Vol.2018-MUS-118 No.10
Vol.2018-SLP-120 No.10
2018/2/20

MIDI2Pose: 鍵盤演奏情報を用いたオンライン演奏動作生成

Bochen Li　前澤 陽[1,a)]

概要：本稿では，人間が演奏した電子鍵盤楽器の押鍵情報に対して，それを演奏した際の関節座標系列を，オンラインで生成する手法について述べる．従来手法では，特定の演奏者に対する動作のクセを獲得することはできず，また，手指運動に直接関係のない全身の運動は獲得できなかった．そこで我々は，押鍵情報と拍節構造の時系列と，任意の関節位置の時系列の対応付けを学習する手法 MIDI2Pose を提案する．評価実験の結果，本手法は学習データに含まれる演奏者の動作を，そうでない演奏者の動作よりも 35%小さな誤差で推定できることが示され，特定奏者に対する演奏動作が学習できることが示された．また，被験者実験では 75%の楽曲において，人間の動作と生成された動作の間に有意差は見られず，提案手法は極端に不自然な動作を生成しないことが示唆された．

MIDI2Pose: Online keyboard performance motion generation from performance data

Bochen Li　Akira Maezawa[1,a)]

1. はじめに

楽器を演奏する任意の音源や演奏データに対して，適切な演奏者の動作シーケンスを生成することは重要である．例えば音楽鑑賞においては，音楽表現を伝達する上で演奏動作は重要である [1, 2]．よって，楽曲の演奏データに対して適切な演奏動作を生成することで，より没入感の高い音楽鑑賞が実現できると考えられる．また，合奏における主従関係を予測する場合，視覚情報も有効であること [3]．よって，演奏動作を計算機が生成できるようになることで，計算機による伴奏システムと人間がより適切に連携できると考えられる．そこで本稿では，演奏シーケンスに対して，適切な骨格の動作シーケンスをオンラインで生成することを考える．

このような用途における演奏動作生成では，3 つの要求仕様がある．第一に，楽曲に合わせて，全身の骨格動作を生成する必要がある．第二に，骨格動作は特定の演奏者が行う動作を模する必要がある．第三に，演奏されている楽曲に含まれる音楽的な文脈に適合した動作を生成する必要がある．従来手法では，動作生成を，運指から定まる手指位置と，身体性に関する制約を用いた逆運動学問題として定式化されていたため，(1) 肩から指先の動作のみを生成対象としているため全身の動作が生成できず，(2) 任意の演奏者に対する動作の特徴を反映させるのが困難であり，(3) 手指位置に反映されないような音楽的な制約が統合できないという問題があった．これらの問題に対処するためには，音楽的な文脈情報と，手指位置の制約となる押鍵位置情報を統合できる必要がある．また，演奏者個人の特性を，すべての関節動作に反映させる必要がある．

そこで，我々は**図 1** に概要を示すような演奏動作生成手法 Pose2MIDI を提案する．Pose2MIDI は任意の楽曲データを演奏しているときのリアルタイムの押鍵情報及び拍節構造を入力とし，対応する骨格動作系列をオンラインで出力するようなモデルである．特定演奏者の演奏データから

[1] ヤマハ株式会社
Yamaha Corporation, Iwata, Shizuoka 438–0942, Japan
[2] ロチェスター大学
University of Rochester
[a)] akira.maezawa@music.yamaha.com

Vol.2018-MUS-118 No.10
Vol.2018-SLP-120 No.10
2018/2/20

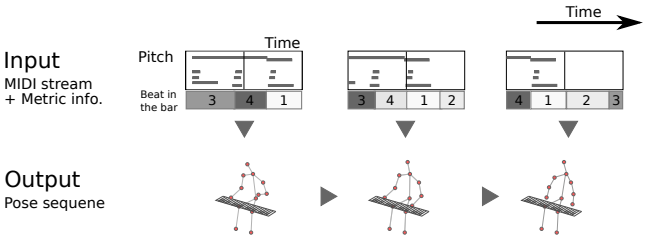

図 1 本手法の概要．鍵盤演奏の押鍵情報時系列から，骨格位置の時系列を生成する．

モデルを学習することで，任意の演奏者における動作の特性を，任意の関節に対して反映させることができる．また，拍節構造といった手指運動には明示的に反映されない要素と，押鍵情報といった手指運動に反映される要素を統合することが可能になる．

