Generation of Dance Movement Sequences Based on Audio Information

Ziyao Meng*

Hebei Zhengzhong Experimental School, Hebei, China
*Corresponding author: 17742646305@163.com

Abstract:

The task of action sequence generation based on audio is a cross-modal generation task, which automatically generates continuous action sequences with similar or consistent time, semantics or emotion with the input information through the input audio signal. With advances in computer vision, as well as digital entertainment, methods that link human speech to digital body movements have made rapid progress. At present, the main methods are based on neural networks and deep learning for multi-modal generation. Through the multi-modal generation method based on computer network, it often faces the problems of low correlation between the generated content and the input content, low overall fluency, and unclear emotional expression. Based on the systematic review of the existing literature, this paper analyzes the mainstream methods in the task of dance movement generation. Finally, the key problems to be solved in this field are discussed, and the future research content and direction are prospected.

Keywords: Action sequence generation; neural networks; deep learning; multi-modal generation; dance movement generation.

1. Introduction

With the rapid development of the Internet, there is a growing demand for the generation of human motion in movies, special effects, computer games, virtual characters and other fields. In the past decades, most of these motion generation tasks have been realized by motion capture systems, and thanks to its mature system, remarkable achievements have been obtained, but the cost is high. The disadvantages such as limited application scenarios have greatly increased the demand for multi-modal generated action sequences in industries such as movies and video games. As an emerging research topic of artificial intelligence, multi-modal audio generation of action sequences has also been more widely studied [1-3].

Multi-modal audio based on action sequence generation is a technology that fuses audio signals with other modal information (such as text, vision or music features) and generates continuous human action sequences highly coordinated with audio semantics, rhythm and emotion through computational models. Its core goal is to establish a cross-modal mapping mechanism to transform the temporal structure (such as beat, pitch), semantic content (such as language meaning, emotional tendency) and context information of sound into natural and fluent motion expressions that meet physical rationality, semantic consistency and Spatio-Temporal fluency, providing key technical support for virtual human interaction, film and television animation and computer games.

The rise of deep learning has brought more attention to the problem of audio-driven 3D gesture generation, as it is a more scalable and versatile way to create gesture systems. For a broader approach, research has turned to probabilistic models. These methods have considerable potential to describe a series of movements, from which different implementation scenarios can be sampled, and it has better generalization potential beyond the scope of available data.

For the task of dance generation, although some researchers have proposed to introduce the influence of music into motion generation in the last century, due to the uncertainty of dance generation and the difficulty of learning the correlation with music, the current research in this multi-modal direction still faces great challenges. The generation method based on neural network usually has more diverse generated content and more accurate actions and has gradually become the mainstream research method in the task of generating action sequences based on audio generation.

However, prior methods do not provide a way to control the global structure of the motions that have been synthesized and tend to focus on local motions that are smooth while ignoring the context or overall themes of the dance. When the target task is to integrate and regulate the motion corresponding to the input music information, it will face more severe challenges. At any given moment, the current dance pose can be succeeded by the emergence of numerous potential poses, and the Spatio-Temporal structure of body motion enhances the complexity of the generated motion. Although it follows a provided audio beat and has a distinctive style, it does not illustrate a rich variety of movements and lacks a globally consistent dance background to constrain and control dance movements.

Firstly, this paper summarizes the methods of dance generation combined with the existing research on audio generation of dance sequences and summarizes the previous methods. Finally, based on the analysis of the existing research, the future research is prospected. Based on the analysis of the literature, this paper summarizes the advantages and disadvantages of the literature methods, puts forward the problems existing in the current research, and provides directions and suggestions for future research.

