7R-01

VR ダンストレーニングのための 複数カメラを用いたモーションキャプチャシステム

江嵜　一優†　　長尾　確‡
名古屋大学　大学院情報学研究科†‡

1. はじめに

バーチャルリアリティは近年、医療現場や工場における教育や安全講習、訓練などで盛んに利用されており、スポーツのトレーニングでも徐々に利用され始めている。従来は専用のスーツを装着してモーショントラッキングを行う OptiTrack[1]等が用いられていた。最近ではカメラによる人物の姿勢推定技術を用いた動作の分析も行われており、デバイス等を装着する必要がないため、より現実に近い状況でトレーニングが可能となる。本研究では、VR を用いてダンスのトレーニングを行うことを目的とし、複数のカメラを用いたリアルタイムの全身トラッキング手法を提案する。

2. 関連研究

画像を用いた人物の姿勢推定技術は多く行われている。Bajpai and Joshi は、機械学習を用いた単眼カメラによる人体骨格検出手法である MoveNet[2]を提案した。彼らは COCO データセット[3]に加え、ヨガ・フィットネス・ダンスなどの大きく姿勢が変化する動画やモーションブラーをデータとして用い、動きが激しい動画での検出精度を向上させた。

複数のカメラを用いることで 3 次元の人物姿勢を推定する手法も研究されている。Iskakov らは、複数の 2D ビューから 3D 情報を結合する学習可能な三角測量法に基づく 3D 姿勢推定の仕組みを提案した[4]。既存の代数的三角測量に加え、信頼度重みを用いた代数的手法と、3 次元畳み込みにより 2D 姿勢推定からの特徴マップを 3D ボリューム上に投影する手法を提案し、最も高い精度を達成した。

3. モーションキャプチャシステム

モーションキャプチャシステムでは複数のカメラを用いて全身のボディトラッキングを行い、VR 内のアバターにリアルタイムに反映させる。まず単眼のカメラ画像から CNN を用いて人物の 2D 姿勢を推定する。それを複数視点で同時に行い、三角測量により 3D 姿勢に変換する。3D 姿勢データは UDP プロトコルにより VR アプリに送信され、FK(Forward Kinematics)[5]を用いて VR アバターの動きに変換される。

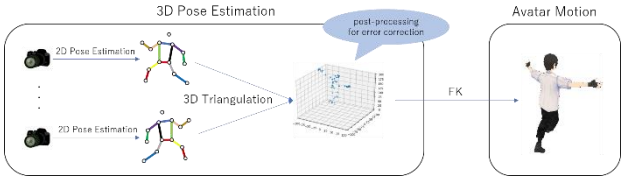

図 1：モーションキャプチャシステムの処理フロー

3. 2D および 3D 人物姿勢推定

2D 姿勢推定では、まずカメラ画像から人物領域を抽出する。224×224 のグレースケール画像を入力として各キーポイントの 2 次元ヒートマップを出力し、重み付き平均を用いてキーポイントの 2D 座標を検出する、2D 姿勢はその後の 3D 姿勢推定に利用されるだけでなく、次のフレームでの人物領域の抽出にも使用される。

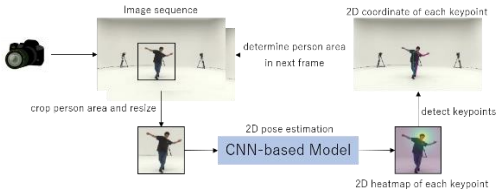

図 2：2D 姿勢推定のフロー

ここでの姿勢推定は VR における全身トラッキングを目的とするため、リアルタイム性が求められる。そのため、推論速度を重視した MobileNet V2 を特徴抽出器として使用し、そのあとに逆畳み込み層を重ねたシンプルなエンコーダ・デコーダモデルを使用する。

次に、推定した 2D 姿勢とカメラパラメータを用いて 3D 姿勢の推論を行う。Iskakov らの手法に倣い、2D 姿勢推定モデルはキーポイントとともに信頼値を出力し、重み付き連立方程式を解くことで 3D 姿勢を推定する。

Multi-Camera Motion Capture System for VR Dance Training
†ESAKI, Kazuhiro (esaki.kazuhiro.m7@s.mail.nagoya-u.ac.jp)
‡NAGAO, Katashi (nagao@i.nagoya-u.ac.jp)
†‡Graduate School of Informatics, Nagoya University

