モバイル端末向け 3 次元人体姿勢推定における時間軸情報の活用

中溝 雄斗[1,a)]　柳井 啓司[1,b)]

概要

近年，VR/AR，スポーツ解析などの普及を受け，リアルタイムにユーザ動作を反映するための基盤技術として 3 次元人体姿勢推定技術の需要が急速に拡大している。しかしながら，既存のモデルの多くは依然として，計算量の課題からモバイル端末への普及はあまり進んでおらず，一部のモバイル志向のモデルに関しては，フレーム単位での姿勢推定を前提としており，その精度に課題が残る。そこで本研究では，時間軸情報を加味したモバイル端末向け 3 次元人体姿勢推定モデルを新たに提案する。提案手法では，従来のフレーム単位での姿勢推定の後段に時間軸を考慮した姿勢リファインメントモジュールを新たに導入することで効率的に時間軸を考慮した姿勢推定を実現した。提案手法はリアルタイム性を維持しつつも，3 次元姿勢推定精度でベースラインを上回る性能を示した。

1. はじめに

3 次元人体姿勢推定は VR/AR，スポーツ解析，遠隔リハビリなどでリアルタイムにユーザ動作を反映する基盤技術として需要が急速に拡大しており，実際にこれらの分野では姿勢推定を活用した商用サービスや製品が年々増加している。一方で，商用化が進む今日においても，3 次元人体姿勢推定モデルの多くは依然として GPU 搭載の PC やクラウドサーバーなどの計算基盤を必要としており，計算量の課題からモバイル端末への普及はあまり進んでいない。

研究としては，特に 2 次元姿勢シーケンスから 3 次元姿勢シーケンスへとリフトする形式のモデルにおいて，時間的文脈を活用した Transformer ベースモデルが大きく進展している。STCFormer [12] は時間方向の自己注意と空間方向の自己注意を並列的に処理する Spatio-Temporal Criss-Cross Attention を提案し，当時の SOTA を達成した。また，KTPFormer [9] はキネマティクスと各関節位置の軌跡の事前知識を明示的に組み込んだ自己注意機構を提案し，関節間の時空間関係を効率的に捉えることで優れた性能を残している。これらの研究は一般により多くのフレームを入力した際に，性能が向上する傾向を示しており，3 次元人体姿勢推定における時間軸情報の有用性を示している。

一方で，モバイル実装を志向した軽量モデルは主に単一フレーム入力にとどまる。MobileHumanPose [4] は MobileNetV2 [11] ベースのバックボーンを採用し，Galaxy S20 GPU 上で 12ms 以内の推論を実現したが，その性能は日常利用には不十分な程度に留まった。ConvNeXtPose [8] も AR フィットネス用途で実用性を示したものの，フレーム独立型であり，その精度は 2022 年時点の PC 向けモデルと同程度であり，依然として改善の余地が残る。

本研究は，時間軸を考慮した高精度性と，モバイル端末上でのリアルタイム推論を両立する新たな 3 次元姿勢推定アーキテクチャを提案し，高性能な時系列処理を必要とする現実環境においても，軽量かつ継続的に動作可能な 3D 姿勢推定を実現することを目的とする。本手法により，従来困難であったオンデバイスでの高精度・低遅延な姿勢推定が可能となり，幅広いモバイルアプリケーションへの応用が期待される。

2. 関連研究

2.1 3 次元姿勢推定

既存の 3 次元人体姿勢推定手法は，主に 1 段階手法と 2 段階手法に分類できる。1 段階手法とは，入力された動画像や画像列から直接 3 次元関節位置を回帰・推定する手法であり，HMR 2.0 [5] などが該当する。この系統の手法は，画像から直接空間情報を抽出するため，エンドツーエンドに最適化可能であるという利点がある。

