SKELETON PLAYS PIANO: ONLINE GENERATION OF PIANIST BODY MOVEMENTS FROM MIDI PERFORMANCE

Bochen Li[1] **Akira Maezawa**[2] **Zhiyao Duan**[1]
[1] University of Rochester, USA [2] Yamaha Corporation, Japan
{bochen.li, zhiyao.duan}@rochester.edu, akira.maezawa@music.yamaha.com

ABSTRACT

Generating expressive body movements of a pianist for a given symbolic sequence of key depressions is important for music interaction, but most existing methods cannot incorporate musical context information and generate movements of body joints that are further away from the fingers such as head and shoulders. This paper addresses such limitations by directly training a deep neural network system to map a MIDI note stream and additional metric structures to a skeleton sequence of a pianist playing a keyboard instrument in an online fashion. Experiments show that (a) incorporation of metric information yields in 4% smaller error, (b) the model is capable of learning the motion behavior of a specific player, and (c) no significant difference between the generated and real human movements is observed by human subjects in 75% of the pieces.

1. INTRODUCTION

Music performance is a multimodal art form. Visual expression is critical for conveying musical expression and ideas to the audience [4,5]. Furthermore, visual expression is critical for communicating musical ideas among musicians in a music ensemble, such as predicting the leader-follower relationship in an ensemble [15].

Despite the importance of body motion in music performance, much work in automatic music performance generation has focused on synthesizing expressive audio data from a corresponding symbolic representation of the music performance (e.g., a MIDI file). We believe that, however, body motion generation is a critical component that opens door to multiple applications. For educational purposes, for example, replicating the visual performance characteristics of well-known musicians can serve as demonstrations for instrument beginners to learn from. Musicologists can apply this framework to analyze the role of gesture and motion in music performance and perception. For entertainment purposes, rendering visual performances along with music audio enables a more immersive music enjoyment experience as in live concerts. For automatic accompaniment systems, appropriate body movements of machine musicians provide visual cues for human musicians to coordinate with, leading to more effective human-computer interaction in music performance settings.

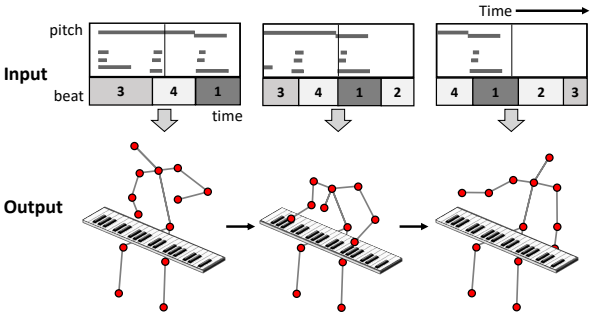


Figure 1. Outline of the proposed system. It generates expressive body movements as skeleton sequences like human playing on a keyboard instrument, given the input of MIDI note stream and metric structure information.

For generating visual music performance, i.e., body position and motion data of a musician, it is important to create an expressive and natural movement of the *whole body* in an online fashion. To consider both expressiveness and naturalness, the challenge is to maintain some common principles in music performance constrained by the musical context being played. Most previous work formulates it as an inverse kinematics problem with physical constraints, where the generated visual performance is limited to hand shapes and finger positions. Unfortunately, this kind of formulation fails to address the two challenges; specifically, (1) it fails to generate the whole body movements that are relevant to music expression, such as the head and body tilt, and (2) it fails to take into account the musical context constraints for generation, which do not contribute to ergonomics.

Therefore, we propose a body movement generation system as outlined in Figure 1. The input is a real-time *MIDI note stream* and a *metric structure*, without any additional indication of phrase structures or expression marks. The MIDI note stream provides the music characteristics and the artistic interpretations, such as note occurrence, speed, and dynamics. The metric structure indicates barlines and beat positions as auxiliary information. Given these the system can automatically generate expressive and natural body movements from any performance data in the

MIDI format. We design two Convolutional Neural Networks (CNN) to parse the two inputs and then feed the extracted feature representations to a Long Short-Term Memory (LSTM) network to generate proper body movements. The generated body movements are represented as a sequence of positions of the upper body joints [1]. The two complementary inputs serve to maintain a correct hand position on the keyboard while conveying musical ideas in the upper body movements. To learn a natural movement, we employ a two-stage training strategy, where the model is trained to learn the joint positions first, then later trained to also learn the body limb lengths.