2. 関連研究

従来，演奏に対する骨格情報の動作系列生成は，手指の座標情報を制約とした逆運動学問題として定式化されている．逆運動学に適切な制約を設けることで，自然な動作を生成したり [4]，パーソナライズされた動作を生成できる [5]．しかし，逆運動学アプローチには 3 つの大きな課題がある．第一に，特定の個人に対する動作の特性は，制約の設計や手動のパラメータチューニングなどにより実現されていたため，特定の個人の動作を獲得することが困難であること．第二に，生成される動作は肩から指先までのみであり，頭部や上体の傾きなどはモデル化されていないこと．第三に，手指位置には直接関係しない，拍節構造といった音楽的な文脈情報を取り入れることができないこと．演奏動作とは音楽的な文脈にも影響されるため [1, 2]，このような文脈情報を取り入れることが好ましい．音楽の文脈を踏まえて適切な押鍵情報を生成する問題設定としては演奏表情付けがあるが [6]，演奏表情付けでは動作生成を対象としていなかった．

3. 手法

本手法では，人間が演奏したピアノ演奏の押鍵情報と小節線から経過した拍数のストリームを入力とし，入力に同期した人間の骨格座標のストリームを一定の遅延を経てから出力する．従来の手指運動の生成手法とは対照的に，細かな手指運動自体はモデル化しない代わりに，楽曲に合った，大まかな全身の演奏動作を生成することを目標とする．

骨格座標としては，ピアノ演奏において重要と思われる，頭部・首・両肩・両肘・両手首の 8 関節の座標をモデル化する．座標は単一のアングルで撮影されたピアノ演奏動画に対する二次元座標とする．以後座標インデックスを $d \in \{1, 2 = D\}$ とし，関節のインデックスを $k \in \{1, \cdots, 8 = K\}$ とする．入力には MIDI の発音司令から得られるノート番号とベロシティを用いる．動作の更新は一定の周期 ΔT で行われ，τ フレームの遅延が生じるものとする．なぜならば演奏では予備動作が含まれるため，押鍵情報を遡って動作を生成する必要があるためだ．

本手法では，局所的な演奏情報から演奏を特徴付けるような低次元データ（「演奏特徴量」と呼ぶ）と，拍節構造を要約したような低次元データ（「拍節構造特徴量」と呼ぶ）を抽出し，これらの特徴量の時系列に基づいて骨格座標系列を生成する．特徴量の手動設計は困難であることと，適切な骨格時系列のモデル化が困難であることから，**図 2** に示すようなニューラルネットワークを用いて，データドリブンに特徴抽出や時系列モデル化を行うことを考える．

3.1 CNN による演奏特徴量抽出

演奏特徴量を抽出するために，ピアノ演奏のストリームから，周期 ΔT でピアノロール $X_{t,n}$ を算出する．ピアノロールとは，時刻 $t\Delta T$ で音高 n が演奏されていた時に $X_{t,n} = 1$ となるようなデータである．次に，各フレーム t において，フレーム $t - 2\tau$ から t までのピアノロールを $2\tau \times N$ 次元の二次元画像と見なし，二層の CNN と全結合層の順で通すことで，フレーム t における，50 次元の演奏特徴量を得る．

演奏特徴量には，現在時刻周辺における手指位置を示していると考えられる．なぜならば，CNN は動作生成において重要な局所的なフレーズと，その発生位置をモデル化するからだ．

