第 28 回日本バーチャルリアリティ学会大会論文集（2023 年 9 月）

深層学習を用いた動作予測結果のアバター表示のための汎用的なデータ形式利用の検討

平井龍之介 [1)]，佐々木大祐 [1)]，小山功太郎 [2)]，上島直登 [2)]，許超舜 [1)]，牧野泰才 [1)]，篠田裕之 [1)]

1) 東京大学 情報理工学系研究科（〒 113-0033 東京都文京区本郷 7-3-1）

2) 東京大学 工学部（〒 113-0033 東京都文京区本郷 7-3-1）

概要: 機械学習を用いて人物の 0.5 秒程度先までの動作を予測する動作予測技術において，対象の動作は一般的にモーションキャプチャ技術により 3 次元骨格情報を計測したものが用いられる．モーションキャプチャが人物の動作を記述する形式は製品によって異なるため，機器から直接出力されるデータを用いた場合複数種のモーションキャプチャの出力を単一の深層学習モデルで取り扱うのは難しい．本論文では，モーションキャプチャから出力されたデータを汎用的な表現に置き換えた上で動作予測を行うことで，モーションキャプチャの種類に依存しない動作予測を実現する手法について検討する．特にデータ形式の違いによる予測性能の差について検証し，動作予測に適した表現形式を検討する．

キーワード： 動作予測，深層学習

1. 序論

1.1 人体の動作予測

人体の動作には慣性や重心，関節の可動域などの影響によりある程度の予備動作や不可欠な重心移動が存在する．例えば，ジャンプの前には身を屈めるようなような動作が必要となる，歩行時には軸足に重心を移動させる必要があるため，8 の字に重心が移動する性質があるなどが挙げられる．この，動作に伴う不可欠な予備動作を入力に，短期的な未来の人体の姿勢を推論する技術を身体動作予測と呼ぶ．

近年，深層学習を用いることで数百 ms 後の身体姿勢を数 cm 単位の精度で予測できることが知られている．堀内らは深層学習によって 0.5 秒後の姿勢情報を推定する手法を提案しており [1]，また板井らは推論された姿勢情報を被験者にフィードバックすることで被験者にもたらされる感覚について論じている [2]．また，上田らは身体動作予測技術を利用することによって，遠隔操作映像の遅れを補償する仕組みを提案している [3]．深層学習を利用した動作予測技術はアニメーションを伴うエンタテインメントや運動支援などの分野に広い応用可能性がある．

1.2 データ形式による人体の姿勢情報の差異

人体の姿勢には各関節の位置のみでなく，関節の回転角などの様々な情報が含まれる．また，人体の骨格構造は基本的に伸長しないため，姿勢によらず骨格により接続された関節間の距離は常に一定であるという制約が存在する．そのような身体姿勢情報を表現する手法は一意ではなく，その表現の詳細度や自由度に応じて複数存在する．

例として，Microsoft の Kinect シリーズでは主に各関節の座標情報によって姿勢を表現する．また，モーショントラッキングソフトウェアにおける汎用的な動作記録データ形式である BVH では，関節の初期座標，初期角度，関節同士の親子関係に加えて各時間における座標変化や角度変化によって姿勢情報を記録する．ゲームエンジンである Unity には，身体重心の座標，角度に加え人体の各関節の可動域を $(-1, 1)$ で正規化し，それを値域とする 95 個のパラメータで姿勢情報を表現した HumanPose というデータ形式が存在する．

これらの姿勢を表現するデータ形式のうち，どれを利用するかは用途に依存し，高精度の姿勢情報が要求される映像制作においては詳細な姿勢情報が記録されたデータ形式を用いる，通信量について考慮する必要があるバーチャルライブ・ストリーミングなどにおいてはポーズあたりのデータ量が少ない関節座標のみのデータ形式を通信し，後述する逆運動学による姿勢推定で不足した情報を補完するなどの手法がとられる．

1.3 情報の差異が動作予測のアバター表示に与える影響

身体動作予測研究においても姿勢情報として様々なデータ形式が利用されている．板井らは姿勢情報として Kinect より取得される 32 点の関節位置情報を用いており，上田らは姿勢情報としてモーショントラッカーより出力される各関節の座標から計算された，各関節が初期位置からなす角度を用いている．

