H-022

MediaPipe を用いた多視点画像による 3D ポーズ推定法の検討

Examination of 3D pose estimation method from multi-view images using MediaPipe

徳田　泰地†　　弘中　哲夫†
Tokuda Taichi　Hironaka Tetsuo

1. はじめに

3D ポーズ推定とは，画像から被写体の姿勢を推定する技術であり，関節や部位の空間的な位置（ランドマーク）を求めることでこれを行う．この技術を用いて推定した姿勢を仮想の身体に連動させることで，仮想の身体を自分の身体のように動かすことが可能になる．この際，手や足などのより自身の身体の細かな部位を反映するため，該当する部位も含めて正確に推定できることが求められる．

正確に被写体のランドマークを推定できる方法として，複数の視点を用いた 3D ポーズ推定法がある．しかし，現在研究されている多くの手法[1][2][3][4][5][6][7][8]は 13～17 個のランドマーク配置が主流となっており，これらのランドマーク配置では手首よりも先を推定することができない．また，体が隠れる状況（オクルージョン）やカメラの画角による見切れに対応するため，多数のカメラを必要とする場合がある．その場合，設置の労力や使用スペースが増大し，使用できる状況が限られてしまう．

よって，本研究では少ないカメラで従来よりも多くのランドマークを正確に推定できる 3D ポーズ推定法の開発を目的とし，上記の目的を達成する方法として，三角測量と Google 社が開発した MediaPipe Pose[9]による 2D ランドマークの推定を組み合わせた 3D ポーズ推定法を提案する．その後，精度の評価を行い，提案手法の有用性を検討する．

2. 提案手法

図 1 に提案手法の概要を示す．まず，入力には，撮影タイミングが同期された 2 つの画像と撮影を行ったそれぞれのカメラの内部パラメータ，外部パラメータを使用する．

そして，2D ランドマーク推定により各画像から画像上の関節の位置を示す 2D ランドマークを推定し，3D ランドマークの復元では推定した 2D ランドマークを実際の空間上の位置へと復元する．次節からは 2D ランドマーク推定と 3D ランドマークの復元の 2 つの処理について詳細に記述する．

2.1 2D ランドマークの推定

前章で述べた問題点のうち，オクルージョンや見切れによる必要カメラ数の増加を解決するため，MediaPipe Pose [9]を用いた 2D ランドマークの推定が有効だと考える．

MediaPipe Pose は手足の末端や一部分が隠れる程度の軽度のオクルージョンに対して正確に推定が行えるため，死角が発生した部分に対してもある程度の精度で推定できることが期待される．また，もう一つの特徴として，オクルージョンが発生したとしても推定を中断しないという特徴があるため，前述の内容と合わせて，対象の部位の隠蔽や画角による見切れが発生すると大幅に精度が低下する，または推定できないという問題点に対応できる可能性が高い．

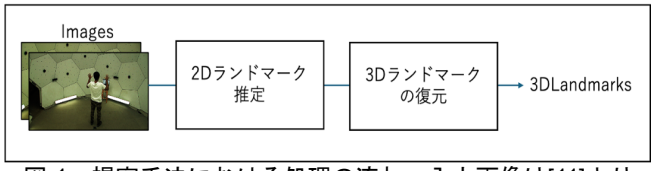

図 1：提案手法における処理の流れ．入力画像は[11]より

こうして推定したランドマーク出力のうち，奥行きとなる Z 成分を除く 2 次元の情報のみを取り出し，2D ランドマークの検出器として使用する．この時，ランドマークの XY 成分はそれぞれ画像の幅と高さで正規化されている．このままでは入力画像のアスペクト比が考慮されないため，画像の幅と高さに各ランドマークをスケーリングする．

2.2 三角測量を用いた 3D ランドマークの復元

図 1 の 3D ランドマークの復元は画像平面と実空間の対応関係から三角測量を用いて画像上の特徴点を実際の位置に復元する処理である．推定された 2D ランドマークとカメラの位置を結んだ直線同士の交点を求めることで 3D 点群の復元を行う．

なお，2 台のカメラでも正確に復元することが理論上可能であるが，ベースライン長や視線角度，対象との距離などの設置条件によって精度が大きく左右されることが知られている[10]．特に，ベースラインが短い場合や，視線方向が平行に近い配置では視差が小さくなり，奥行き方向の精度が著しく低下する可能性がある．したがって，実装上は十分な視差を得るためのカメラ配置が重要となる．

3. 評価

本章では当手法の精度を評価する方法とその結果を述べる．まず，実験方法から説明し，その実験結果と入力との関連性を述べる．

3.1 評価方法

本研究では CMU Panoptic データセット[12]の中からある 2 つの動画を使用する．動画は指定された動きを行う 2 人の男性が撮影されており，一人が指定された動きを行った後，交代して同じ動きを行なっている．これを様々な角度から撮影した動画から，図 2 のように動画全体で見切れやオクルージョンが発生していない 2 つの動画を用いる．

†広島市立大学大学院情報科学研究科 Graduate School of Information Sciences, Hiroshima City University

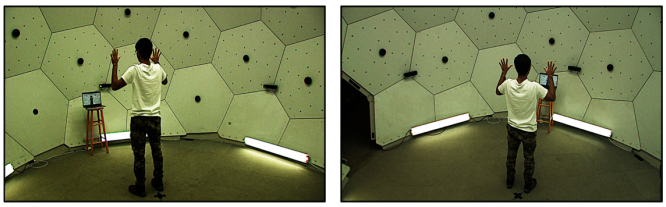

図 2：入力する 2 つの動画の 1 フレーム[11]．視線方向が交わっており，ベースライン長も短くない動画ペアを使用

評価は精度と視認性の二つを軸に行う．方法としては，二つの動画から 1 フレームを取り出し，当手法の 3D ポーズ推定を行う．得られた 3D ランドマークとデータセット内の正解の 3D ランドマークを同一プロット内で可視化し，目視によりそれぞれの評価を行う．

