GrooveNet: Real-Time Music-Driven Dance Movement Generation using Artificial Neural Networks*

Omid Alemi
School of Interactive Arts + Technology
Simon Fraser University
oalemi@sfu.ca

Jules Françoise
School of Interactive Arts + Technology
Simon Fraser University
jfrancoi@sfu.ca

Philippe Pasquier
School of Interactive Arts + Technology
Simon Fraser University
pasquier@sfu.ca

ABSTRACT

We present the preliminary results of GrooveNet, a generative system that learns to synthesize dance movements for a given audio track in real-time. Our intended application for GrooveNet is a public interactive installation in which the audience can provide their own music to interact with an avatar. We investigate training artificial neural networks, in particular, Factored Conditional Restricted Boltzmann Machines (FCRBM) and Recurrent Neural Networks (RNN), on a small dataset of four synchronized music and motion capture recordings of dance movements that we have captured for this project. Our initial results show that we can train the FCRBM on this small dataset to generate dance movements. However, the model cannot generalize well to music tracks beyond the training data. We outline our plans to further develop GrooveNet.

KEYWORDS

Human Movement, Dance, Music, Dance Generation, Audio-Driven Movement generation

ACM Reference format:
Omid Alemi, Jules Françoise, and Philippe Pasquier. 2017. GrooveNet: Real-Time Music-Driven Dance Movement Generation using Artificial Neural Networks. In *Proceedings of SIGKDD 2017 Workshop on Machine Learning for Creativity, Halifax, Nova Scotia, Canada, August 2017 (ML4Creativity 2017),* 6 pages.
https://doi.org/10.475/123_4

1 INTRODUCTION

Generating human movement remains one of the most challenging problems in computational modelling: movement is continuous, highly dimensional, and fundamentally expressive. Recognizing and generating everyday movements such as walking demands the development of elaborate models that can capture the coordination of a large set of joints [1, 27].

In this paper, we address the problem of movement generation for the case of dance, a creative activity that best illustrates the complexity and expressiveness of human movement. Dancing involves complex cognitive and sensorimotor processes: it requires both fine motor control and equilibrium, accurate timing and rhythmic synchronization, memory, and imagery, as well as aesthetic qualities [3]. The way we dance in response to music depends on the genre of both dance and music, the expertise of the dancer, and their interpretation of the music in real-time. Beyond mere synchronization processes, there is no well-established relationship between movements features and musical features, except in simplified cases such as sound tracing [6, 20]. As a result, generating dance movements from music is a highly non-linear and time-dependent mapping problem.

We investigate how machine learning can capture the cross-modal dependencies between synchronized sequences of musical features and movement parameters. We present *GrooveNet*, a system for real-time music-driven dance generation that uses artificial neural networks to learn the relationships between audio features and motion capture data.

The generation of human-like creative movement is necessary for a wide range of applications, spanning animation, gaming, virtual reality, and virtual characters. Our primary field of application is artistic and aims to explore the possibilities of computer-generated movement for creative purposes. Our music-to-dance generative system will be used in a public interactive installation allowing the audience to affect the movements of a dancing avatar by playing their own music. The avatar will be rendered with non-realistic visualizations of human movement through a holographic display.

GrooveNet relies on a machine learning model trained on a set of recordings of dance movements performed with dance music. We formulate this problem as learning the effects of one sequence on another sequence, in which both sequences are defined along a relatively dense time dimension (e.g., as opposed to text). The model is trained on synchronized sequences of dance and music features, in order to generate new movements from a new music track.

The development of GrooveNet faces a number of challenges related to the task of learning how dance movements are coordinated with music. First, as we already discussed, the mapping between audio features and movement parameters is highly non-linear, and it requires the model to learn and embed complex temporal structures. Second, there are no publicly available datasets that contain synchronized pairs of dance and music in the form of motion capture data and raw audio features. Therefore, we have recorded a dataset of four dance performances, raising about 23 minutes of synchronized movement and audio data. As a result, the model should learn efficiently from a relatively small dataset. Third, our main application consists of a public interactive installation allowing the audience to provide their own music to a dancing avatar. This

*__Complementary Material:__ http://omid.al/groovenet-material-ml4c

ML4Creativity 2017, August 2017, Halifax, Nova Scotia, Canada

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

demands that movements are generated in real-time from an audio stream and that the algorithm generalizes to new, and possibly unheard, music sequences.

2 RELATED WORK

In this section, we review movement generation techniques that rely on machine learning with a particular focus on dance movements and audio-driven approaches.

2.1 Machine-Learning-Based Movement Generation

A variety of machine learning models are used for learning and generating human movement in the form of motion capture data. They range from dimensionality reduction techniques [24, 27], to Hidden Markov Models [5], Gaussian Processes [28], and neural networks [8, 17, 26].

Dimensionality reduction techniques can capture the underlying correlations behind the joint rotations representing the postures in motion capture data [24, 27]. However, such techniques require pre-processing steps such as sequence alignments and fixed-length representation of the data, which limit their application to real-world dance data. Most importantly, their inability to directly model the temporality of the movement data is critically limiting for movement generation. Gaussian Process Latent Variable Models (GPLVMs) [28] can efficiently generalize over the variations in human movement, but they are limited by needing heavy computational and memory resources, which makes them unsuitable for real-time generation. Hidden Markov Models (HMMs) overcome the limitations of the two aforementioned families of models [5], but provide a limited expressive power in terms of capturing the variations in the data.

Neural networks provide a better expressive power than HMMs and depending on their size and architecture they can generate new samples in real-time. Convolutional Autoencoders have shown promising results in generating motion capture data with offline control [17]. Factored Conditional Restricted Boltzmann Machine (FCRBM), with its special architecture that is designed to support controlling the properties of the generated data, has shown to be able to generate movements in real-time, and learn a generalized space of the movement variations [1, 26]. In addition, Recurrent Neural Networks (RNNs), and in particular Long Short-Term Memory RNNs (LSTM-RNNs) are used to learning and generation movements [8] in an unsupervised and uncontrolled manner.

GrooveNet employs a similar machine learning model as the one used by Taylor and Hinton [26] and Alemi et al. [1]. As our initial experiments showed that it is more challenging to train RNNs on our small dataset, we have yet to employ RNN-based architectures such as the seq-to-seq model [2, 25], which is a more relevant architecture for our task of sequence-to-sequence mapping.

2.2 Dance Movement Generation

Many of the existing machine-learning-based movement generation techniques have been applied to dance. Hidden Markov Models and their extensions have been applied to the synthesis of dance movements [5, 22]. In particular, Wang et al. trained Hierarchical Hidden Markov Models with non-parameteric output distributions (NPHHMM) on motion capture data containing ballet walk, ballet roll, disco, and complex disco [29]. Another approach relies on dynamical systems modelling to capture the dynamics of dance movements. Li et al. [19] used Linear Dynamical Systems (LDS) to learn and generate dance movements. They train their model on 20 minutes of dance movement of a professional dancer, performing mostly disco. Their model automatically learns motion textons, representing local movement dynamics. The intuition behind the textons is that each complex movement sequence consists of simple repetitive patterns. For example, a dance sequence might consist of repeated moves such as spin, hop, kick, and tiptoeing. The approach allows for real-time synthesis and provides a number of ways to generate movements, such as key-framing and noise-driven generation.

Recently, artificial neural networks have been successfully applied to the synthesis of dance movements. Donahue et al. [11] focused on generating choreographies, represented as step charts that encode the timing and position of steps, for the Dance Dance Revolution game. They used LSTMs to generate a new step chart, given a raw audio track. Their method, however, is limited to the generation of sequences of discrete step indicators rather than continuous movements. Crnkovic-Friis and Crnkovic-Friis [8] used Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNNs) to learn and generate choreography. They trained the model on 6 hours of contemporary dance data captured using Microsoft Kinect. This approach does not provide any methods for controlling the generation and does not accompany any music.