3.1.1 CNN による拍節構造特徴量

拍節構造特徴量を抽出するため，現在時刻周辺での拍節構造を低次元ベクトルで表すことを考える．そこで，各フレーム t に対して，そのフレームが小節上の何拍目を弾いているかを求め，1 拍目の場合 1 番目の要素，小節線の 1 拍前の場合 2 番目の要素，それ以外の場合は 3 番目の要素が 1 となり，それ以外が 0 となるような 3 次元のベクトル c_t を算出する．次に，各フレーム t において，フレーム $t - 2\tau$ から t までのベクトルを纏めたものを $2\tau \times 3$ 次元の二次元画像と見なし，CNN と全結合層を経ることで，フレーム t における，10 次元の拍節構造特徴量を得る．

3.2 LSTM による骨格動作生成

骨格動作の生成のため，演奏特徴量と拍節構造特徴量を入力とした時系列モデルを考える．動作においては骨格位置における時間軸上での連続性が重要であるため，これらの特徴量を入力とした 2 層の LSTM を構築する．LSTM の出力ベクトルを全結合層に与える，フレーム t における関節 k の座標 d の成分 $y_{t,k,d}$ を得る．

このように各フレームにおいて，長さ 2τ のピアノロール $\boldsymbol{x}$ 及び拍節情報 $\boldsymbol{c}$ から，関節座標 y を出力するネットワークを　$y(\boldsymbol{x}, \boldsymbol{c}|\theta)$ と表す．ここで，θ はネットワークのパラメータである．

Vol.2018-MUS-118 No.10
Vol.2018-SLP-120 No.10
2018/2/20

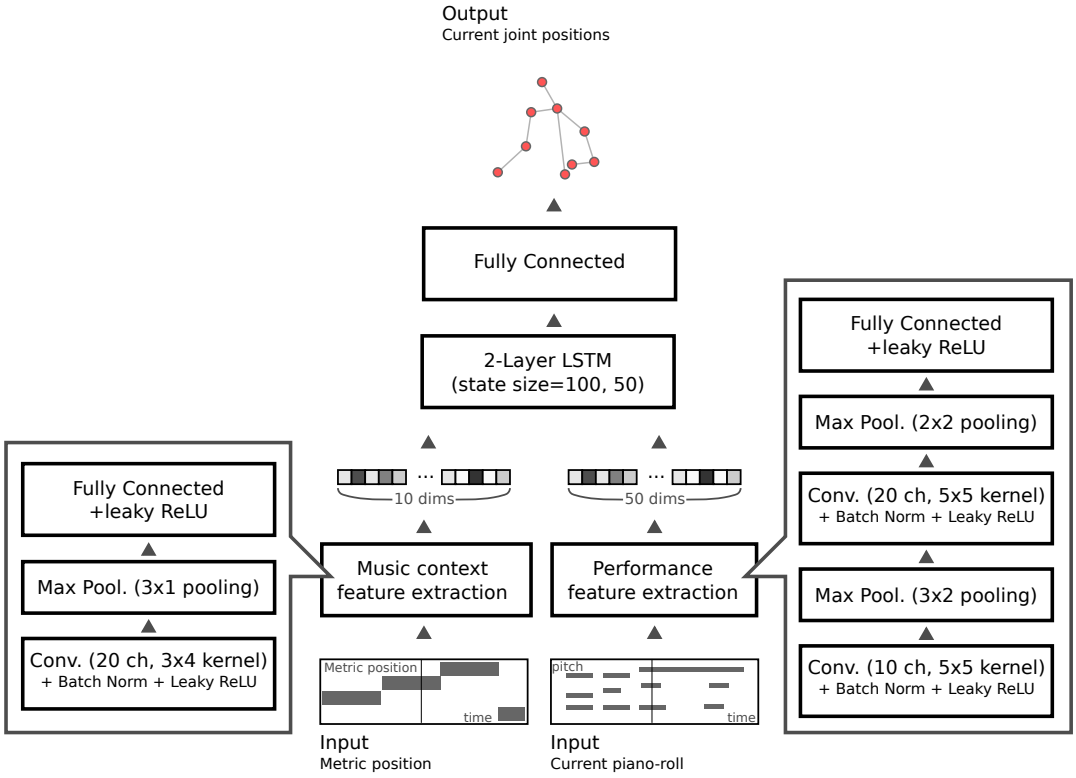

図 2　ネットワークの構成図.