精度の評価は 3 段階に分かれ，正解データに極めて近い場合を評価 A，ズレは確認できるが，全体として妥当な場合を評価 B，これら以外で大きくズレが発生している場合を評価 C とする． 対象としては，全身，右手，左手の 3 つについて行った．

また，オクルージョンの影響を確かめるため，各フレームにおいて対象の部位の視認性も評価した．以下の定義で 3 段階に分けて評価を行う．対象の部位を両方で確認できる場合は FV(Fully Visible)，片方のカメラのみ確認できる場合は PV(Partially Visible)，両方のカメラで確認できない場合（8 割以上隠れている場合を含む）を NV(Not Visible) とする．この評価は右手，左手でそれぞれ行う．

正解のデータが 137 フレーム目から存在したため，137 フレーム目から上記を 100 フレームごとに行い，計 90 フレームのうち評価が不可能な 6 フレームを除いた 84 フレームの評価を行なった．

表 1：全身，右手，左手の精度評価結果

	全身		右手		左手	
評価	フレーム数[枚]	確率[%]	フレーム数	確率	フレーム数	確率
A	60	71.4	39	58.9	38	59.5
B	19	22.6	33	25.6	35	29.2
C	5	6.0	12	15.5	11	11.3

表 2：精度・オクルージョン評価のフレーム数

	オクルージョン					
精度	右手			左手		
	FV	PV	NV	FV	PV	NV
A	27	12	0	26	11	0
B	11	18	4	21	14	0
C	1	5	6	1	7	3

3.2 結果

表 1 は全身，右手，左手の各評価のフレーム数とその割合を記述したものである．結果によると，評価 A のフレームは約 71%を占め，評価 B のフレームも含めると 90%以上のフレームが正解データに近い推定ができていることが分かる． 一方，左手，右手については 60%近くが正解データに極めて近く，評価 B を含めると 8 割以上が正確に推定できている可能性が高いことが確認できる．

表 2 は右手左手のオクルージョン度合いと精度の各評価の内訳を示す表である．これによると，オクルージョンの評価 FV のフレームは半数近くが評価 A となっており，評価 B のフレームも加えると，ほとんどはデータセットに近い推定ができる可能性が高い．一方，評価 PVの場合では，評価 A の割合が下がったが，約 86%の該当フレームが精度面で評価 A または評価 B となっていた．これらのフレームは，手は片方のカメラで見えていないが腕は両方のカメラで視認できており，上記のような場合に評価 A となる傾向を確認した．

4. 考察・まとめ

本研究では，少ないカメラで従来よりも多くのランドマークを正確に推定できる 3D ポーズ推定法の開発を目的とし，上記の目的を達成する方法として，三角測量を用いた 3D ランドマークの復元に MediaPipe Pose による 2D ランドマークの推定を組み合わせた 3D ポーズ推定法を検討した．

目視評価ではあるが，結果から全身と両手ともに約 80%の確率で正確に推定ができていることから，従来よりも多い数のランドマークを正確に推定できる可能性があることが分かる．

また，オクルージョンの具合が評価 PV であるフレームも腕まで見えていれば評価 A となることから，このような軽度のオクルージョンが発生している画像に対しても大きく精度が落ちず合理的な推定が可能である可能性が高い．このような軽度なオクルージョンに対する堅牢性は撮影時の死角を補完し，実験のように最低限のカメラでも正解データに近い推定ができる可能性を示唆した．

今後は，より厳密な数値評価を行う他，より実用性を高めるために 3 台以上での利用や，外部パラメータの推定を組み込みたい．

参考文献

[1] Jinbao Wang, et al., ”Deep 3D human pose estimation: A review,” Computer Vision and Image Understanding, 2021.
[2] Amal El Kaid, et al., ”A Systematic Review of Recent Deep Learning Approaches for 3D Human Pose Estimation,” Journal of Computer Vision, 2023.
[3] K. Iskakov, et al., “ Learnable triangulation of human pose, ” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2019.
[4] Tao Wang, et al., ”Direct multi-view multi-person 3d human pose estimation,” Advances in Neural Information Processing Systems, 2021.
[5] Xiaoyue Wan, et al., ”View consistency aware holistic triangulation for 3d human pose estimation,” Computer Vision and Image Understanding 236, 103830, 2023.
[6] Muhammed Kocabas, et al., ”Self-Supervised Learning of 3D Human Pose using Multi-view Geometry,” arXiv preprint arXiv:1903.02330, 2019.
[7] Jinbao Wang, et al., ”Markerless body motion capturing for 3D character animation based on multi-view cameras,” arXiv:2212.05788, 2022.
[8] Chung, Jen-Li, Lee-Yeng Ong, and Meng-Chew Leow, "Comparative Analysis of Skeleton-Based Human Pose Estimation" Future Internet 14, no. 12: 380. https://doi.org/10.3390/fi14120380, 2022.
[9] Valentin Bazarevsky, et al., ”BlazePose: On-device Real-time Body Pose tracking,” arXiv preprint arXiv:2006.10204, 2020.
[10] 公益財団法人 画像情報教育振興協会 (CG-ARTS) 奥富 正敏ら,” ディジタル画像 処理 [改訂第二版],” 2022.
[11] Hanbyul Joo, et al., ”Panoptic Studio: A Massively Multiview System for Social Interaction Capture,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 41, Issue 1, 2019.