H-007

上半身と下半身の単一画像からの 3D 人物姿勢の推定に関する研究
A Study on 3D Human Pose Estimation from Single Images of Upper and Lower Body

細越　一希†　　プリマ　オキ ディッキ アルディアンシャー†
Kazuki Hosogoe　　Oky Dicky Ardiansyah Prima

1. はじめに

近年の深層学習の発展に伴い、単一画像からの 3D 人物姿勢推定に関する研究[1] [2] [3]が盛んに行われており、これらの技術の利用により、スポーツやリハビリにおける動作解析が容易になることが期待されている。当該技術により、対象者の体に器具を装着せずに通常の単眼カメラのみで 3D 姿勢を推定できるが、撮影環境や撮影角度によっては一部の関節を撮影できない場合がある。既存の 3D 人物姿勢推定は主に全身の姿勢推定を対象としており、一部の関節の画像情報が欠落することによって関節全体に対する 3D 姿勢推定精度の低下を招く可能性を指摘できる。そこで本研究では、単一画像に含まれる関節を上半身と下半身に分け、それぞれの部位に特化した 3D 人物姿勢推定モデルを実験的に構築し、関節全体と一部関節のそれぞれで構築したモデルによる推定精度の違いを検証する。

2. 3D 人物姿勢推定モデルの提案

本研究における全関節と一部関節の 3D 人物姿勢推定モデルでは、2D 姿勢データと 3D 姿勢データを入力と出力とする教師データをもとに、2D 姿勢データから 3D 姿勢データを推定するニューラルネットワークモデル(以後、これを「2D・3D 変換モデル」と呼ぶ)を構築する。図 1 に、各入力データに対応した 2D・3D 変換モデルを示す。当該変換モデルは 3D Pose Baseline[1]を採用し、He ら(2015)の初期値[4]により設定した重みを用いて入力データの次元を 1024 に拡張した後、バッチ正規化(Batch Normalization)や正規化線形関数(ReLU: Rectified Linear Unit)、ドロップアウト(Dropout)の一連の処理を 2 段階で実行し、3D 姿勢推定の結果を出力する。

2.1 概要

本実験では、Human 3.6M データセット[5] [6]をもとに、2D・3D 変換モデル I～III(図 1 参照)を構築し、各モデルに対する精度を検証する。Human 3.6M では、各々の姿勢データが 32 関節で構成されているが、本実験において主要な 16 関節のみを使用する。ここで、16 関節の内から、9 関節を上半身、そして 7 関節を下半身の関節とする。

2.2 分析方法

既知の 3D 姿勢データと各モデルから推定された 3D 姿勢データを比較し、両者の違いを分析する。ここで、形状を維持しながら両者の姿勢データを一致させるために、プロクラステス分析を用いる。プロクラステス分析では、比較対象の片方の形状を基準とし、もう片方の形状に対して並進、回転、一様なスケーリングを行い、誤差が最小になるように形状を一致させる。これにより、カメラの姿勢や位置の違いを吸収し、両者の姿勢データの形状の違いのみを比較することが可能となる。最後に、両者の姿勢データから各々の関節座標の対を抽出し、全フレームの各関節座標の対から各関節における二乗平均平方根誤差(RMSE: Root Mean Square Error)を求める。