Controlling ML-based Movement Generation. There are a number of different methods to control the qualities of the movements generated by a machine learning model: 1) train a separate model for each realization of a movement quality, 2) use parametric statistical distributions to capture the variations of movement [15, 31], and 3) design machine learning models specifically to accommodate the task of controlling the generation process [26].

For GrooveNet, we employ the generation controlling approach based on the FCRBM architecture proposed by Taylor and Hinton [26]. However, our work differs from the aforementioned neural-network-based dance generation approaches in that we attempt to generate continuous dance movements controlled by a given audio track.

2.3 Audio-Driven Movement Generation

Speech-Driven Synthesis. Many approaches to movement generation for virtual avatars rely on audio signals to guarantee that the synthesized gestures are consistent with other modalities. In general, the input audio is a speech signal that drives the generation of movements of the lips [30], eyebrows [10], head [16], or hands [18]. In this case, the goal is to ensure that the motor behavior is realistic and consistent with both the content and the expression of the input speech utterances.

Most approaches rely on probabilistic models such as Hidden Markov Models (HMMs) and its extensions [10, 16, 30] or Hidden Conditional Random Fields [18]. Chiu and Marsella [7] proposed Hierarchical Factored Conditional Restricted Boltzmann Machines (HFCRBM) to learn and generate gestures, controlled by the prosody of speech. Using a set of training data that includes motion capture recordings of gestures accompanied with the voice recordings of the actors (represented by pitch, intensity, and correlation), the

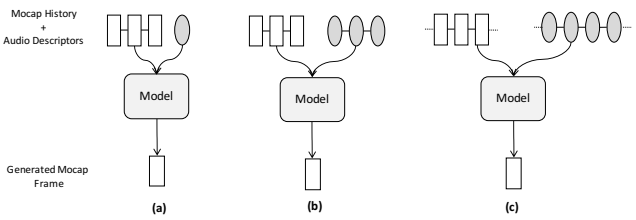

Figure 1: The mapping approaches: (a) one-to-many, (b) synchronized many-to-many, and (c) unsynchronized many-to-many. Each rectangle represents a single mocap frame and each ellipse represents a single audio descriptor frame. Connected frames represent consecutive frames.

model learns the relationship between the prosody of speech and the movement. The model then generates novel gestures given a new set of voices.

Music-Driven Dance Generation. While music-driven dance generation also considers cross-modal sequence-to-sequence mapping, it is important to underline the complexity of music-to-dance mapping. While in speech the acoustic and motion signals are often generated by the same underlying process, the relationships between music and movement in dance are far more complex and arbitrary. They depend on the genre and context, the expertise and personal characteristics of the performer, and they present a complex hierarchy of temporal structures, spanning from the short-term synchronization of gestures to the beat to long-term evolutions of the dance patterns.

Ofli et al. introduced an audio-driven dancing avatar using HMM-based motion synthesis [21]. For training, their approach requires movement to be manually annotated into specific patterns (or dance figures) synchronized with the beats. For the generation, the audio is segmented using beat detection, and the recognition of the patterns from Mel-Frequency Cepstral Coefficients is used to select the motion patterns to generate. Their approach was further extended to include unsupervised analysis of the dance patterns [22]. Ofli et al. describe three types of models: *musical measure models* and *exchangeable figures models*, which respectively represent many-to-one and one-to-many associations between musical patterns and dance figures, as well as *figure transition model* that capture the intrinsic dependencies of dance figures. Yet, one of the main limitation of these approaches lies in the synthesis approach, that relies on the classification of the input musical patterns, and therefore gives few opportunities for generating novel movement patterns.

GrooveNet differs from the aforementioned music-driven dance generation approaches in that it does not rely on classification or segmentation of the audio signal. The rationale behind GrooveNet is to allow the model to learn a continuous cross-modal mapping from the audio information to movement data in an unsupervised manner, as opposed to a supervised classification-based approach in which one would limit the generalizability of the model by restricting the mapping to a set of pre-defined patterns.

3 PROPOSED APPROACHES

In this paper, we aim to learn the relationships between low-level audio features and movement parameters for the continuous synthesis of full-body movements, without any supervision imposed on the mapping. With GrooveNet, we are investigating several directions to address this problem. This involves the pipeline of the system, the choice of a suitable machine learning model, and different methods for representing the audio data.

