# Dance2Music: Automatic Dance-driven Music Generation

**Gunjan Aggarwal**[1] **Devi Parikh**[1,2]
[1]Georgia Tech, [2]Facebook AI Research
gunjan10@gatech.edu, parikh@gatech.edu

## Introduction

Dance and music typically go hand in hand. The complexities in dance, music, and their synchronisation make them fascinating to study from a computational creativity perspective. While several deep learning [4; 5] and search [8] based solutions have been proposed to generate dance for a given music, the other direction of automatically generating music for a dance is still relatively unexplored. This capability could have several creative expression and entertainment applications.

We present some early explorations in the direction of automatically generating music for a given dance. Several works use sensors to control music generation through hand movements and eye gaze [7; 10; 6]. In contrast, our setup has a lower barrier to entry (it uses just a camera) and can also generate music post hoc (from a dance video).

We present a search-based *offline* approach that generates music after processing the entire dance video and an *online* approach that uses a deep neural network to generate music on-the-fly as the video proceeds. We quantitatively evaluate our *online* approach on 45 videos. We also present a strong baseline and compare our approaches on 10 dance clips via human studies. We have integrated our *online* approach in a live demo! A video of the demo can be found here: https://sites.google.com/view/dance2music/live-demo.

## Approach

Our goal is to generate music that goes well with an input dance. We hypothesize that if notes are (dis)similar when the associated poses in the dance are (dis)similar, the music will feel synced with the dance. Our approach maximizes this correlation between poses and notes.

**Offline Approach** In our offline approach, we generate music for an existing dance video. Given a sequence of $N$ dance frames, and a hyper-parameter $K$, we generate a music sequence of length $\frac{N}{K}$. $K$ – the *music interval* – denotes the number of dance frames after which we play a musical note. Based on an initial qualitative assessment, in our experiments we set $K$ to be 6. Recall that our aim is generate music, such that the generated music is maximally correlated with the dance.

**Correlation.** We bring dance and music to a unified similarity matrix representation and then compute the Pearson correlation between the two matrices vectorized. Let $D$ be the dance similarity matrix, and $M$ be the music similarity matrix. We define them as $D[i][j] = \text{cosine\_similarity}(pose[i], pose[j])$ and $M[i][j] = \frac{\text{abs(note[i]-note[j])}}{4.0}$. Here pose[i] denotes the 36 dimensional pose representation extracted from OpenPose [2] for the $i^{th}$ dance frame, and note[i] represents the note played at $i^{th}$ music interval. We use 5 piano notes in the C major pentatonic scale to compose our music. As the search space is exponential in the number of notes (e.g., $5^{60}$ for a 12-second video), an exhaustive search is infeasible. We use beam search to find the sequence with the highest correlation with the dance.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

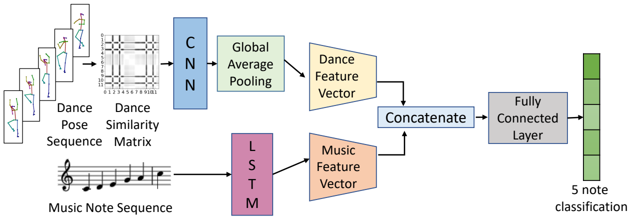


Figure 1: Neural network architecture for note prediction. Our network gets as input the local history of both dance similarity matrix and note sequences up to time $t$ and predicts the note for $t + 1$.

### Online Approach

The above approach cannot be run in real-time due to beam search – the note to be produced at a given time $t$ is not determined until the computation for future notes has been done. To enable real-time note generation, we model the problem using neural networks. We run the *offline* approach on a set of videos to collect paired dance and music data. Using this data, we train a neural network to take in as input the local history of the dance poses and generated notes (corresponding to past 60 frames), and produce as output the note to be played next. We model this as a 5-way classification task (to generate one of the 5 notes). Figure 1 shows the model architecture. For generating music on-the-fly as dance frames come in, we start with the first note being E4, and then sample iteratively from our trained model. The note predicted at time $t$ along with local history of both generated notes and dance poses is fed as an input for the note prediction at time $t + 1$.

