VIRTUALCONDUCTOR: MUSIC-DRIVEN CONDUCTING VIDEO GENERATION SYSTEM

Delong Chen, Fan Liu, Zewen Li, Feng Xu*

College of Computer and Information, Hohai University, China
fanliu@hhu.edu.cn

ABSTRACT

In this demo, we present the *VirtualConductor*, a system that can generate conducting video from a given piece of music and a single user's image. First, a large-scale conductor motion dataset is collected and constructed. Then, we propose an Audio Motion Correspondence Network (AMCNet) and adversarial-perceptual learning to learn the cross-modal relationship and generate diverse, plausible, music-synchronized motion. Finally, we combine 3D animation rendering and a pose transfer model to synthesize conducting video from a single given user's image. Therefore, any user can become a virtual conductor through the *VirtualConductor* system.

***Index Terms*—** Adversarial learning, orchestral conductor, audio motion correspondence

1. INTRODUCTION

In recent years, deep learning has shown its advantages in learning discriminative feature representations [1] and learning high-quality generation [2] from massive data. As a notable research line in this field, learning the cross-modal mapping from sound to human motion has drawn a lot of attention. Various types of applications, including speech gesture generation and musical gesture generation (dancing and instrument playing), have been developed in recent years. But researchers pay little attention to the motion generation of an orchestral conductor. Moreover, there is not a large-scale conductor motion dataset currently available. Therefore, we build a system to make the first attempt towards music-driven conductor motion generation and realize a virtual conductor.

To build a large-scale conductor motion dataset, we first collect concert performance video recordings, then extract conductor motion by pose estimation [3]. Meanwhile, different types of audio features, including MFCC, spectral centroid, spectral bandwidth, onset envelope, estimated tempo, and predominant local pulse, are extracted. Finally, the constructed dataset consists of conductor motion data and aligned music features in a total of 40 hours.

However, modeling conductor motion still has several challenges. First, the conductor motion is highly complicated because it conveys various types of information, including tempo, strength, and emotion. Meanwhile, the generated motion should be closely synchronized with music. Moreover, because of different conducting styles, mapping music to conductor motion is a one-to-many task, which is difficult to learn by standard mean squared error (MSE) regression. In this demo, based on the constructed dataset, we propose the *VirtualConductor* system to tackle the above difficulties. We use a combination of MSE loss, pose perceptual loss, and adversarial loss to train the motion generator. In this way, the generated motion can be simultaneously diverse, plausible, and synchronized to music. Finally, by combining 3D animation rendering and pose transfer [4] module, the system can generate conducting video from given music and a single user's image. In the following sections, we will introduce our system in detail.

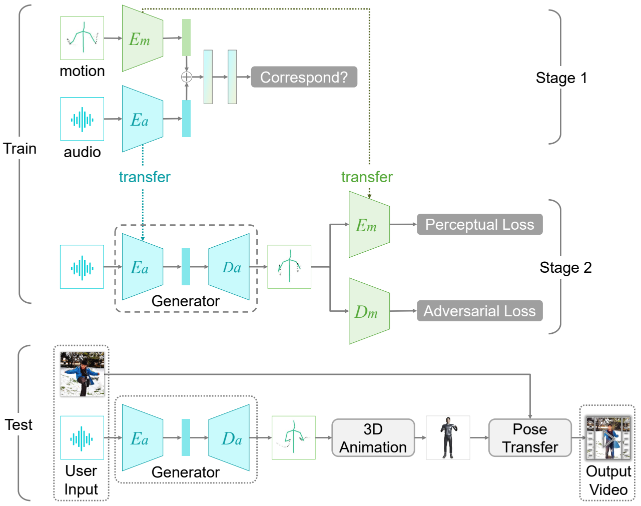

Fig. 1. The pipeline of presented demo *VirtualConductor*.