Pipeline. Three strategies for mapping audio data to motion capture (mocap) data are illustrated in Figure 1: (a) one-to-many mapping, (b) synchronized many-to-many mapping, and (c) unsynchronized many-to-many mapping. While in all approaches the model takes a sequence of mocap frames as input, they differ in how the audio descriptors are involved. In a one-to-many mapping approach, the audio descriptor at time t together with the input mocap history determines the generated mocap frame at time t. In a synchronized many-to-many mapping, for each mocap frame, there exist an audio descriptor. The model takes a sequence of mocap history corresponding to the frames at the time interval of $[t - N, t - 1]$, and a sequence of audio descriptors, with the same length as the history, and generates the output mocap frame at time t. In an unsynchronized many-to-many mapping, the model takes the mocap history and the audio descriptor sequences that have different lengths, and it is up to the model to determine their temporal correlations. In this paper, we present a model following the one-to-many mapping approach.

Machine Learning Model. The two machine learning models that we are employing are Factored Conditional Restricted Boltzmann Machines (FCRBMs) and Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN). FCRBM has shown to be a suitable choice for movement generation, in particular to allow for a fine control over the generated movements [1, 26] and because it can generalize over the space of variations. The LSTM-RNN is powerfull to model time series with complex temporal structures, and was shown efficient for controlled character and hand-writing generation [14], as well as uncontrolled dance movement generation [8].

Our initial experiments show that compared to the LSTM-RNN, it is easier to train the FCRBM on real-valued, continuous data. Also, FCRBM works better on smaller training sets, and it is faster during generation. While FCRBM works better for the one-to-many mapping approach, the LSTM-RNN is more suitable for many-to-many approaches. As our initial experiments have not been successful with the LSTM-RNN yet, in this paper, we only report the experiments using FCRBM.

Audio Representation. With respect to representing audio data, we follow two different approaches: 1) feature extraction and 2) feature learning. We describe our approach to audio feature extraction in Section 4. For feature learning, we have recently started training GrooveNet on audio features based on the temporal embeddings from a WaveNet-style auto-encoder [12], which is trained on raw audio from musical instrument sounds. In this paper, we only report the results from training GrooveNet with the extracted features.

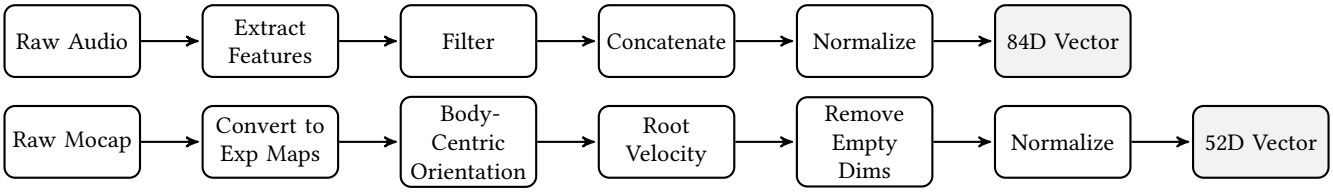

Figure 2: Overview of GrooveNet's data processing pipeline for the audio and movement modalities.

4 DATASET AND FEATURE EXTRACTION

Few motion capture datasets include dance movement data. To our knowledge, no dataset of synchronized music and motion capture data is currently available online. We created a dataset containing four performances of a dancer. The music tracks are made by Philippe Pasquier and Philippe Bertrand at the Robonom sound studio in France using the *StyleMachine lite* from the Metacreative Technologies company[1]. We used three of the generated songs that belong to the genre of electronic dance music, with a regular tempo varying between 125 and 135 beats per minute.

The dancer's movements were captured using a 40-cameras Vicon optical motion capture system. The motion capture data was post-processed and synchronized with the audio data. The resulting dataset contains about 23 minutes of motion capture data recorded at 60 frames-per-second, giving a total of 82151 frames. We recorded a total of four sequences, with two sequences dancing to the first song, and two sequences dancing to the second and the third songs.

