超広角カメラを用いたウェアラブル姿勢推定システムの構築

平野 稔祐 * Dong-Hyun Hwang† Chen-Chieh Liao* 小池 英樹 *

概要.

身体装着型モーションキャプチャは固定カメラの設置場所を考慮せず使用できるという利点がある．先行研究である MonoEye は胸部に超広角カメラを取り付けるだけの姿勢推定を可能とし，社会的受容性の高い身体装着型モーションキャプチャ手法を提案した．しかし，既存のハードウェア構成では，ポータブル性に欠けるためリアルタイムでの処理ができないという問題点があった．そこで本研究では，MonoEye を基本として，姿勢推定モデルの軽量化とポータブルデバイスの設計に取り組み，ウェアラブル姿勢推定システムを構築する．TensorRT と量子化により軽量化した姿勢推定モデルを用いて，MonoEye と同等の精度で 6.1 倍高速化し，最大 47FPS での姿勢推定を実現した．

1 はじめに

モーションキャプチャは，インタフェースや CG アニメーション，スポーツ科学など様々な応用分野で活用されている．特に，身体装着型カメラを用いたモーションキャプチャは，カメラの設置場所を考慮する必要がなく，空間や環境の制約を受けずに利用可能である．

身体装着型カメラから装着者の姿勢を推定する試みは様々な手法が提案されてきた [1, 3, 4, 5]．その一つとして，胸にカメラを装着する MonoEye[1] がある．MonoEye は，280 度の視野角を持つ超広角カメラを用いて，魚眼画像に映り込む手足と顔から姿勢推定を行うシステムである．胸部にカメラを装着するため，頭部にカメラを装着する手法 [3, 4, 5] と異なり，カメラの装着に帽子や眼鏡を必要としない利点がある．

MonoEye のネットワークはリアルタイム性能を考慮して設計された．しかし，持ち運び可能な機器としては実現されていない．アプリケーションへの組み込みには，リアルタイムな応答性能とウェアラブル化が求められる．持ち運び可能で省電力なデバイス上で，十分な速度で推論を行うにはさらなる高速化が必要である．そこで本研究では,MonoEye モデルの高速化とハードウェアの設計を行い，リアルタイム推論とウェアラブル化を実現する．

2 実装

図 1 にウェアラブル姿勢推定システムの全体像を示す．MonoEye は PyTorch を用いて実装されており，それを構成する 3 つのモデル BodyPoseNet, HeadPoseNet，CameraPoseNet がそれぞれ，体の姿勢，頭部方向，カメラの向きを推論する． 本システムでは，以前のシステムと同様にこれら 3 つのモデルを用いて姿勢推定を行う．姿勢推定モデルの軽量化と並行処理，システムのチューニングを行い，リアルタイムな姿勢推定を実現した．

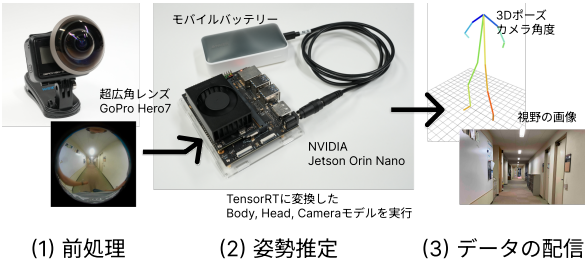

図 1. システム全体像

図 2. システムを装着した様子

推論は，NVIDIA Jetson Orin Nano 8GB(以下 Jetson) の GPU 上で行う．魚眼画像は，アクションカメラ (GoPro Hero 7 Black) に超広角レンズ (Entaniya M12-280) を取り付け，キャプチャボード経由で取得する．駆動に必要な電力は USB PD 対応のモバイルバッテリーから供給した．図 2 のように，カメラは胸部，Jetson とモバイルバッテリー

* 東京工業大学 情報理工学院
† NAVER Cloud

表 1. TensorRT へ変換したモデルの予測結果と速度 Body:MPJPE(mm)，Head:MAE(°)，Camera:MAE(°)

	MonoEye dataset			Realworld	Inference Time(ms)			
	Body	Head	Camera	Body	Body	Head	Camera	FPS
MonoEye	35.14	2.58	2.92	175.71	114.0	79.8	39.5	7.8
TensorRT(fp16)	35.15	2.58	2.92	175.71	15.9	4.7	2.4	**47.6**
TensorRT(fp16+int8)	100.03	3.24	4.46	188.70	11.5	4.1	2.1	53.0

は背中に装着する．デバイス全体で 1.2kg，約 2 時間連続して駆動することができる．

Jetson 上で高速に推論するために，PyTorch のモデルを TensorRT へと変換する．TensorRT とは，機械学習モデルを最適化して NVIDIA GPU 上での推論を高速化する開発キットである．TensorRT への変換には，trtexec[1]を用いた．また，モデルの軽量化と高速化のため，量子化を行う．8 バイト整数 (int8) と半精度浮動小数点数 (fp16) による量子化を比較し，精度と速度のバランスから fp16 を選んだ．