また, 特に動画入力に焦点を当てたものとしては, IVT [10] などが該当し，これらは Transformer ベースの自己注意機構により，フレーム列から時空間の文脈情報を効果的に抽出することで優れた推定性能を達成している。しかしながら，これらの手法は通常，大規模なモデルサイズと高い計算コストを伴い，モバイル端末でのリアルタイム推論には適さない。

[1] 電気通信大学
[a)] nakamizo-y@mm.inf.uec.ac.jp
[b)] yanai@cs.uec.ac.jp

一方で，2 段階手法とは，まず画像から 2 次元の関節位置を推定し，その後，得られた 2 次元姿勢のシーケンスを 3 次元空間にリフティングするアプローチである。この方式は，OpenPose [3] などの既存の 2 次元姿勢推定モデルの性能を活用できることから，システムの柔軟性と実装容易性の点で利点がある。STCFormer [12] は時間方向の自己注意と空間方向の自己注意を並列的に処理する Spatio-Temporal Criss-Cross Attention を提案し，単純に時空間方向を交互に参照する注意機構と比較して，効率的に高い性能を達成した。また，KTPFormer [9] はキネマティクスと各関節位置の軌跡の事前知識を明示的に組み込んだ自己注意機構である Kinematics Prior Attention と Trajectory Prior Attention を導入することで，関節間の時空間関係を効率的に捉え，優れた性能を残している。

2 段階手法の精度は 2 次元人体姿勢推定機に大きく依存するほか，人間による 2 次元姿勢表現からの 3 次元復元自体が高度なタスクであることを踏まえると，2 次元姿勢表現を介した段階間の接続が情報のボトルネックとなるといえる。本研究では 2 段階手法の全体構造を参考にしつつ，中間表現としての 2 次元関節座標を経由せず，画像から直接 3 次元姿勢を推定したうえで，それを時間的に補正する形式を採用する。

2.2 モバイル端末向け 3 次元姿勢推定

モバイル端末上でのリアルタイム推論を志向した 3 次元姿勢推定モデルとしては，1 段階手法がいくつか提案されている。Google によるオープンソースの MediaPipe フレームワークでは，BlazePose [2] に基づいた 3 次元姿勢推定モデルが導入されており，heatmap・offset・回帰を組み合わせる独自の構造により，高速かつ精度の高い推定を実現している。

また，Choi ら [4] は，軽量なバックボーンを採用することでモデルの推論速度とサイズを最適化し，モバイル端末上での応用を想定した設計を行っている。Hwang ら [6] は，精度向上のために知識蒸留を活用し，小型モデルに大規模モデルの知識を蒸留する学習手法を提案した。加えて，近年では ConvNeXtPose [8] が提案されており，ConvNeXt ベースの軽量な CNN バックボーンとシンプルな回帰ヘッドを組み合わせることで，モバイル端末上でも実用的な精度と速度を両立する設計となっている。特に，AR フィットネスや動作解析といった実応用を意識し設計されており，実験によりその有用性を示している。

これらの手法はいずれも，単一フレームから直接 3 次元姿勢を推定するワンステージ構造であり，リアルタイム処理が可能という強みを持つ。一方で，逐次的に到着する映像フレームを時間的に統合しながら処理を行う軽量なアーキテクチャは，既存研究では提案されていない。既存手法の多くは，空間情報の抽出に重点を置いており，時間的文脈の活用については，スムージングなどの後処理に留まっているのが現状である。本研究では，こうした課題を踏まえ，時間情報をリアルタイムに活用可能でありながら，モバイル環境でも実行可能な軽量モデルの設計を目的とする。

3. 提案手法

3.1 概要

提案手法では，ConvNeXtPose [8] をベースとしたフレーム単位で動作する初期 3 次元姿勢モジュールと，時系列情報および事前知識を活用して姿勢を洗練する Transformer ベースの 3 次元姿勢リファインメントモジュールからなる 2 段階構成を採用する。提案手法の概要図を図 1 に示す。

