

身振り空間の三次元解析に向けた姿勢推定手法 Pose Estimation Method for 3D Analysis of Gesture Space

家永 直人[†], 関根 和生[‡]
Naoto Ienaga, Kazuki Sekine

[†] 筑波大学, [‡] 早稲田大学
University of Tsukuba, Waseda University
ienaga@iit.tsukuba.ac.jp

概要

身振り研究において三次元計測は重要であるが、モーションキャプチャ (MoCap) には限界がある。本研究では、4つの人物姿勢推定手法 (HPE) を、MoCap と比較することで身振りの三次元推定精度を検証した。結果、2台のカメラと HPE による手法がもっとも高精度であり、誤差は 50mm 程度であった。また、クラウド上で利用可能な三次元解析ツールも提案する。本手法は MoCap の代替手段として有望である。

キーワード: 身振り、三次元計測、人物姿勢推定

1. はじめに

身体の動きは三次元空間で行われるにもかかわらず、身振り (ジェスチャー) は主に二次元平面上で測定や解析されてきた。例えば、身振りが行われる空間 (身振り空間) の標準的なモデルは、身体前方を複数の領域に分割する二次元平面として定義されている (McNeill, 1992)。光学式モーションキャプチャ (MoCap) は、身体動作を三次元計測するための一般的な手法であるが、非常に高価であることや、計測に必要な再帰性反射マーカの装着が話者 (特に子ども) の自然な身振りの生成を妨げることがある。

画像から人体のキーポイント (手首、肘、肩といった関節などの注目点) 位置を推定する技術である人物姿勢推定 (Human Pose Estimation, HPE) が、非常に高精度になってきている。そこで本研究では、HPE によるキーポイントの三次元位置推定精度を MoCap と比較し、HPE が MoCap の代替となるかを検証する。

これまで研究では、歩行、走行、跳躍などの比較的単純な動作における比較が多かった (Needham et al., 2021)。より複雑な動作では、投球やポップダンスといった速い三次元動作における比較も行われた (Nakano et al., 2020; Labuguen et al., 2020)。しかし、身振り研究の観点からの検証は不十分である。また手指は、手首、肘、肩といった追跡しやすい他の関節に比べてあまり検証されていない。本研究では従来研究と異なり、身振り研究において主要な解析対象となる

ことが多い手指を含む上半身の 8 点のキーポイントの精度を、現実的なシナリオで検証する。さらに、機械学習の専門家でなくとも使いやすいオープンソースの HPE による三次元解析ツールを提案する。

2. データセット

10名の成人 (21~24 歳、平均 22.1 ± 0.83 歳、男性 4 名) が実験に参加した。全員が日本語を母語とした。本研究は筑波大学の倫理委員会の承認を受けて実施され (審査承認番号 2023R737、2025R005)、各参加者から書面によるインフォームド・コンセントを得た。

参加者は約 1 分間のアニメ映像 (Tweety, Warner Bros.) を 2 回視聴した後、前方のカメラに向かって、できるだけ多くの身振りをういながら映像の内容を説明した。アニメ映像は 8 本であった。参加者は MoCap (OptiTrack Flex 3) と、3 台のビデオカメラ (Panasonic HC-VX992MS, 参加者の正面に 1 台、斜め前方に 2 台) で撮影された (図 1)。3 つの動画は手動で同期された。すべての動画は解像度 1920×1080 ピクセルで、30FPS で録画された。キャリブレーションへの影響を避けるためにオートフォーカス機能を無効にし、ピントは参加者の位置に合わせた。

3. 計測手法

本研究では、精度は高いが実装が煩雑な RTMPose (以下、RTM, version 1.1.0, Jiang et al., 2023) と、実装が容易で GPU も不要な MediaPipe Pose (以下、MP, version 0.9, Lugaresi et al., 2019) を選択した (図 2)。RTM 単体では三次元位置を推定できないため、キーポイントの二次元位置から三次元位置を推定する MocapNET version2 (Qammar and Argyros, 2019) を併用した (以下、RTM1)。なお、MP は単体で三次元推定可能である (以下、MP1)。これらの手法には参加者の正面に設置されたカメラの動画が使用された。なお、RTM1 と MP1 ではスケールを決定できないため、各参加者の推定された肘と肩の距離が、メジャーで計測された実測値と一致するようにした。

