Synthesizing Realistic Human Dance Motions Conditioned by Musical Data using Graph Convolutional Networks

João P. M. Ferreira[1][*], Renato Martins[1,2], Erickson R. Nascimento[1]

[1]Departament of Computer Science
Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil

[2]INRIA
Sophia Antipolis, France

{joaoferreira,erickson}@dcc.ufmg.br, renato.martins@inria.fr

Abstract. *Learning to move naturally from music,* i.e.*, to dance, is one of the most complex motions humans often perform effortlessly. Existing techniques of automatic dance generation with classical CNN and RNN models undergo training and variability issues due to the non-Euclidean geometry of the motion manifold. We design a novel method based on GCNs to tackle the problem of automatic dance generation from audio. Our method uses an adversarial learning scheme conditioned on the input music audios to create natural motions. The results demonstrate that the proposed GCN model outperforms the state-of-the-art in different experiments. Visual results of the motion generation and explanation can be visualized through the link:* `http://youtu.be/fGDK6UkKzvA`*.*

1. Contributions

We list as main contributions of this master's thesis: *i)* A new graph convolutional network (GCN) architecture to synthesize human motion considering the non-Euclidean geometry of the human body skeleton manifold in the synthesis; *ii)* A generative conditional strategy to relate the music style information with the motion generation and that provides to control and change the generated motion style over time; *iii)* A novel multimodal dataset comprising audio, videos, and extracted visual motion data from dancing actors to three different music styles. The code and data are made publicly available to the community at the project website[1].

2. Publications

The results of this thesis were published into the international journal publication in the *Computers & Graphics* [Ferreira et al. 2020]. This journal has been recently ranked 4 in the top Computer Graphics publications[2]. We also would like to highlight that the student also contributed as co-author to two related publications on transferring human motion between videos: one to the international conference *WACV* [Gomes et al. 2020] and one to the *International Journal of Computer Vision (IJCV)* [Gomes et al. 2021].

[*] This work relates to a M.Sc. thesis. Authorship order: student, co-advisor, and advisor.

[1] https://verlab.github.io/Learning2Dance_CAG_2020/

[2] https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng_computergraphics

3. Methodology

3.1. Introduction

One of the enduring grand challenges in computer graphics is to create plausible animations to virtual avatars. Synthesizing human motion through learning techniques is becoming an increasingly popular approach to alleviating the requirement of new data capture to produce animations. Realistic human motion synthesis has a myriad of applications such as in graphic's animation for entertainment, robotics, and in graphic rendering engines with human crowds. However, generating such animations is often a cumbersome and time consuming task, that demands experienced professionals. Learning to move naturally from music, *i.e.*, to dance, is one of the more complex motions humans often perform effortlessly. Modeling such movements has been often relegated to motion capture systems. In this context, synthesizing human dance motions through learning techniques is becoming an increasingly popular approach to alleviating the requirement for new data capture to produce dance animations [Lee et al. 2019, Li et al. 2020, Huang et al. 2021, Li et al. 2021]. Yet, these recent approaches are challenged by the rich spatiotemporal motion distribution of possible moves to a same music style, and by the non-Euclidean geometry of the motion manifold structure.

This thesis proposes an automatic learning-based dance generation approach, using audio information (*e.g.*, music) as input, that considers both the variability and non-Euclidean geometry of the human motion during the generation. In summary, we aim to synthesize human motions from music regarding three main aspects: *i)* The synthesized motions should be plausible (as realistic as possible) to users; *ii)* The generated motions must preserve the main dance/music style characteristics; *iii)* It must generate motions with variability to different music styles. We consider all these aspects with a carefully designed learning-based approach trained with an adversarial regime. Both the non-Euclidean and temporal constraints of the human body motion are handled by the proposed spatial-temporal convolutional graph network model, trained using music audio and visual dance samples.

3.2. Related Work

In the last few years there has been substantial improvements in human pose estimation from images in both $2D$ or $3D$ spaces. An remarkable example in $2D$ is [Cao et al. 2019]. Deep fake methods to generate realistic virtual avatars are also increasingly achieving better results as shown in works such as [Wang et al. 2018, Gomes et al. 2020]. Following the same trend, human motion generation methods have been proposed in the last few years aiming to synthesize realistic human motions to music [Lee et al. 2019, Li et al. 2020, Huang et al. 2021, Li et al. 2021]. The recent seminal work of [Kipf and Welling 2017] introduced GCNs, that proved to be a more suitable architecture to deal with the human body structure, with state-of-the-art results in several tasks such as in human action recognition [Yan et al. 2018].

Despite these advances, partially achieved by the possibilities given by convolutional neural networks (CNN), both the human body structure and motion space variability are yet not well modeled by existing motion generation approaches. Thus, designing adapted architectures is a central challenge in properly modeling the human motion variability in the dance context.

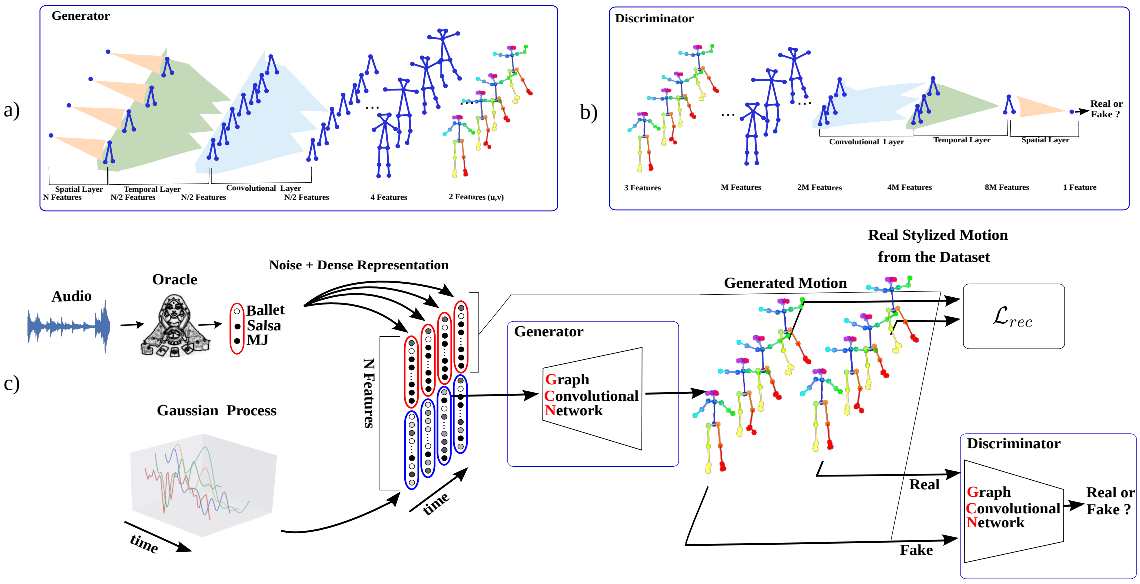

Figure 1. Motion generation model and adversarial training strategy.

3.3. Method

Our method has four main steps: *i)* Input audio classification into its corresponding music style; *ii)* The combination of the music style with a sampled Gaussian noise with temporal coherence into a single latent vector; *iii)* The synthesis of a sequence of $2D$ human poses temporally coherent from the latent vector; *iv)* and finally, the animation of virtual avatars using the generated motions. The architecture and training protocol for our methodology is illustrated in Figure 1.

In the first step of our methodology, we classify an input audio using a one dimensional CNN. We use as backbone the architecture of SoundNet proposed by [Aytar et al. 2016]. Then the result of the audio classification is transformed in a dense representation of the audio class. This dense representation is then used to control the motion generation. In other words, the dance class of the output motion is defined by the audio classification step. In the second step, we create a dense representation of the dance class and a temporal coherent Gaussian noise sampled with a Gaussian Process. The combination of the sampled Gaussian noise with the dense representation of the audio class is the input of our generative network.

