第 29 回日本バーチャルリアリティ学会大会論文集（2024 年 9 月）

上方から観察可能な 3 点の二次元特徴量を利用した歩行予測

Gait prediction using two-dimensional features at three points that can be observed from above

日高祐哉 [1)]，高田賢太 [2)]，杉本証 [2)]，牧野泰才 [1)]，篠田裕之 [1)]
Yuya Hidaka, Kenta Takada, Sho Sugimoto, Yasutoshi Makino, Hiroyuki Shinoda

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: これまで深層学習を用いて人間の歩行先を予測する技術が研究されており，そのときに必要となる特徴量は比較的少なくて済むことが示されている．本研究ではその特徴量の中でも，上方から観測できる胸部および両足首の二次元特徴量のみから，深層学習を用いて人間の進行方向を推定する手法を提案する．円を描きながら運動している人間の座標情報を教師データとし，全結合層で正則化を加えた学習を行うことによって，0.5 秒間の歩行データから 0.3 秒後の人間の位置を予測できることを確認した．

キーワード： 動作予測，深層学習

1. 序論

近年，機械学習を用いて身体動作を予測する研究が盛んに行われている．堀内らは深層学習を用いて 0.5 秒後の姿勢情報を予測できることを示しており [1]，Wu らの研究では，深層学習を用いて人間のポーズを予測し，新しい格闘技のトレーニングシステムを提案している．[2].

本研究では，胸部と足首との間の距離が歩行時に大きく変動しないのであれば，それを一定とみなすことで，上部からの画像でも同様に予測が可能になるのではないかと考えた．そこで，深層学習を用いて上方からでも観察できる胸部および両足首の 2 次元特徴量のみを用いて，人間の位置を予測する手法を提案する．これを利用すると，予測に必要な特徴点が少ないために計測と推定に要する時間が短くなるほか，オクルージョンに対して強い予測が可能になるのではないかと考えられる．例えば部屋の中を自由に移動しながら VR 体験をするような場合，環境やアバターの描画をスムーズに行うことを考えると，人の移動方向を事前に知ることが有用である．Kinect などのように深度センサを用いて 3 次元の骨格情報を利用して歩行の予測などは王らや渡部らの研究によって可能であることが示されているが [3] [4]，室内で複数人が移動をしているような状況や，ソファや机などの物体がある状況でのオクルージョンの影響により，必ずしも 3 次元の情報を全て利用できるとは限らない．部屋上方にカメラを設置し，頭部と両足首の 3 点で歩行方向の予測が可能であれば，オクルージョンの影響の少ない予測が実現できる．本研究では，3 点の 2 次元座標を利用した予測の可能性を検証するために，正則化を施した深層学習を行った．その結果，3 点の 2 次元座標情報を利用した予測について，学習に利用したものと同じ動作をするデータセットにおいて 36 mm 程度，異なる動作をするデータセットにおいて 74 mm 程度の平均二乗誤差で推定が行えることが示された．

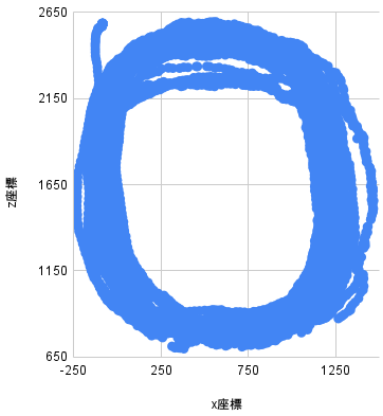


図 1: 学習に用いる歩行の様子（単位は mm）

2. 方法

2.1 学習データの生成

本研究では 0.5 秒間の歩行データから 0.3 秒後の人間の座標情報を予測するためのデータセットを生成した．学習に用いるデータは図 1 のように直径約 1.5 m の円を描くように運動している様子を Kinect を用いて 30 fps で撮影したものである．まずはこのデータから学習に用いる胸部，両足首の座標情報を抽出する．この際，鉛直成分も取り除くことで x 方向，z 方向の 2 次元データとしている．次に，15 フレーム分を切り出して入力用のデータとし，9 フレーム先，すなわち切り出し始めから 23 フレーム先のデータを切り出して正解データとした．その後，オフセットのために切り出された全 16 フレーム分のそれぞれの座標から 15 フレー