3.3　位置誤差と骨格誤差の最小化に基づくパラメータ推定

ネットワークのパラメータ θ を得るため，教師データである押鍵情報から算出されたピアノロールのデータ $\boldsymbol{X} = \{\boldsymbol{x}_t \in \mathcal{R}^{2\tau\times N}\}$，拍節情報 $\boldsymbol{C} = \{\boldsymbol{c}_t \in \mathcal{R}^{2\tau\times 3}\}$ と対応する骨格情報 $\boldsymbol{Y} = \{\boldsymbol{y}_t \in \mathcal{R}^{K\times D}\}$ の対応関係を学習する．学習においては，関節位置の誤差を表す目的関数 $J_p(\boldsymbol{y}, \boldsymbol{c}, \boldsymbol{x}, \theta)$ と，パラメータに対する罰則を加算した，次のような目的関数を最小化することを考える：

$$J(\theta) = \sum_t J(\boldsymbol{y}_t, \boldsymbol{c}_t, \boldsymbol{x}_t, \theta) + \beta|\theta|^2. \qquad (1)$$

ただし β は重み付け係数である．ここで，関節位置の誤差を表す目的関数 J_t は次のように表す:

$$J_t(\boldsymbol{y}, \boldsymbol{c}, \boldsymbol{x}, \theta) = \sum_n |y(\boldsymbol{x}, \boldsymbol{c}|\theta) - \boldsymbol{y}|. \qquad (2)$$

本稿では目的関数の最小化には，確率的勾配降下法の一種である ADAM [7] を用いる．

4.　評価実験

本手法を評価するため，生成された動作の精度と自然さを評価した．そこで，手指の位置制約に直接寄与しない要素の有効性を検証し（実験 1），データドリブンに学習することで特定の個人の動作における特徴が獲得できるかを検証し（実験 2），得られた生成結果と実演奏の動作の自然さを主観評価を通じて比較した（実験 3）．

以降の実験においては，ADAM を 15 エポック実行したのち，評価データでの誤差が最小となるモデルを選択した．

4.1　データセット

ピアノ演奏に対する押鍵情報と，演奏に同期した関節位置系列のデータセットを用意した．関節位置は，定位置から撮影されたカメラから見たときの二次元座標を用いた．データセットでは，男女一名ずつ，計 2 名のピアニストがそれぞれ異なる 8 曲を演奏し（合計 16 曲），各曲はそれぞれ 3 から 5 テイクずつ，最低でも 3 種類のニュアンスで演奏するよう指示した上で収録された．また，正解データの関節位置系列は 1 テイクを通して平均 0，標準偏差 1 となるよう正規化した．

4.2　実験 1 - 拍節構造の有効性

この実験では，手指位置に直接関係のある押鍵情報に加え，手指位置とは無関係だが音楽的に重要な拍節構造を併用することに対する効果を検証した．

4.2.1　実験条件

まず，1 曲を除く全ての楽曲データを用いて 2 種類のモデルの学習を行った．具体的には，提案手法に加え，提案手法の LSTM に拍節構造特徴量を入力しないものが学習された．すなわち，前者では押鍵情報と拍節構造を考慮するのに対して，後者では押鍵情報のみを考慮して骨格を生成するよう学習される．次に，学習に用いられなかった楽曲に対して，それぞれのモデルで骨格データを生成した．これら 2 パターンの生成骨格データと正解骨格データとの平均絶対誤差（Mean Absolute Error; MAE）を評価した．