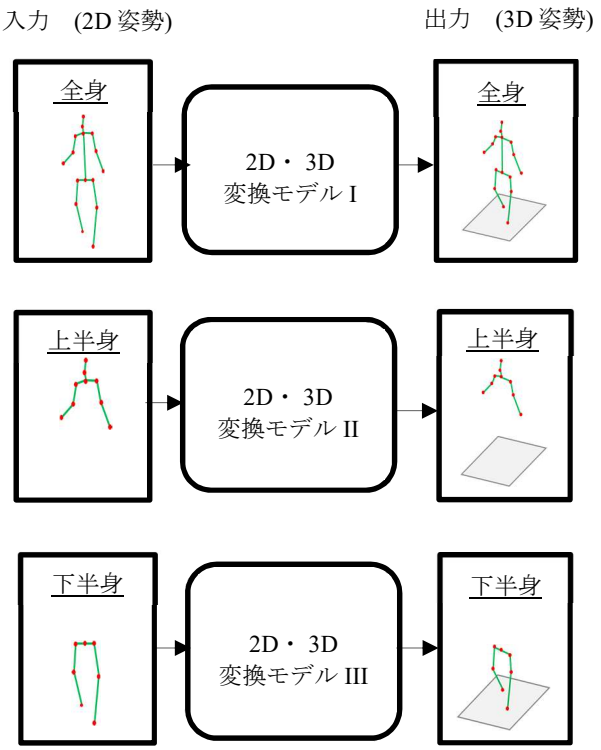


図 1　全関節と一部関節の 3D 人物姿勢推定モデル

3. 結果

Human 3.6M から取得した 5 人分のデータ、計 1,687,744 フレームの 2D 姿勢と 3D 姿勢データを利用して、2D・3D 変換モデル I～III を構築した。その際、エポック数を 200、学習率を 0.001、ミニバッチサイズを 64 とした。また、コードの実装には TensorFlow 1.15.0 を使用した。

3.1 Human 3.6M を使用した精度検証

本検証では、2D・3D 変換モデルの構築に使用されなかった Human3.6M の 548,800 フレーム分の姿勢データをモデルの入力データや正解データとして使用し、各 2D・3D 変換モデルによる推定の精度を検証した。表 1 に各モデルによる姿勢推定の誤差を示す。上半身用モデル(モデル II)と全身用モデル(モデル I)による上半身の姿勢推定の精度を比較した際、上半身用モデルを使用した場合の方が RMSE の平均は約 12.48mm 小さくなった。また、下半身の姿勢推定

†岩手県立大学大学院ソフトウェア情報学研究科
Graduate School of Software and Information Science, Iwate Prefectural University

表 1　Human3.6M を使用した精度検証

関節	RMSE (mm)		
	全身	上半身	下半身
Hip	37.86		15.77
Hip (Right)	46.95		32.41
Knee (Right)	57.85		51.49
Foot (Right)	77.14		59.31
Hip (Left)	43.34		30.42
Knee (Left)	59.60		50.44
Foot (Left)	81.67		63.75
Spine	34.05	51.24	
Thorax	33.76	27.71	
Head	53.58	45.65	
Shoulder (Left)	41.60	37.39	
Elbow (Left)	65.52	45.13	
Wrist (Left)	100.92	69.94	
Shoulder (Right)	42.08	35.43	
Elbow (Right)	69.36	46.64	
Wrist (Right)	100.08	69.50	
平均(上半身)	60.11	47.63	
平均(下半身)	57.77		43.37

表 2　RGB-D カメラを使用した精度検証

関節	RMSE (mm)		
	全身	上半身	下半身
Hip	82.25		33.18
Hip (Right)	87.35		51.11
Knee (Right)	98.30		71.47
Foot (Right)	118.14		68.15
Hip (Left)	89.77		53.08
Knee (Left)	103.21		74.46
Foot (Left)	121.96		67.48
Spine	98.41	94.70	
Thorax	108.57	80.70	
Head	145.74	161.43	
Shoulder (Left)	98.66	86.46	
Elbow (Left)	114.90	115.98	
Wrist (Left)	279.36	175.68	
Shoulder (Right)	100.44	97.14	
Elbow (Right)	116.46	150.86	
Wrist (Right)	314.09	202.82	
平均(上半身)	152.96	129.53	
平均(下半身)	100.14		59.85

