

Visual-IMU 融合型モーションキャプチャにおける 関節別重みづけ手法の提案

2311045 大塚拓哉 成見研究室

1 はじめに

モーションキャプチャは、人の動きを三次元の関節位置として計測する技術である。ゲームやアニメーション制作など様々な分野で利用が増えてきており、需要が高まっている。

本研究では、カメラ映像を用いるマーカーレス光学式と、身体に装着した慣性計測装置 (IMU) を用いる慣性式を扱う。マーカーレス光学式は導入しやすい一方、身体の影響や照明変化によって精度が低下する。慣性式は画像上の遮蔽に強いが、センサが少ないと全身姿勢が曖昧になり、データ欠損や装着方向のずれも誤差の原因となる。そこで、カメラと IMU の異なる長所を組み合わせる Visual-IMU 融合が研究されている [1, 2]。

既存手法の RobustCap は、各関節のカメラデータから得た信頼度を平均し、全身共通の融合判断に用いる [2]。しかし、片腕だけが隠れた場合には、脚などのカメラ情報は正常に得られる。全身共通の信頼度では、このような局所的な劣化が起きた関節と、正常に観測できている関節を個別に扱えない。

本研究では、カメラ側と IMU 側の信頼度を関節ごとに求め、関節別に融合比率を変える。これにより、各関節で信頼度の高い方式の推定結果を強く反映し、局所的に劣化した関節の誤差を減らしながら、正常な関節の精度を維持することを目的とする。

2 関連技術

マーカーレス光学式は、カメラ映像から人物の関節位置を推定して姿勢を求める方式である。身体に専用のマーカーを取り付ける必要がなく、必要な機材も比較的少ない。

この方式では、映像上の人物から頭、肩、肘、手首などの二次元関節位置を検出し、その並びから三次元姿勢を推定する。一方、関節がほかの身体部位に隠れる場合や、照明によって人物の見え方が変化する場合などは、正しい位置を検出にくい。

慣性式は、身体に装着した IMU から向きと加速度を取得して姿勢を求める方式である。センサーが身体や衣服で隠れても計測が続けられるが、センサを装着していない部位は限られた情報から補う必要がある。

IMU (慣性計測装置) は、加速度センサとジャイロセンサを組み合わせた小型のセンサである。加速度センサは重力を含む動きの変化を、ジャイロセンサは回転の速さを計測

する。

本研究では、頭部、骨盤、左右の腕、左右の脚に相当する身体 6 か所へ IMU を装着する構成を想定する。

GRU は、過去の情報を保持しながら、時間的に連続するデータを処理するニューラルネットワークである [3]。人の動きやセンサの計測値は時間とともに連続して変化するため、本研究ではカメラ側と IMU 側の姿勢推定に GRU を用いる。

信頼度は、GRU が推定した関節位置をどの程度信用できるかを表す値である。

3 関連研究

複数視点画像と IMU を組み合わせる手法として、Gilbert ら [1] は、視覚情報と IMU から求めた姿勢を最終段で融合する手法を提案した。Huang ら [4] は、DeepFuse を提案し、視覚の三次元ボリュームへ IMU から得た骨方向を組み込んだ。Bao ら [5] は、FusePose を提案し、視覚と IMU の情報を人体のつながりを表す運動学空間で融合した。

単眼画像と少数の IMU を組み合わせる手法として、Pan ら [2] は、RobustCap を提案した。RobustCap は、各関節のカメラデータから得た信頼度を平均し、全身共通の融合判断に用いる。Bao ら [6] は、単眼映像と少数の IMU を組み合わせる Hybrid 3D Human Pose Estimation を提案した。Chen ら [7] は、SparseNet を提案し、カメラ側と IMU 側の姿勢を確率分布として推定して、その不確実性を考慮して融合した。したがって、信頼度や不確実性に応じて融合の重みを変える考え方自体には先行例がある。

4 提案手法

提案手法は、カメラ用 GRU、IMU 用 GRU、関節別融合の三つの処理から構成する。(図 1)

同じ動作でも、遮蔽が起きた関節ではカメラ側の信頼度が低下するが、IMU 側の信頼度は高い場合がある。関節ごとの状態を表す信頼度を融合判断に利用することが、本研究の基本的な考え方である。

カメラ側では、映像から事前に抽出した二次元関節位置、検出信頼度、欠測情報を入力する。IMU 側では、身体 6 か所に装着した IMU から得られる向き、加速度、欠測情報を入力する。

検出信頼度と欠測情報を与えることで、遮蔽や検出失敗をカメラ用 GRU が識別できるようにする。IMU 用 GRU には

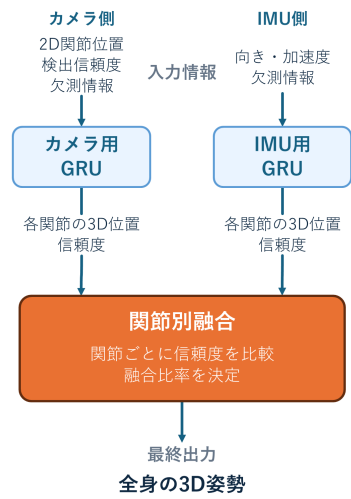


図 1 提案システムの概略

各センサの計測値と欠測情報を与える．各 GRU は，過去の入力から引き継ぐ内部状態を用いて，時間的な動きのつながりから全身姿勢を推定する．

カメラ用 GRU と IMU 用 GRU は，それぞれ全身の各関節について，三次元位置と信頼度を推定する．二つの GRU を分けることで，カメラと IMU で性質の異なる情報を個別に処理する．学習時には，三次元位置の推定誤差が大きい出力ほど信頼度が低くなるようにする．各 GRU が全関節の三次元位置を出力するため，二つの推定結果を同じ関節ごとに比較できる．また，信頼度も関節ごとに出力することで，全身を一つの信頼度で判断しない．

