Dance with a Robot: Encoder-Decoder Neural Network for Music-Dance Learning

Baijun Xie
bdxie@gwu.edu
The George Washington University
USA

Chung Hyuk Park
chpark@gwu.edu
The George Washington University
USA

ABSTRACT

This late-breaking report presents a method for learning sequential and temporal mapping between music and dance via the Sequence-to-Sequence (Seq2Seq) architecture. In this study, the Seq2Seq model comprises two parts: the encoder for processing the music inputs and the decoder for generating the output motion vectors. This model has the ability to accept music features and motion inputs from the user for human-robot interactive learning sessions, which outputs the motion patterns that teach the corrective movements to follow the moves from the expert dancer. Three different types of Seq2Seq models are compared in the results and applied to a simulation platform. This model will be applied in social interaction scenarios with children with autism spectrum disorder (ASD).

CCS CONCEPTS

• **Computer systems organization** → **Robotics**; • **Computing methodologies** → *Neural networks*; • **Human-centered computing** → Collaborative interaction.

KEYWORDS

Robotics; neural networks; Seq2Seq; encoder-decoder

ACM Reference Format:
Baijun Xie and Chung Hyuk Park. 2020. Dance with a Robot: Encoder-Decoder Neural Network for Music-Dance Learning. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20 Companion), March 23–26, 2020, Cambridge, United Kingdom.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3371382.3378372

1 INTRODUCTION

Music can be the soul of dance. The basic components of music, such as beat, rhythm, and flow can be the natural source of dance choreography. Since dances can improve social skills, the combination of music and dance can also play effective roles in social interactions. Therefore, a music-to-dance synthesis mapping algorithm could be beneficial in applications targeting music-based dance teaching, social-behavioral learning, and communication skills development in social interactions. In this study, we propose a robot interaction system with Seq2Seq learning neural network structure [7] designed to target several social interaction scenarios, such as social interactions with children with ASD [2, 6].

2 SYSTEM OVERVIEW

The Seq2Seq model has drawn considerable attention in recent years which is developed by [7]. Intuitively, in Figure 1, the Seq2Seq model is based on a recurrent neural network (RNN) architecture, which aims to address the learning of challenging mapping between the multi-modal input and output sequences. The RNN has shown to be effective in complex motion learning by robotic systems [5, 9], so we anticipate our architecture to be effective in learning human dance skills. Our learning architecture utilizing the Seq2Seq model comprises two parts: the encoder model for processing input sequences and reducing dimensionality and the decoder model for output sequence generation. In this study, we sample the dataset from the expert dancer in a teaching video with music. We map the input movement sequences from the video and the music features from the music of the video, to the output sequence that are the continuous motion features from the dancer by using the Seq2Seq model. In the teaching phase, we train the model by using the motion features from the expert dancer, while, in the teaching phase, the trained encoder and decoder models correct the user's motion input. At here, the expert dance is recorded by video and the system user is the experiment participant. Finally, the model's outputs are applied to a robot platform, NAO, for demonstration for social interactions. The proposed system can run in real-time with NVidia RTX 2080 Ti graphics card.

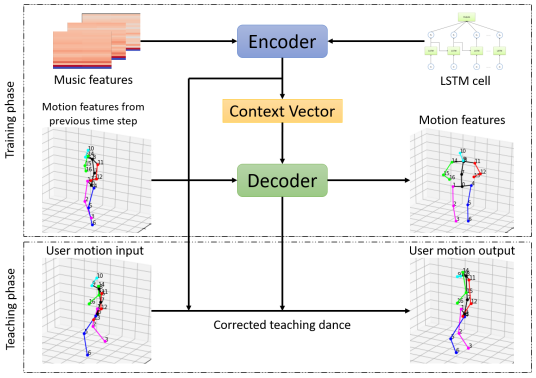

Figure 1: Seq2Seq model with Encoder and Decoder.

2.1 Encoder for Input Sequence Processing

For the musical feature to provide as an input to our training network, we use the Mel-frequency cepstral coefficients (MFCCs) [3],

HRI '20 Companion, March 23–26, 2020, Cambridge, United Kingdom

ACM ISBN 978-1-4503-7057-8/20/03.
https://doi.org/10.1145/3371382.3378372

which is a widely used feature set in audio analysis. MFCCs are extracted from the music of the dance video via an open-source library [4]. As can be seen in Figure 1, the encoder accepts a sequence of music features as inputs and outputs a fixed-length context vector that can represent the input sequence. Both encoder and decoder have the temporal neural network structure, comprising a Long Short-Term Memory (LSTM) module.

2.2 Decoder for Output Sequence Generation

The decoder, another LSTM module, treats the context vector generated by the encoder as the initial states. The teacher forcing scheme is used during training, where the model uses the ground-truth target sequence from a prior time step as input. At here, the target sequence is the sequence of motion features in 3D. The extracted motion features based on the human body 2D keypoints detection algorithm, OpenPose [1]. The final human pose 3D estimation is realized by employing the approach developed by [8]. During teaching sessions, the overall Seq2Seq model receives the music features sequences as inputs and the output becomes 3D human pose sequences decoded from the context vector from the encoder and the motion vector from a novice user, allowing our network to provide continuing and yet corrective movements.