4.1 Audio Data Representation and Feature Extraction

Our goal is to generate movement in real-time from an audio stream. To that end, the audio signal must be represented by a sequence of features that describe the acoustic properties of the music continuously, with a similar temporal density as the movement data.

For each audio file, we extracted a set of low-level features at the same framerate as the motion capture data. We used a standard set of features described in the music information retrieval literature [4, 23], including low-level features (RMS level, Bark bands), spectral features (energy in low/middle/high frequencies, spectral centroid, spectral spread, spectral skewness, spectral kurtosis, spectral rolloff, spectral crest, spectral flux, spectral complexity), timbral features (Mel-Frequency Cepstral Coefficients, Tristimulus), melodic features (pitch, pitch salience and confidence computed with the YIN algorithm [9], inharmonicity, dissonance). The features were computed using the Essentia open-source library [4], with a window size of 66.7 ms and a hop size of 16.7 ms. The feature sequences were filtered with a finite impulse response (FIR) low-pass filter with a cutoff frequency of 5 Hz, in order to guarantee a smooth evolution of the audio descriptor that matches the time scale of dance movements. The resulting sequences are synchronized with the motion capture data and contain 84 dimensions (Figure 2-top).

4.2 Motion Capture Data Representation

The original motion capture data uses a skeleton with 30 joints, resulting in 93 dimensions including the root position, with their rotations represented in Euler angles. The data is recorded at 60 frames-per-second. We converted the Euler angle representations to exponential maps [13] to avoid loss of degrees-of-freedom and any discontinuities. We removed the empty and fixed dimensions of the data. We also replaced the root's global orientation with its rotation velocity along the axis that is perpendicular to the floor plane and replaced the root's global translation with the 2-dimensional velocity of the root as projected on the floor plane. The resulting dataset contains 90151 frames, each represented by a 52-dimensional vector (Figure 2-bottom).

[1] https://metacreativetech.com

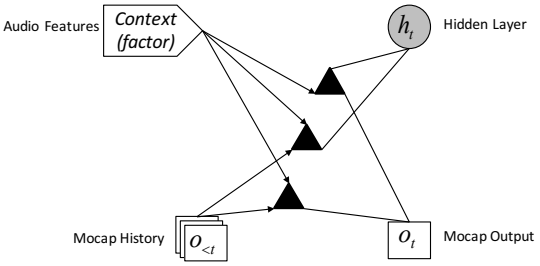

Figure 3: The architecture of a Factored CRBM with audio features fed into its context unit and mocap feature to its output/visible units.

5 THE MACHINE LEARNING PROCESSES

5.1 Factored Conditional Restricted Boltzmann Machines

We use a Factored Conditional Restricted Boltzmann Machine (FCRBM) [26], shown in Figure 3, as the underlying machine learning model of this version of GrooveNet. FCRBM is an energy-based generative model that learns to predict its output given a sequence of input data, modulated by its context data. Using a set of three multiplicative gates, the values of the context unit modulate the weights between the condition units (history data), the hidden units, and the output visible units. This arrangement allows the context data to directly, and in a non-linear way, control the network's output by manipulating the networks energy landscape.

FCRBM supports a multi-dimensional discrete or continuous context variable, which allows this model to capture and represent different qualities and semantics of human movement. Furthermore, one can interpolate or extrapolate the context values in order to create new movements that did not exist in the training data. In GrooveNet, we feed the audio features to this context unit to allow the model to learn the relationship between the audio features and the dynamic processes behind the movements in the training data.

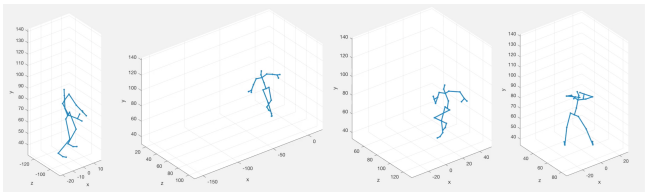

Figure 4: Some still frames from the generated movement patterns.