異なるデータ形式の姿勢情報の間には，表現形式の違いのほかに，その表現が含む情報量にも差異が存在する場合がある．例えば，各関節の座標で構成される姿勢情報からは各関節がどのように捻られているかなどの情報を直接知ることはできない．姿勢情報を 3DCG アバターによって可視化する場合，先に述べたような関節の角度などの情報が必要となるため，姿勢情報を各関節の座標として表現している場合にはアバターによる可視化の際に何らかの手法を用いて不足する姿勢情報を補完する必要がある．このよう

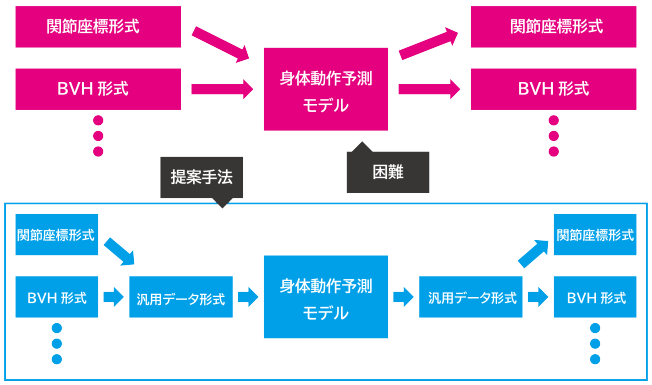

図 1：本研究の概略図．異なるデータ形式の姿勢情報に対する動作予測を単一のモデルで行うのは困難である（上部），汎用データ形式を経由して動作予測を行う（下部）

な姿勢情報データの間の情報量の差異を補完する手法として，関節間の距離や関節の可動域といった制約を考慮した上で対象とする関節の座標を目的の座標に近づけるような最適化問題を解くことで欠けた姿勢情報を決定する逆運動学（Inverse Kinematics, IK）と呼ばれるものが存在する．

表現の手法も保持する情報量も異なるような姿勢情報のデータ形式に対し，単一の深層学習モデルで身体動作予測を実行するのは難しい（図 1 上部）．多数のデータ形式を汎用的なデータ形式に変換した後に深層学習モデルに入力し推論結果を元のデータ形式へと再変換する仕組み（図 1 下部）があれば，単一のモデルを様々な用途へと利用することが出来るようになるため，そのような仕組みの開発は身体動作予測技術を実社会で応用するにあたっては重要な課題であると考えられる．本論文では，ゲームエンジンである Unity が提供する HumanPose と呼ばれるデータ形式を利用し，汎用的なデータ形式を経由した身体動作予測を行った際，元のデータ形式を保持したまま身体動作予測を行った場合と比べて予測結果にどのような変化が生じるのかについて実験を行い，その結果について議論する．

2. 汎用的なデータ形式を利用した身体動作予測

本章では前章の内容を踏まえ，汎用的な表現を利用した身体動作予測の意義と，深層学習に適した姿勢情報の表現の持つ特性について述べる．

前章で述べたように，現在身体姿勢データの表現には多くのデータ形式が利用されており，そのすべてに対応可能な深層学習モデルを構築するのは難しい．そのため，人体動作予測を実社会上に応用するには，これらの身体姿勢データを汎用的なデータ形式に変換した上で深層学習モデルによる動作予測を実行，その後任意のデータ形式に復元する処理が必要となる．

適切な汎用データ形式を策定し，それを経由することで身体動作予測が様々なデータ形式で利用可能となれば，単一の身体動作予測の学習済みモデルを運動補助やバーチャルリアリティにおける身体動作ストリーミングのクオリティ向上など様々な用途で利用できるようになる．よって，身体動作予測技術を実社会で応用するにあたり，上述の汎用データ形式を用いた学習の実現は重要な課題である．

深層学習に用いる汎用データ形式は，学習を効率的に行うため，データの構成に自由度がなく，人体の骨格などからくる制約を表現に取り込めるものが望ましい．また，より多岐に渡る用途に利用可能とするため，アバターによる可視化が可能な情報量を持ったデータ形式であることが望ましい．