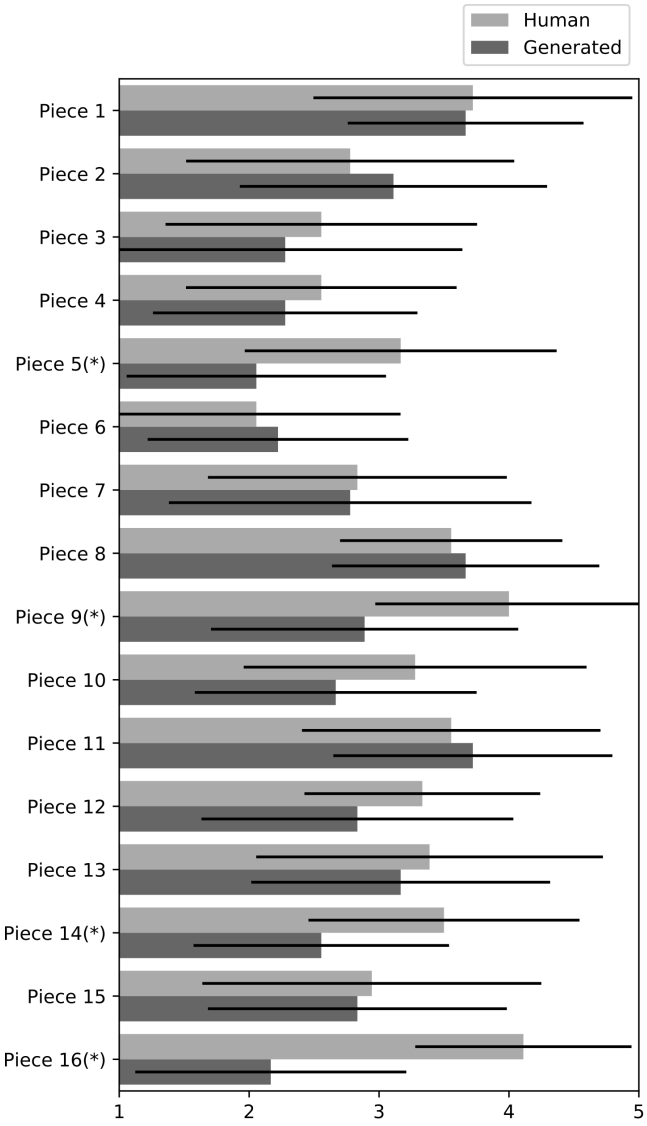

図 3 主観評価の平均値（バーは標準偏差）．「Human」は実演奏データ，「Generated」は生成されたデータに対する主観評価．楽曲名の隣にアスタリスクがあるものは Wilcoxon Signed Rank Test で有意差が見られたもの（p=0.05）．

4.2.2 実験結果と考察

拍節構造特徴量を用いない場合，MAE は 0.180 であった．また，拍節構造特徴量を用いた場合，MAE は 0.173 であった．このことから，拍節構造をモデル導入することが有効であることが示される．

拍節構造は手指位置とは無関係であるが，音楽的には重要な特徴である．よって，演奏動作生成においては，身体制約を表す要素だけでなく，音楽表現上重要な特徴を併用することが重要であることが示唆される．

4.3 実験 2 - 個人の演奏動作の獲得

この実験では，学習データに含まれる特定個人の演奏動作に特化したモデルを，本手法は学習できるかを検証した．

4.3.1 実験条件

本学習データを収録した 2 名の演奏者 A と B に対し，まず，A が演奏した楽曲のうち，1 曲を除く全楽曲を用いて学習を行った．次に，A が演奏した楽曲のうち，学習に用いられなかった楽曲から生成された関節座標系列と，正解の動作系列との MAE を評価した．次に，A の動作を用いて学習された同モデルを用いて，演奏者 B が演奏した楽曲から生成された関節座標系列と，正解となる B の関節座標系列との MAE を評価した．