2. RELATED WORK

There has been work on cross-modal generation, mostly for speech signals tracing back to the 1990s [1], where a person's lips shown in video frames are warped to match the given phoneme sequence. Given the speech audio, similar work focuses on synthesizing photo-realistic lip movements [14], or landmarks of the whole face [6]. Some other work focuses on the generation of dancers' body movements [9, 12] and behaviors of animated actors [11].

Similar problem settings for music performances have been rarely studied. When the visual modality is available, the system proposed in [8] explores the correlation between the MIDI score and visual actions, and is able to target the specific player in an ensemble for any given track. Purely from the audio modality, Chen et al. [3] propose to generate images of different instrumentalists in response to different timbres using cross-modal Generative Adversarial Networks (GAN). Regarding the generation of videos, related work generates hand and finger movements of a keyboard player from an MIDI input [17] through inverse kinematics with appropriate constraints. All of the above-mentioned works, however, do not model musicians' creative body behavior in expressive music performances.

Given the original MIDI score, Widmer et al. [16] propose to predict three expressive dimensions (timing, dynamics, and articulations) on each note event using a Bayesian model trained on a corpus of human interpretations of piano performances. It further gives a comprehensive analysis of computer's creative ability in generating expressive music performances, and proves that certain aspects of personal styles are identifiable and even learnable from MIDI performances. Regarding to the expressive performance generation in visual modality, Shlizerman et al. [13] propose to generate expressive body skeleton movements and adapt them into textured characters for pianists and violinists. Different from our proposed work, they take the input of audio waveforms rather than MIDI performances. We argue that MIDI data is a more scalable format to carry context information, regardless of recording conditions and piano acoustic characteristics. And most of piano pieces have the sheet music in MIDI format, which can be aligned with a waveform recording.

[1] We do not generate lower body movements as they are often paid less attention by the audience.

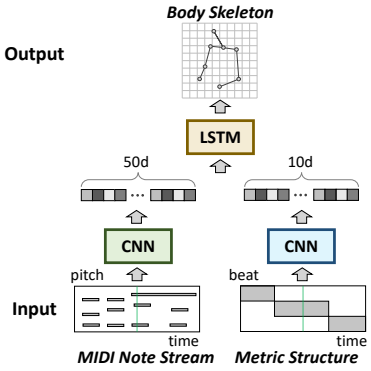


Figure 2. The proposed network structure.

3. METHOD

The goal of our method is to generate a time sequence of body joint coordinates, given a live data stream of note events from the performer's actions on the keyboard (MIDI note stream), and synchronized metric information. We seek to create the motion at 30 frames-per-second (FPS), a reasonable frame-rate to ensure a perceptually smooth motion. In this section, we introduce the technical details of the proposed method, including the network design and training conditions. We first use two CNN structures to parse the raw input of the MIDI note stream and the metric structure, and feed the extracted feature representations to an LSTM network to generate the body movements, as a sequence of upper-body joint coordinates forming a skeleton. The network structure is shown in Figure 2.

3.1 Feature Extraction by CNN

In contrast to traditional methods, our goal is to model expressive body movements that are associated with the keyboard performance. In this sense, the system should be aware of the general phrases and the metric structure in addition to each individual note event. Instead of designing hand-crafted features, we use CNNs to extract features from the raw input of the MIDI note stream and the metric structure, respectively.

3.1.1 MIDI Note Stream

We convert the MIDI note stream into a series of two-dimensional representations known as the *piano-roll matrix*, and for each of them extract a feature vector ϕ_x as the *piano-roll feature*.

