I-022

人の関節回転角系列からの身体中心位置の推定
Estimating Root Trajectory of Human Motion from Joint Rotation Sequence

木佐 省吾[1)]　遠藤 結城[1)]　栗山 繁[1)]
Shogo Kisa　Yuki Endo　Shigeru Kuriyama

1　はじめに

3DCG アニメーションにおける人のモーションは，姿勢を表す関節の回転角と身体中心の 3 次元位置からなる時系列データで表現される．この姿勢と位置の変化は人体構造上矛盾の無い正しい動きとなる数値関係を有する必要がある．例えば，既存のモーションデータを再利用する際に骨格を他人に変更するだけでも不自然に足が浮いたり滑ったりする．また，モーション作成手法の 1 つであるキーフレームアニメーションにおいては，姿勢の変化に合わせた位置の変化が自然になるように製作者が手付けしている．このような問題は，姿勢から位置を推定することで解決できるが，運動学に基づく方法では先に足の接地状態を考慮する必要があり，走り等同時に足が浮く動作に対しては正確に位置を推定できない．

そこで本稿では，姿勢情報から身体中心位置を正確に推定する手法を深層学習を用いて初めて実現する．深層学習手法として，近年自然言語処理分野で高い性能を発揮している Transformer [1] で使用される Self-Attention をモーションデータに適用した上で，見た目上不自然になる高周波信号の抑制を図る改良を施す．さらに，モーションキャプチャ（MoCap）データを使用して，時系列データに適用可能な基本的な深層学習モデルと比較し，提案モデルの有効性を検証する．

2　基礎となる手法

Transformer は自然言語処理の機械翻訳タスクを解く深層学習モデルであり，Recurrent Neural Network (RNN) や Convolutional Neural Network (CNN) を用いず，Attention によって時系列データを処理する．

Transformer 内で使用される Attention は，入力を $Query(Q)$, $Key(K)$, $Value(V)$ とする，式 (1) で表される Scaled Dot-Product Attention である．

$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \qquad (1)$$

ここで，d は Q, K, V のチャネル数である．さらに，この Scaled Dot-Product Attention に対して並列化を行った Multi-Head Attention が Transformer では使用されている．

Self-Attention は上述の Attention のうち，Q, K, V 全てが同じ前層の出力から与えられるものである．Self-Attention はある時刻の出力を算出する際に，入力された特徴量の全ての時刻を同時に参照できるため，畳み込みより広い範囲の特徴を一度に参照できる．また，時間方向に伝播させる必要がある RNN よりも効率的に依存関係を学習できる．ただし，Attention では RNN のように時系列の順序を考慮できないため，Positional Encoding を特徴量に加算することによって順序情報を追加している．

1)　豊橋技術科学大学，Toyohashi University of Technology

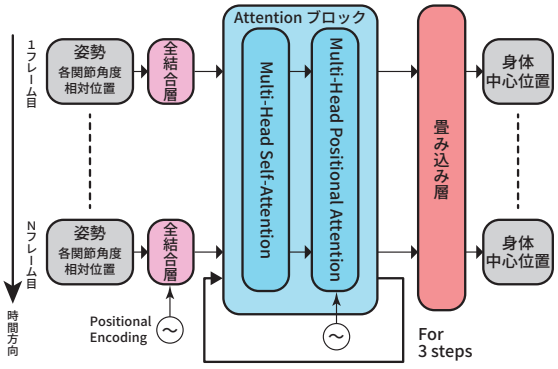


図 1: Self-Attention を用いた提案モデル．

3　提案モデル

前章の通り，長時間の特徴量の依存関係を学習しやすい Self-Attention は，自然言語と同じ時系列データであるモーションデータにおいても高い性能を発揮すると予想される．ゆえに本稿では，Self-Attention を用いてモーションデータを扱う学習モデルを提案する．