We design a graph convolutional network (GCN) to generate the human motions. The synthesis of temporally coherent motion is done by the GCN model trained with an adversarial strategy. Generative adversarial networks (GANs) are the state of the art in generative models, which training protocol consists of two neural networks competing against each other. One network (generator) tries to generate realistic samples of a distribution, and the other network (discriminator) aims to distinguish between the samples generated by the generator network, and the samples from the dataset, *i.e.* the real ones. In general, GANs aim to transform a noise into a sample of a distribution, often the noise has lower dimensionality than the distribution samples. The full network is mainly composed of three types of layers: temporal and spatial upsampling operations, temporal and spatial downsampling operations; and graph convolutions. We use a combination of a classical

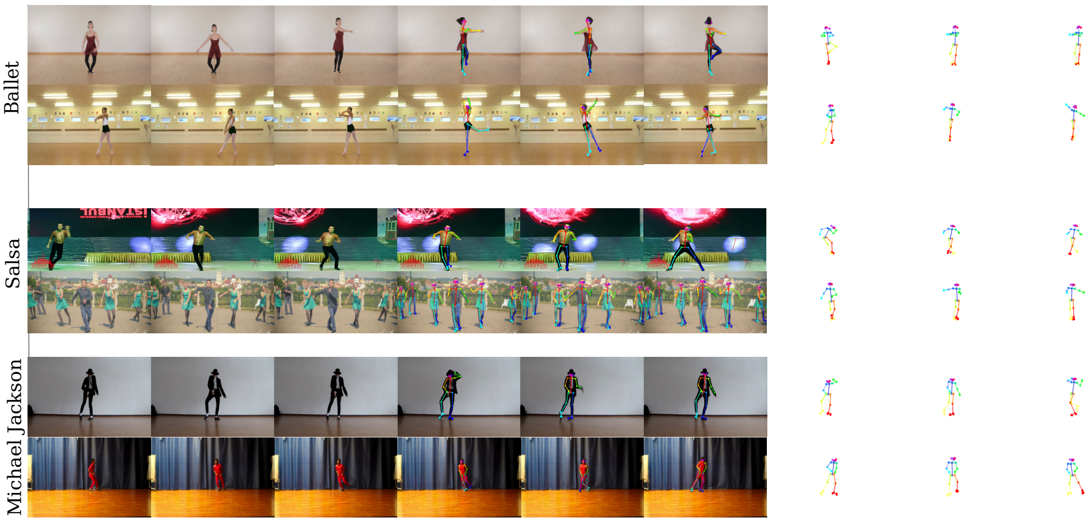

Figure 2. Visualizations of some samples from our multimodal dataset.

conditional GAN loss term with a motion reconstruction term to create our training loss function.

Finally, after training the generator network, we can synthesize a sequence of $2D$ human poses to be used by a deep-fake method, *e.g.*, vid2vid [Wang et al. 2018] to create video sequences of virtual avatars performing the synthesized motions. We highlight that any other method of virtual avatar animations could be used in this step, we choose vid2vid for its simplicity and since its widely used in the community. At the end of our proposed pipeline, we are able to generate realistic video sequences of virtual actors dancing, some results and the pipeline can be seen in the video available at: `https://youtu.be/fGDK6UkKzvA`.

3.4. Audio-Visual Dance Dataset

Due the lack of high-quality dataset specific for the task of automatic human dance generation, we build a new multimodal dataset with human selected samples (with high quality annotation) from publicly available videos on internet of three dance styles. We split the samples into training and evaluation sets that contain multimodal data for the following music/dance styles: Ballet, Michael Jackson, and Salsa. These two sets are composed of two data types: visual data from careful selected parts of publicly available videos of dancers performing representative movements of the music style and paired audio data. An illustration of some samples in the dataset can be seen in Figure 2.

3.5. Experiments & Results

We evaluate our method against the state-of-the-art approach presented in the work of [Lee et al. 2019], hereinafter named D2M. We also conducted a user study with 60 users. As discussed in [Ferreira et al. 2020], we observed that most users could not distinguish between the real and generated motions in the user study. Moreover, we present the results for the most commonly used metric for the evaluation of generative models (Fréchet inception distance – FID) in Table 1. This metric measures the gap between the distributions of generated motion samples to real motions. Our approach presented

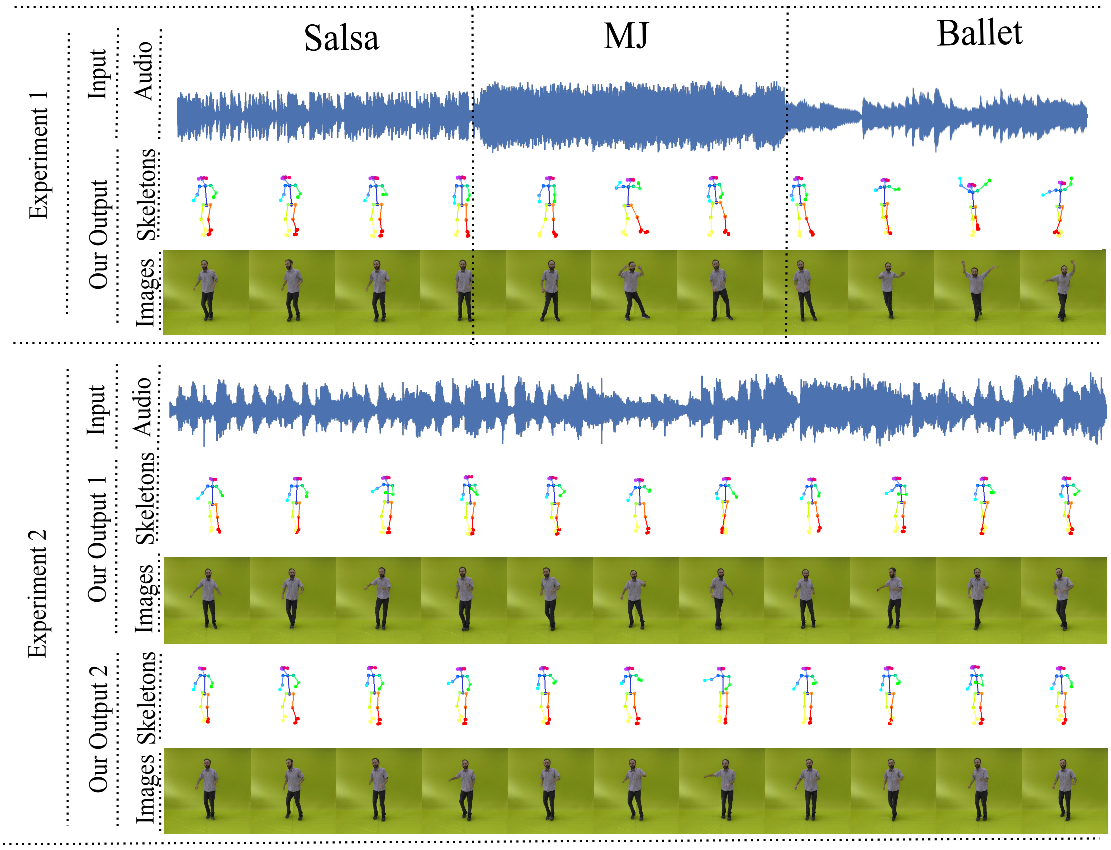

Figure 3. Motion generation experiments with the proposed approach.