次に，A と B を入れ替えて上記の実験を行った．つまり，B の演奏で学習したモデルに対して，B の演奏と A の演奏を入力したときの関節座標系列を生成し，それぞれ正解となる B と A と関節座標系列との間の MAE を評価した．

4.3.2 実験結果と考察

同一の演奏者により学習と評価を行う場合，2 名の MAE に対する平均は 0.170 であり，学習と評価に使う演奏者が異なる場合，2 名の MAE に対する平均は 0.269 であった．学習と評価に用いられる演奏者が同一の場合は，同一でない場合と比べて誤差が小さいことから，本手法では，特定の演奏者における動作上の特性もしくは身体上の特性を，適切に獲得できていることが示された．

この結果からは，このような誤差の違いが骨格の違いといった身体的な特徴に依るものか，動作による演奏表現の相違によるものかは断定できない．特に，演奏者の性別が異なることから体格もやや異なるため，骨格の違いによる関節座標の違いは大きな要因の一つとなる．とはいえ，演奏者の動作をデータドリブンに獲得できることが示されており，明示的に順運動学を記載することなく，特定の個人の特性に適合できることが分かる．

4.4 実験 3 - 生成された動作の主観評価

この実験では，正解データの骨格位置系列と生成された骨格位置系列の自然さを比較した．

4.4.1 実験条件

まず，データセットに含まれる各曲に対する演奏データから，ランダムに選定された 15 秒の演奏データを抽出した．次に，それぞれの演奏データに対して (1) 正解データから得られた骨格位置と (2) 演奏データに対して，MIDI2Pose により骨格位置を生成したもののそれぞれをアニメーションとして合成した．アニメーションは，**図 4** に示すように，生成された関節座標位置を直線で結んだ，スケルトン状のデータを，学習データを収録したカメラアングルで映したピアノ画像の上にオーバーレイした．なお，ここでは拍節構造特徴量は用いなかった．また，特定の楽曲に対する動作を生成する際は，その楽曲以外の全演奏データで学習を行った．

このような合計 32 パターンのアニメーションをランダムな順番で被験者に提示した（被験者数 18）．各アニメーションに合わせて，アニメーションに対応する演奏データ

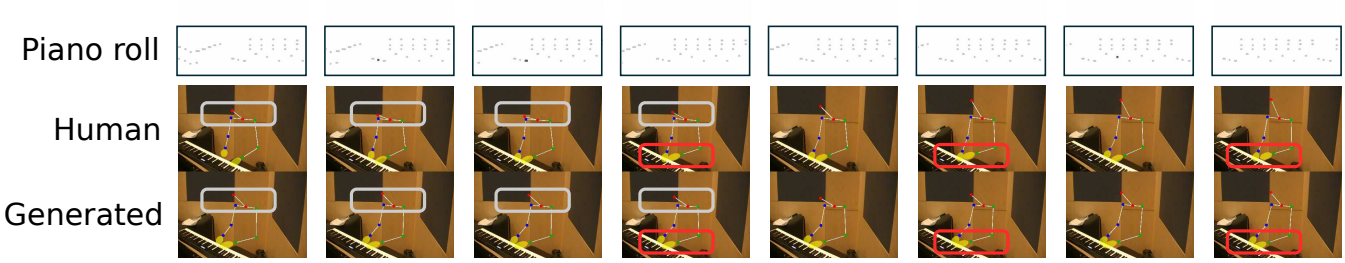

図 4　特に評価結果が低かった生成結果の例（Piece 16）．手首の関節位置は実演奏に近いものの，実演奏の方が手首が俊敏に動いたり（赤色の枠），フレーズ境界における頭部の動きがオーバーである（灰色の枠）．