2. Method Analysis

The generation method based on neural network usually has more diverse generated content and more accurate actions and has gradually become the mainstream research method in the task of generating action sequences based on audio generation. Fukayama et al. proposed a probabilistic framework for generating choreography of dance movements in 2015, so that the generated dance movements could combine the logic before and after music and be more in line with realistic dance movements [4]. In 2016, Crnkov-Friis et al. proposed a new system named chor-rnn based on recurrent neural network (RNN) to generate novel choreography sequences of the choreography style represented in the corpus, using deep recurrent neural network, which can understand and generate the choreography style, grammar, and to some extent understand the semantics [5]. Tang et al. proposed a Long Short-Term Memory (LSTM) autoencoder model in 2018 is used for extraction of mapping from motion feature and acoustic. The model can study how dancers adjust their local joint posture and movement rhythm to express changes in musical emotion and rules for selecting actions in choreography.

To ensure that the emotion of the music is reflected in the synthesized dance [6]. In recent years, multi-modal action generation has been further developed: Aristidou et al. proposed a music-driven dance synthesis framework, which divided the task of generating dance movements into three parts: Choreography Level, Motif Level and Pose Level [7].

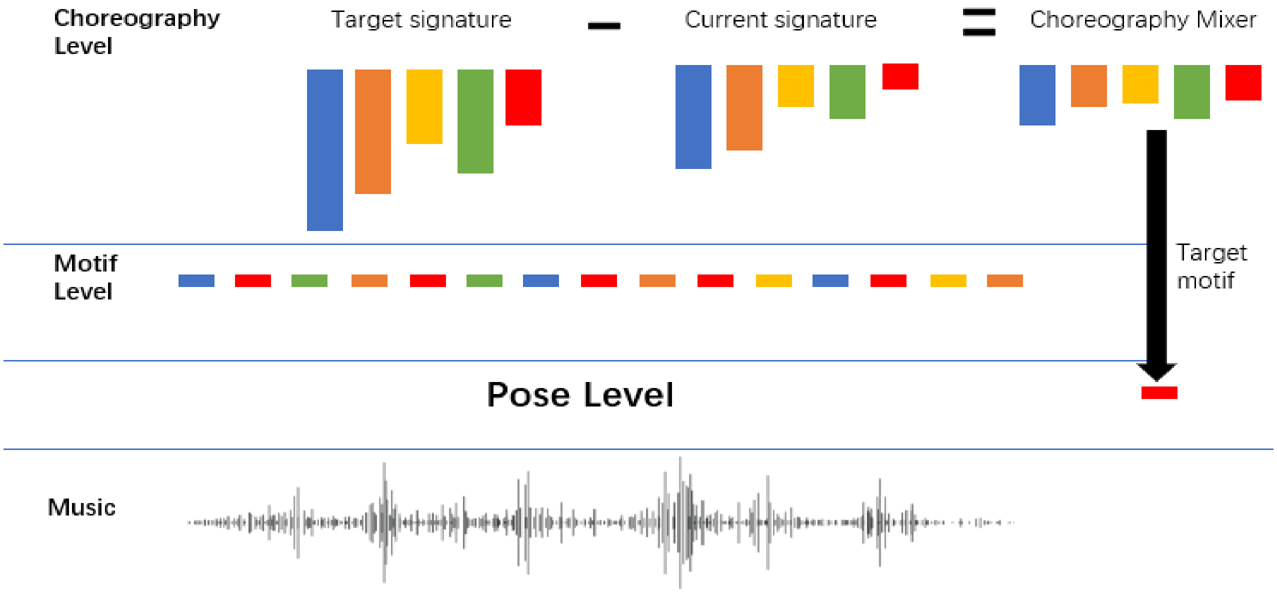

Fig. 1 Three Level process presentation

As shown in Fig.1, the operation process and relationship between the three levels are shown: Firstly, the input audio information is transformed into rhythm coding through Choreography Level, so that the action sequence generated in the next stage is correlated with the style and rhythm of the input audio. The generation of dance movements is driven by selecting continuously matched movement themes and creating appropriate sequences. Then, through Motif Level, action words are defined as a set of multi-continuous poses to represent the local changing of the pose, and a smaller dimension is used to map action words extracted from a set of similar action sequences into a common feature space to create a more compact embedding way. By using Pose Level to ensure that the generated movements can follow the beat, the movement structure is controlled under the condition of the dance theme. By using Self-conditioned Long Short-Term Memory (acLSTM) network to generate motion frames, the extracted motion words were embedded into short spatio-temporal sequences conditioned on audio with features, and then the motion words were converted into motion topics. Finally, the dance movements were generated based on the audio information.