近年の HPE の進展にもかかわらず、単視点のみか

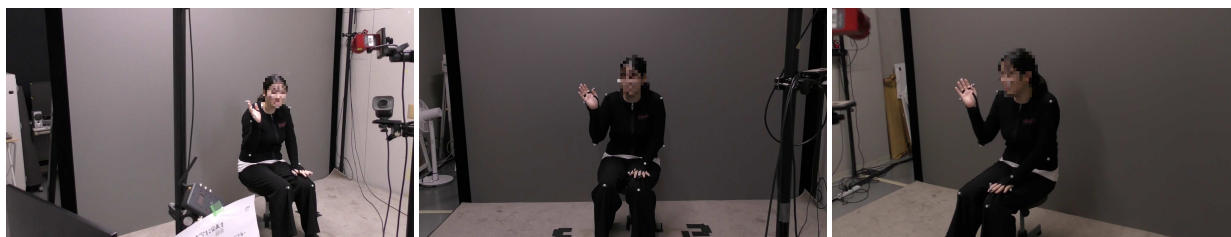


図 1: 3 台のビデオカメラで撮影された参加者の様子。

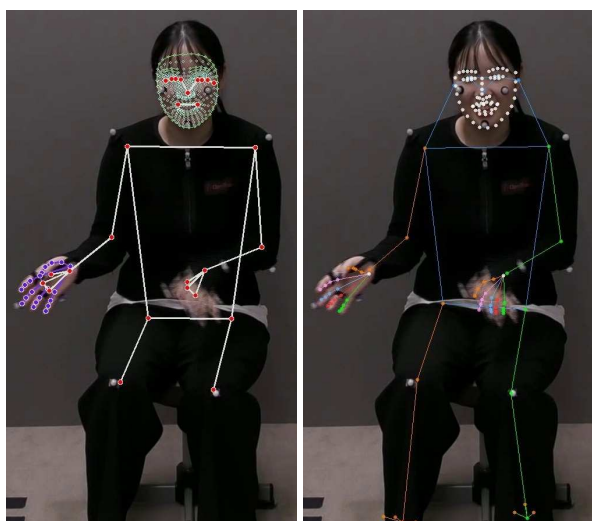


図 2: MP (左) と RTM (右) の例。

らの三次元推定の精度は十分ではない。そのため、2 台のカメラの動画から得たキーポイントの二次元位置を用いて、三角測量により三次元位置を推定する手法も実装した。カメラキャリブレーションには、一般的なチェッカーボードパターンを用いた (Zhang, 2000)。カメラキャリブレーションとは、カメラのレンズ特性と、カメラ間の相対的な位置と姿勢を推定する手法である。以下、三角測量との組み合わせを RTM2、MP2 とする。これらには、参加者の斜め前方に設置された 2 台のカメラの動画を用いた。

MoCap のため、参加者の爪先、膝、親指と中指の先端、小指の付け根、手首、肘、肩、頬、顎に計 19 個の再帰性反射マーカを装着した (参加者 ID1 のみ、頬の代わりに耳たぶにマーカが装着された)。19 か所の内、身振り研究において主に対象とされる上半身の 8 個のキーポイント (左右の親指と中指の先端、手首、肘) を選んだ。各 HPE 手法において、完全に同一位置を表すキーポイントが存在しない場合は、もっとも近い位置のキーポイントを代わりに使用した。各手法で追跡に失敗したキーポイントは分析から除外した。

