H-006

CNN-LSTM を用いた 2D Human Motion Prediction

須藤雅基　　張善俊
神奈川大学 大学院 理学研究科 理学専攻　神奈川大学 情報学部 計算機科学科

1．はじめに

近年、画像処理を高速化することへの重要性は高まっている。例えばユーザーのモーションを認識してなんらかのインタラクティブな処理を行う際には高速な処理、すなわち低遅延であることが重要となる。特に AR・VR においてユーザーのモーションを元に画像処理を行う際には低遅延性が重要となる。先行研究においても LSTM を用いて人物動作を予測しプリレンダリングすることで、VR デバイスに投影する映像の処理遅延を解消する試みが行われている[1]。

2．目的

我々は計算量が少なく軽量な姿勢予測モデルを構築し、オフラインのモバイル端末上で動作させることで、計算資源の少ない端末上での姿勢予測の実用性を検証することを目的とする。

そのため本研究ではまず CNN-LSTM を利用した 2 次元動作予測モデルを構築し、パラメータを変更しつつ、精度と実行時間の変化を検証する。

3．処理フロー

本研究で用いる姿勢予測モデルは RGB 画像を入力として受け取り、BB tracker[4] を用いて画像をクリッピングする。次にクリッピング画像から CNN で関節点を抽出し、関節点と周辺の画素を元に lattice optical flow を計算する。最後に LSTM に連続した 5 フレームを入力し、出力として 30 fps の動画の場合 15 フレーム後、すなわち 0.5 秒後の姿勢座標の予測値を出力する。

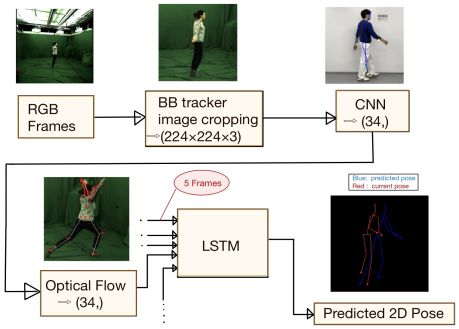

図 1．CNN-LSTM 推論時処理フロー

図 2．ボクシング動作姿勢予測例

4．実験

本研究では 4 種類の LSTM モデル (optical flow 入力情報有無、1 層 LSTM or 2 層 LSTM)のモデルの精度と推論時間を

	Optical flow 有	Optical flow 無
2 層 LSTM	58.7 %	56.6 %
1 層 LSTM	61.3 %	56.2 %

表 1 . Human3.6m に対する精度 (PCKh@0.5)

	Optical flow 有	Optical flow 無
2 層 LSTM	23.4 %	21.7 %
1 層 LSTM	26.4 %	25.1 %

表 2 . 自作データに対する予測精度 (PCKh@0.5)

	Optical flow 有	Optical flow 無
2 層 LSTM	42.6[ms]	41.0[ms]
1 層 LSTM	39.5[ms]	38.8[ms]

表 3 . Average Inference Time[ms]

調べる。

CNN モデルは ImageNet で事前学習済みの ResNet-50 モデルに対して関節点を推定するために AIST Dance Video データセットを用いて転移学習を行う。Human3.6M データセットと自作データセット(ボクシング、反復横跳び等の動作を収録)で精度比較を行う。

5. 実験結果

実験結果を表 1, 2, 3 にまとめる。LSTM による姿勢予測について、optical flow 有のモデルは精度が高くなった。その一方、二層 LSTM モデルを使用した場合は一層 LSTM と比べて精度が低下した。また、表 3 より AIT はモデルのパラメータの数に比例して増加していることが確認できる。

6. まとめ・今後の課題

本研究の CNN モデルは推論に BB tracker を用いる。これにより画面内のどの位置に対象人物がいてもしかし実用上を考えた場合は予測した関節点が一時的に間違った場所を指してしまうと、BB が逆効果となり推論を妨げてしまう状況が目立った。

また、本研究内ではモバイル端末上での CNN-LSTM の精度について検証することはできなかったものの、AR, VR 分野では本研究で検証した予測モデルを用いることで処理遅延を減少させることができると期待される。今後はモバイル端末上における本モデルの実用性を検証する。

7. 参考文献

[1]: Ripan Kumar Kundu et al., “A Study on Sensor System Latency in VR Motion Sickness”

[2]: Erwin Wu et al., ”FuturePose - Mixed Reality Martial Arts Training using Real-time 3D Human Pose Forecasting with a RGB Camera”, IEEE Winter Conference on Applications of Computer Vision(2019)

[3]: DUSHYANT MEHTA et al., “ VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera” ACM TOG (SIGGRAPH 2017)

[4]: Ruilong Li et al., “AI Choreographer: Music Conditioned 3D Dance Generation with AIST++”, ICCV (2021)

[5]: Dushyant Mehta et al., “Monocular 3D Human Pose Estimation In The Wild Using Improved CNN Supervision”, 3DV (2017)

[6]: Catalin Ionescu et al., ”Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments”