IS1-015

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

電気通信大学　中溝 雄斗　柳井 啓司

はじめに

近年では広域な時系列情報を考慮することで、高精度な3次元人体姿勢推定を可能にしたモデルが登場している

しかし、モバイル向けでリアルタイム推論可能なモデルでは、時系列情報は活用されておらず、その性能は低いままである

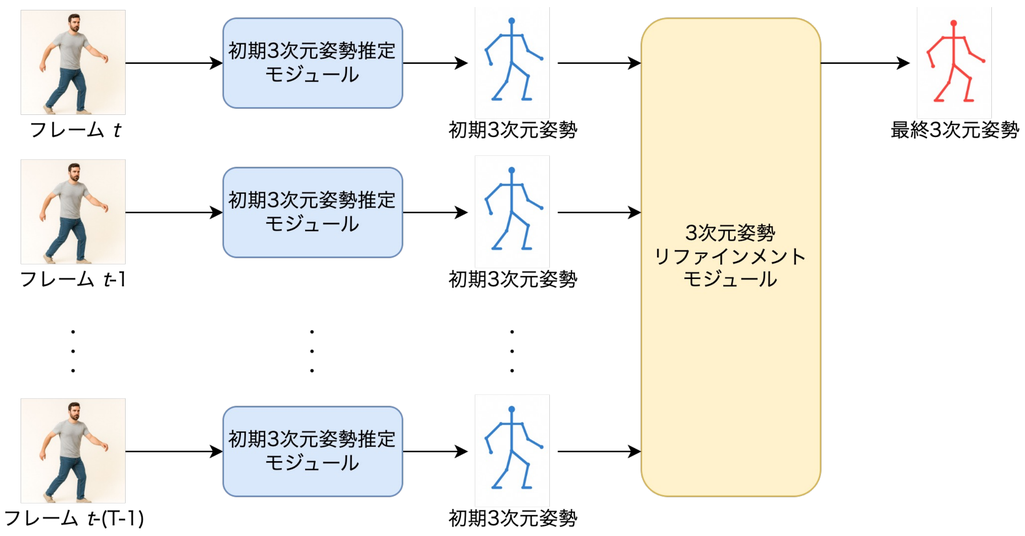
目的

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の両立

手法

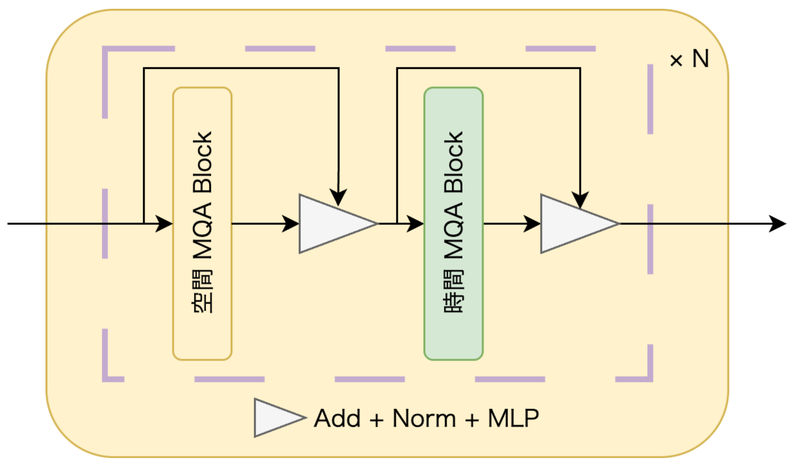
モデル構造

提案手法はConvNeXtPose [1]を初期3次元姿勢推定モジュール (IPEM) として、後段に時間軸を考慮し、推定姿勢を改良する3次元姿勢リファインメントモジュール (PRM) を追加して構成される



※ 初期3次元姿勢推定モジュールの構造はConvNextPoseに準拠するが、高速化のため内部のsoft-argmax処理過程のsoftmaxを2次テイラー展開近似している

3次元姿勢リファインメントモジュールは時間方向および空間方向それぞれのMulti-Query Attention (MQA) によって構成される



各MQAでは各アテンションマップに対して、KTPFormer [2] を参考に構築した関節的もしくは時間的構造を表現する隣接行列によるスケーリングを導入することで、事前知識を導入している

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

※ Attnは元々のsoftmax適応前のアテンションマップ、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