To prepare the piano roll, the MIDI note stream input is sampled at 30 frames-per-second (FPS) to match the target frame rate. This quantizes the time resolution into the unit of 33 ms, as a video frame. Then for each time frame t we define a binary piano-roll matrix $\mathbf{X} \in \mathbb{R}^{128\times 2\tau}$, where element (m, n) is 1 if there is a key depression action at pitch m (in MIDI note number) and frame $t - \tau + n - 1$, and 0 otherwise. We set $\tau = 30$. The key depression timing is quantized to the closest unit boundary. Note that the sliding window covers both past τ frames and future $\tau - 1$ frames, and the note onset interval in $\mathbf{X}$ captures enough

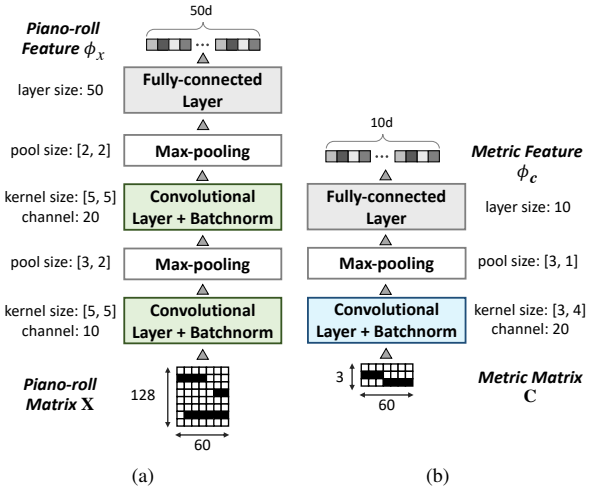


Figure 3. The CNN structures and parameters for feature extraction from the (a) MIDI note stream and (b) metric structure information.

information for motion generation to "schedule" its timing. Looking into the future is necessary for the generation of proper body movements, which is also true for human musicians: to express natural and expressive body movements, a human musician should either look ahead on the sheet music, or be acquainted with it beforehand. Later in Section 3.2 we will introduce in which cases we can avoid the potential delays in real-time applications.

We then use a CNN to extract features from the binary piano-roll matrix $\mathbf{X}$, as CNNs are capable of capturing local context information. The design of our CNN structure is illustrated in Figure 3.a. The input is the piano-roll matrix $\mathbf{X}$ and the output is a 50-d feature vector ϕ_x as the piano-roll feature. There are two convolutional layers followed by max-pooling layers, and we use leaky rectified linear units (ReLU) for activations. The kernel spans 5 semitones and 5 time steps, assuming that the whole body movement is not sensitive to detailed note occurrence. Overall, it is thought that in addition to generating expressive body movements, the MIDI note stream constrains the hand positions on the keyboard.

3.1.2 Metric Structure

Since the body movements are likely to correlate with the musical beats, we also input the metric structure to the proposed system to obtain another feature vector. This metric structure indexes beats within each measure, which is not encoded in the MIDI note stream. The metric structure can be obtained by aligning the live MIDI note stream with the corresponding symbolic music score with explicitly-annotated beat indices and downbeat positions.

Similar to the MIDI note stream feature, we sample them with the same FPS and window length, and, at each frame t, define the metric information as a binary *metric matrix* $\mathbf{C} \in \mathbb{R}^{M \times 2\tau}$, with $M = 3$. Here, element (m, n) is a one-hot encoding of the metric information at frame $t - \tau + n - 1$, where the three rows correspond to downbeats, pick-up beats, and other positions, respectively. We then build another CNN to parse the metric matrix $\mathbf{C}$ and obtain a 10-d output vector ϕ_c as the *metric feature*, as illustrated in Figure 3.b.

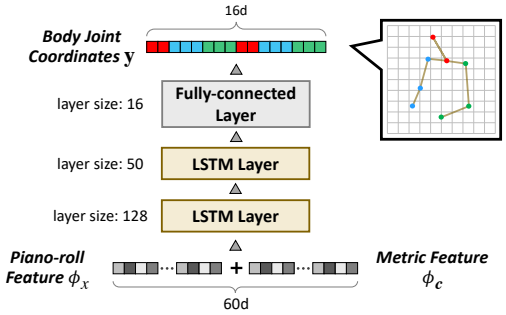


Figure 4. The LSTM network structure for body movement generation.