4. 実験

姿勢推定モデルの学習にはダンスに特化した動画データベースである AIST データセット[6]およびそれを拡張した AIST++[7]を用いた。AIST データセットは30名のダンサーによるダンスを各9視点から撮影した約 1010 万枚の画像を含むデータベースである。AIST++はそれに対して 2D および 3D キーポイントアノテーション（COCO フォーマット）とカメラパラメータを提供しており、3D キーポイントのアノテーションを持つ既存のデータセットとしては最大かつ最も豊富なものである。

2D 姿勢推定の学習は、損失関数としてピクセル毎の平均二乗誤差(MSE)を用い、評価指標として各キーポイントに対して正解座標と推定座標の誤差（L2 ノルム）の平均である MPJPE を用いた。同じ 2D 姿勢推定のモデルである MoveNet を比較対象として用いた。

表 1：2D 姿勢推定モデルの学習結果

Method	MPJPE (pixel)	Execution time (ms)
MoveNet	4.9	25
提案手法	8.8	9

表 1 に示すように、MoveNet と比較して精度は 4.9pixel から 8.8pixel と誤差が大きくなったが、推論時間は 25ms から 9ms に向上した。

3D 姿勢推定モデルの学習は、損失関数、評価指標ともに MPJPE を用いた。比較対象として、同じデータセットを用いて学習し、かつ効率性を重視した手法である DeciWatch[8]を用いた。

表 2：3D 姿勢推定モデルの学習結果

Method	MPJPE (mm)
DeciWatch	67.2
提案手法	50.8

表 2 に示すように、提案手法の方がより高い精度で 3D 姿勢推定を行うことができた。

5. VR アバターへの適用

3D 姿勢推定ではフレーム毎に推論を行うのでノイズが含まれる。動きを自然に見せるため、直前の N フレームを用いて移動平均を適用した。また姿勢推定には時間を要するため、遅延が生じる。そこで、タイムスタンプを用いて過去の姿勢から現在の姿勢を等速運動により予測することで、遅延を制御可能にした。これらはパラメータとしてユーザが調整することができる。

VR アバターはボーン構造で表現されており、ボーンの長さなどの骨格情報とボーンの回転でアバターの姿勢を表現する。一方でカメラによる姿勢推定は各関節点の位置座標を推定するため、これを VR アバターへ適用するために FK を用いて回転情報へ変換する。

図 3：VR アバターへの適用

6. おわりに

3D 姿勢推定技術の研究は盛んに行われており、高精度に 3D 姿勢の推定が可能になってきた。一方で、VR 用トラッキングとして利用するためにはリアルタイム性は必要不可欠である。本研究では精度と推論速度のトレードオフを考慮し、実用可能なモーションキャプチャシステムを作成した。今後はこのシステムを用いて VR 内でダンストレーニングの実験を行い、利用データやユーザからのフィードバックに基づいてシステムの改善を行っていく予定である。

参考文献

[1] OptiTrack, Motion Capture Systems, https://www.optitrack.com/
[2] R. Bajpai and D. Joshi, "MoveNet: A Deep Neural Network for Joint Profile Prediction Across Variable Walking Speeds and Slopes," IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-11, 2021, Art no. 2508511, doi: 10.1109/TIM.2021.3073720.
[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays,
Pietro Perona, Deva Ramanan, C. Lawrence Zitnick & Piotr Dollár (2014), Microsoft COCO:
Common Objects in Context, arXiv:1405.0312, https://cocodataset.org.
[4] Iskakov, Karim, Burkov, Egor, Lempitsky, Victor and Malkov, Yury, Learnable Triangulation of Human Pose, International Conference on Computer Vision (ICCV), 2019.
[5] Paul, Richard. Robot manipulators: mathematics, programming, and control: the computer control of robot manipulators. MIT Press, Cambridge, Massachusetts. 1981. ISBN 978-0-262-16082-7.
[6] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, Masataka Goto, AIST Dance Video Database: Multi-genre, Multi-dancer, and Multi-camera Database for Dance Information Processing, Proceedings of the 20th International Society for Music Information Retrieval Conference, pp. 501-510, 2019
[7] Ruilong Li and Shan Yang and David A. Ross and Angjoo Kanazawa, AI Choreographer: Music Conditioned 3D Dance Generation with AIST++, ICCV, 2021.
[8] Zeng, Ailing, Ju, Xuan, Yang, Lei, Gao, Ruiyuan, Zhu, Xizhou, Dai, Bo, Xu, Qiang, “DeciWatch: A Simple Baseline for 10x Efficient 2D and 3D Pose Estimation”, European Conference on Computer Vision, 2022.