3.2 初期 3 次元姿勢推定モジュール

本研究では，1 段階目の 3 次元姿勢推定モジュールとして ConvNeXtPose [8] をベースとした軽量な画像入力型の 3 次元姿勢推定モデルを採用しており，単一フレーム $I_t \in \mathbb{R}^{H\times W\times 3}$ から直接，3 次元関節位置 $\hat{P}_t \in \mathbb{R}^{J\times 3}$ を推定する。ConvNeXtPose は ConvNeXt 系の CNN をバックボーンとし，特徴マップを深さ分離畳み込みによりアップサンプリングした後，関節ごとの 3D 座標を回帰する軽量な 1 段階構成のモデルである。この設計により，一定程度の推定精度を達成しながらも，FLOPS およびモデルサイズを抑え，モバイル端末上でのリアルタイム推論が可能となっている。本研究では，この ConvNeXtPose の XS 構成を採用している。XS 構成は，バックボーンの各ステージにおけるチャネル幅を [40, 80, 160, 320]，層構成を [2, 2, 6, 2] とし，アップサンプリング部に出力チャネル数 128 の深さ分離畳み込み層と Bilinear2D アップサンプリング層からなるブロックを 2 つ重ねた構造を用いた構成である。

また，本モジュールではモバイル環境での高速動作を考慮し，関節ごとの 3D 座標を中間表現から得る soft-argmax 処理の過程において，softmax 内に含まれる $\exp(x)$ の計算を，以下の 2 次テイラー展開近似により実装している。

$$\exp(x) \approx 1 + x + \frac{1}{2}x^2 \tag{1}$$

従来の $\exp(x)$ を伴う softmax 処理と比較して，簡潔な演算構造を持つため，モバイル向けの GPU などでの推論との親和性が高い。この近似により，指数演算を回避しながら softmax の正規化効果を保つことができ，モバイル環境での高速処理に寄与している。

3.3 3 次元姿勢リファインメントモジュール

本研究では，2 段階目の処理として，空間および時間方向の依存関係を活用し，推定された T フレーム分の 3 次元姿勢系列 $\hat{P}_{t-(T-1)}, ..., \hat{P}_t$ からより洗練された最終フレームの 3 次元姿勢 P_t を推定する，Transformer ベースのリファインメントモジュールを導入し，構造的整合性および時間