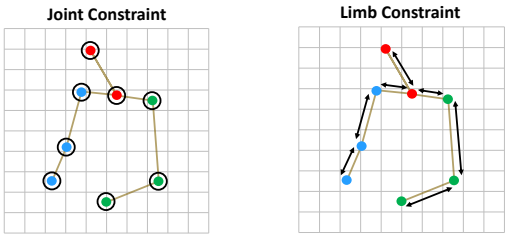
3.2 Skeleton Movement Generation by LSTM

To generate the skeleton sequence, we apply the LSTM network, which is capable of preserving the temporal coherence of the output skeleton sequence while learning the pose characteristics associated with the MIDI input. The input to the LSTM is a concatenation of the piano-roll feature ϕ_x and the metric feature ϕ_c, and the output is the normalized coordinates of the body joints $\mathbf{y}$. Since musical expression of a human pianist is mainly reflected through upper body movements, we model the x- and y- visual coordinates of K joints in the upper body as $\mathbf{y} = \langle y_1, y_2, \cdots, y_{2K} \rangle$, where K is 8 in this work, corresponding to nose, neck, both shoulders, both elbows, and both wrists. The first K indices denote the x-coordinates and the remaining denote the y-coordinates. Note that all the coordinate data in $\mathbf{y}$, for each piece, are shifted such that the average centroid is at the origin, and scaled isotropically such that the average variance along x- and y-axis sums to 1. The network structure is illustrated in Figure 4. It has two LSTM layers, and the output layer is fully-connected to get the 16-d vector approximating $\mathbf{y}$ for the current frame. The output skeleton coordinates are temporally smoothed using a 5-frame moving window. We denote the predicted body joint coordinates, given $\mathbf{X}$, $\mathbf{C}$ and network parameters θ, as $\hat{\mathbf{y}}(\mathbf{X}, \mathbf{C}|\theta)$.

Since the LSTM is unidirectional, the system is capable of generating motion data in an online manner, with a latency of 30 frames (i.e., 1 second). However, feeding the pre-existing reference music score (after aligned to the live MIDI note stream online) to the system enables an anticipation mechanism like human musicians, which makes it applicable in real-time scenarios without the delay.

3.3 Training Condition

To train the model, we minimize, over θ, the sum of a loss function $J(\mathbf{y}, \mathbf{C}, \mathbf{X}, \theta)$ evaluated over the entire training dataset. The loss function expresses a measure of discrepancy between the predicted body joint coordinates $\hat{\mathbf{y}}$ and



(a) (b)

Figure 5. The two constraints applied during training.

the ground-truth coordinates $\mathbf{y}$.

We use different loss functions during the course of training. In the first 30 epochs, we simply minimize the Manhattan distance between the estimated and the ground-truth body joint coordinates with weight decay:

$$J(\mathbf{y}, \mathbf{C}, \mathbf{X}, \theta) = \sum_k |\hat{y}_k(\mathbf{X}, \mathbf{C}|\theta) - y_k| + \beta\|\theta\|^2, \quad (1)$$

where k is the index for the body joints and $\beta = 10^{-8}$ is a weight parameter. We call this kind of loss the *body joint constraint* (see Figure 5.a). After 30 training epochs, we add another loss to ensures that not only the coordinates are correct but also consistent with the expected limb lengths:

$$J(\mathbf{y}, \mathbf{C}, \mathbf{X}, \theta) = \sum_k |\hat{y}_k(\mathbf{X}, \mathbf{C}|\theta) - y_k| + \sum_{(i,j)\in E} |\hat{z}_{ij}(\mathbf{X}, \mathbf{C}|\theta) - z_{ij}| + \beta\|\theta\|^2, \quad (2)$$

where $z_{ij} = (y_i - y_j) + (y_{K+i} - y_{K+j})$ is the displacement between two joints i and j on a limb (e.g., elbow-wrist), $E = \{(i, j)\}$ is the set of possible limb connections (i, j) of a human body. We call the added term the *body limb constraint* (see Figure 5.b). This is similar to the geometric constraint as described in [10]. There are 7 limb connections in total, given the 8 upper body joints. We then train another 120 epochs using the limb constraint. We use the Adam [7] optimizer, which is a stochastic gradient descent method, to minimize the loss function.

