MUSIC-TO-BODY-MOVEMENT GENERATION WITH RECURRENT NEURAL NETWORK AND REINFORCEMENT LEARNING

Hsien-Lun Chen
Department of Computer Science
University of Washington, WA
peter05152425@gmail.com

Li Su
Institute of Information Science
Academia Sinica, Taiwan
lisu@iis.sinica.edu.tw

ABSTRACT

This paper presents a preliminary reinforcement learning model in music-to-body movement generation which combines LSTM based music-to-motion generation with a DDPG-based skeleton tracking model to calibrate the movement of the bowing hand for a virtual violinist. Comparing to simply applying the LSTM generation, the DDPG integrated model smooths the overall motion of the bowing hand and achieves a larger range of motion.

1. INTRODUCTION

Animating musicians in music performance has been one of the most challenging task in character animation. To capture the essence of the musical performance, animators need to have insights of both the animation principals of movement and instrumental techniques. Therefore, motion capture and pose estimation are widely used in the animation industry to accelerate the production pipeline of music performance animation. In recent years, using deep learning techniques such as recurrent neural networks (RNN) [1], Transformers [2], and generative adversarial networks (GAN) [3] to generate musicians' body movement relying simply on the conditioned music audio input has become a new alternative to the existing approaches.

In this paper, we extend the work of Audio-to-Body Dynamics (A2B) [1], a long-short-term memory (LSTM) RNN that is trained on violin recital videos to generate 2-D body movement from the given audio content, from generating two dimensional skeletons to three dimensional ones.

However, the skeleton generated from the A2B model does not show clear movement when obvious bowing movement were made by the violinist. Therefore, we apply Deep Deterministic Policy Gradient (DDPG), a reinforcement learning technique widely applied in motion generation which learns from experience [4], on the bowing hand to enhance the range of movements over continuous action spaces.

2. METHOD

2.1 Dataset

The dataset used in this research is a subset of [2], which is collected from the selected violin solo videos performed by violin major students from the Taipei National University of the Arts. In total 10 violinists performed the same repertoire of 14 classical violin solo pieces ranging from Baroque to Romanticism eras. We employed the pose estimation method [5] to obtain the 3-D skeletons, which contains the time-varying positions of 15 body joints with a frame rate of 30 fps. The audio recordings were sampled at 44.1 kHz. During training, thirteen music pieces performed by the fourth violinist, total of 80,203 frames, were chosen. For evaluation, the performance of J. S. Bach's *Partita No. 2 in D minor* (BWV 1004) from the first violinist was chosen to generate the video in the supplement material using our proposed model.

2.2 The A2B model

The A2B model [1] is an end-to-end model that generates the 2-D skeleton movement of music performance from audio content being performed. It contains an RNN network with the LSTM cell with hidden state of 200 units, trained with back propagation with time steps of 400, time delay of 24ms. After encoding the audio features, a fully connected layer is added to enhance the performance and decode the PCA coefficients of body keypoints. In this paper, we revised the A2B model such that it can output 3-D skeletons.

2.3 The DDPG model

DDPG [4] is a model-free off-policy algorithm that consists of a critic network $Q(s, a|\theta^Q)$ and an actor network $\mu(s|\theta^\mu)$ with weights θ^Q and θ^μ. In each training episode, the actor would propose an action given a state, and the critic would evaluate the action from the actor given that state. Moreover, in each episode, the overall transition (state, action, reward, and new state) would be stored in a buffer, so the later learning would be sampled from a random minibatch of N transitions from the buffer. At the end of each episode, the critic is updated by minimizing the distance between ground truth y_i and the output of critic, represented as the loss function $\mathcal{L}$ in Eqn (1):

$$\mathcal{L} = \frac{1}{N} \sum_i \left(y_i - Q(s_i, a_i|\theta^Q) \right)^2, \qquad (1)$$

Further, the actor policy is updated with the sampled policy gradient computed with Eqn (2):

$$\nabla_{\theta_\mu} J \approx \frac{1}{N} \sum_i \nabla_a Q(s, a|\theta^Q)\big|_{s=s_i, a=\mu(s_i)} \nabla_{\theta_\mu} \mu(s|\theta^\mu)\big|_{s_i} . \quad (2)$$

In this work, y_i is the ground truth skeleton labels obtained from the pose estimation tool, the state vector s_i is the 3-D positions of the right hand and right elbow, the possible action vectors include the euler rotation angle of x, y, and z directions for the right hand and right elbow, and the reward is the negative of the sum of the l_2 distance from the predicted right hand and elbow to the training data.

2.4 Training and Generation Process

Firstly, given the MFCC audio features and the 3D skeleton data of the performance of the violists, an initial model based on the LSTM network proposed in A2B is trained to generate the movement of a virtual violinist. The generated skeleton would be divided into three parts - torso, right hand (bow hand), and left hand. Meanwhile, DDPG is also trained on the right hand with the same set of dataset, while each frame of the skeleton movement is taken as one episode for the DDPG.