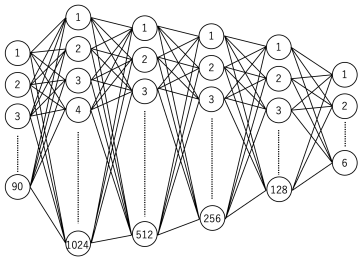


図 2: 使用したニューラルネットワークの構造

ム目の座標を引き，これらを 1 セットの教師データとした．つまり時刻 t での胸部と両足首の x, z 座標を成分に持つベクトルを $\boldsymbol{r}_t$ とおくと，学習に用いるデータは

$$\boldsymbol{r}_t - \boldsymbol{r}_{t+15}, \boldsymbol{r}_{t+1} - \boldsymbol{r}_{t+15}, \ldots, \boldsymbol{0} \tag{1}$$

からなる入力と

$$\boldsymbol{r}_{t+24} - \boldsymbol{r}_{t+15} \tag{2}$$

からなる正解データで構成されている．このオフセット処理によって各データが最終フレームの相対座標で表されるようになり，歩行のスケールやカメラからの絶対座標に左右されにくい学習を進めることができるようになる．よって入力用のデータは胸部，両足首の 3 点の横方向および奥行き方向の 15 フレーム目からの相対座標で 90 次元，正解データは 24 フレーム目の胸部，両足首の 3 点の横方向および奥行き方向の 6 次元となる．

2.2 ニューラルネットワークの設定

本実験で使用したニューラルネットワークは中間層 4 層から成る全結合層の DNN であり，前述の通り入力層のニューロン数が 90，出力層のニューロン数が 6 となっている（図 2 参照）．全ての層において活性化関数は ReLU を用いており，出力の結果と正解データとの誤差は MSE によって評価する．この際，過学習の影響を低減するためにに正則化項を加えた学習を行った．正則化の方法は L^2 正則化であり，損失関数を式で表すと

$$\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2 - \lambda \sum_{i=1}^{5} \|W_i\|_F^2 \tag{3}$$

となる．ここで，式 (3) の第一項は計算結果と真値との二乗ノルムを計算しており，第二項は伝搬に用いる行列のフロベニウスノルムの二乗 (各成分の二乗和) を計算している．λ は正則化項の重みを決定づけるパラメータであり，本研究では $\lambda = 3 \times 10^{-3}$ とした．学習率，エポック数，バッチサイズはそれぞれ 10^{-3}，3000，10 でオプティマイザーは Adam である．

2.3 評価方法

学習モデル作成後，本研究では図 4 のような二つのデータについて予測の検証を行った．

- 学習に用いたデータ同様，円を描くようにして歩行しているデータ（テストデータ 1）

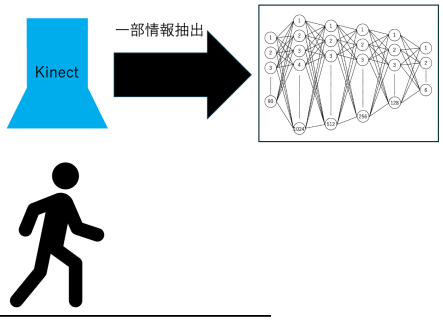
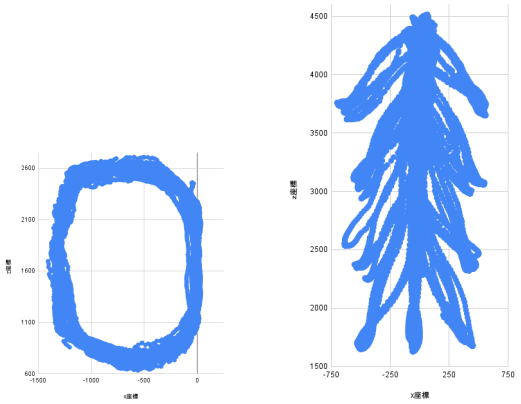