3 評価

3.1 実験

モデル軽量化後の精度と，システムの処理速度を評価する．精度の評価は，MonoEye 合成データのテストセットからランダムに選んだ 1 万枚と，マーカーレスモーションキャプチャ（Qualisys）で撮影した男性 2 人の実世界データ 6870 枚を用いて行う．評価指標として，BodyPoseNet では MPJPE(mm)，HeadPoseNet と CameraPoseNet では MAE(°) を用いた． 実世界データは MPJPE を計算する前に Procrustes 解析 [2] を適用し，推定結果の座標系を真値と一致させた．

処理速度は，モデルの推論速度とシステムの処理時間を計測する．GPU のウォームアップを除いて，システムを 30 秒動作させ平均を計算した．

3.2 結果

実験結果を表 1 に示す．TensorRT 変換によって推論時間が短縮されている．従来の PyTorch モデルでは，Jetson 上で BodyPoseNet の推論に 114.0ms 必要であった．一方，本システム TensorRT(fp16) では 15.9ms で推論できる． fp16 と int8 を組み合わせた量子化では 11.5ms とさらに短縮された．

fp16 への量子化によって，モデルは 277.05MB から 139.87MB へと軽量化された．fp16 を用いて量子化しても精度の変化は見られなかった． 一方で int8 と fp16 を組み合わせた量子化では，実世界データで 7.4%精度が低下した．int8 では MonoEye のモデルと比較して 1/4 の精度で計算を行うため，誤差が生まれたと考えられる．

以上の結果から，今回のシステムでは fp16 での量子化を採用した．fp16 での量子化によって，精度を維持しながら 47FPS 以上の速度で処理することができる．47FPS は，60FPS 以上で撮影が可能な光学式のモーションキャプチャに劣るが，日常的な動作をキャプチャするには十分であると考えられる．

4 アプリケーション

MonoEye の軽量化によって Jetson の計算リソースには余裕があるため，姿勢推定と並列に他の軽量な機械学習モデルを動かすことが可能である．使用者の姿勢から行動を推定し周囲の危険を検知してフィードバックするようなアプリケーションへの応用が考えられる．

ウェアラブル姿勢推定システムの応用例として，図 3 に Activity Highlighter を提案する．Activity Highlighter は，超広角レンズに映る情報を元に，ユーザーの活動を記録するライフロガーである．物体検出モデルを併用して，姿勢と周辺に映る物体の記録，検索が可能である．

図 3. ライフログアプリケーション

5 まとめ

本研究では，TensorRT による MonoEye モデルの軽量化と Jetson Orin Nano を使ったポータブルデバイス設計により，ウェアラブル姿勢推定システムを実現した．これにより，日常生活やスポーツの場など，リアルタイム性の求められる場面での MonoEye の活用が期待できる．

[1] https://github.com/NVIDIA/TensorRT/tree/main/samples/trtexec

謝辞

本研究は JST CREST JPMJCR17A3 の支援を受けている．

参考文献

[1] D.-H. Hwang, K. Aso, Y. Yuan, K. Kitani, and H. Koike. MonoEye: Multimodal Human Motion Capture System Using A Single Ultra-Wide Fisheye Camera. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20, p. 98–111, New York, NY, USA, 2020. Association for Computing Machinery.

[2] D. G. Kendall. A Survey of the Statistical Theory of Shape. *Statistical Science*, 4(2):87–99, 1989.

[3] H. Rhodin, C. Richardt, D. Casas, E. Insafutdinov, M. Shafiei, H.-P. Seidel, B. Schiele, and C. Theobalt. EgoCap: Egocentric Marker-Less Motion Capture with Two Fisheye Cameras. *ACM Trans. Graph.*, 35(6), dec 2016.

[4] D. Tome, P. Peluse, L. Agapito, and H. Badino. xR-EgoPose: Egocentric 3D Human Pose from an HMD Camera. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 7728–7738, 2019.

[5] W. Xu, A. Chatterjee, M. Zollhofer, H. Rhodin, P. Fua, H. Seidel, and C. Theobalt. Mo2Cap2: Real-time Mobile 3D Motion Capture with a Cap-mounted Fisheye Camera. *IEEE Transactions on Visualization & Computer Graphics*, 25(05):2093–2101, may 2019.

未来ビジョン

本研究では，モデルの軽量化とシステムのチューニングによって胸装着型モーションキャプチャのウェアラブル化を実現した．推論に使用している Jetson Orin Nano の計算リソースにはまだ余裕がある．画像の前処理とモデルの更なる最適化によって，現状の 47FPS を超えて 60FPS 近い高速で安定したキャプチャを実装できるだろう．

今後はモデルの精度向上と，実践的なアプリケーションへの活用を目標としている．MonoEye は特徴量抽出に ResNet101 を用いている．改良として，MobileViT など軽量な Transformer ベースのアーキテクチャの導入や時系列データの利用が考えられる．ワイヤレスでの映像取得と軽量なアーキテクチャの採用によって，スマートフォンのような，より身近なデバイス上で推論することも期待できる．

身体装着型モーションキャプチャはカメラの設置場所を考慮する必要がないという利点がある．例えば，カメラの設置しきれない広大な空間や移動の多い場面，複数人が同時に活動してカメラの死角が多い場面，カメラの設置が難しい狭い空間でその強みを活かせると考えている．特に，胸装着型カメラは，カメラ装着のためにメガネや帽子を使う必要がなく，一つのカメラで身体と同時に前方の環境も撮影できるという特徴がある．障害物の検知や危険予測，ライフログアプリケーションの応用など，更なる活用に取り組んでいきたい．