Aiming at the problem of how to synthesize dance movements from audio through adversarial training and Graph Convolutional Network (GCN), Ferreira J P et al. proposed a new architecture for managing audio data to synthesize movements [8]. The sound signal was encoded by CNN architecture to extract music style. By combining musical style and spatio-temporal latent vectors to tune the GCN architecture and train it through adversarial training to predict changes between 2D arthrosis positions, providing reasonable movements while preserving features of different dance styles.

Since gestures are highly individual, indecisive, and often difficult to be accurately defined by speech, and dance movements are often synchronized with music structures and have high randomness, Alexanderson et al. adjusted the DiffWave architecture to model 3D posture sequences, and used Conformer to replace dilated convolution. We achieve improved modeling capabilities in audio-driven human motion generation tasks, and pioneer a new method based on probabilistic principles to combine and transform dance styles. In addition, a classifier-free guidance mechanism is introduced to control the generated dance movement style, so as to realize the adjustable goal of the expression strength of dance style [9].

Zhang Yue proposed a problem based on the need for a large number of frames when using the method of adding motion to generate dance sequences. This paper proposes a new intelligent dance generation network TG-dance based on TransGAN, which can generate long-term and high-quality dance sequences in the future by using a small number of action frames [10]. The network uses the generative adversarial network as the framework, proposes a new idea of multi-level extension of action sequence, and designs a multi-level action encoder combining Transformer and upsampling layer.

In view of the above methods, this paper analyzes and summarizes their advantages and disadvantages, as shown in Table 1. In the existing research, the data set used by most methods is relatively single, covering limited types of dances, resulting in differences in the types of dance movements generated by each method, and usually there are few types of dances that can be generated, which makes it difficult to evaluate the dance sequences generated by different systems under the same standard. In addition, most of the generated dance movements are from the movements within the dataset, limiting the diversity of the generated dance movements.

Table 1. Comparison of advantages and disadvantages of different models

Method	Merit	Deficiency
Aristidou et al.	1): Ability to generate long-term and temporally consistent actions 2): It can synchronize the generated action word sequence with the audio	1): The dataset lacks richness of actions 2): The data used for training is affected by errors during acquisition 3): There are more outliers when clustering all motion words to the nearest topic
Ferreira J P et al.	1): The generated movements are highly similar to the real dance movements	1): The type of dance used in training is relatively single 2): The animation framework used is relatively single
Alexanderson S et al.	1): Able to control the intensity of dance style performance 2): Multiple diffusion models are combined	1): Two kinds of movement (gesture, dance) are not combined 2): Failure to jointly control different parameters in the input information
Zhang Yue	1): It can better obtain relevant information between dance and music 2): Able to predict long dance sequences	1): The type of dance used in training is relatively single 2): Mainly based on Transformer module, it cannot meet the requirement of generating the same or similar dance movements when the same music segment appears

3. Problem Analysis and Suggestions

Although great progress has been made in the research topic of generating action sequences, when using diffusion model to generate action sequences, the generation speed is often too slow to meet the needs of real-time interaction. The traditional autoregressive model accumulates errors when iteratively generating long sequences, resulting in motion distortion or rhythm misalignment. The feature set used at training time has low latitude and is limited to rapidly changing musical features (no display indications about bars, choruses, starts, or ends), thus failing to allow high-level structure in the input to emerge; The training data used by some methods have artifacts caused by acquisition errors, and some actions are not fully covered, which affects the quality of generated actions. The acoustic features such as volume, melody and instrument are not fully fused. For example, the samples of small dances such as national dance and improvised dance are insufficient, resulting in the scarcity of data. Some methods lack the artistic creativity of the choreographer, and cannot independently generate the choreography movements of the original dance, resulting in the mediocrity of the generated movements. In the generation, the influence of environmental factors around the character is not considered, so that the environment of the generated action is single and the environment variables are lack, which leads to the lack of support for multi-character interaction or scene.