図 3: 実験の外観．Kinect は地面に設置していたが，上方から観察可能な特徴量を用いているため，便宜的に上方に設置している図としている．



(a) テストデータ 1　　**(b)** テストデータ 2

図 4: 評価に用いた歩行データ

- 基本前方（図中下方）へ進んでいるが，一定の時間間隔で斜め前にランダムに方向を変えて歩行する，最大 5 歩分のデータ（テストデータ 2）

これらのデータのうち，最初の 2000 フレームを切り出して真値と予測値を可視化，さらに定量的な評価を行うためにRMSE を求めた．テストデータ 1 は円を描いていることから滑らかに進行方向を変化させながら動くデータ，テストデータ 2 は止まっている状態や突然斜め前に動き出すところがあることから，動きの変化が大きいモデルとみなすことができる．

3. 結果と考察

テストデータ 1 に対して学習モデルを適用すると，結果は図 5 のようになった．横軸が時間，縦軸が変位を表し，それぞれ x,z 座標を表す．z 方向は Kinect に対して奥行方向であり，テストデータ 2 の進行方向に対応する．青いラインが実際の 0.3 秒後の座標値，オレンジが 0.5 秒間のデータを利用した予測値を表す．したがって両者が一致しているほど予測精度が高いことを示す，RMSE は 36.44 mm であった．学習に用いた教師データと類似した歩行方法であったので，非常に良い精度で予測できたと考えられる．足首の

グラフを見ると進行方向が変化する点で真値との差が大きくなってしまっているが，これは進行方向を変化させるという動作には足首の寄与が大きく，学習モデルが 15 フレーム前からの情報だけでは決定することが困難であることを示していると考えられる．しかし，胸部に関しては真値との差が大きくなっていないことから，滑らかに動くモデルに関して身体の軸となる部分に関しては正確に予測できていると考えられる．

テストデータ 2 に対して学習モデルを適用すると，結果は図 6 のようになり，RMSE は 74.46 mm であった．テストデータ 1 と比べると誤差は大きくなってしまっているが，歩行のスケールに対しては良い精度で予測できていると考えることができる．一方でテストデータ 1 に比べると胸部の座標の誤差が全体的に少し大きく，特に進行方向が変わる時刻での誤差が大きくなっていることが確認できる．

テストデータ 2 に関して，予測値が真値から大きく外れてしまっている部分を解析するため，図 6 から 500〜750 フレーム目を抽出したものが図 7 である．500 フレーム目から数えて 50 フレーム目までの真値を見ると，x 座標も z 座標も値が変化していないことから静止しており，50 フレーム目あたりから斜め前に向けて動き出していることが確認できる．このことから，動き出しの状態に関しては予測精度が落ちてしまうことが分かるが，150 フレーム目あたりで真値との大きな誤差が解消されていることも確認できる．これらのことから，この学習モデルは初動の誤差からの修正が優れていると考えることができる．初動に関しては，人間の意思のみで決まるものであり過去の歩行情報との因果関係が存在しないため予測することは原理的に不可能であると考えられるが，初動の誤差の解消を少しでも早くすることは望まれるので今後の課題となる．また，図 7 の 125 フレーム目あたりから確認できるが，急に停止する動作に対しても予測しきれていないことが確認できるが，急停止も初動と同様に前の動きと因果関係がないため，原理的に予測不可能であると考えられる．