関節別融合では，各関節についてカメラ側と IMU 側の信頼度を比較し，信頼度が高い側を強く反映するように融合比率を決める．例えば，左手首のカメラ情報だけが遮蔽によって劣化した場合，左手首では IMU 側を強く反映する．一方，カメラから正常に観測できる関節では，カメラ側の推定結果を引き続き利用する．最後に，関節別に融合した結果から全身の三次元姿勢を出力する．

この処理では，信頼度の大小だけで一方を完全に選択するのではなく，二つの推定結果を信頼度に応じた比率で組み合わせる．そのため，両方式の情報を利用しながら，局所的に信頼度が低下した情報の影響を小さくできる．

5 評価方法

今後，公開されている TotalCapture データセット [8] を用いてモデルを実装し，評価実験を行う．評価では，提案する関節別の融合と，RobustCap のように全身共通の信頼度を用いる融合を比較する．

比較には，カメラだけで姿勢を推定する方式，IMU だけで姿勢を推定する方式，二つの方式を一定の比率で融合する方式も用いる．これにより，二つのセンサを組み合わせる効

表 1 使用予定のデータセット [8]

データセットの内容	
計測データ	8 視点映像，13 個の IMU，Vicon 正解姿勢
被験者	男性 4 人，女性 1 人
動作の種類	4 種類
1 動作辺りのデータ量	約 3,000～5,000 フレーム

果と，関節ごとに融合比率を変える効果を分けて確認する．また，身体の一部でカメラ情報または IMU 情報を人工的に劣化させた条件を用意する．ここで劣化した情報とは，データの欠損やセンサーの装着ズレを想定している．劣化した関節では誤差を小さくできるか，正常な関節では精度を維持できるかを分けて確認する．これにより，局所的な劣化に対して関節別に融合比率を変える効果を調べる．

全身の平均誤差だけでは，劣化した関節が改善しても，正常な関節の精度低下を見落とす可能性がある．そこで，劣化した関節と正常な関節を分けて評価し，目的どおりに影響を局所化できているかを確認する．

6 おわりに

本研究では，局所的な劣化に強い Visual-IMU 融合方法の確立を目指し，関節ごとの信頼度に基づく重みづけ手法を提案する．カメラ用 GRU と IMU 用 GRU が出力する三次元関節位置と信頼度を用い，信頼度の高い方式を関節ごとに強く反映することで，局所的な劣化が生じた関節の誤差低減と，正常な関節の精度維持を目指す．

現在までに，関連研究の調査，研究方針の決定，モデルの基本設計を行った．今後は二つの GRU と関節別融合処理を実装し，データセットを用いて，局所的な劣化が起きた場合の姿勢推定精度を評価する．

参考文献

[1] Andrew Gilbert, Matthew Trumble, Charles Malleson, Adrian Hilton, and John Collomosse. Fusing visual and inertial sensors with semantics for 3D human pose estimation. *International Journal of Computer Vision*, Vol. 127, pp. 381–397, 2019.

[2] Shaohua Pan, Qi Ma, Xinyu Yi, Weifeng Hu, Xiong Wang, Xingkang Zhou, Jijunnan Li, and Feng Xu. Fusing monocular images and sparse IMU signals for real-time human motion capture. *ACM Transactions on Graphics*, Vol. 42, No. 6, pp. 204:1–204:10, 2023.

[3] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555, 2014.

[4] Fuyang Huang, Ailing Zeng, Minhao Liu, Qiuxia Lai, and Qiang Xu. DeepFuse: An IMU-aware network for real-time 3D human pose estimation from multi-view image. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 429–438, 2020.

[5] Yiming Bao, Xu Zhao, and Dahong Qian. FusePose: IMU-vision sensor fusion in kinematic space for parametric human pose estimation. arXiv preprint arXiv:2208.11960, 2022.

[6] Yiming Bao, Xu Zhao, and Dahong Qian. Hybrid 3D human pose estimation with monocular video and sparse IMUs. arXiv preprint arXiv:2404.17837, 2024.

[7] Xiaodong Chen, Wu Liu, Qian Bao, Xinchun Liu, Ruoli Dai, Yongdong Zhang, and Tao Mei. Motion capture from inertial and vision sensors. arXiv preprint arXiv:2407.16341, 2024. arXiv:2407.16341v4, revised 2026-04-03.

[8] CVSSP. TotalCapture Dataset. <https://cvssp.org/data/totalcapture/>. 参照日：2026 年 9 月 29 日.