創作ノート

姿勢推定を用いたサウンドインタラクションの検討
A study on Human Pose Estimation Based Sound Interaction

公文 太一, 小坂 直敏
Taichi KUMON, Naotoshi OSAKA
東京電機大学
Tokyo Denki University

概要

本研究では、単一カメラによる舞踊動画に対して、機械学習による姿勢推定を用いて、サウンドインタラクションを行うシステムを検討する。本システムでは、機械学習の実行環境である RunwayML を用いて、人間の動きに対してリアルタイムに姿勢推定を行い、頭や四肢の位置情報を取得している。得られた情報に基づいて、音楽合成において発音、 デチューン、LPF(Low Pass Filter)、FM(Frequency Moderation)、AM(Amplitude Moderation) などを変化させている。また姿勢推定や、音響の情報に基づいた映像生成も行う。人間の舞踊と機械学習による姿勢推定を用いることにより、創造的なサウンドインタラクションや AI の活用を実現する。

本研究は、ステージパフォーマンスやメディアアート、オンラインでのパフォーマンスなどへの応用が考えられる。

1. はじめに

1.1. 身体とサウンドインタラクション

音楽と舞踊の身体動作は、時間軸上で進行していく芸術表現という点において、密接に関連している。コンピュータや、各種センサデバイスを用いて音楽と身体を関連付けた作品は多数制作されている。これらの作品で、身体動作の計測に利用されてきたデバイスとして、身体に直接取り付けるモーションキャプチャ型のセンサや、kinect など非接触型のセンサがある。

ここでは、2 次元の舞踊動画に対して機械学習により、姿勢推定を行うことによるサウンドインタラクションを検討する。この方法では、単純な 2 次元画像から身体姿勢を得ることができる。これにより、一般的なカメラを利用して、サウンドインタラクションが可能となる。この方法では、深度センサの測定距離の問題や、パフォーマがセンサを身に着ける際の物理的な制約が無く、より幅広い舞踊表現やサウンドインタラクションが可能となる。

また、一般的なカメラ入力を姿勢推定に用いることにより、コストが下がることや、設置の容易さ、姿勢推定を簡単に行えることなどから、2020 年以降急速に拡大しているオンラインでのパフォーマンスでも応用ができる。

身体を用いた作品では身体動作と音の直感的対応が重要である。ここでは、身体動作に基づいた音色の表現を実現するため、周波数領域で変化を中心とした、サウンドインタラクションの検討を行う。

1.2. 関連研究・作品

身体動作と音楽を関連付けた作品は過去にも多数制作されている。*L'homme transcende*(後藤, 2009) はダンサーがセンサを搭載したボディスーツを装着することにより音響表現を行っている。また、*Hypnoid* (後藤, 2015) は kinect を用いた、身体とコンピュータグラフィックス、音響のインタラクティブな音楽作品となっている。

morphecore(Rhizomatiks, 2020) は舞踊のキャプチャデータに基づいて動く CG オブジェクトに対して脳の情報のデコーディング技術を用いて変化を加えることにより、CG 空間上で制約のない身体表現を実現した舞踊作品である。

AI と音楽や人間の関連に基づいた作品として *Duet with YOO*（Yamaha 博報堂 I-Studio, 2018）がある。この作品では AI によって人間のピアノ演奏を解析し合奏を行う、パーティクルによる人型のオブジェクトが演奏を模した動きをすることなどにより、AI の演奏参加を映像表現によって伝える作品である。