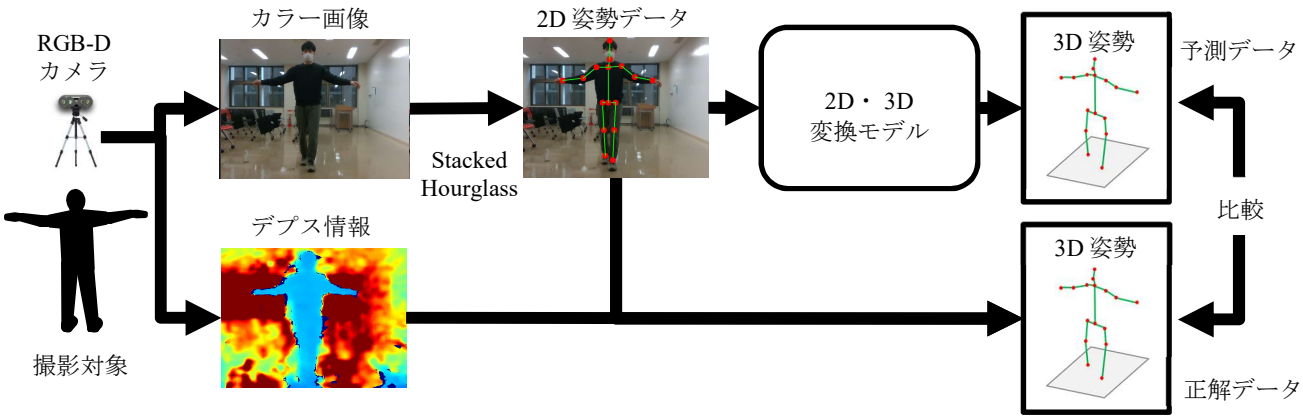


図 2　RGB-D カメラを用いた入力データと正解データの生成

の精度においては、全身用モデル(モデル I)を使用した場合と比較して下半身用モデル(モデル III)を使用した場合の方が RMSE の平均は約 14.40mm 小さくなった。

3.2 RGB-D カメラを使用した精度検証

本検証では、RGB-D カメラを用いて 13,940 フレーム分の姿勢データを生成し、それらをモデルの入力データや正解データとして使用し、各 2D・3D 変換モデルによる推定の精度を検証した。図 2 に、データセットの作成手法を示す。本手法では、RGB-D カメラから取得したカラー画像に対して 2D 姿勢推定手法 Stacked Hourglass[7]を適用して、得られた 2D 姿勢データを入力データとした。また、各関節の奥行情報をデプス情報から収集し、Stacked Hourglass で得られた 2D 姿勢データと組み合わせて 3D 姿勢データを作成し、これを検証用の正解データとする。本検証では、RGB-D カメラとして Intel RealSense Depth Camera D435 を使用し、撮影対象が立っている状況や座っている状況、歩行している状況を、解像度 640×480、フレームレート 30fps で撮影した映像を使用した。

表 3　RGB-D カメラを使用した精度検証における予測データと正解データ

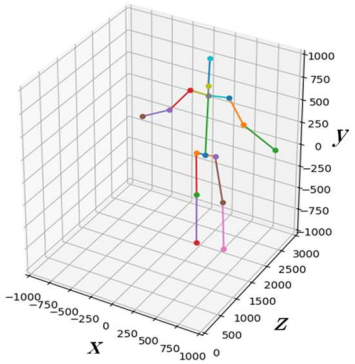
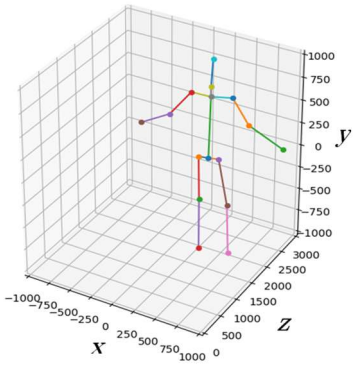
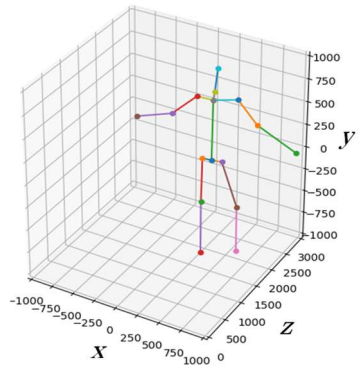
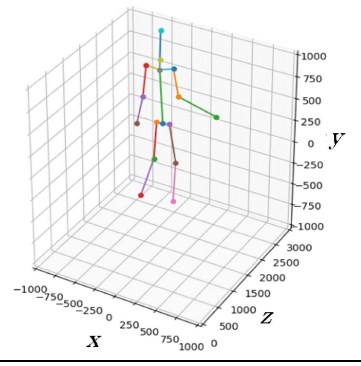
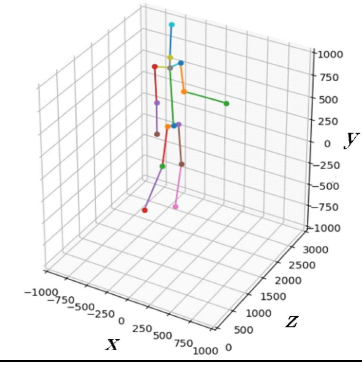
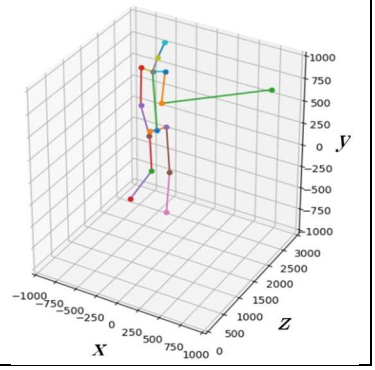
	予測データ	予測データ （プロクラテス分析後）	正解データ
誤差が小さい予測の例			
誤差が大きい予測の例			