学習

学習は2段階に分けて実施する

- 1段階目：IPEMのみを対象に、フレーム単位の入力で学習
- 2段階目：PRMのみを対象に、シーケンス単位の入力で学習

損失は以下のように設定

- 1段階目：関節位置に対するL1損失
- 2段階目：関節位置に対する加重L2損失と関節移動速度に対するL2損失の重み付け和

実験

1. Human 3.6M (Protocol 2) [3] による評価

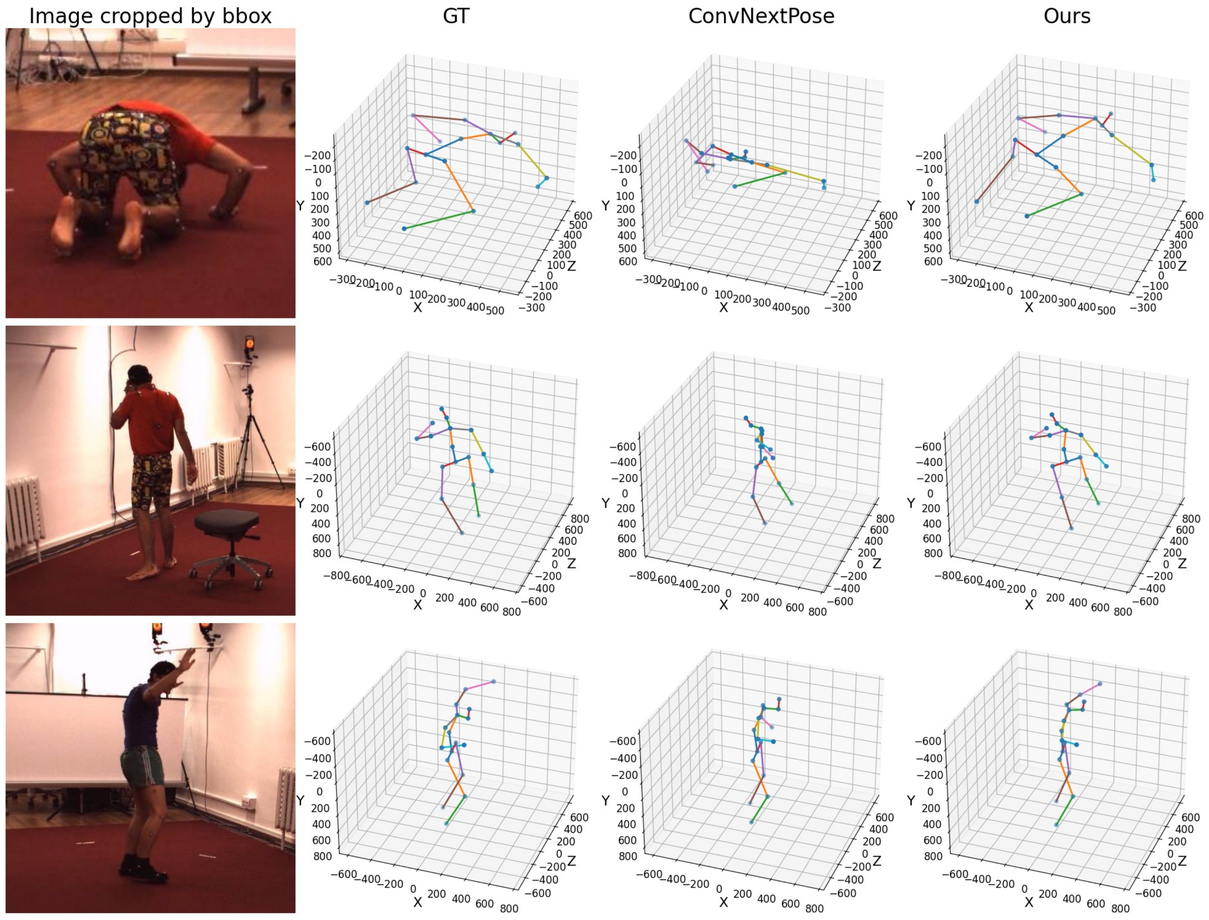
ConvNeXtPoseと比較して、Mean Per Joint Position Error (MPJPE)とLatencyの両方を改善

手法	MPJPE / mm (↓)	Latency / ms (↓)
ConvNeXtPose	58.31	29.35 ± 0.34
ConvNeXtPose (Taylor近似)	58.93	**10.12 ± 0.29**
Ours	**54.83**	18.35 ± 0.36

※ LatencyはSamsung Galaxy S20のGPU上で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30fpsで入力が与えられる設定で1000回推論を実行し、その平均Latencyである。また、論文記載の結果では、空間方向のTransformerにおいて位置埋め込みが適切に行われていない実装上の誤りが確認されたため、本結果では修正したモデルで再度評価を行った。

2. 定性的な評価

特に深くしゃがんだ際の画像や人体を側面から捉えた画像など遮蔽の強い入力画像において、大幅に推定精度を改善している



3. アブレーション実験

空間的、もしくは時間的MQAの一方のみの場合、精度の改善は非常に限定的であったため、時空間を総合的に考慮し、リファインメントすることが重要であることが示された

手法	MPJPE / mm (↓)
ConvNeXtPose	58.31
ConvNeXtPose (taylor近似)	58.93
+ 空間的MQA	57.67
+ 時間的MQA	58.86
+ 時間的MQA & 空間的MQA (Ours)	**54.83**

※ 空間的、もしくは時間的MQAの一方のみで実験を行う際には、モデル全体の表現能力の差を低減するため、それぞれMQA Blockの層数を2倍にして実験を行った

おわりに

モバイル端末でのリアルタイム推論と時間軸を考慮した推論を両立

- 時間軸を考慮し、姿勢を補正する軽量なPRMを導入した
- ConvNeXtPoseをモバイル端末GPU上でのLatencyとHuman 3.6Mによる評価における精度の両面で上回った
- 特に遮蔽の強い入力画像の際に精度を向上させた

今後の課題

- モジュール間の接続部が情報ボトルネックとなっていると考えられるため、バイパス機構によるボトルネックの解消

[1] Nguyen Hong Son, et al. "Convnextpose: A fast accurate method for 3d human pose estimation and its ar fitness application in mobile devices." IEEE Access 11, 2023.
[2] Peng Jihua, et al. "Ktpformer: Kinematics and trajectory prior knowledge-enhanced transformer for 3d human pose estimation." CVPR 2024.
[3] Ionescu Catalin, et al. "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments." IEEE transactions on pattern analysis and machine intelligence, 36, 7, 2013.

miru 2025.07