を一定のベロシティ（打鍵強度）で演奏された音楽音響信号を同時に提示した．被験者は動画の自然さを「明らかに生成データ」「おそらく生成データ」「分からない」「おそらく実演奏データ」「明らかに実演奏データ」の 5 段階で評価した．被験者は 20 代から 50 代で，17 人が楽器演奏経験者であった（ピアノ演奏経験者 15 名）．

4.4.2　実験結果と考察

各楽曲に対する正解データと生成データの平均評価値とその標準偏差を図 **3** に示す．各曲に対して Wilcoxon Signed Rank Test を行ったところ，16 曲中 12 曲では有意差が見られなかった（$p = 0.05$）．このことから，生成データと実演奏データの違いは著しくは違うものとして認識はされないことが示唆される．

特に評価の低かった Piece 16 の生成結果と実演奏データを図 4 に示す．このデータは約 130 BPM で八分音符を右手と左手で交互に演奏しており，実演奏データには俊敏な手首の動きが見られる．一方生成結果ではこのような特徴が現れていない．また，フレーズの境界で実演奏データでは体勢が大きく前のめりになるのに対して，生成結果ではこのような特徴が得られておらず，平坦な動作になっている．他にも特に統計的有意差があったデータでは，生成データの動作が鈍っていることが多かった．そのため，より俊敏な動作を生成させるために，より多くの文脈情報を併用する必要があると考えられる．

5.　おわりに

本稿では，入力された押鍵情報に対応する骨格位置系列を生成する手法 MIDI2Pose を提案した．押鍵情報と骨格位置の対応付をデータドリブンに学習することで，手指位置の制約や逆運動学といった事前情報を使わずに，多くの曲で自然な演奏動作が生成されることが確認された．また，拍節構造といった，手指の制約に関連しないが音楽上重要な要素を併用することでより精度の高い演奏動作生成が可能になることが示された．本手法により音楽的な表現を反映でき，任意の演奏者を模倣できる演奏動作生成が可能になる．また，オンライン生成を行うことで，こういった自然な動作生成を，計算機による合奏システムといった実時間音楽システムにも統合できるようになる．

今後の課題としては，より多くの音楽的な文脈情報の活用，より多くの関節情報の推定，三次元関節座標の推定，推定結果に基づく CG 合成などが挙げられる．

参考文献

[1] Jane W. Davidson. Visual Perception of Performance Manner in the Movements of Solo Musicians. *Psychology of Music*, 21(2):103–113, 1993.
[2] Sofia Dahl and Anders Friberg. Visual Perception of Expressiveness in Musicians' Body Movements. *Music Perception: An Interdisciplinary Journal*, 24(5):433–454, 2007.
[3] Chia Jung Tsay. The vision heuristic: Judging music ensembles by sight alone. *Organizational Behavior and Human Decision Processes*, 124(1):24–33, 2014.
[4] 山本 和樹, 上田 悦子, 末永 剛, 竹村 憲太郎, 高松 淳, and 小笠原 司. ピアノ演奏における自然な手指動作 CG の自動生成. **日本バーチャルリアリティ学会論文誌**, 15(3):495–502, 2010.
[5] 高井 康太, 千葉 広大, 藤村 武史, 平田 純也, 合田 竜志, 巳波 弘佳, and 長田 典子. ピアノ演奏 CG アニメーションの自動生成：演奏モーションのヒューマナイズと GPU レンダリング (学生研究発表会). **映像情報メディア学会技術報告**, 35.8:73–76, 2011.
[6] Gerhard Widmer, Sebastian Flossmann, and Maarten Grachten. YQX Plays Chopin. *AI Magazine*, 30(3):35, 2009.
[7] Diederik P. Kingma and Jimmy Lei Ba. Adam: a Method for Stochastic Optimization. *International Conference on Learning Representations 2015*, pages 1–15, 2015.