図 1 に示す提案モデルにおいて，入力は固定長 N の身体の各関節角度（クォータニオン表現）と順運動学による身体中心に対する相対位置の姿勢系列とし，出力は身体中心位置とする．関節相対位置を入力に含めることにより提案モデルは様々な人の身長に対応した身体中心位置の推定が可能である．

構造に関しては推定する際に入力の全系列が揃っているため，Transformer のような自己回帰は必要ない．よって，非自己回帰型 Transformer [2] のデコーダ側のみを基に構築する．ただし，フレーム単位毎の全結合層はカーネルサイズ 1 の畳み込みと等価であり，高周波成分の特徴のみが処理されやすい．これは自然言語の様な離散的な出力では問題になりにくいが，身体中心位置の遷移という連続的な出力では高周波成分の強調に繋がる恐れがある．これを回避するために，非自己回帰型 Transformer の Attention 層の後の全結合層を取り除く．さらに，最終層には畳み込み層を使用し，高周波成分の低減を図る．

本タスクでは入力の動作の違いと骨格の違いに対する 2 種類の汎化性能が必要となる．したがって，Universal Transformer [3] と同様に Attention ブロックを重み共有して繰り返すことにより，RNN のような帰納バイアスを取り入れ，汎化性能の向上を図る．予備実験において，パラメータ数が同じになる Attention ブロックが 1 個のみの場合や，重み共有無しで 3 回繰り返した場合よりも高い性能を発揮することを確認した．

正規化として Layer Normalization (LN) [4]，活性化関数として Leaky ReLU (LReLU) を導入した．

4　実験

4.1　データセット

データセットには我々が取得した歩きや座り，起き上がり動作といった様々な動作の MoCap データを使

表 1: 各モデルの平均二乗誤差とパラメータ数.

モデル	平均二乗誤差	パラメータ数
RNN	10.49	616,067
CNN	9.46	167,746,819
提案モデル	**7.73**	**153,475**

用する．本実験では，ルート関節を含む 24 個の関節を 120fps で計測された MoCap データを 60fps にダウンサンプリングし，16 フレームずつずらしながら 256 フレーム分切り出したものを入力とする．このとき，入力に用いるルート関節向きと各関節角度の総自由度は 96 自由度，ルート関節を除く関節相対位置の総自由度は 69 自由度であるため，入力は 256 × 165 次元となる．また，出力に用いる身体中心位置は 3 自由度であるから，出力は 256 × 3 次元となる．ただし，身体中心位置の水平面に関してはデータ毎の初期位置が原点になるように変更している．

学習には男性 7 名，女性 5 名（うち子供 1 名）の 435 個の MoCap ファイルから切り出した 119,150 個のデータを使用する．テストには学習データとは別の男性 2 名の 92 個の MoCap ファイルから切り出した 21,134 個のデータを使用する．

4.2 比較モデル

比較モデルとして，時系列データに適用可能な 2 種類の基本的な深層学習モデルを用いる．入出力は提案モデルと同一である．

1 つ目は，RNN を用いたモデルである．提案モデルの Attention ブロックを双方向の Gated Reccurent Unit (GRU) [5] に置き換え，重み共有無しで 3 回繰り返す．加えて，最終層の畳み込み層をフレーム毎の全結合層に置き換え，Positional Encoding も取り除いた．

2 つ目は，CNN を用いたモデルである．構造はチャネル連結のスキップ結合を持つ Pix2Pix [6] を基にし，時系列データに適用できるよう 1 次元畳み込みとした．また，正規化には LN，活性化関数には LReLU を用いた．