3. 実験

本章では，汎用的なデータ形式を経由した身体動作予測を行った際，元のデータ形式を保持したまま身体動作予測を行った場合と比べて予測結果にどのような影響が生じるのかについて実験を行う．本論文では Azure Kinect によって記録された各関節の座標によって表現される動作データおよびそれを汎用データ形式として Unity の持つ HumanPose データ形式に変換したものについて身体動作予測を行うことで，データ変換が予測結果に及ぼす影響について調べる．

動作データのデータ形式として HumanPose データ形式を採用した理由として，以下の二点を挙げる．

- HumanPose データ形式はあらかじめパラメータの個数が決定されているため，データの構成に自由度が無く，関節の可動域を $(-1, 1)$ で表現するものである．骨格からくる制約が暗にデータ形式に組み込まれているため，深層学習に用いるデータ形式として適していると考えられる．
- 現在 Unity は映像制作からバーチャルリアリティ・コンテンツの制作まで広く使われており，モーションデータに関しても多くのデータ形式に対応している．HumanPose への変換も容易であるため，実用上の観点からも HumanPose を経由した身体動作予測は有用であると考えられる．

以下に述べる 2 種類の身体動作予測システムについて，フレームレート 30 の動作データ列に対し，15 フレームの連続する部分動作データ列を入力とし，その 9 フレーム後の姿勢を推論するタスクを実行させ，真値との差異を比較する．なお，姿勢情報を HumanPose データ形式に変換するには関節の角度などを含んだ詳細な姿勢データが必要なため，両者の比較は IK によって角度情報を含んだ詳細な姿勢情報に変換されたものに対して行われる．両者のモデルの差異について，図 2 に模式的に示している．

変換なし条件 Azure Kinect により記録された関節座標データ（3 次元，25 点）を入出力として深層学習モデルに身体動作予測を学習させる．出力された予測結果を IK を利用して角度情報を含んだ詳細な姿勢データに変換する．

変換あり条件 Azure Kinect により記録された関節座標データを IK を利用して詳細な姿勢データに変換後，それを HumanPose（重心座標（3 次元）と，関節情報（1 次元）95 点）に変換する．変換後の HumanPose データ形式を入出力として深層学習モデルに身体動作予測を学習させる．

図 2: 各条件の概略図．変換ありの条件では，事前に汎用データ形式（HumanPose）に変換した後で，身体動作予測を行う．

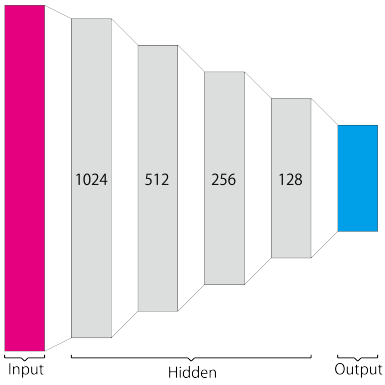

図 3: 実験に用いたネットワークの概要図．隠れ層に記載された数字は各層の次元を示す．入出力の次元は変換なしとありの場合で異なる．また，変換あり，Tanh 層追加の場合は，Output の一部（HumanPose の関節の曲がり具合に相当するパラメータ）については Tanh を求める処理が加わる．

本実験ではデータセットとして Azure Kinect で記録された関節のうち，鼻や耳といった微細な部位（NOSE 1 点、EYE, EAR, CLAVICLE の左右 2 点ずつ）を除いた 25 点の関節座標情報によって構成される歩行動作 108 動作から抽出された 11136 点のデータを 8 : 1 : 1 の割合で分割し，それぞれを学習用，評価用，テスト用のデータとして用いた．実験に用いた深層学習モデルは板井ら [2] と同じく 5 層の全結合層から構成されるもので，入出力のみをそれぞれのパラメータ数に変更したものを用いた（図 3）．また，損失関数には最小二乗誤差を用いた．この際，変換なし条件の入出力次元数はそれぞれ 1125, 75，変換あり条件の入出力次元数はそれぞれ 1470, 98 であった．また，深層学習モデルに姿勢情報の持つ平行移動不変性を学習させるためにデータ拡張としてデータをランダムな値分オフセットさせた．

変換ありの場合について，ネットワークの出力を HumanPose の値域に合わせるために，変換ありのネットワークの末尾に Tanh 層を加えたものについても，同様の実験を行った．