1.3. 目的

本研究では、機械学習による姿勢推定を用いて身体、音響、映像のインタラクションを行うシステムの検討

を目的とする。音響においては、舞踊の感情表現と対応した音響合成し、直感に合うサウンドインタラクションを目指す。

2. システムの構成

本システムでは、主に 3 種類のシステムを使用し全体としては音響と映像、身体のインタラクションを行う。システムの構成と相関図を図 1 に示す。

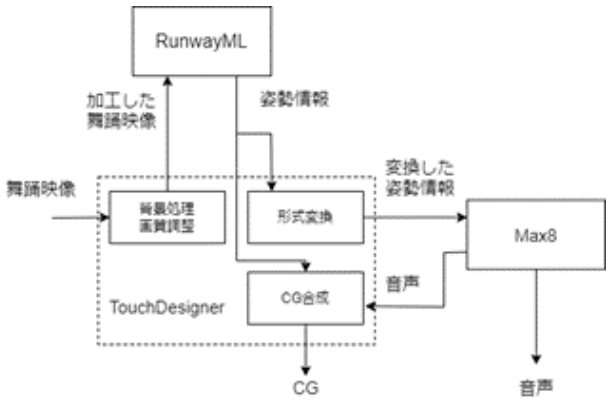

図 1: システムの構成

2.1. 姿勢推定

姿勢推定には RunwayML を用いる。RunwayML は RunwayAI,inc が開発を行っている機械学習ツールで簡単に機械学習ライブラリを利用できるシステムである。

ここでは TensorFlow で構築された姿勢推定モデルである Posenet を実行するために利用している。Posenet の実行には軽量な CNN (Convolutional neural network) である MobileNetV2 が使用されている。このモデルを利用ことにより実時間での姿勢推定を行うことができる。

Posenet では体の主要関節や目、鼻など 17 箇所の位置を 2 次元のデータとして推論する（図 2)。 実行結果は json 形式で OSC（Open Sound Control） (wright , 1997) による出力によりデータ処理のシステムに送信される。

2.2. 映像処理、姿勢情報処理、映像生成

TouchDesigner によって構築されたシステムでは、以下の機能を実行している。

- 舞踊映像の入力を受けつけ、それを RunwayML による姿勢推定が安定するよう加工し送信
- RunwayML から送信された姿勢情報を Max8 で扱いやすいよう形式変換を行い Max8 に送信
- 姿勢情報、音声情報に基づいた映像の生成

本システムでは録画済みの映像、カメラによる入力両方に対応している。どちらを使用するかの切り替えや、必要に応じての拡大や差分画像を利用した簡易な背景処理を施した映像を RunwayML に送信している。

json 形式で受信する姿勢情報を、適切な形式に変換し、同システム内の映像処理部分と音響処理用の Max8 に送信している。Max8 へのデータ送信には OSC を用いている。

映像生成部では、検出した各体の部位からパーティクルが噴射している。また周波数情報に基づいて地面のメッシュを合成している。また信号波形を可視化している。

2.3. 音響処理

音響処理には Max8 を用いている。本システムの Max パッチは大きく 2 つに分かれており、1 つは TouchDesigner のシステムによって処理された姿勢情報を OSC を用いて受信するパッチである。もう一つは、実際に音響合成を行うパッチである。

図 2: Posenet で取得されるポイント (TensorFlow, 2018) より転載

3. サウンドインタラクションの実装

3.1. インタラクションの方針

下道らの研究 (下道,2019) によると、自動で音楽が進行していく状況で、身体動作によって和音の転回形

を変化させるシステムと音楽が進行していく状況で身体動作によって音色を変化させるシステムが体験者の評価が良かったと報告されている。

この研究を参考に、ここでは、推定した姿勢情報を、周波数領域での変化に適用することを目指した。周波数領域での変化に関連する音響合成にはフィルタ、FM、AM などが考えられる。今回はフィルタ、FM のほかに基準となる周波数に対して、同様の波形で僅かに異なる周波数の信号を重ねるデチューンと呼ばれる手法を使用した。

身体動作との対応として、舞踊の振りの大きさと、身体にかかる負荷を音色により表現することを目指している。FM、フィルタ、デチューンは舞踊の振りと関連していて、振りが大きくなると、より派手な音色に変化する。

体の負荷として重心のずれを考えた。体に負荷がかかる場合に警告音のような印象の発振器が発音することにより負荷を表現している。以上の機能を実装することにより、舞踊の感情表現と対応して直感にあったサウンドインタラクションを実現している。

3.2. 使用する身体情報

今回使用している RunwayML 内で実行されている姿勢検出に使用している Posenet では、17 の身体部位の位置情報を推論している。本研究では、音響の生成には、鼻（頭）、両手首、左腰、右腰の平均を使った腰、両足の位置情報を利用している。

3.3. 実装

以下、Max ８内で実装しているサウンドインタラクションシステムの実装について解説する。

RunwayML による推論結果は 30fps で送信される。このデータを音響表現のパラメータとして利用する場合変化が断続的になるため、フレームごとの値を線形補間している。