The task that the neural network is modelling is difficult because it has access to partial information. For the *offline* approach, the note generated at any time is determined based on not just past but also future notes and frames. However, our neural network doesn't have access to future notes and frames (thus enabling the "online" use case).

**Baseline** Inspired by [1] that creativity is a combination of quality (value) and surprise (novelty), we design a baseline that generates a sequence of notes that is both in sync with the dance and unpredictable. We select a random note when the similarity between poses across a music interval is below a threshold, otherwise continue playing the previous note. We also compared to other (weaker) baselines optimized for only quality and only surprise but do not mention here for space constraints. We also experiment with adding chords to the music domain. Examples of music composed by our various algorithms can be found here: `https://sites.google.com/view/dance2music`.

## Results

We perform our experiments using the AIST dataset [9] and videos from [3]. We assume a single dancer in the video. All videos are ∼12 seconds long.

**Automatic metrics.** We start by evaluating how well the neural network in our *online* approach mimics the *offline* approach. We train the model on 455 videos from the AIST dataset, and evaluate it on 45 videos. It achieves a test accuracy of 73.5% at the task of predicting the next note accurately (that is, match the note produced by the *offline* approach). Note that performance of predicting the most common note would have an accuracy of 21.7%.

**Human Study.** Next we conduct human studies to compare our approaches to the strong baseline presented earlier. We use 10 videos – 8 from AIST and 2 from [3]. This gives us 10 video pairs for each comparison type: *offline* vs. baseline and *online* vs. baseline. For each comparison we show subjects a pair of videos, both contain the same dance but the music is generated by the different approaches. We ask subjects: *Which music composition goes better with the dance?* Additional details around the diversity of subjects can be found in Appendix section.

For *offline* vs. baseline, *offline* was preferred 77% of the times (23 out of 30), which is statistically significant. Subjects preferred the *online* approach over the baseline 70% of the time (21 out of 30), which is also statistically significant. This, combined with the automatic metrics reported above where the *online* approach mimics the *offline* approach well, suggests that our neural network model is a promising direction for generating music on-the-fly that goes well with a live dance.

## Ethical Considerations

The pose estimation might fail for certain body shapes missing from the dataset on which it was trained. Thus, our system might not produce music that goes well with the dance for certain body shapes. Also pentatonic scale, although found in most of the world's music, is more popular in certain parts of the world and might not sound entertaining/pleasing to people in other places.

## References

[1] M. Boden. *The Creative Mind.* Abacus, London, 1992.

[2] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE PAMI*, 43(1), 2019.

[3] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros. Everybody Dance Now. In *ICCV*, 2019.

[4] R. Huang, H. Hu, W. Wu, K. Sawada, M. Zhang, and D. Jiang. Dance Revolution: Long-Term Dance Generation With Music Via Curriculum Learning. *arXiv preprint arXiv:2006.06119*, 2020.

[5] H.-Y. Lee, X. Yang, M.-Y. Liu, T.-C. Wang, Y.-D. Lu, M.-H. Yang, and J. Kautz. Dancing To Music. *arXiv preprint arXiv:1911.02001*, 2019.

[6] R. Morales-Manzanares, E. F. Morales, R. Dannenberg, and J. Berger. SICIB: An Interactive Music Composition System Using Body Movements. *Computer Music Journal*, 25(2), 2001.

[7] T. Nawaz, M. S. Mian, and H. A. Habib. Infotainment Devices Control By Eye Gaze And Gesture Recognition Fusion. *IEEE Transactions on Consumer Electronics*, 54(2), 2008.

[8] P. Tendulkar, A. Das, A. Kembhavi, and D. Parikh. Feel The Music: Automatically Generating A Dance For An Input Song. *arXiv preprint arXiv:2006.11905*, 2020.