各身体部位とその名称を図 4 に示す．上述の手法によって学習を行ったモデルに対し，重心に相当する骨盤（Hips）の絶対座標の推定誤差に加え，姿勢をよく表現する身体の末端

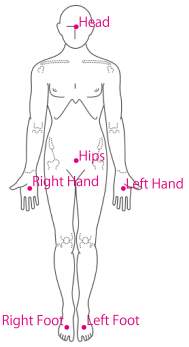

図 4: 比較対象とした身体部位の名称と位置

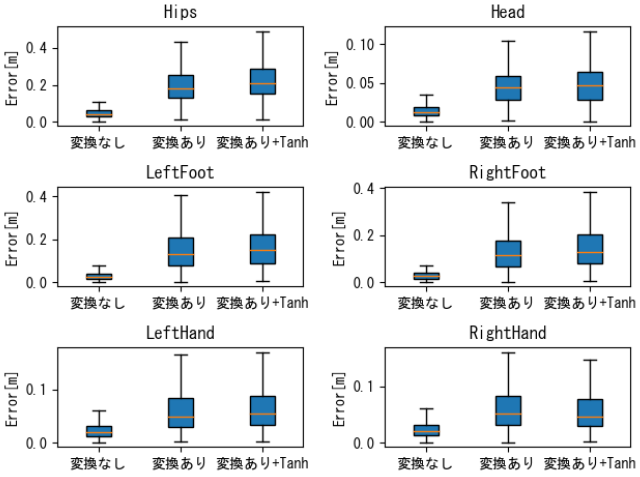

図 5: 各条件について，各身体部位ごとの誤差についての箱ひげ図．Hips については絶対座標，その他については Hips からの相対座標について推定誤差の平均と分散について算出した．

部分として頭部（Head），左足（Left Foot），右足（Right Foot），左手（Left Hand），右手（Right Hand）に注目し，各部位の Hips からみた相対座標の推定誤差について比較を行った．推定誤差は学習データと同様の環境で記録された歩行データ 72 動作を用いて算出した．変換なし，変換あり，変換あり (末尾に Tanh 層) の各条件について，各部位に注目した際の誤差に関する箱ひげ図を図 5 に示す．

まず，全体の傾向として変換あり条件の方が座標における誤差が大きくなっていることがわかる．テストデータのうちの一つについて，Hips の座標の真値とそれぞれの条件下で Hips の座標の推論結果を図 6 に示す．変換あり条件では，姿勢情報のみならず，身体の重心の推論にも誤差が生じていることがわかる．

4. 考察

これらの誤差の拡大の一因として考えられるものに，深層学習モデル自体の表現力の低さが挙げられる．今回用いた深層学習モデルは全結合層のみを用いた単純なものであったため，モデルの表現力が低く，無関係なパラメータの変

