International Journal of Multimedia Data Engineering and Management
Volume 12 • Issue 1 • January-March 2021

Multimodal Dance Generation Networks Based on Audio-Visual Analysis

Lijuan Duan, Beijing University of Technology, China
Xiao Xu, Beijing University of Technology, China
Qing En, Beijing University of Technology, China

ABSTRACT

3D human dance generation from music is an interesting and challenging task in which the aim is to estimate 3D pose from visual and audio information. Existing methods only use skeleton information to complete this task, which may cause jittering results. In addition, due to lack of appropriate evaluation metrics for this task, it is difficult to evaluate the quality of the generated results. In this paper, the authors explore multi-modality dance generation networks through constructing the correspondence between the visual and the audio cues. Specifically, they propose a 2D prediction module to predict future frames by fusing visual and audio features. Moreover, they propose a 3D conversion module, which is able to generate the 3D skeleton from the 2D skeleton. In addition, some new human dance generation evaluation metrics are proposed to evaluate the quality of the generated results. Experimental results indicate that the proposed modules can meet the requirements of authenticity and diversity.

KEYWORDS

3D Pose, Audio-Visual, Classification, Dance Generation, LSTM, Metrics, Mixture Density Networks, Multimodal, Skeleton, VAE

INTRODUCTION

Dance is a performing art that uses the body to perform graceful or difficult movements in which rhythmic movements are the main means of expression through music. An excellent dance performance requires a professional choreographer, which maybe time-consuming and expensive. Although it can ensure a high degree of completion, the resulting choreography can only be used for single music, which has great limitations. Therefore, how to learn creative choreography by capturing repetitive motion connections and intrinsic characteristics of dance is meaningful. In this paper, we aim to explore multi-modality dance generation networks through constructing the correspondence between the visual and the audio cues.

Generating dance from music is a challenging generative task. Firstly, to keep dance and music in synchronization, the resulting dance movements must follow given musical style and beats. Secondly, dance is diverse in nature, that is, the dancing posture can follow every possible movement. Thirdly, the spatial structure of body movement in dance will lead to high complexity, so it's challenging to search for the connections between movements.

To address the above challenges, GrooveNet (Alemi et al., 2017) firstly investigates a variety of audio to movement mapping methods for describing audio information. These methods can provide some line of thought for solving the problem of the task. However, due to the small training dataset,

DOI: 10.4018/IJMDEM.2021010102

their proposed models lack universal applicability. After that, some methods (Tang et al., 2018; Yalta et al., 2019) also convert the task into a music to dance mapping problem, in which the audio information is input into models through a series of feature extraction and then turned into the required skeleton information. We believe that it's difficult to learn the inner connection of the movements by learning the relationship of the audio than directly.

Based on the above analysis and human dance in reality, three basic observations motivate our research. (i) Important features picked out from audio information have sufficient expressive power. (ii) A series of movements from dancer are inherently coherent. (iii) Some experience is needed to coordinate music and dance movements reasonably. Therefore, the task is no longer a mapping of music to dance, but a fusion of the two based on movement in our proposed model.

In this paper, we take music and initial 2D human skeleton as input. We then combine multiple modal information and use LSTM (Long Short-Term Memory) and MDN(Mixture Density Networks, (Richter, 2003)) to generate 2D skeleton prediction sequence and image prediction sequence of human body. Moreover, convolution blocks are used to generate 3D skeleton from 2D skeleton, thereby obtaining the output dancing movements. Finally, new evaluation metrics are proposed for evaluating the outputs of different modalities and methods.

The contributions of this paper are as follows:

- Novel multi-modality dance generation networks are proposed through constructing the correspondence between the visual and the audio cues;
- We propose new evaluation metrics of human dance generation, based on which the generation results of different modalities and methods are interpretable.

RELATED WORK

Multimodal Deep Learning

Each source or form of information can be called a modality. For example, human senses include touch, hearing, sight, and smell; the medium of information includes voice, video, text, etc.; various sensors include radar, accelerometer, etc. With the development of deep learning in recent years, many research hotspots have emerged in the combination of multimodal learning and deep learning. It can be roughly divided into the following directions: Representation, which uses the complementarity between multiple modalities to eliminate the redundancy between modalities, so as to learn better feature representation (Kiros et al., 2014; Mroueh et al., 2017); Translation, which converts the information of one modal into the information of another modal (Peng et al., 2016; Antol et al., 2015); Alignment, to find the correspondence between different modal information branches from the same instance (Meutzner et al., 2017; Neverova et al., 2015); Fusion,to combines the information of multiple modals for classification or regression tasks (Bahdanau et al., 2014; Karpathy & Fei-Fei, 2015). The above tasks are the most basic multimodal deep learning tasks, and also point out the direction for future development.

Audio-Visual Task

As we all know, a complete information source contains visual, audio and other modal information, which not only contains common information, but also contains complementary information. If the information between the modalities is fully integrated, the corresponding information will be better understood. In recent years, there have been many audio-visual tasks. For example, the task of judging audio-visual consistency, how to determine whether the sound and vision are consistent (Arandjelovic & Zisserman, 2017; Korbar et al., 2018); Use auditory information to help classify visual tasks (Naranchimeg et al., 2018); Audio-visual information can be used to locate the sound source in the image (Arandjelovic & Zisserman, 2018; Senocak et al., 2018); "Cocktail problem",