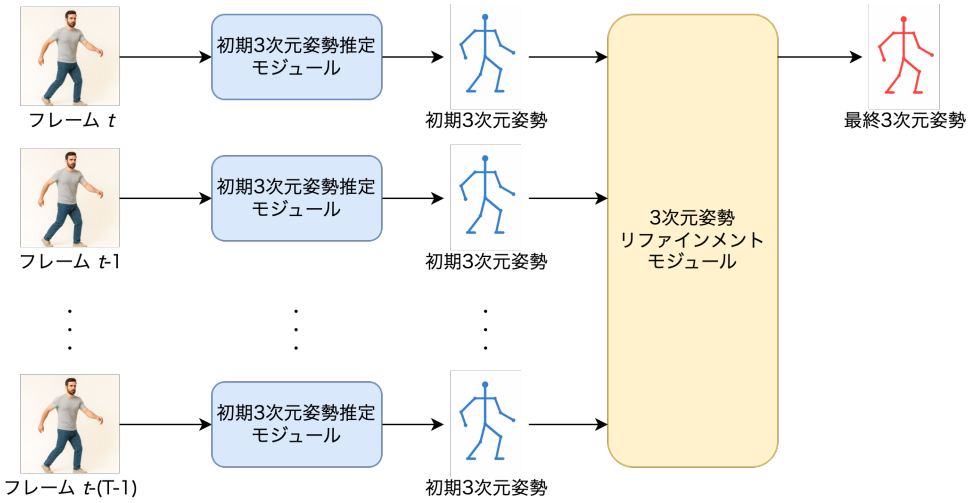

図 1 提案手法の概略図

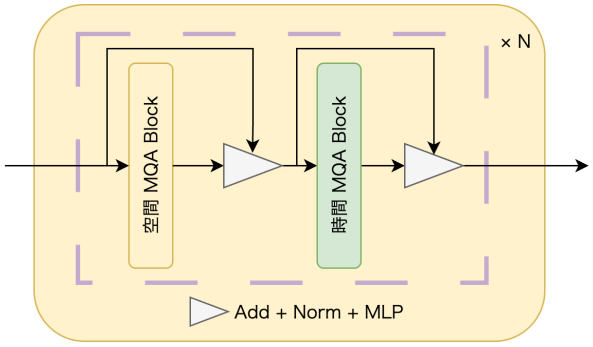

図 2 3 次元姿勢リファインメントモジュールの概略図

的一貫性の向上を図る。このモジュールの概略図を図 2 に示す。

このモジュールは，空間方向・時間方向それぞれに独立な自己注意機構を適用する Spatio-Temporal Transformer 構成を採用しており，各ブロックには Multi-Query Attention (MQA) を導入することで計算効率を高めている。また，時間方向の注意処理においては，過去のキー・バリューを保持する KV キャッシュ機構を導入しており，逐次入力に対するリアルタイム推論を実現可能としている。

加えて，提案手法では各アテンションマップに対して，関節構造および時間的隣接性を考慮したスケーリング項を導入することで，追加層を導入することなく事前知識の組み込みを実現している。

$$\mathrm{Attn}' = \mathrm{Attn} \times (1 + \lambda \cdot a \cdot (\mathrm{Adj}_1 + \mathrm{Adj}_2)) \qquad (2)$$

ここで，Attn は通常の Softmax 適用前の注意マップ，Adj_1 は関節構造あるいは時間的近接性に基づく固定の隣接行列，Adj_2 は学習可能な自由接続行列，a はスカラーの学習可能な重み係数，λ はスケーリングを制御するハイパーパラメータである。このスケーリングは空間方向・時間方向の両方のアテンションマップに適用されており，身体構造に基づく関節間の依存性や，時間的に隣接するフレーム間の連続性が注意重みに効果的に反映される設計となっている。このスケーリング設計は，KTPFormer [9] におけるグラフ畳み込みの構成から着想を得ており，本研究ではこれを自己注意機構に導入することで，空間的および時間的な依存関係の柔軟なモデリングを実現している。

本研究では，Transformer の各ブロックにおいて埋め込み次元 256，ヘッド数 8 を採用し，時間方向および空間方向の層数 N は $N = 4$，スケーリング係数 λ は $\lambda = 0.5$ と設定した。

3.4 学習

提案手法では，学習を以下の 2 段階に分けて実施する。まず，第 1 段階では，初期 3 次元姿勢推定モジュールのみを対象に，フレーム単位で学習を行う。この学習では，soft-argmax 処理を通じて得られた関節位置に対する L1 損失を損失関数として用いる。学習設定は ConvNeXtPose [8] に準拠しており，2 次元および 3 次元の混合姿勢データセットを用いて学習を行う。

続いて，第 2 段階では，初期モジュールの重みを固定したうえで，3 次元姿勢リファインメントモジュールのみを学習する。学習設定は KTPFormer [9] に準拠しており，損失関数には関節ごとに重みを付与した加重 L1 損失に加え，時間的一貫性を評価する速度誤差を用いる。学習には 3 次元姿勢データセットを用い，構造的整合性と時間的な滑らかさの向上を図る。

4. 実験

4.1 実験設定

本研究では，以下の 2 つの観点から提案手法を評価した。一つ目が 3 次元姿勢推定精度である。従来手法に倣い，Mean Per Joint Position Error (MPJPE) を用いて評価した。MPJPE は推定された各関節位置と真値とのユー