4. 結果と考察

各キーポイントごとに MoCap と HPE の三次元位置のユークリッド距離を算出し、参加者ごとに平均

値が計算された。統計解析には、Microsoft EXCEL と HAD を用いた (Shimizu, 2016)。

4 つの手法間でユークリッド距離に差があるか検定するために、一要因分散分析を実施した (表 1)。各キーポイントの平均ユークリッド距離を従属変数、手法の種類を独立変数 (参加者内要因) とした。結果、すべてのキーポイントにおいて手法の主効果が認められ、すべての p 値は 0.001 未満であった。Bonferroni の t 検定の結果 ($p < .05$)、すべてのキーポイントにおいて、MP2 と RTM2 は、MP1 と RTM1 よりも有意に誤差が小さいことが示された。一方で、MP と RTM の誤差は全体として類似していた。しかし、MP は特に顔や手において RTM より推定に失敗することが多く、RTM の方が優れている印象であった。

実験結果から、単眼手法 (MP1・RTM1) は大きな誤差を示すことがわかった (図 3)。これは奥行き方向の誤差に起因すると考えられる。一方、両眼手法 (MP2・RTM2) は高精度であった。なお、MoCap で測定されたマーカの位置と、HPE で推定されたキーポイントの位置は完全に一致しない点に注意されたい (図 2)。マーカは骨の位置を基準に装着されたが、HPE は身体内部にある点を検出するためである。また、マーカは皮膚 (身体の外側) にしか取り付けられない。例えば、マーカの位置は身体の向きによって変化するが、HPE のキーポイントの推定位置は向きによってほとんど変化しない。これらを考慮すると、50mm 程度の誤差は高精度であると我々は結論付けた。

本研究結果は、一般的な動画と計算資源が MoCap の代替になり得ることを示している。カメラは MoCap よりも遥かに安価であり、スマートフォンの普及により、もっとも手軽に利用できる機器となっている。設置や撤去が容易であり、再帰性反射マーカを装着する必要がない点も、身振りの研究者にとって利便性が高い。HPE は 100 以上の多数のキーポイントも検出可能であり、多くの研究において十分である。

5. 身振り空間の三次元可視化ツール

本研究で検証した HPE の実用性を高め、身振り研究に携わる多くの研究者がその成果を手軽に活用

表 1: 一要因分散分析の結果。N 列は比較に使われたキーポイントの総数。

キーポイント	N	F	η^2	多重比較
右親指	35,684	81.86***	0.90	MP2<MP1, RTM2<MP1, MP2<RTM1, RTM2<RTM1
左親指	46,095	51.27***	0.85	MP2<MP1, RTM2<MP1, MP2<RTM1, RTM2<RTM1
右中指	51,679	78.11***	0.90	MP2<MP1, RTM2<MP1, MP2<RTM1, RTM2<RTM1
左中指	55,477	62.26***	0.87	MP2<MP1, RTM2<MP1, MP2<RTM1, RTM2<RTM1, MP2<RTM2
右手首	128,650	262.32***	0.97	MP2<MP1, RTM2<MP1, MP2<RTM1, RTM2<RTM1, RTM1<MP1, RTM2<MP2
左手首	132,269	123.63***	0.93	MP2<MP1, RTM2<MP1, MP2<RTM1, RTM2<RTM1
右肘	127,956	102.51***	0.92	MP2<MP1, RTM2<MP1, MP2<RTM1, RTM2<RTM1
左肘	142,859	131.20***	0.94	MP2<MP1, RTM2<MP1, MP2<RTM1, RTM2<RTM1, MP1<RTM1

