized sensor data obtained with the measured and predicted data from the sensor data generator until the issues have been resolved. Otherwise, the motion estimator receives the measured sensor data to accurately and continuously reconstruct the new character poses. In our examples, we show that our system outperforms the previous approach without the sensor data generator and we believe that it can be considered a handy and robust mocap system.

CCS CONCEPTS

• **Computing methodologies** → **Animation**; **Deep Learning**.

KEYWORDS

Human Motion, Low-cost Mocap System, Deep Learning

ACM Reference Format:
Seong Uk Kim, Hanyoung Jang, and Jongmin Kim. 2020. A Robust Low-cost Mocap System with Sparse Sensors. In *Proceedings of SA '20 Posters*. ACM, New York, NY, USA, 2 pages.

1 INTRODUCTION

Mocap systems are now widely used as a tool for efficiently producing virtual characters in many VFX and gaming companies and are also utilized to analyze human movement for clinical and rehabilitation purposes. An optical mocap system is a sufficiently accurate method of recording human motions. However, it is very expensive and imposes tedious workloads when setting up a capture volume, as multiple cameras should be well positioned around the desired capture volume to stably detect the trajectories of several markers placed at key locations on the actor's body. Meanwhile, the current commercially available sensor-based mocap systems are less expensive and easy to use when compared to the optical mocap system but dozens of sensors are generally needed to compute a full body character pose. In addition, most of them are highly dependent on conventional Inverse Kinematics (IK) schemes, thereby often producing unnatural motions owing to insufficient reflection of details of real human movement. Also, the sensors that provide positions and rotations potentially suffer from noise, drift, and lost issues over time. Therefore, our goal is to establish a robust deep learning-based low-cost mocap system with a minimal number of sensors for recording a faithful full-body character.

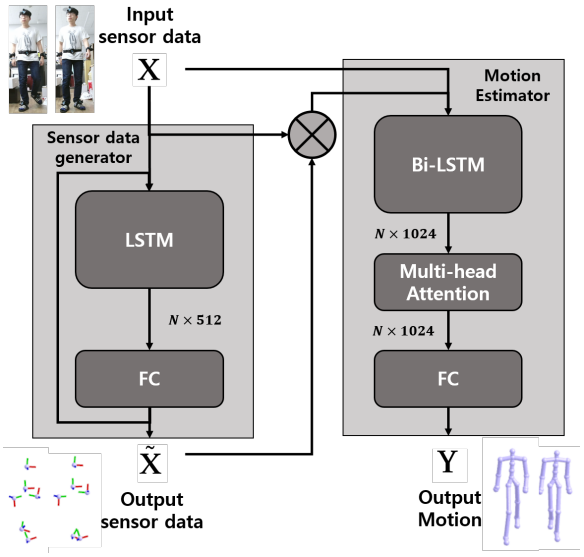

Figure 1: System overview: We independently train the sensor data generator and motion estimator and use them to reconstruct human motion from sparse sensors.

Our system is an interchangeable neural network system consisting of a motion estimator and a sensor data generator, as shown in Figure 1. The motion estimator is based on a Bidirectional Recurrent Neural Network (BRNN) trained using a large amount of mocap data to reconstruct human motion from six sensors attached on the actor's body. Since we utilize the BRNN for the past and future observations, our system leads to better performance for training sequential data than other networks. We also embed a multi-head attention mechanism [Vaswani et al. 2017] into our networks for selectively weighting hidden variables for better estimation of output motions and to achieve fast learning. A sensor data generator based

SA '20 Posters, November 17-20, 2020, Daegu, Republic of Korea

on the auto-conditioned Recurrent Neural Network (acRNN) [Zhou et al. 2018] is employed for future sensor data prediction from the observed sequences. When the measured sensor data are invalid, the motion estimator feeds the new sensor data computed with the measured and predicted data from the sensor data generator until the issue of invalid data is resolved. Otherwise, the measured sensor data are used as the motion estimator input to efficiently and continuously reconstruct character animation.

2 OUR APPROACH

We use the CMU mocap data to train our system ([CMU 2013]). In the preprocessing step, all characters from the CMU mocap data are retargeted to the template body configuration by solving the numerical IK, and each character consists of 31 joins in total. We use six sensors placed on the end-effectors and the root joint. The length of each motion clip and sensor data sequences n is 240. Training mocap data provide both input and output data for our neural network system. We extract the positions and rotations of the joints from them.