2. SYSTEM DESIGN AND IMPLEMENTATION

2.1. Audio Motion Correspondence Learning

We first design an AMCNet to learn the correspondence between audio and motion. As shown in Fig.1, the AMCNet consists of a music encoder E_a, a motion encoder E_m, and fuse layers. The features extracted by two encoders are concatenated and passed to fuse layers. The AMCNet output a

978-1-6654-4989-2/21/$31.00 ©2021 IEEE

possibility in the range of (0,1), indicating whether the input music frame and motion frame correspond. The loss function of the AMCNet is shown in Eq.1, where x_i and y_j are the random scaled correspond pairs, while x_i and y_k are not correspond pairs. The trained feature extractors of AMCNet have the following properties: the extracted music feature is closely synchronized to the target motion, and the extracted motion feature is scale-invariant.

$$\begin{aligned} L_{AMC} =& \frac{1}{N}\sum_{i,j=0}^{N}(\text{AMCNet}(x_i, y_j) - 1)^2 + \\ & \frac{1}{N}\sum_{i,k=0}^{N}(\text{AMCNet}(x_i, y_k) - 0)^2 \end{aligned} \tag{1}$$

2.2. Adversarial-Perceptual Learning

Due to it is difficult to learn the one-to-many mapping by using standard MSE loss, we relax the constraint of MSE by combining adversarial loss and perceptual loss. The adversarial loss enables the model to approximate the distribution of the real conductor motion and generate a realistic motion sequence, while the perceptual loss ensures the generated motion conforms to music. As shown in Fig.1, we use the learned motion encoder E_m of AMCNet to calculate the perceptual loss, and transfer the music encoder E_a into the motion generator G. At the same time, a discriminator D_m with Lipschitz constraint is set up to guide the generator towards the real motion distribution. The loss function of the motion generator G is shown in Eq.2

$$\begin{aligned} L_G =& \lambda_{mse}\frac{1}{M}\sum_{i=0}^{M}(G(x_i) - y_i)^2 + \\ & \lambda_{per}\frac{1}{M}\sum_{i=0}^{M}(E_m(G(x_i)) - E_m(y_i))^2 + \\ & \lambda_{adv}\frac{1}{M}\sum_{i=0}^{M}D(G(x_i)) \end{aligned} \tag{2}$$

3. DEMONSTRATION

The *VirtualConductor* requires a music file and a single user's image as input. The system extracts the music feature and generates a motion sequence. The motion sequence is subsequently rendered to the 3D avatar, and the user's conducting video by pose transfer module [4]. The final video output results are respectively shown in Fig. 2. The demo system is implemented in Pytorch with an NVIDIA 2080Ti GPU. It can generate conductor motion of complete Beethoven Symphony No.5 in 1.431 seconds and synthesize conducting video in about 13 fps.

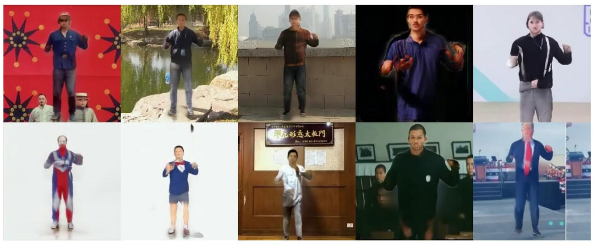

Fig. 2. Demo results generated by the *VirtualConductor*.

4. CONCLUSION

In this demo, we make the first attempt towards music-driven conductor motion generation. First, a large-scale conductor motion dataset is constructed. Then, we propose a 2-stage model which includes audio motion correspondence learning and adversarial perceptual learning. Experimental results show that the cross-modal relationship between music and motion is effectively learned. Finally, taking a music file and a single user's image as input, the *VirtualConductor* system can generate diverse, plausible, music-synchronized conducting video and enable anyone to become a conductor.

5. ACKNOWLEDGEMENT

This work was partially funded by the Natural Science Foundation of Jiangsu Province under Grant No. BK20191298, Fundamental Research Funds for the Central Universities under Grant No. B200202175.

6. REFERENCES

[1] Zechao Li, Jinhui Tang, and Tao Mei, "Deep collaborative embedding for social image understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2070–2083, 2019.

[2] Jinhui Tang, Jing Wang, Zechao Li, Jianlong Fu, and Tao Mei, "Show, reward, and tell: Adversarial visual story generation," *ACM Trans. Multim. Comput. Commun. Appl.*, vol. 15, no. 2s, pp. 54:1–54:20, 2019.

[3] H. Fang, S. Xie, Y. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2353–2362.

[4] Wen Liu, Wenhan Luo Lin Ma Zhixin Piao, Min Jie, and Shenghua Gao, "Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis," in *The IEEE International Conference on Computer Vision (ICCV)*, 2019.