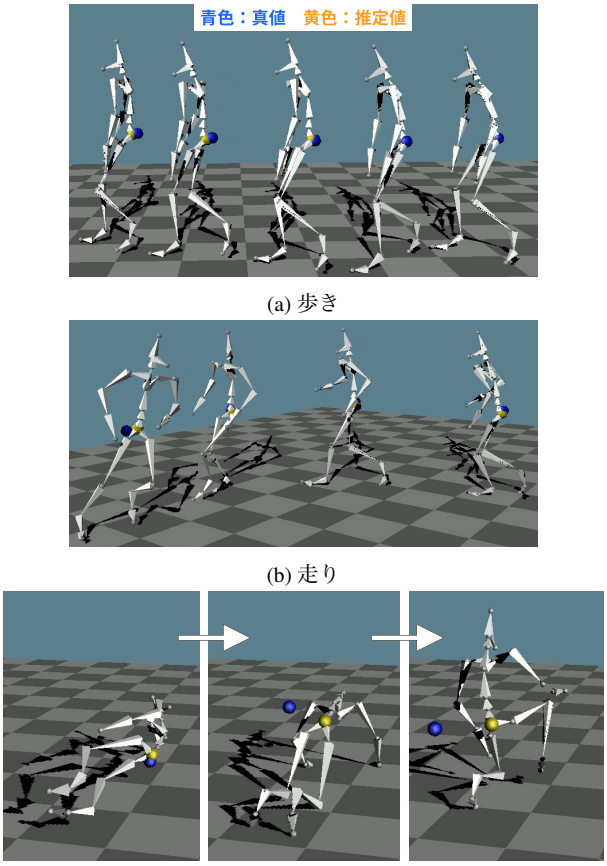
4.3 学習設定

最適化には 3 手法全てにおいて Adam 最適化を学習率 = 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$ で使用した．ただし，RNN を用いたモデルには勾配クリッピングを 5.0 で適用し，勾配爆発が起きた時点で学習を打ち切った．また，バッチサイズは 64，エポック数は 300 回（RNN を用いたモデルは 123 回で打ち切り）とした．

また，実装には TensorFlow を使用し，学習リソースには産総研の AI 橋渡しクラウド (ABCI) の NVIDIA Tesla V100 × 1 を利用した．

4.4 推定結果

各モデルの推定値と真値との平均二乗誤差を表 1 に，提案モデルの可視化結果を図 2 に示す．表から，提案モデルの誤差が最も小さいことがわかる．これは，Self-Attention の特徴である全ての時刻の特徴量を効率的に参照しつつ，重み共有によって汎化性能も確保できたためだと考えられる．可視化結果も単純な歩きや走り動作では良好である．加えて，提案モデルは必要なパラメータ数も少なく軽量である．ただし，可視化結果を比較すると CNN モデル以外では推定中心位置の不自然な振動が見られやすく，CNN モデルに対する見た目上の



(a) 歩き

(b) 走り

(c) 起き上がり

図 2: 提案モデルによる推定結果の可視化.

品質の劣化が感じられる．これは CNN モデルでは時間軸方向に次元数を縮小・拡張しながら畳み込みを適用したことにより，幅広い周波数領域の特徴量を処理し，高周波信号を抑制できたためと考えられる．

5 おわりに

本稿では，姿勢情報から身体中心位置を深層学習に基づいて推定する手法を提案し，Self-Attention を用いた提案モデルが他の基本的な深層学習モデルより最も誤差が小さく，単純な動作では品質を確保できることを確認した．ただし，提案モデルにおける高周波信号の抑制は不十分であったため，今後は CNN モデルのように見た目上の不自然な振動をさらに抑制することを目指す．また，起き上がりといった複雑な動作では足や手の滑りといった不自然さが取り除けていない．この複雑な動作でも正確に推定できるよう改良を検討する．

参考文献

[1] A. Vaswani *et al*. Attention is all you need. In *NeurIPS*, pp. 5998–6008. 2017.

[2] J. Gu *et al*. Non-autoregressive neural machine translation. In *ICLR*, 2018.

[3] M. Dehghani *et al*. Universal transformers. In *ICLR*, 2019.

[4] J. Ba, R. Kiros, G. E. Hinton. Layer normalization. *CoRR*, Vol. abs/1607.06450, , 2016.

[5] K. Cho *et al*. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*, pp. 1724–1734, 2014.

[6] P. Isola *et al*. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.