断続的発音にされる高周波の正弦波発振器は体の傾きによって制御されている。体の傾きの検出には、頭と腰の x 座標の差を使用している。頭と腰の x 座標の差が 0.1 以上になると発振器が発声する。

図 3 はコード進行担当している部分のブロックダイアグラムである。図 3 の D1 D4 はデチューンに使用されるブロックであり、その構成を図 4 示す。ここでは、基準の周波数を合成する、のこぎり波発振器とデチューン用ののこぎり波発振器 2 基の 3 基を 1 組としそれを 4 組用いてコードを演奏している。

体の広げ方が大きくなるとデチューンの深さがより深くなり、複雑な音色となる。体の広げ方の計算には、頭の高さと左右の足のうち低いほうの足の高さの差、左右の手首の距離を利用してデチューンの深さを決めている。

同様に、ローパスフィルタのカットオフ周波数も、体の広げ方によって決定されている。体の広げ方が大きくなるとフィルタが開いていく。

フィルタと、デチューンにより、体を大きく使うとより派手な音色になる。

シンセベルの音色では、FM のパラメータを身体姿勢によって変化させている。FM の変調周波数を両手首の距離、変調指数を高く上げている側の手首の絶対位置によって決定している。

表 1 に今回実装したサウンドインタラクションの機能と参照している身体姿勢の情報、パラメータの値の範囲をまとめる。

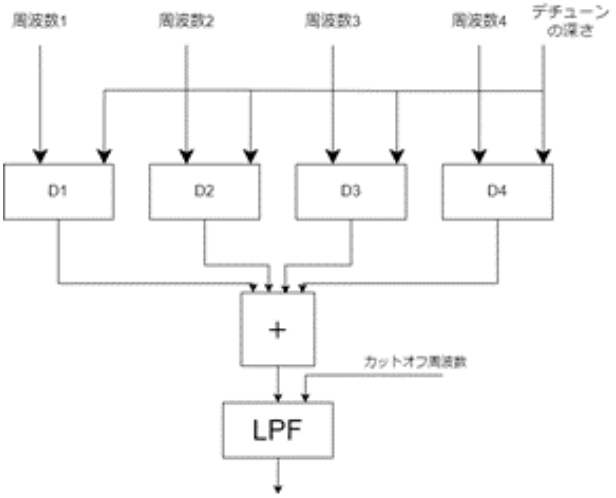

図 3: コード用ブロックの構成

表 1: 身体動作と音合成パラメータの対応

機能	身体動作	パラメータ	値
デチューン	体の広がり	深さ	0-40[Hz]
LPF	体の広がり	カットオフ	0-8000[Hz]
FM	両手首の距離	変調周波数	0-1500[Hz]
FM	高い方の手首の位置	変調周波数	100-600[Hz]
正弦波発振器	体の広げ方傾き	発音/停止	傾きが 0.1 以上で発音

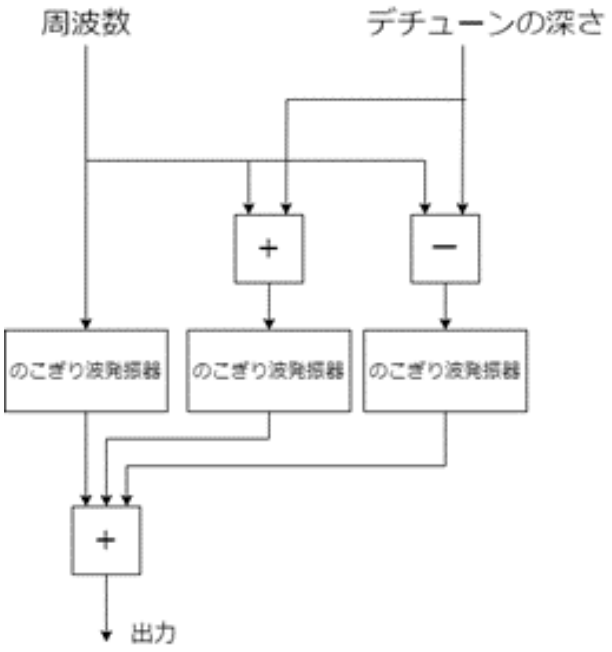

図 4: デチューン用発振器の構成

4. 他のセンサとの比較

4.1. 単一カメラ入力による姿勢推定を用いたサウンドインタラクションの利点

今回制作したシステムは、単一のカメラと画像処理技術、機械学習による姿勢を利用している。ボディスーツ型のセンサや、kinect などカメラ型のセンサを用いる場合には専用のハードウェアをパフォーマンスが行われる現場で使用する必要がある。また。パフォーマの動作範囲や動きの自由度に制限が加わる。