The input of the motion estimator and sensor data generator $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is composed of positions and rotations and the sensor data generator output is $\tilde{\mathbf{X}} = \{\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_n\}$. The output of the motion estimator is represented as the joint rotations $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$. The Long Short-Term Memory (LSTM) is used for the motion estimator and sensor data generator, and fully connected layers are also used for generating the output motion. Multi-head attention is employed with the BRNN for the motion estimator. When the measured sensor data are invalid, the new sensor data $\bar{\mathbf{X}}$ are obtained by independently interpolating the sensor positions and rotations. The interpolated $\bar{x}_t^i$ at the i-th sensor of the t-th frame is defined as $\bar{x}_t^i = \tilde{\mathbf{x}}_{T'} \otimes (\mathbf{x}_{t+T} \oslash \tilde{\mathbf{x}}_{T'})^{\alpha_t^i}$, where α_t^i is an ease-in/ease-out function with $0 \leq \alpha_t^i \leq 1$ and $T' = t + T/2$ (see Figure 2 for more details). We would like to refer the reader to [Lee 2008] for the mathematical notations. The computed $\bar{\mathbf{X}}$ is entered into the motion estimator and the character animation $\mathbf{Y}$ is then obtained.

We train the network in such a way as to minimize the prediction and smoothness loss functions. Note that the sensor data generator and the motion estimator are trained independently. The $\mathcal{L}_2$ norm distance is defined as the error between the predicted and training motion and we also regularize the network parameters.

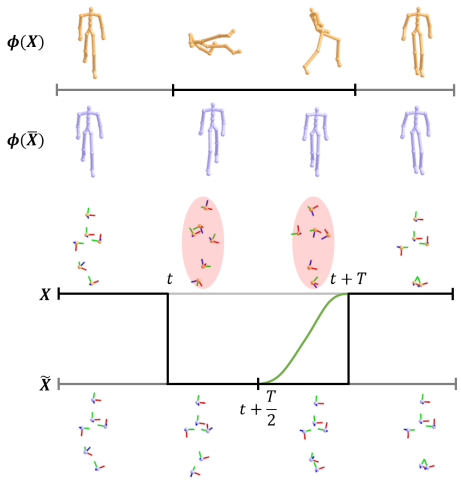

Figure 2: The sensors depicted in the highlighted red circles suffer from noisy issues. The resulting motion $\phi(\tilde{\mathbf{X}})$ is generated in a period of time ranging from t to $t+T/2$ where t is the time when those issues take place. Here, $\phi(\cdot)$ represents the motion estimator and $T = 120$ is a maximum length for stably generating future sensor data. We start to interpolate between $\tilde{\mathbf{x}}_{t+T/2}$ and $\mathbf{x}_{t+T}$ for obtaining the interpolated sensor data $\bar{\mathbf{X}}$. The resulting motion $\phi(\bar{\mathbf{X}})$ is then produced in a period of time ranging from $t + T/2$ to $t + T$. We linearly interpolate the sensor positions and perform spherical linear interpolation on sensor rotations.

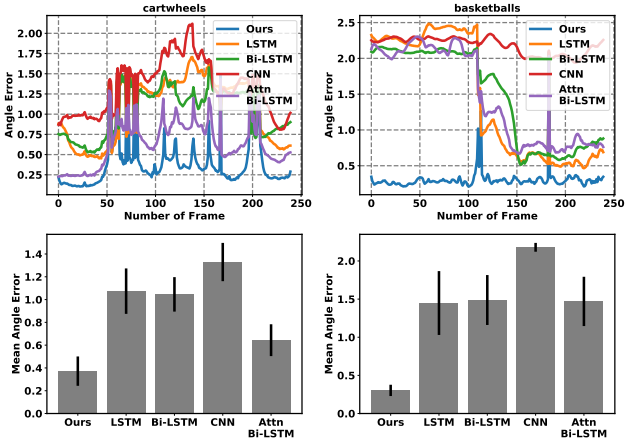

Figure 3: Quantitative performances of four different neural networks relative to ours. Our network accurately estimates the human motion from the test mocap data set.

3 RESULTS

Our system effectively generates believable character animation from sparse sensors owing to the well-established deep learning framework. To achieve better accuracy, the ability to intelligently synthesize the new sensor data using the measured and predicted data from the sensor data generator when the sensor data are not valid is a key feature of the proposed system. We evaluated the reproduction errors in terms of mean and standard deviation with other deep learning architectures (Figure 3) and showed that the proposed system outperforms the others.

REFERENCES

CMU. 2013. Carnegie-Mellon Motion Capture Database. http://mocap.cs.cmu.edu/.

Jehee Lee. 2008. Representing rotations and orientations in geometric computing. *IEEE Computer Graphics and Applications* 28, 2 (2008), 75–83.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

Yi Zhou, Zimo Li, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. 2018. Auto-conditioned recurrent networks for extended complex human motion synthesis. In *International Conference on Learning Representations*.