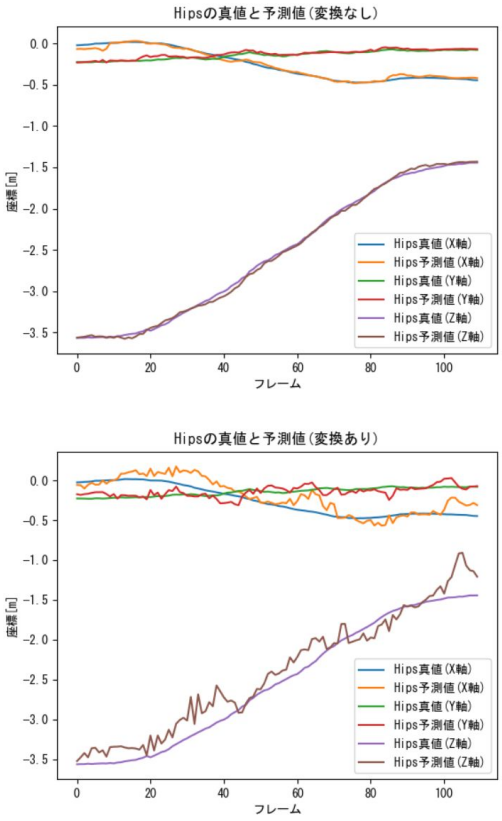

図 6: Hips の座標の真値とそれぞれの条件下での Hips の座標の推論結果

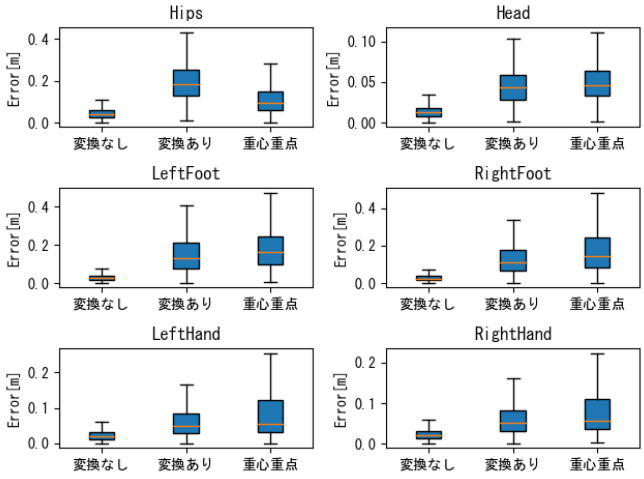

図 7: 図 5 と同様の手法で推定誤差を計算した箱ひげ図．重心重点ラベルは変換あり，重心推定の部分の重みを 10 倍の条件で学習させたもの．

化に別のパラメータが影響されてしまう場合があった．今回変換なし条件に比べて変換あり条件のデータは次元数が大きいため，そのような表現力の低さが精度に強く表れた可能性がある．

追加実験として変換あり条件について，損失関数のうち重心を推定する部分の重みを 10 倍にしたところ，重心の推定精度は向上したものの，末端部分の精度は低下した（図 7]）．このことからも，本実験で用いた深層学習モデルの表現力は重心推定と姿勢推定を共に高精度で行うには不足である可能性が示唆された．

これとは別の要因として，センサの計測誤差による不自然な姿勢の発生がある．Azure Kinect によって取得された関節座標データの一部は Azure Kinect のカメラに写っていない，ほかの部位にきわめて近い位置にあるなどの要因によって誤った位置に配置されている場合があった．このような大きな誤差を持った関節座標データを含んだ姿勢情報について HumanPose への変換を行うと，本来 $(-1, 1)$ で表現されるべき HumanPose の各関節の曲がり具合が，その範囲から逸脱してしまう場合があった．実際，Tanh によって値域を明示的に制限した場合において，微小に誤差が増大しているのは上に述べたような要因によるものであると考えられる．

5. 結論

本論文ではより多様な用途へと応用可能な身体動作予測モデルを生成するため，汎用的なデータ形式を用いて予測を行う手法を検証した．歩行動作の予測について，関節座標による姿勢表現を用いた高精度な予測と比較し，汎用的なデータ形式を用いることによって予測誤差が増加することが確認された．今後はこれらの予測誤差を低減し，関節座標データ形式などと同程度にする手法を検討したい．

本論文では単一のデータ形式から汎用データ形式への変換のみを取り扱ったが，データの汎用性について議論するのであれば，多数のデータ形式からの変換についても検討が必要であると考える．本論文では 5 層の全結合層という単純な深層学習モデルを用いて身体動作予測を行ったが，実社会で応用可能な精度で身体動作予測を実現するには，Transformer など，より高度な深層学習モデルを利用する必要がある．実際の環境で利用されるような複雑な深層学習モデルに関しても，汎用的なデータ形式が与える影響について議論することが今後の課題であると考える．

謝辞 本研究は科研費 21H03479 の支援を受けて行われた．

参考文献

[1] Yuuki Horiuchi, Yasutoshi Makino, and Hiroyuki Shinoda. Computational foresight: Forecasting human body motion in real-time for reducing delays in interactive system. In *Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces*, pp. 312–317, 2017.

[2] 板井俊樹, 牧野泰才, 篠田裕之. 動作予測情報を利用したアバター動作変調による身体感覚操作. ロボティクス・メカトロニクス講演会 2023, pp. 2A2–G17, 2023.

[3] 上田樹, 宍戸英彦, 北原格. 深層学習による運動予測を用いた遠隔操作映像の時間補償. 研究報告コンピュータビジョンとイメージメディア (CVIM), Vol. 2020, No. 39, pp. 1–8, 2020.