Here we propose to combine the two kinds of constraints in our training epochs. The body limb constraints are important because the loss of joint positions are minimized *independently of each other* in the body joint constraint. Figure 6 demonstrates several generated skeleton samples on the normalized plane, where the limb constraint is not applied in the following 120 epochs. Limb constraint adds dependencies between the loss among different joints, encouraging the model to learn a natural movement that considers the consistency of limb lengths. We only use this constraint at later epochs, however, because the body joint constraint is an easier optimization problem; if we optimize with body limb constraints from the very beginning, the training sometimes fails and remains a state of what seems a local optima, perhaps because the loss function wants to minimize the body joint errors but the gradient must pass through regions where the

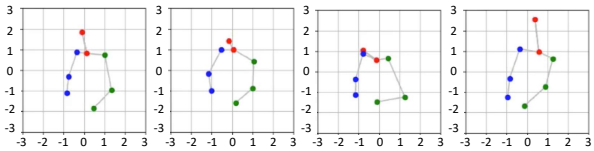


Figure 6. Several generated unnatural skeleton samples without the limb constraint.

limb constraint increases. In this case, the arrangements of the body joints tend to be arbitrary and not ergonomically reasonable.

4. EXPERIMENTS

We perform objective evaluations to measure the accuracy of the generated movements, and subjective evaluations to rate their expressiveness and naturalness.

4.1 Dataset

As there is no existing dataset for the proposed task, we recorded a new audio-visual piano performance dataset with synchronized MIDI stream information on a MIDI keyboard. The dataset contains a total of 74 performance recordings (3 hours and 8 minutes) of 16 different tracks (8 piano duets) played by two pianists, one male and one female. The two players were respectively assigned the primo and the secondo parts of 8 piano duets. Each player then played the 8 tracks multiple times (1-7 times) to render different expressive styles, e.g., normal, exaggerated, etc. At each time the primo and secondo are recorded together to ensure enough visual expressiveness on the players for interactions. The key depression information (pitch , timing, and velocity) is automatically encoded into the MIDI format by the MIDI keyboard. For each recording, the quantized beat number and the downbeat positions were annotated by semi-automatically aligning the MIDI stream and the corresponding MIDI score data. The camera was placed on the left-front side of the player and the perspective was fixed throughout all of the performances. The video frame rate was 30 FPS. The 2D skeleton coordinates were extracted from the video using a method based on OpenPose [2]. The video stream and the MIDI stream of each recording were manually time-shifted to align with the key depression actions. Note that we extract the 2D body skeleton data purely from computer vision techniques instead of capturing 3D data using motion sensors, which makes it possible to use the massive online video recordings of great pianists (e.g., Lang Lang) to train the system.

4.2 Objective Evaluations

We conduct two experiments to assess our method. Since there is no similar previous work to model the players' whole body pose from MIDI input, we set different experimental conditions for the proposed model as baselines and compare them. First, we investigate the effect of incorporating the metric structure information, which is likely to be relevant for expressive motion generation but does not directly affect the players' key depression actions on

the keyboard. Second, we compare the performance of the network when training on a specific player versus training on multiple players. To numerically evaluate the quality of the system output, we use the mean absolute error (MAE) between the generated and the ground-truth skeleton coordinates at each frame.

4.2.1 Effectiveness of the Metric Structure

The system takes as the inputs the MIDI note stream and the metric information. Here we investigate if the latter one can help in the motion generation process, by setting a baseline system that takes the MIDI note stream as the input, ignoring the metric structure by fixing ϕ_c to 0. We evaluate the MAE of the two models, using piece-wise leave-one-out testing over all the 16 tracks.

Results show that adding the metric structure information into the network can decrease the MAE from **0.180** to **0.173**. The unit is in the scale of the normalized plane, where the length of an arm-wrist limb is around 1.2 (see Figure 6). The result is significant because it not only demonstrates that our proposed method can effectively model the metric structure, but also that features that are not indirectly related to physical placement of the hand *does* have an effect on expressive body movements. Although our dataset for evaluation is small, we argue that overfit should not exist since the pieces are quite different.