[9] S. Tsuchida, S. Fukayama, M. Hamasaki, and M. Goto. AIST Dance Video Database: Multi-Genre, Multi-Dancer, And Multi-Camera Database For Dance Information Processing. In *ISMIR*, 2019.

[10] O. Vaidya, K. Jadhav, L. Ingale, and R. Chaudhari. Hand Gesture Based Music Player Control In Vehicle. In *IEEE I2CT*, 2019.

# Appendix: Additional Technical Details

## Dataset

We perform our experiments using the AIST dataset [9] and videos from [3]. The former consists of street dance performed by different people across 10 different genres. It has each dance video captured from 9 angles; we use the ones with front facing camera view. The latter consists of short YouTube videos where a single subject dances in front of a static camera. All videos are ∼12 seconds long.

## Overview of offline approach

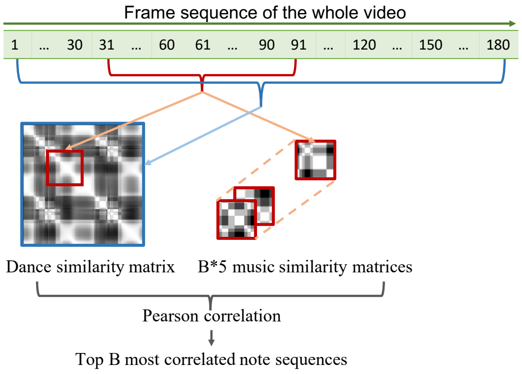
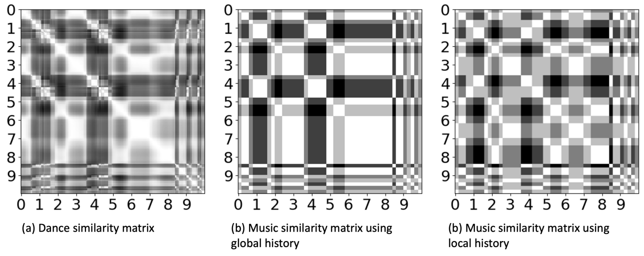

Frame sequence of the whole video
1 ... 30 31 ... 60 61 ... 90 91 ... 120 ... 150 ... 180
Dance similarity matrix
B*5 music similarity matrices
Pearson correlation
Top B most correlated note sequences

Figure 2: An overview of our *offline* music generation approach. At each time step, we look at the dance similarity matrix corresponding to a past local window (highlighted in red) and sort each of the $B * 5$ local candidate note sequences using Pearson correlation. The top $B$ correlated sequences are kept for the next steps of beam search, and the process repeats till notes have been generated for the entire video.

## Local history over global in online approach

We find that using the local history allows us to capture more minute details in the dance, while using the global history results in more "flat" music (see Figure 3).



(a) Dance similarity matrix

(b) Music similarity matrix using global history

(b) Music similarity matrix using local history

Figure 3: Example dance and music similarity matrices for our approaches. The axis labels indicate the time (in secs). Note that the music produced by using global history (b) looks flat (identical notes played for long duration of time) compared to using local history (c).

## Neural network training pipeline

Figure 1 shows the model architecture. We process the past dance similarity matrix via 6 CNN layers, each of kernel size 3 and filters 64, 128, 128, 256, 512 and 32 respectively. Max pooling is done after the first and third convolution and ReLU activation is used after each CNN layer. Global average pooling is performed to get the dance feature representation. The past music note sequence is fed as an input to an LSTM with hidden dimension 32 and the resulting features are concatenated with the dance features. Three fully connected (FC) layers of sizes 512, 256 and 128, each followed by ReLU activation are applied over the concatenated features. Finally an FC layer of size 5 is applied to get the final logits. The network is trained with Adam optimizer for 200 epochs with a learning rate of 2e-4. The dance similarity matrix and the music sequence are padded to ensure that each data point in a batch is of the same size.

### Human study participants distribution

18 subjects (11 male, 7 female) voluntarily participated in the study. Their ages range from 16 to 49 years. Each subject evaluated up to 3 video pairs for each comparison type.