Due to the particularity of dance movements, there is no unified standard for measurement and comparison after the generation of dance movements. There are few evaluation methods of dance movements and lack of standardized evaluation.

Therefore, future research can focus on the following aspects:

1) By expanding the dataset, adding dance types and music data, creating a variety of style datasets and adapting them in the model.

2) The music features are subdivided, and acoustic features such as volume, melody and instrument are taken into account when extracting information features, so that the generated action is more consistent with the music melody.

3) In the process of generation, it is combined with the emotion module to strengthen the characteristics of the emotional performance of the dance.

4) Reinforcement learning, combined with AI, to autonomously explore combinations of dance movements, not limited to the movements in the dataset.

5) Further study the diffusion model to shorten the time of dance generation to achieve real-time interaction.

6) Consider the interaction of multiple characters or the influence of surrounding environment and objects to improve the variability of dance.

7) Establish comprehensive evaluation criteria, such as

fluency, cultural coverage, audio information matching and other factors.

4.Conclusion

This paper generates the whole dance movement based on audio information, introduces the basic process of the representative previous work, and analyzes the limitations of the existing methods. The main problems are as follows: The ability of real-time interaction is poor, there is no comprehensive evaluation standard, the innovation of generated movements is insufficient, the richness of data samples is low, and the influence of the environment of the generated dance is less considered. Finally, the summary and future work directions are prospected. Seven directions for reference are proposed in this paper: Expand the data set, add dance types, subdivide music features, strengthen the combination with AI, strengthen the learning ability of the model, consider the influence of the environment of the character, improve the variability of the dance, improve the diffusion model, shorten the generation time, and establish a unified and comprehensive evaluation standard. In conclusion, the research and analysis work in this paper can improve the important reference significance for dance and gesture generation.

References

[1] Fan Rukun, Xu Songhua, Geng Weidong. Example-based automatic music-driven conventional dance motion synthesis. IEEE transactions on visualization and computer graphics, 2011, 18(3): 501-515.

[2] Ferstl Y, Neff M, McDonnell R. Multi-objective adversarial gesture generation, Proceedings of the 12th ACM SIGGRAPH Conference on Motion, Interaction and Games. 2019: 1-10.

[3] Hu Hao, Liu Changhong, Chen Yong, et al. Multi-scale cascaded generator for music-driven dance synthesis, 2022 International Joint Conference on Neural Networks (IJCNN). IEEE, 2022: 1-7.

[4] Fukayama S, Goto M. Music content driven automated choreography with beat-wise motion connectivity constraints. Proceedings of SMC, 2015: 177-183.

[5] Luka C, Louise C. Generative choreography using deep learning. arXiv preprint arXiv:1605.06921, 2016.

[6] Tang Taoran, Jia Jia, Mao Hanyang. Dance with melody: An lstm-autoencoder approach to music-oriented dance synthesis, Proceedings of the 26th ACM international conference on Multimedia. 2018: 1598-1606.

[7] Aristidou A, Yiannakidis A, Aberman K, et al. Rhythm is a dancer: Music-driven motion synthesis with global structure. IEEE transactions on visualization and computer graphics, 2022, 29(8): 3519-3534.,

[8] Ferreira J P, Coutinho T M, Gomes T L, et al. Learning to dance: A graph convolutional adversarial network to generate realistic dance motions from audio. Computers & Graphics, 2021, 94: 11-21.

[9] Alexanderson S, Nagy R, Beskow J, et al. Listen, denoise, action! audio-driven motion synthesis with diffusion models. ACM Transactions on Graphics (TOG), 2023, 42(4): 1-20.

[10] Zhang Yue, Research on the Intelligent Generation Method of Dance Motions Based on Deep Learning, Shanghai: Shanghai University, 2023.