2.3 Robot Platform

The robot platform we used for simulation is NAO, a humanoid robot developed by Softbank Robotics. We also used the Choregraphe as our simulation environment, which allows us to monitor and control the robot. After the decoder generates the sequence of human body pose keypoint features, the corresponding rotation angles for every joint are computed and mapped to NAO via joints' angles control provided by the robot platform.

3 RESULTS

3.1 Model Training Performance

Three different types of Seq2Seq models have been compared in this study: Vanilla LSTM, Bidirectional LSTM, and LSTM with self-Attention. The time-series data, music features, and motion features sequences data were segmented, and every data sample has 20-time steps for feeding into the LSTM neural network. The data were randomly split. 80% of the data were used for training and 20% of the data were used for validation. During training, all the hyper-parameters were set to be the same for comparison. RAMSprop optimizer was used because it suits for training the recurrent neural network. Since the distance between the model motion features output and origin motion features is expected to be minimized, the loss function during training we used is mean squared error. The final results are reported based on the loss, root mean squared (RMSE) and R-squared (R^2). As can be show in Table 1, the bidirectional LSTM outperform other two models, whose validation loss is 0.179, RMSE is 0.1341 and R^2 is 0.9420.

3.2 User Study

A user study was conducted on one researcher in the Assistive Robotics and Telemedicine (ART-Med) lab at the George Washington University (user study approved by the Institutional Review Board (GW IRB 111540)). This study will be applied to human-robot social interaction scenarios and recruit more participants, such as children with ASD, in the future. The participant was asked to learn the dance from the video tutorial at first. The sessions were video-recorded for post-processing with OpenPose [1], which could effectively detect movements during the experiment. The trained model accepted both the music features from the video and the participants' motion features as input, and the outputs were the motion features for robot-assisted instructions.

Table 1: Models Performance Comparison

Model Type	Val. Loss	Val. RMSE	Val. R^2
Vanilla LSTM	0.180	0.1341	0.9417
Bidirectional LSTM	0.179	0.1337	0.9420
LSTM with Attention	0.204	0.1422	0.9340

3.3 Simulation

In the simulation part, as mentioned in Sec. 2.3, the model output is applied to the NAO robot platform. Figure 2 (a) shows the sequence of human pose keypoints from the dancer on the video, Figure 2 (b) shows the sequence of human pose keypoints from the user, and Figure 2 (c) shows the motions mapping on the robot.

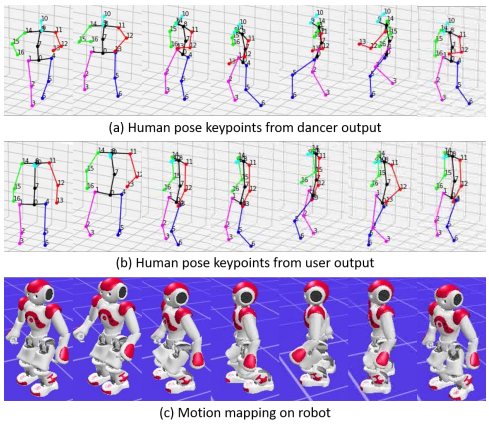

Figure 2: The simulation of human pose keypoints and the mappings on robot.

4 DISCUSSION AND CONCLUSION

In conclusion, the Seq2Seq model was found to be applicable for the mapping between multimodal features. In this study, we used three encoder-decoder structures to map from music features to motion features. All of these three models have the ability to learn the correlation between the current data and the previous part of the data. The weakness of these models is that they could be time-consuming for training and hard to synchronize. Although the results show an outstanding performance in learning with our architecture, it should be noted that only one dancing video was used at this stage. For future work, we will enrich our dataset and elaborate on the robot control schemes for richer social interaction.

REFERENCES

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7291–7299.

[2] Hifza Javed, Myounghoon Jeon, Ayanna Howard, and Chung Hyuk Park. 2018. Robot-assisted socio-emotional intervention framework for children with Autism Spectrum disorder. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 131–132.

[3] Beth Logan et al. 2000. Mel Frequency Cepstral Coefficients for Music Modeling.. In *ISMIR*, Vol. 270. 1–11.

[4] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python.

[5] Chung Hyuk Park, Jae Wook Yoo, and Ayanna M Howard. 2010. Transfer of skills between human operators through haptic training with robot coordination. In *2010 IEEE International Conference on Robotics and Automation*. IEEE, 229–235.

[6] Sara M Scharoun, Nicole J Reinders, Pamela J Bryden, and Paula C Fletcher. 2014. Dance/movement therapy as an intervention for children with autism spectrum disorders. *American Journal of Dance Therapy* 36, 2 (2014), 209–228.

[7] I Sutskever, O Vinyals, and QV Le. 2014. Sequence to sequence learning with neural networks. *Advances in NIPS* (2014).

[8] Denis Tome, Chris Russell, and Lourdes Agapito. 2017. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2500–2509.

[9] Jasmin Velagic, Nedim Osmic, and Bakir Lacevic. 2008. Neural network controller for mobile robot motion control. *World Academy of Science, Engineering and Technology* 47 (2008), 193–198.