*** $p < .01$ 。すべてのキーポイントにおいて、df1 = 3, df2 = 27。

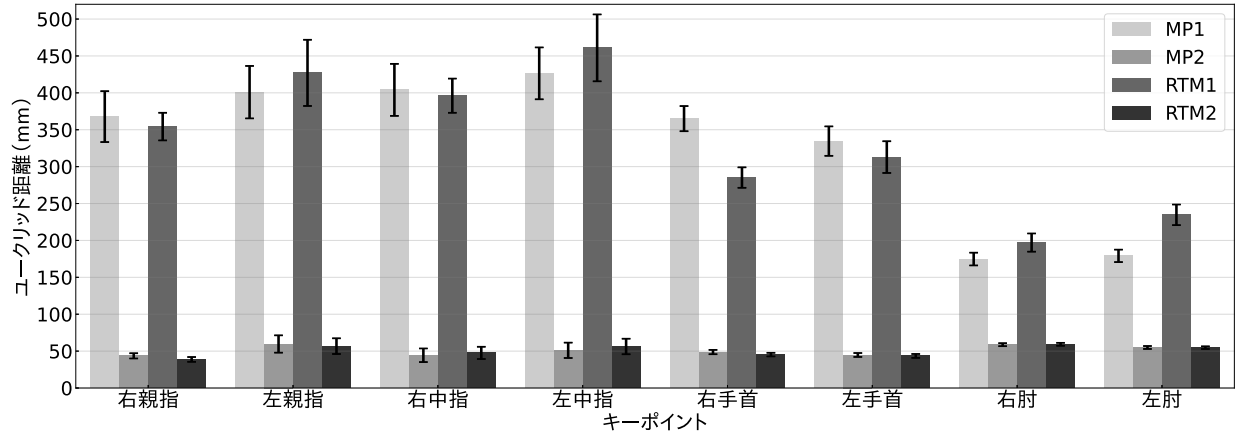


図 3: MoCap と HPE 間の平均ユークリッド距離。エラーバーは 10 名の参加者の標準誤差。

できるよう、クラウドベースの身振り空間三次元可視化ツールを開発した。本ツールは専門的なプログラミング知識や高性能な専用マシンを必要とせず、標準的なコンピュータと Web ブラウザさえあれば、誰でも無償で利用できることを最大の特徴とする。ツールは Google Colaboratory (Colab) 上で動作する 2 つの Notebook から構成され、GitHub 上で公開されている (<https://github.com/naotoienaga/Colab-3D-Motion-Analysis>)。本ツールはカメラキャリブレーションと、姿勢推定と可視化の 2 段階で構成される。以下に簡単に利用方法を説明する。

5.1 カメラキャリブレーション

ユーザーはまず、1 つ目の Notebook を実行してキャリブレーション用ボードの画像を生成し、印刷する。次に、そのボードを動かしながら、2 台のカメラで同時に撮影する。撮影した動画を同期し、Google Drive (Drive) 経由で Colab に読み込ませ、再度 Notebook

を実行するだけで、キャリブレーションの結果が自動で算出、保存される。

5.2 姿勢推定と可視化

キャリブレーション完了後、ユーザーは同じカメラの設定と配置のまま、解析対象となる人の動作を撮影する。撮影した動画を Drive にアップロードし、2 つ目の Notebook を実行する。HPE により動画の各フレームから人のキーポイントが検出され、三角測量に基づいて三次元座標が計算される。全ての処理が完了すると、手首の移動軌跡と身振り空間の三次元プロットが描画される (図 4)。

本ツールはオープンソースであるため、ユーザーが必要に応じてプログラムを自由に改変できる。本ツールは、これまで高価な機材や専門知識を要した身振りの三次元解析を、誰もが手軽に実施できる機会を提供することで、当該分野の研究の裾野を広げることに貢献するものである。