On the other hand, we also observe that even without the metric structure information, the system output is still reasonable by learning the music context from the MIDI note stream. This setting broadens the use scenarios of the proposed system, such as when the MIDI note stream is from an improvised performance without corresponding metric structure information. Nevertheless, including a reference music score is beneficial for the system not only because it improves the MAE measure, but it also enables an anticipation mechanism to favor real-time generation without potential delays.

4.2.2 Training on A Specific Player

In this experiment, we evaluate the model's performance when fixing the same player for training and testing. Now the experiments are carried out on the two players separately. We first divide the dataset into two subsets, each obtaining the 8 different tracks performed by the two players respectively. On each subset we use the leave-one-out testing for the 8 tracks and calculate the MAE between the generated and ground-truth coordinates of body skeletons. The average of the MAE from the two subsets is **0.170**. Comparing the MAE of 0.173 in Section 4.2.1 and the MAE of 0.170 in this experiment, we see that training on a generic model only on a target player is slightly better than training over different players. This slight improvement may not be statistically significant. The marginal difference also suggests that even when trained on multiple players as in Section 4.2.1, the system is capable of remembering the motion characteristic of each player.

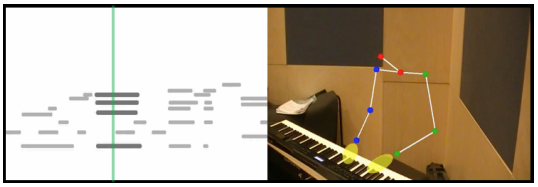


Figure 7. One sample frame of the assembled video for subjective evaluation.

4.3 Subjective Evaluation

Although the objective evaluation using MAE reflects the system's capability of reproducing the players' body movements on a new MIDI performance stream, this measure is still limited. There can be multiple creative ways on body motions to expressively interpret the same music, and the ground-truth body motion is just one possibility. In addition, from MAE we cannot infer the naturalness of the generated body movements, which is even more important than simply learning to reproduce the motion. In this section, we conduct subjective tests to evaluate the quality of the generated body movements, addressing both expressiveness and naturalness. The strategy is to mix the ground-truth body movements with the generated ones and let the testers to tell if each sample is real (ground-truth from human) or fake (generated).

4.3.1 Arrangements

In the subjective evaluation, we mix the two players together and cross-validate on the 16 tracks, as in Section 4.2.1. Here we do not add the metric structure input because positive feedbacks on the generation results purely from the keyboard actions will promise broader use cases of the system, i.e., improvised performance without a reference music score.

From the generated skeleton coordinates, we recover them to the original pixel positions on real video frames using the same scaling factor when normalizing the ground-truth skeleton before training. Then we generate an animation showing body joints as circles and limb connections as straight lines on the background environment image taken by the camera from the same perspective. In the same generated video, we also render a dynamic piano-roll that covers a rolling 5-second segment around the current time frame together with the synthesized audio. For a fair comparison, instead of using the original video recordings of real human performances, we generate human body skeletons by repeating the same process using the ground-truth skeletal data. Figure 7 shows one sample frame of the assembled video as a visualization.

We arrange 16 pairs of the generated and ground-truth skeleton motions on all the 16 tracks, and randomly crop a 10-second excerpt from each one (excluding several chunks containing long silence parts or page turning motions). This results in 32 video excerpts. We shuffle the 32 excerpts before showing them to subjects for evaluation.