表 1　3 次元姿勢推定精度および GPU レイテンシの評価結果

	MPJPE (mm) ↓															GPU レイテンシ (ms) ↓
Method	Dir.	Disc.	Eat	Greet	Phone	Pose	Pur.	Sit	SitD.	Smoke	Photo	Wait	Walk	WalkD.	WalkT.	Avg.	Latency
ConvNeXtPose	**53.25**	57.86	52.15	57.48	60.75	**52.33**	53.65	68.30	78.75	57.11	66.68	53.10	44.92	61.99	50.27	58.31	29.72 ± 0.29
ConvNeXtPose*	55.25	59.87	49.29	58.73	61.22	53.86	55.39	68.77	76.34	57.53	67.36	54.16	44.71	64.65	51.06	58.93	$\mathbf{10.33 \pm 0.32}$
Ours	53.33	**57.44**	**46.51**	**55.68**	**58.29**	52.61	**52.86**	**66.22**	**72.79**	**54.45**	**64.18**	**52.01**	**39.40**	**58.97**	**42.57**	**55.71**	18.51 ± 0.36

クリッド距離の平均で定義される。二つ目は Android 端末上での GPU レイテンシーである。ConvNeXtPose に倣い，Samsung Galaxy S20 の GPU 上でレイテンシを TensorFlow Lite を用いて測定した。実運用を想定し，アプリケーション上で 30 fps の設定で連続 1 000 回推論を実行し，その平均レイテンシにより評価した。

4.2　学習設定

4.2.1　初期 3 次元姿勢推定モジュール

初期 3 次元姿勢推定モジュールの学習は，ConvNeXtPose の設定に準拠して実施した。使用された入力画像サイズは 256×256，バッチサイズは 128，学習エポックは 70，最適化には AdamW を使用し，重み減衰係数は 0.1 に設定した。学習率は 4×10^{-3} を採用し，最初の 5 エポックをウォームアップ期間として学習率を線形に増加させた。その後はハーフサイクルコサインスケジューリングに従って，4×10^{-6} まで徐々に減衰させた。学習データとしては，3 次元姿勢データセットである Human3.6M [7] および 2 次元姿勢データセットである MPII [1] を用いた混合データセットを使用した。Human3.6M は，屋内環境において被験者の多様な動作を高精度なモーションキャプチャ装置で記録した大規模な 3 次元姿勢データセットであり，3D 姿勢推定における標準ベンチマークとして広く用いられている。本研究では，そのうち被験者 1, 5, 6, 7, 8 のデータを学習に，9, 11 を評価に用いた。また，MPII は日常生活における多様な姿勢を含む 2 次元姿勢アノテーション付きの画像データセットであり，本研究ではポーズの多様性や視点のばらつきへの汎化能力を高めるために活用された。

4.2.2　3 次元姿勢リファインメントモジュール

3 次元姿勢リファインメントモジュールの学習では，初期 3 次元姿勢推定モジュールの重みを固定し，3 次元姿勢リファインメントモジュールのみを対象として学習を行った。入力系列長 T は，$T = 30$ とし，50fps の映像から 5 フレーム間隔でサンプリングした 30 フレームを入力とした。学習には Human3.6M を使用し，初期 3 次元姿勢推定の学習と同様に被験者 1, 5, 6, 7, 8 を訓練に，被験者 9, 11 を評価に用いた。バッチサイズは 64，学習エポック数は 10 とした。最適化には AdamW を用い，学習率は 6.4×10^{-4}，重み減衰係数は 0.1 に設定した。学習率のスケジューリングには，指数減衰にウォームアップを組み合わせた手法を採用しており，最初の 0.5 エポックはウォームアップ期間としてステップ単位で線形に学習率を増加させた後，残りの期間ではエポック単位で指数関数的に減衰させた。