Table 1. Quantitative values of Fréchet inception distance (FID).

	FID[1]		
Dance Style	D2M	Ours	Real
Ballet	20.20 ± 4.41	**3.18 ± 1.43**	2.09 ± 0.58
MJ	**4.38 ± 1.94**	8.03 ± 3.55	5.60 ± 1.42
Salsa	12.23 ± 3.20	**4.29 ± 2.38**	2.40 ± 0.75
Average	12.27 ± 7.27	**5.17 ± 3.33**	3.36 ± 1.86

[1] *Better closer to 0.*

overall the smallest FID distances for the considered dance styles than the state-of-the-art competitor.

Qualitative results of our methodology are shown in Figure 3. Experiment 1 (top) shows the ability of the method to generate different sequences with smooth transitions from one given input audio composed of different music styles. The Experiment 2 (bottom) illustrates the capability to generate two different motion sequences from the same given input music. In turn, the proposed method can change the dance generation style over time (as shown in the Experiment 1), when the auditory data changes as well. It also has the capability to synthesize, from a same audio sample, as many motion sequences as desired (as shown in the results shown in Experiment 2).

4. Conclusions

This master thesis addressed the problem of automatic human motion generation in the dance context. We propose a new GCN architecture trained with an adversarial strategy, to generate new dance motions, which in turn are used to animate and produce video sequences of virtual human avatars. A novel multimodal music-dance dataset is built to allow the dance learning to different styles, which are made publicly available to the community. Our method can generate diverse motion sequences for the same input audio, which is a limitation of existing methods using the audio directly as input. Moreover, the proposed approach allows changing the dance style over time using the music style conditioning mechanism. Experimental evaluations showed that our generated motions are more realistic than the state-of-the-art competitor, with results even comparable to the real motions in a user study.

References

Aytar, Y., Vondrick, C., and Torralba, A. (2016). Soundnet: Learning sound representations from unlabeled video. In *Advances in neural information processing systems*.

Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., and Sheikh, Y. A. (2019). Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ferreira, J. P., Coutinho, T. M., Gomes, T. L., Neto, J. F., Azevedo, R., Martins, R., and Nascimento, E. R. (2020). Learning to dance: A graph convolutional adversarial network to generate realistic dance motions from audio. *Computers & Graphics*.

Gomes, T. L., Martins, R., Ferreira, J., Azevedo, R., Torres, G., and Nascimento, E. R. (2021). A shape-aware retargeting approach to transfer human motion and appearance in monocular videos. *International Journal of Computer Vision*.

Gomes, T. L., Martins, R., Ferreira, J., and Nascimento, E. R. (2020). Do as I do: transferring human motion and appearance between monocular videos with spatial and temporal constraints. In *IEEE Conference on Applications of Computer Vision (WACV)*.

Huang, R., Hu, H., Wu, W., Sawada, K., and Zhang, M. (2021). Dance revolution: Long sequence dance generation with music via curriculum learning. In *ICLR 2021*.

Kipf, T. N. and Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *ICLR*.

Lee, H.-Y., Yang, X., Liu, M.-Y., Wang, T.-C., Lu, Y.-D., Yang, M.-H., and Kautz, J. (2019). Dancing to music. In *Advances in Neural Information Processing Systems*.

Li, J., Yin, Y., Chu, H., Zhou, Y., Wang, T., Fidler, S., and Li, H. (2020). Learning to generate diverse dance motions with transformer. *arXiv preprint arXiv:2008.08171*.

Li, R., Yang, S., Ross, D. A., and Kanazawa, A. (2021). Learn to dance with aist++: Music conditioned 3d dance generation. In *eprint arXiv: 2101.08779*.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Liu, G., Tao, A., Kautz, J., and Catanzaro, B. (2018). Video-to-video synthesis. In *Conference on Neural Information Processing Systems*.

Yan, S., Xiong, Y., and Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI Conference on Artificial Intelligence*.