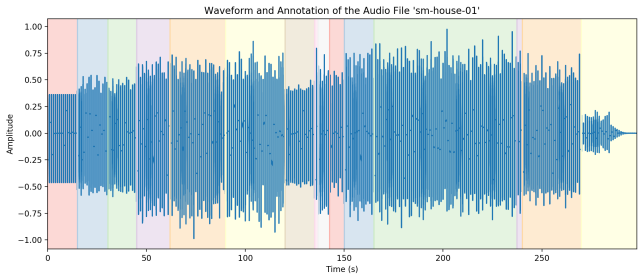

Figure 5: An example of manual segmentation of the first song used for preliminary experiments in generating movement patterns.

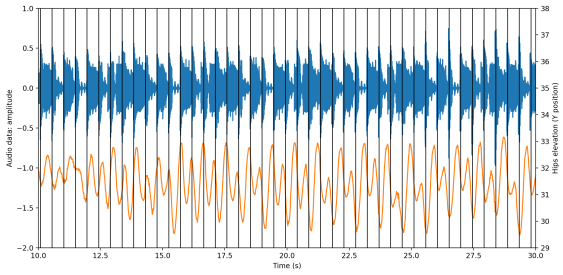

Figure 6: Visualization of the hips' position along the vertical axis (bottom) and the audio amplitude (top).

5.2 Learning

We train the model to predict the next motion capture frame at time t, given a recent history of the motion capture frames $[t-N, t-1]$, where N is the order of the model, representing the number of past motion capture frames to include in the prediction. The prediction is modulated by a single frame of audio features at time t, fed to the context unit.

5.3 Generation

The generation is done using an iterative sampling process. The model predicts one frame of the movement at a time, given previous frames of movement and the audio features. It then uses the newly generated frame of movement as part of the input for predicting the next frame and continues the generation process.

6 PRELIMINARY RESULTS AND DISCUSSION

In this section, we present the preliminary results of GrooveNet. All the animated outputs of the results are available on the accompanying webpage of the paper at: http://omid.al/groovenet-material-ml4c/.

6.1 Learning and Generating Movement Patterns

We start our investigation by learning the individual dance patterns that exist in our training data to see if the model can generate the patterns independently of the audio data. To this end, we manually segment the dance sequences based on the main parts of the song, as illustrated in Figure 5. This annotation allows us to assess where the model can effectively encode consistent dance patterns. We then use a one-hot encoding scheme to label each pattern. Once the model is trained, we can then generate each pattern by feeding the desired label to the context unit of the FCRBM.

The results show that the FCRBM is able to learn the patterns from a very small training set (only one mocap sequence of about 4 minutes). The model generates the same pattern continuously and repetitively as long as the same label is given to it while changing the label will cause the model to transition to another pattern. Still frames from the generated patterns for the first track are shown in Figure 4. The videos of the generated patterns are also available in the supplementary material.

6.2 Dancing with Training Songs

In the previous experiment, we assessed the ability of the model to encode independent dance patterns from the manual annotation of the dataset. We now consider a fully unsupervised approach, where the model is trained with the entire dataset composed of four performances without any additional annotation. Our goal is to evaluate whether the model can learn the mapping between the audio features and the movement parameters on longer sequences, with a larger corpus of music and dance. As the first step in this direction, we assess the movements generated by the model using as input data the songs already 'heard' during the training.

The results, as presented in the supplementary materials, show that the FCRBM is able to generate dance patterns consistent with the training set. Furthermore, the model captured the synchronization patterns between the rhythmic structure of the song and the generated movement (Figure 6). While the generated dance movements are plausible, we can note that the movements are at times jerky and can present artifacts such as foot sliding. The novelty in the generated movements remains to be further investigated.

6.3 Dancing with Unheard Songs

We evaluate the generalizability of the model, testing its performance on the songs that were not included in the training data. The results, as shown on the accompanying website, show that the FCRBM is not generalizing beyond the songs that exist in the training data.

6.4 Computational Performance

A model with 500 hidden units, 500 factors, and an order of 30 past frames consists of 1,452,720 number of trainable parameters, and it takes on average 0.0115 seconds to generate each frame on an

Intel(R) Core(TM) i7-4850HQ CPU at 2.30GHz. This is fast enough to generate the movements at 60 frames-per-second in real-time.