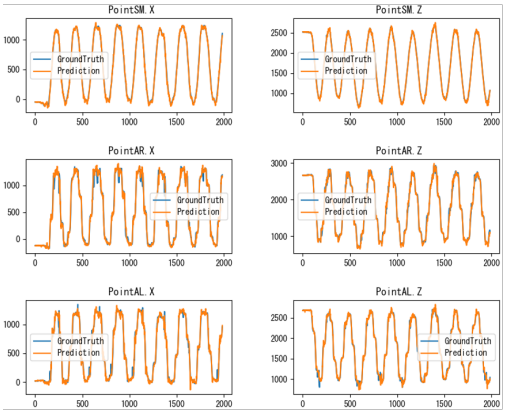


図 5: テストデータ 1 での結果．SM が胸部，AR，AL がそれぞれ右足首，左足首を指す

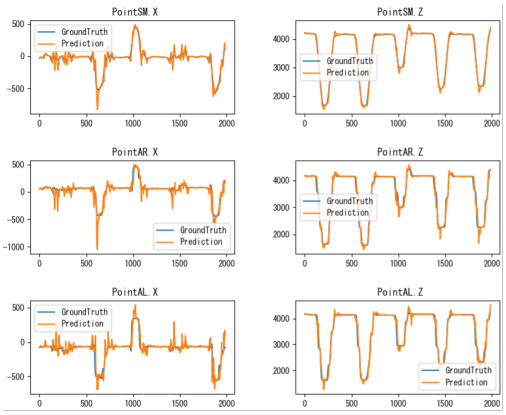


図 6: テストデータ 2 での結果

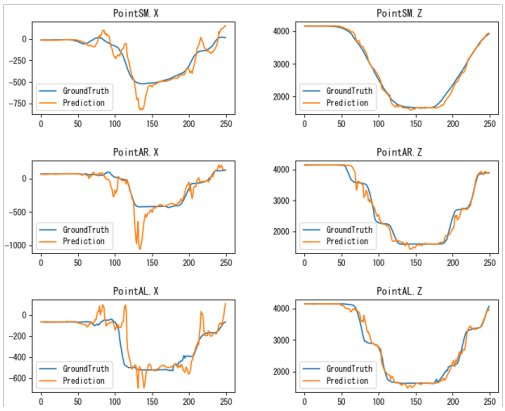


図 7: テストデータ 2 の詳細 (500〜750 フレーム)

4. 結論

本論文では，深層学習を用いて上方から観察できる特徴量のみで人間の歩行を予測できるかについて検証した．正則化を用いた深層学習を行うことよって，良い精度で歩行を予測できる汎用性の高いモデルを作成できていることを確認した．今後は歩行速度による予測精度の変化や，歩行以外の様々な運動に対する予測精度の変化などといったことを調べることによって学習モデルの頑健性を定量的により詳しく調べることを検討したい．

本実験では全結合層により予測モデルを生成したが，歩行という時系列情報を含むデータに対しては RNN や LSTM といった時系列を考慮する深層学習モデルを利用することも考えられる．実際，LSTM を用いて姿勢推定を行う手法も提案されている [5]．これらと比較し，より精度の高いモデルを選択していくことが必要である．

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*, ISS '17, pp. 312–317, New York, NY, USA, 2017. Association for Computing Machinery.

[2] Erwin Wu and Hideki Koike. Futurepose - mixed reality martial arts training using real-time 3d human

pose forecasting with a rgb camera. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1384–1392, 2019.

[3] Ansheng Wang, Yasutoshi Makino, and Hiroyuki Shinoda. Machine learning-based human-following system: Following the predicted position of a walking human. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4502–4508, 2021.

[4] 渡部慎太郎, 牧野泰才, 篠田裕之. 人間の歩行動作予測に貢献する重要な身体部位の解明. ロボティクス・メカトロニクス講演会講演概要集 2023, pp. 2A1–H28. 一般社団法人 日本機械学会, 2023.

[5] Huseyin Coskun, Felix Achilles, Robert DiPietro, Nassir Navab, and Federico Tombari. Long short-term memory kalman filters: Recurrent neural estimators for pose regularization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.