We recruit 18 subjects from Yamaha employees, who are in their 20's to 50's, all with rich experience in musical

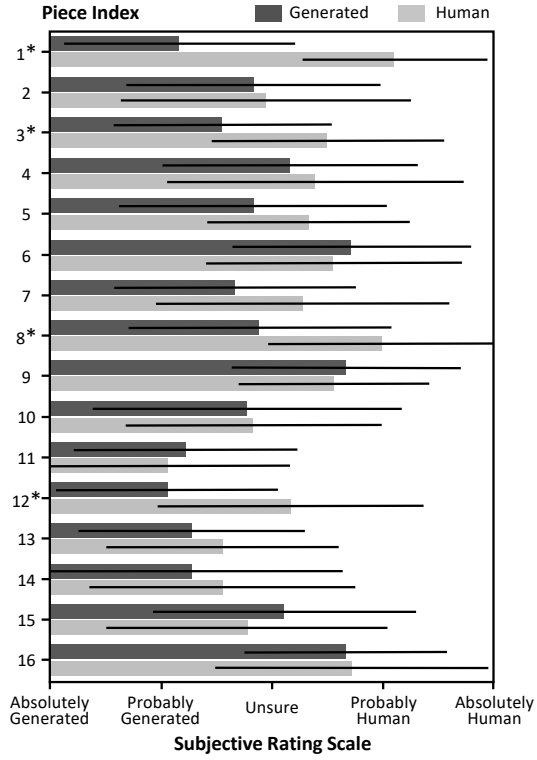


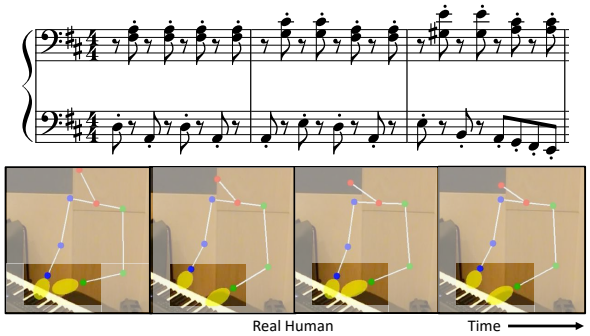
Figure 8. Subjective evaluation on expressiveness and naturalness of the generated and human skeleton performance videos. The tracks with significant different ratings are marked with "*".

acoustics or music audio signal processing. 17 subjects have instrument performance experiences (15 on keyboard instruments). This guarantees that most of them have a general knowledge of how a human pianist performance may look like based on a given MIDI stream, considering different factors such as hand positions on the keyboard according to pitch height, dominant motions for leading onsets, etc. Based on expressiveness and naturalness they rated the videos on a 5-point scale: absolutely generated (1), probably generated (2), unsure (3), probably human (4), and absolutely human (5).

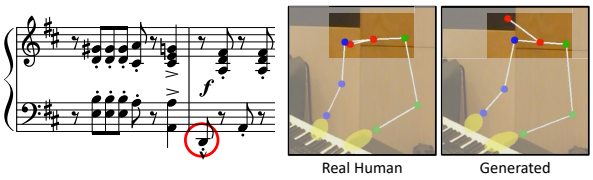
4.3.2 Results

Figure 8 shows the average subjective ratings as bar plots and their standard deviations as whiskers. A Wilcoxon signed rank test on each piece shows that no significant difference is found in 12 out of the 16 pairs ($p = 0.05$). This suggests that for $3/4$ of the observation videos, the generated body movements achieve the same level of expressiveness and naturalness as the real human videos.

In Figure 8, the pieces with significant differences in the subjective ratings between generated and real human videos are marked with "*". On the 1st piece, we observe an especially significant difference. Further investigation reveals that this piece is in a fast tempo (130 BPM), where the eighth notes are alternatively played by the right and left hand with an agile motion, as shown in Figure 9.a. The generated performance lacks this kind of dexterity. In addition, the physical body motions from the human players are distinct and exaggerated around the phrase boundaries, but the generated ones tend to create more conservative motions. Figure 9.b gives an example, where in the real human's performance the head moves forward extensively on the leading bass note (marked in red), whereas the generated one does not. Another observed drawback is the improper wrist positioning of a resting hand; a random position is often predicted in these cases. This is because the left/right hand information is not encoded in the MIDI file, and when only one hand is used, the system does not know which hand to use and how to position the other hand. Generally speaking, the generated movements that are rated significantly lower than real human movements tend to be somewhat dull, which might provide the subjects a cue to discriminate between human and generated movements. We present all of the generated videos online [2].