4.3　結果

提案手法，ベースラインである ConvNeXtPose[8]，そして ConvNeXtPose 内の soft-argmax 処理に用いられる Softmax 関数を 2 次テイラー近似した ConvNeXtPose*の評価結果を表 1 に示す。提案手法は GPU レイテンシを 18.51 ± 0.36 ms に留め，30fps でのリアルタイム推論を実現しつつ，MPJPE を 2.60 mm 改善した。また，ConvNeXtPose*と ConvNeXtPose を比較すると，Softmax 関数の近似により，MPJPE が 0.62 mm 悪化しているものの，GPU レイテンシが 19.39 ms 改善しており，Softmax の推論速度への悪影響が確認された。さらに，提案手法と ConvNeXtPose*を比較すると，Softmax 関数の 2 次テイラー近似により生じた精度低下を，提案手法では回復させており，MPJPE は ConvNeXtPose*比で 3.22 mm 改善されている。GPU レイテンシについても，提案手法は ConvNeXtPose*より 8.18 ms の増加にとどめつつ，元の ConvNeXtPose と比較して依然として 11.21 ms の短縮を維持している。これにより，本手法は 30fps のリアルタイム推論環境下において，推定精度改善と高速処理を同時に実現することが確認された。

5.　まとめ

本研究では時間軸情報を加味したモバイル端末向け 3 次元人体姿勢推定モデルを提案した。提案手法では従来のフレーム単位で姿勢を推定するモデルの後段として，Transformer ベースの姿勢リファインメントモジュールを導入することで効率的に時間軸情報を考慮した 3 次元姿勢推定を実現している。これにより，提案手法はベースである ConvNeXtPose と比較して，30fps でのリアルタイム推論を実現しつつ，3 次元姿勢推定精度の改善を達成した。

謝辞

本研究の初期において，ご助言とご示唆を賜りました Deep Patel 氏に深く感謝申し上げます。また，同時期に計算資源を提供してくださった NEC Laboratories America Inc. の皆様にも，心より御礼申し上げます。

参考文献

[1] Andriluka, M., Pishchulin, L., Gehler, P. and Schiele, B.: 2D Human Pose Estimation: New Benchmark and State of the Art Analysis, *Proc. of IEEE Computer Vision and Pattern Recognition* (2014).

[2] Bazarevsky, V., Ivan, G., Karthik, R., Tyler, Z., Fan, Z. and Matthias, G.: BlazePose: On-device real-time body pose tracking, *arXiv:2006.10204* (2020).

[3] Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S. and Sheikh, Y. A.: OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, *Proc. of IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 172–186 (2019).

[4] Choi, S., Choi, S. and Kim, C.: MobileHumanPose: Toward real-time 3D human pose estimation in mobile devices, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 2328–2338 (2021).

[5] Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa*, A. and Malik*, J.: Humans in 4D: Reconstructing and Tracking Humans with Transformers, *Proc. of IEEE International Conference on Computer Vision*, pp. 14783–14794 (2023).

[6] Hwang, D.-H., Kim, S., Monet, N., Koike, H. and Bae, S.: Lightweight 3d human pose estimation network training using teacher-student learning, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 479–488 (2020).

[7] Ionescu, C., Papava, D., Olaru, V. and Sminchisescu, C.: 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, pp. 1325–1339 (2014).

[8] Nguyen, H. S., Kim, M., Im, C., Han, S. and Han, J.: ConvNeXtPose: a fast accurate method for 3D human pose estimation and its AR fitness application in mobile devices, *IEEE Access*, Vol. 11, pp. 117393–117402 (2023).

[9] Peng, J., Zhou, Y. and Mok, P.: KTPFormer: Kinematics and trajectory prior knowledge-enhanced transformer for 3d human pose estimation, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 1123–1132 (2024).

[10] Qiu, Z., Qiansheng, Y., Jian, W. and Dongmei, F.: IVT: An end-to-end instance-guided video transformer for 3d pose estimation., *Proc. of ACM International Conference Multimedia*, pp. 6174–6182 (2022).

[11] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L.-C.: MobileNetV2: Inverted residuals and linear bottlenecks, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 4510–4520 (2018).

[12] Tang, Z., Qiu, Z., Hao, Y., Hong, R. and Yao, T.: 3d human pose estimation with spatio-temporal criss-cross attention, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 4790–4799 (2023).