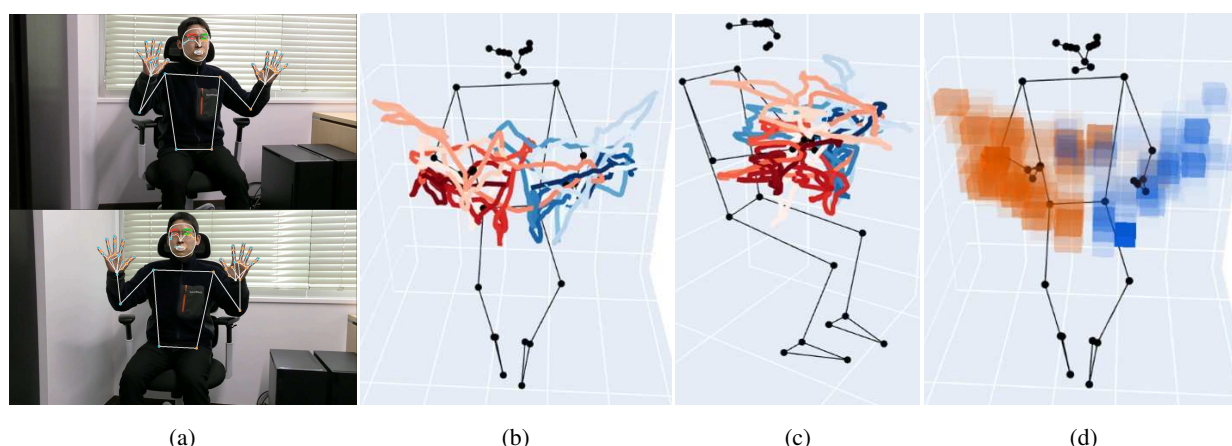


図 4: 三次元解析ツールの動作例。(a)2 台のカメラによる入力動画。(b) 手首の移動軌跡。話者のキーポイントの平均位置（黒い骨格）と共に、両手首（右手は赤、左手は青）が三次元空間内で描いた軌跡が描画される。軌跡の色は時間の経過と共に濃くなる。(c) 手首の移動軌跡を別の視点から見た図。プロットはマウス操作で自由に回転でき、さまざまな角度から観察できる。(d) 身振り空間。両手首が空間内で占有した時間を、色の濃淡を持つボクセルで表現する。身振りが特に頻繁に行われた空間がわかる。

6. 結論

本研究では人物姿勢推定手法によるキーポイントの三次元推定が、モーションキャプチャに代わる手法として有望であることを示した。特に非常に高精度な計測が必須ではない場面では、設置や運用が容易で、マーカーが不要な人物姿勢推定手法の利点は大きい。今後はカメラの台数や配置の最適化など、より実用的なシステム構築に向けた検討が必要である。また、これまでの二次元解析により確立されてきた知見を、三次元解析により再検証することも意義深いであろう。

謝辞

本研究は 2023 年度電気通信普及財団研究調査助成金ならびに JSPS 科研費 JP25K21528 の助成を受けたものです。この場を借りて深く御礼申し上げます。

文 献

- Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. RtmPose: Real-time multi-person pose estimation based on mm-pose. *arXiv preprint arXiv:2303.07399*, 2023.
- Rollyn T Labuguen, Salvador Blanco Negrete, Tonan Kogami, Wally Enrico M Ingco, and Tomohiro Shibata. Performance evaluation of markerless 3d skeleton pose estimates with pop dance motion sequence. In *International Conference on Informatics, Electronics & Vision and International Conference on Imaging, Vision & Pattern Recognition*, pages 1–7. IEEE, 2020.
- Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris

McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

David McNeill. *Hand and mind: What gestures reveal about thought*. University of Chicago press, 1992.

Nobuyasu Nakano, Tetsuro Sakura, Kazuhiro Ueda, Leon Omura, Arata Kimura, Yoichi Iino, Senshi Fukushima, and Shinsuke Yoshioka. Evaluation of 3d markerless motion capture accuracy using openpose with multiple video cameras. *Frontiers in sports and active living*, 2:50, 2020.

Laurie Needham, Murray Evans, Darren P Cosker, Logan Wade, Polly M McGuigan, James L Bilzon, and Steffi L Colyer. The accuracy of several pose estimation methods for 3d joint centre localisation. *Scientific reports*, 11(1):20673, 2021.

Ammar Qammaz and Antonis A Argyros. Mocapnet: Ensemble of snn encoders for 3d human pose estimation in rgb images. In *BMVC*, page 46, 2019.

Hiroshi Shimizu. An introduction to the statistical free software had: Suggestions to improve teaching, learning and practice data analysis (*Japanese*). *Journal of Media, Information and Communication*, 1:59–73, 2016.

Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334, 2000.