After both A2B and DDPG are trained, we first generate the skeleton movement of a virtual violinist given the MFCC features of the testing data. Then, we apply DDPG on the right hand of the generated skeleton to calibrate the bowing movement.

3. RESULT

Figure 1 shows the skeleton sequences (frontal view, depth in the z-axis) generated by [1] (denoted as A2B), A2B with DDPG (denoted as A2B-DDPG), and the ground truth skeleton sequence. The output skeleton movement of A2B, DDPG integrated A2B, and the ground truth are visualized in Figure 1.The input music is an excerpt of performance of J. S. Bach's *Partita No. 2 in D minor* .

Preliminary observations indicate that as the bowing hand is moving upward, the DDPG-integrated model matches the ground truth better in the direction of movement, while the A2B model stays relatively fixed. The supplementary video also shows that despite having some small spikes in the movement of the bowing hand, its range of movement is enhanced and become more fluent throughout the whole animation.

4. CONCLUSION

We introduced a integration of the LSTM based audio to body movement generation network and DDPG algorithm to extend the range of movement of the bow hand of the virtual violinist. For the future work, we will fine tune the DDPG algorithm to stabilize the predicted motion, implement end-to-end integration of the A2B and DDPG, and apply the generated skeleton movements to 3D character models to observe the movements.

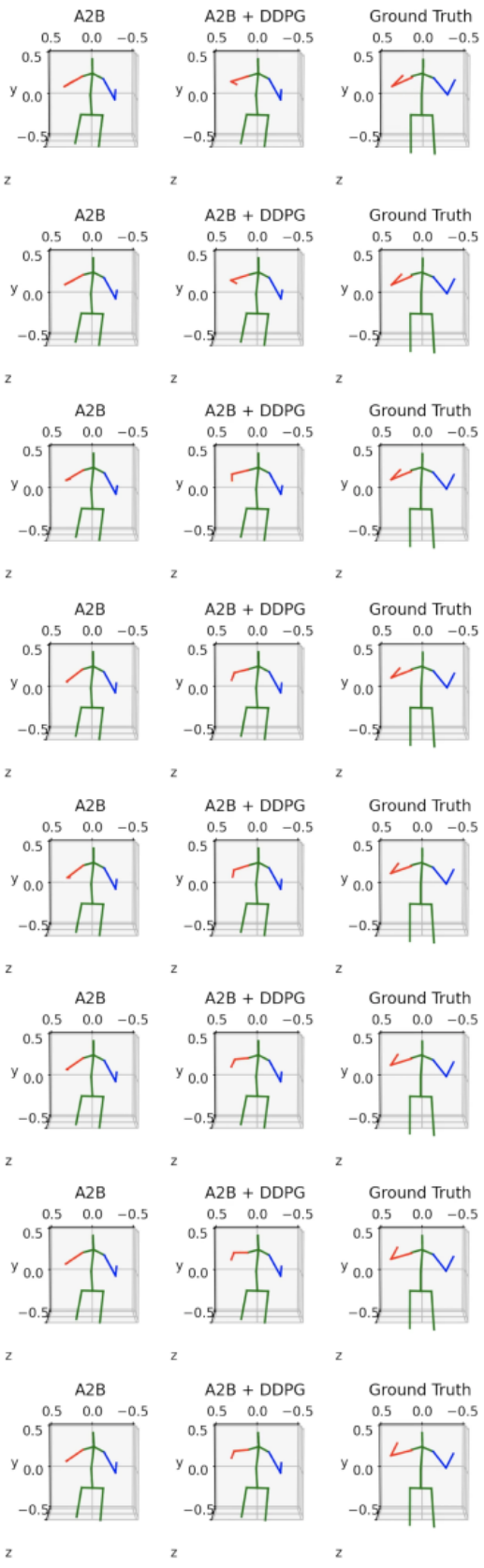

Figure 1. Illustration of the A2B, DDPG integrated A2B, and ground truth animation. For better illustration, the torso is rendered in green, the right hand is in red, and the left hand is rendered in blue respectively.

5. REFERENCES

[1] Eli Shlizerman, Lucio Dery, Hayden Schoen, and Ira Kemelmacher-Shlizerman, “Audio to Body Dynamics” *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[2] Hsuan-Kai Kao and Li Su, “Temporally Guided Music-to-Body-Movement Generation,” in *Proceedings of the 28th ACM International Conference on Multimedia.*, pp. 147–155, October 2020.

[3] G. Sun, Y. Wong, Z. Cheng, M. S. Kankanhalli, W. Geng and X. Li, "DeepDance: Music-to-Dance Motion Choreography With Adversarial Learning," in *IEEE Transactions on Multimedia,* vol. 23, pp. 497-509, 2021

[4] David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, Martin Riedmiller, “Deterministic Policy Gradient Algorithms,” in *Proceedings of the 31st International Conference on Machine Learning,* PMLR 32(1):387-395, 2014.

[5] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli, “3D human pose estimation in video with temporal convolutions and semi-supervised training, ” in *IEEE Conference on Computer Vision and Pattern Recognition*, 7753–7762. 2019