7 CONCLUSION AND ROAD MAP

We presented the initial results from our GrooveNet project, in which we address the generation of dance movements in real-time from musical audio. This problem involves learning a cross-modal mapping between acoustic features and movement data and generating dance movements from a new audio sequence.

We are investigating multiple audio-to-movement mapping approaches, machine learning models (FCRBM and LSTM-RNN), and description methods of the musical information (features extraction vs. feature learning). Among these, we presented the results of training an FCRBM on extracted audio descriptors that are used in the audio signal processing community.

Our preliminary analysis shows that our model can learn and generate basic dance movements, independent of the audio data. In addition, it can learn and generate movements based on the song that it is trained with. However, the model currently falls short in generalizing beyond those songs in the training data and highly overfits. We believe that this is mainly due to our small and sparse training data set.

To further develop GrooveNet, we plan on following a number of directions: 1) capturing more synchronized dance and music data; 2) taking a semi-supervised approach and pre-train a network on more motion capture data of dance moves that do not accompany any music. Together with a pre-trained network that has learned audio embeddings, we hope that the model becomes more robust to 'unheard' songs; and 3) experiment more with LSTM-RNNs, seq-to-seq style architectures, and many-to-many mappings.

ACKNOWLEDGMENTS

This project is partially funded by the the Social Sciences and Humanities Research Council of Canada (SSHRC) and the Natural Sciences and Engineering Research Council of Canada (NSERC). We would like to thank Gemma Crowe for her dance performance and Rick Overington at the Emily Carr University Motion Capture Studio for his help in capturing the training data.

REFERENCES

[1] Omid Alemi, William Li, and Philippe Pasquier. 2015. Affect-expressive movement generation with factored conditional Restricted Boltzmann Machines. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 442–448.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR* abs/1409.0473 (2014).

[3] Bettina Bläsing, Beatriz Calvo-Merino, Emily S. Cross, Corinne Jola, Juliane Honisch, and Catherine J. Stevens. 2012. Neurocognitive control in dance perception and performance. 139, 2 (2012), 300–308.

[4] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, Perfecto Herrera, Oscar Mayor, Gerard Roma, Justin Salamon, José R Zapata, and Xavier Serra. 2013. Essentia: An Audio Analysis Library for Music Information Retrieval.. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR'13)*. Curitiba, Brazil, 493–498.

[5] Matthew Brand and Aaron Hertzmann. 2000. Style machines. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. 183–192.

[6] Baptiste Caramiaux, Patrick Susini, Tommaso Bianco, Frédéric Bevilacqua, Olivier Houix, Norbert Schnell, and N Misdariis. 2011. Gestural Embodiment of Environmental Sounds : an Experimental Study. In *Proceedings of the International Conference on New Interfaces for Musical Expression*.

[7] C.C. Chiu and Stacy Marsella. 2011. How to Train Your Avatar: A Data Driven Approach to Gesture Generation. In *Proceedings of the 10th international conference on Intelligent Virtual Agents*. Springer Berlin Heidelberg, Berlin, Heidelberg, 127–140.

[8] Luka Crnkovic-Friis and Louise Crnkovic-Friis. 2016. Generative Choreography using Deep Learning. *CoRR* abs/1605.06921 (May 2016).

[9] Alain De Cheveigné and Hideki Kawahara. 2002. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America* 111, 4 (2002), 1917–1930.

[10] Yu Ding, Mathieu Radenen, Thierry Artieres, and Catherine Pelachaud. 2013. Speech-driven eyebrow motion synthesis with contextual Markovian models. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 3756–3760.

[11] Chris Donahue, Zachary C Lipton, and Julian McAuley. 2017. Dance Dance Convolution. (March 2017). arXiv:1703.06891

[12] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck, Karen Simonyan, and Mohammad Norouzi. 2017. Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders. (April 2017). arXiv:1704.01279

[13] F Sebastian Grassia. 1998. Practical Parameterization of Rotations Using the Exponential Map. *Journal of Graphics Tools* 3, 3 (1998), 29–48.