(a) The agile fashion in left-right hand alternative playing is not learned.



(b) The exaggerated head nodding on the leading bass note (in red mark) is not learned.

Figure 9. The two typical failure cases.

5. CONCLUSION

In this paper, we proposed a system for generating a skeleton sequence that corresponds to an input MIDI note stream. Thanks to data-driven learning between the MIDI note stream and the skeleton, the system is capable of generating natural playing motions like a human player with no explicit constraints on the physique or fingering, reflecting musical expressions, and attuning the generated motion to a particular performer.

For future work, we will apply more music contextual features to generate richer skeleton movements, and extend our method to the generation of 3D joint coordinates. Generating textured characters based on these skeletons is another future direction.

[2] http://www.ece.rochester.edu/projects/air/projects/skeletonpianist.html

6. ACKNOWLEDGEMENT

This work is partially supported by the National Science Foundation grant 1741472.

7. REFERENCES

[1] Christoph Bregler, Michele Covell, and Malcolm Slaney. Video rewrite: Driving visual speech with audio. In *Proceedings of the ACM Conference on Computer Graphics and Interactive Techniques*, pages 353–360, 1997.

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *Proceedings of the International Conference on Conputer Vision and Pattern Recognition (CVPR)*, 2017.

[3] Lele Chen, Sudhanshu Srivastava, Zhiyao Duan, and Chenliang Xu. Deep cross-modal audio-visual generation. In *Proceedings of the ACM International Conference on Multimedia Thematic Workshops*, pages 349–357, 2017.

[4] Sofia Dahl and Anders Friberg. Visual perception of expressiveness in musicians body movements. *Music Perception: An Interdisciplinary Journal*, 24(5):433–454, 2007.

[5] Jane W Davidson. Visual perception of performance manner in the movements of solo musicians. *Psychology of Music*, 21(2):103–113, 1993.

[6] Sefik Emre Eskimez, Ross K Maddox, Chenliang Xu, and Zhiyao Duan. Generating talking face landmarks from speech. In *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation (LVA-ICA)*, 2018.

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pages 1–5, 2015.

[8] Bochen Li, Karthik Dinesh, Zhiyao Duan, and Gaurav Sharma. See and listen: Score-informed association of sound tracks to players in chamber music performance videos. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017.

[9] Zimo Li, Yi Zhou, Shuangjiu Xiao, Chong He, and Hao Li. Auto-conditioned recurrent networks for extended complex human motion synthesis. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[10] Guanghan Ning, Zhi Zhang, and Zhiquan He. Knowledge-guided deep fractal neural networks for human pose estimation. *IEEE Transactions on Multimedia*, 20(5):1246–1259, 2017.

[11] Ken Perlin and Athomas Goldberg. Improv: A system for scripting interactive actors in virtual worlds. In *Proceedings of the ACM Annual Conference on Computer Graphics and Interactive Techniques*, pages 205–216, 1996.

[12] Ju-Hwan Seo, Jeong-Yean Yang, Jaewoo Kim, and Dong-Soo Kwon. Autonomous humanoid robot dance generation system based on real-time music input. In *Proceedings of the IEEE International Conference on Robot and Human Interactive Communication*, pages 204–209, 2013.

[13] Eli Shlizerman, Lucio Dery, Hayden Schoen, and Ira Kemelmacher-Shlizerman. Audio to body dynamics. 2017. Available: `https://arxiv.org/pdf/1712.09382.pdf`.

[14] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 36(4), 2017.

[15] Chia-Jung Tsay. The vision heuristic: Judging music ensembles by sight alone. *Organizational Behavior and Human Decision Processes*, 124(1):24–33, 2014.

[16] Gerhard Widmer, Sebastian Flossmann, and Maarten Grachten. YQX plays chopin. *AI magazine*, 30(3):35–48, 2009.

[17] Kazuki Yamamoto, Etsuko Ueda, Tsuyoshi Suenaga, Kentaro Takemura, Jun Takamatsu, and Tsukasa Ogasawara. Generating natural hand motion in playing a piano. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3513–3518, 2010.