表 2 に各モデルによる姿勢推定の誤差を示す。上半身用モデル(モデル II)と全身用モデル(モデル I)による上半身の姿勢推定の精度を比較した際、上半身用モデルを使用した場合の方が RMSE の平均は約 23.43mm 小さくなった。また、下半身の姿勢推定の精度においては、全身用モデル(モデル I)を使用した場合と比較して下半身用モデル(モデル III)を使用した場合の方が RMSE の平均は約 40.29mm 小さくなった。

また、表 2 から Head や Wrist における検出精度が低い傾向にある事が読み取れた。予測データと正解データを可視化した所、表 3 に示すように正解データの一部において、明らかに不自然な座標にある関節が見られた。これは、Head や Wrist の座標は人物と背景の境界付近の座標を取得しており、2D 姿勢データとデプス情報を照らし合わせる際、若干の位置のズレにより背景のデプス情報が使用されてしまっていることが原因と考えられる。

4. おわりに

本研究では、3D 人物姿勢推定において部位別にモデルを構築する手法を提案した。そして実証実験より、上半身と下半身の両方において、全身の関節で学習したモデルと比較して、一部の関節で学習したモデルの方が 3D 姿勢推定の誤差が小さく、部位別モデルによる単一画像からの 3D 姿勢推定の有用性を確認した。

今後は、正解データ生成において明らかに正常でないデータの除去手法の検討や、上半身や下半身以外の部位での検証を行い、検証の確度の向上を目指す。また、2D・3D 変換モデルのネットワークを再検討し、より精度の高い姿勢推定を実現したい。更に、本研究の発展として、一部の関節の 3D 姿勢推定の結果を用いた他の関節座標の復元や、2D 姿勢推定で得られた予測の信頼度が高い関節のみを使用した 3D 姿勢推定の有効性を検証したい。

参考文献

[1] J. Martinez, R. Hossain, J. Romero, and J. J. Little, “A simple yet effective baseline for 3d human pose estimation”, Proceedings of the IEEE International Conference on Computer Vision (2017)
[2] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, “3D human pose estimation in video with temporal convolutions and semi-supervised training”, IEEE Conference on Computer Vision and Pattern Recognition (2019)
[3] D. Mehta et al. , “XNect: Real-time Multi-person 3D Human Pose Estimation with a Single RGB Camera”, arXiv preprint arXiv:1907.00837 (2019)
[4] K. He, X. Zhang, S. Ren, J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification”, Proceedings of the IEEE international conference on computer vision (2015)
[5] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, “Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.36, No.7 (2014)
[6] C. Ionescu, F. Li, and C. Sminchisescu, “Latent Structured Models for Human Pose Estimation”, International Conference on Computer Vision (2011)
[7] A. Newell, K. Yang, and J. Deng, “Stacked hourglass networks for human pose estimation”, European conference on computer vision (2016)