一方で、今回使用しているシステムでは、2 次元動画から姿勢検出が行えるため、パフォーマが web カメラのみの環境でも、zoom や YouTube などを通じて動画の転送をすることにより、オンラインでのパフォーマンスにも対応可能である。この点により、オンラインでのパフォーマンスや VR 空間上での作品が増えている中で有効であると考える。

またパフォーマに加わる物理的な制約が少なく、より自由な舞踊による表現が可能である。

4.2. 深度センサやモーションキャプチャと比較した際の問題点

各種モーションキャプチャデバイスや、kinect をはじめとした物理的なセンサと比較した際、奥行値を測定することが困難な点は、豊かな表現を実装するための欠点であるといえる。特に、多チャンネルオーディオを利用するインスタレーション作品など、空間上での音響表現を実現するためには、奥行値を含めた 3 次元データの利用が重要である。このような空間上での音響表現を主体とする作品のインタラクションの構築には、不向きであるといえる。

5. まとめ

時間軸上で進行していく音楽に対して、舞踊の姿勢情報に基づいたインタラクションを行い、音色と舞踊の感情表現を対応させることによる、直感に合うサウンドインタラクションを目指して制作を行った。

作成したシステムでは、姿勢推定の実行、映像生成、データ処理、音響合成をそれぞれ特化したシステム間を OSC によって連携することにより全体を構成している。音響合成においては、体の広げ方、傾け方と対応の良い音合成パラメータを割り振り、その値を定量的に明らかにした。それらに基づいて、周波数領域での変化を実装した。

機械学習による姿勢推定では、奥行値を取得できず、3 次元的な空間表現には向かない。空間音響を重要とする作品の制作には、kinect など深度センサを備えたデバイスの方が有効である。

今後の発展として、映像を利用した機械学習による姿勢推定を用いた、オンラインでのパフォーマンスなどへの応用を期待することができる。

6. 参考文献

Runway AI, Inc. Runway https://runwayml.com/ (accessed February 27, 2022).

TensorFlow,TensorFlowBlog, “Real-time Human Pose Estimation in the Browser with TensorFlow.js” https://medium.com/tensorflow/real-time-human-pose-estimation-in-the-browser- (accessed February 27, 2022).

Matthew Wright (1997) “Open Sound-Control: A New Protocol for Communicating with Sound Synthesizers” https://opensoundcontrol.stanford.edu/files/1997-ICMC-OSC.pdf(accessed February 27, 2022).

下道 雄太、入江 英嗣、坂井 修一 (2019)「身体姿勢を用いた直感的サウンドインタラクションの検討」https://ipsj.ixsq.nii.ac.jp/ej/?action=pages_view_main&active_action=repository_view_main_item_detail&item_id=197828&item_no=1&page_id=13&block_id=8 (accessed February 27, 2022).

7. 参考作品

Ali Nikrang, 2019. Mahler-Unfinished https://ars.electronica.art/futurelab/en/projects-mahler-unfinished/ (accessed February 27, 2022).

後藤 英, 2009. *L'homme transcende*. http://suguru.goto.free.fr/Contents2/L'hommeTranscende/L'hommeTranscende-j.html (accessed February 27, 2022).

後藤 英, 2015. *Hypnoid* http://suguru.goto.free.fr/Contents2/L'hommeTranscende/L'hommeTranscende-j.html (accessed February 27, 2022).

Rhizomatiks, 2020. *Morphcore*. https://rhizomatiks.com/work/morphecore/ (accessed February 27, 2022).

Yamaha, 博報堂 I-Studio, 2018. *Duet with YOO* https://www.yamaha.com/ja/about/ai/duetwithyoo/ (accessed February 27, 2022).

8. 著者プロフィール

公文太一 (Taichi KUMON)

2000 年栃木生まれ、2019 年東京電機大学未来科学部情報メディア学科入学、現在同学科音メディア表現研究室に所属。