[14] Alex Graves. 2013. Generating Sequences With Recurrent Neural Networks. (Aug. 2013). arXiv:1308.0850

[15] Dennis Herzog, Volker Krueger, and Daniel Grest. 2008. Parametric Hidden Markov Models for Recognition and Synthesis of Movements. In *Proceedings of the British Machine Vision Conference*. 163–172.

[16] GO Hofer. 2009. *Speech-driven animation using multi-modal hidden Markov models*. PhD Dissertation. University of Edimburgh.

[17] Daniel Holden, Jun Saito, and Taku Komura. 2016. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)* 35, 4 (July 2016), 138–11.

[18] Sergey Levine, Philipp Krähenbühl, Sebastian Thrun, and Vladlen Koltun. 2010. Gesture controllers. *ACM Transactions on Graphics* 29, 4 (jul 2010), 1.

[19] Yan Li, Tianshu Wang, and Heung-Yeung Shum. 2002. Motion texture: a two-level statistical model for character motion synthesis. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*. ACM, 465–472.

[20] K Nymoen, B Caramiaux, M Kozak, and J Tørresen. 2011. Analyzing Sound Tracings - A Multimodal Approach to Music Information Retrieval. In *ACM Multimedia, MIRUM 2011*.

[21] Ferda Ofli, Yasemin Demir, Yücel Yemez, Engin Erzin, a. Murat Tekalp, Koray Balcı, İdil Kızoğlu, Lale Akarun, Cristian Canton-Ferrer, Joëlle Tilmanne, Elif Bozkurt, and a. Tanju Erdem. 2008. An audio-driven dancing avatar. *Journal on Multimodal User Interfaces* 2, 2 (sep 2008), 93–103.

[22] F. Ofli, E. Erzin, Y. Yemez, and A. M. Tekalp. 2012. Learn2Dance: Learning Statistical Music-to-Dance Mappings for Choreography Synthesis. *IEEE Transactions on Multimedia* 14, 3 (June 2012), 747–759.

[23] G Peeters. 2004. *A large set of audio features for sound description (similarity and classification) in the CUIDADO project*. Technical Report.

[24] Ali-Akbar Samadani, Eric Kubica, Rob Gorbet, and Dana Kulić. 2013. Perception and Generation of Affective Hand Movements. *International Journal of Social Robotics* 5, 1 (2013), 35–51.

[25] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. *CoRR* abs/1409.3215 (2014).

[26] Graham W Taylor and Geoffrey E Hinton. 2009. Factored Conditional Restricted Boltzmann Machines for Modeling Motion Style. In *Proceedings of the th Annual International Conference on Machine Learning ICML*. Montreal.

[27] Joëlle Tilmanne and Thierry Dutoit. 2010. Expressive gait synthesis using PCA and Gaussian modeling. In *Proceedings of the Third international conference on Motion in games*. Springer-Verlag, Berlin, Heidelberg, 363–374.

[28] Jack M Wang, David J Fleet, and Aaron Hertzmann. 2007. Multifactor Gaussian process models for style-content separation. In *Proceedings of the 24th international conference on Machine learning*. ACM, 975–982.

[29] Yi Wang, Zhi-Qiang Liu, and Li-Zhu Zhou. 2005. Learning hierarchical nonparametric hidden Markov model of human motion. In *Proceedings of 2005 International Conference on Machine Learning and Cybernetics, 2005.* IEEE, 3315–3320.

[30] E Yamamoto, S Nakamura, and K Shikano. 1998. Speech-to-lip movement synthesis based on the EM algorithm using audio-visual HMMs. In *Proceedings of the Second Workshop on Multimedia Signal Processing*. 2–5.

[31] Takashi Yamazaki, Naotake Niwase, Junichi Yamagishi, and Takao Kobayashi. 2005. Human Walking Motion Synthesis Based on Multiple Regression Hidden Semi-Markov Model. In *Proceedings of the 2005 International Conference on Cyberworlds*. IEEE, 445–452.