Vol.2021-AVM-112 No.8
2021/2/26

モーションキャプチャデータの関節回転角における ノイズ除去のための損失関数の検討

秋田健太† 松村誠明‡ 山本奏‡ 鶴野玲治†

概要：本研究では、マーカーレスモーションキャプチャデータに重畳するノイズ除去を目的とし、関節回転角が時系列方向に滑らかに連続して変化するという特徴を反映させるため、時系列方向の微分損失を含めて計算する損失関数を設計した。この結果、従来手法(関節回転角の平均二乗誤差)と同様のノイズ除去効果を発揮しつつ、特に学習の早期 epoch にて約 1%の損失削減効果が確認された。

キーワード：マーカーレスモーションキャプチャ，損失関数，微分損失

A Study of Loss Function for Noise Reduction of Joint Rotation Angular on Markerless Motion Capture Data

KENTA AKITA† MASAAKI MATSUMURA‡
SUSUMU YAMAMOTO‡ REIJI TSURUNO†

1. はじめに

3D モデルにモーションを付加する際､人のモーションをキャプチャして付加する方法が広く利用されている。モーションキャプチャの手法としては、光学式・磁気式・慣性式など様々な手法が存在するが、中でも複数台のカメラで撮影した映像から被写体の 2 次元姿勢(特徴点の 2 次元座標)[1, 2]を行い、三角測量により特徴点の 3 次元点座標を推定した後､IK 設定した骨格構造を各特徴点の 3 次元座標にフィッティングさせて 3 次元姿勢を得る手法[3]は、スタジオ等でキャプチャする必要がなく、スポーツなどのモーションをキャプチャする際に有効である。しかしながら、被写体がカメラから遠方にいる場合など、2 次元姿勢推定にて推定誤差が大きくなると、モーションにノイズが重畳してしまう。

本研究では、上記モーションに重畳するノイズの除去にディープニューラルネットワーク(DNN)を用いた手法を用い、その損失計算において時系列方向の微分損失を含めることによるノイズの除去効果を検証したので報告する。

2. 関連研究・課題抽出

モーションキャプチャデータに重畳しノイズを除去する手法として、時系列の振幅に対してローパスフィルタなどを適用する方法が存在するが、これらは適切にパラメータを設定しなければ、被写体の本来の動きまで鈍ってしまう等の影響が表れやすい。そのため、近年は DNN を用いた手法が提案されている。

Holden[4]は光学式モーションキャプチャデータに重畳したノイズをフィードフォワードネットワークで除去する手法を提案した。Mall ら[5]はノイズが重畳しやすい環境においてキャプチャされたモーションデータのノイズを再帰型ニューラルネットワークにより除去する手法を提案した。

これら DNN を用いる手法では、被写体の動きの特徴を学習した上で適切にノイズ除去を行うよう学習モデルが最適化されるため、前述のローパスフィルタのように過度に動きが鈍るような現象を抑制することができる。しかし、学習モデルの最適化は正解データ(ノイズが含まれていないモーションデータ)との誤差を所定の損失関数にて推定値との誤差が最小になるよう反復計算で学習するため、学習に多くの時間を要する。

3. 提案手法

モーションキャプチャデータにおける関節回転角は時系列方向に滑らかに連続して変化する特徴を持っているが、結果として得たい回転角のみの誤差を損失とする場合、時系列方向の特徴を考慮できていない。関節にノイズが重畳する際、回転角に対して強い速度変化・加速度変化が生じることが考えられるため、これらを利用すれば効率的にノイズ除去ができることが予想される。そこで本研究では損失関数内で時系列方向(フレーム)に対して微分を行い、正解データとの差を取る下式の損失を提案する。

$$loss = \mathbb{E}[\|G(x) - y\|] + \alpha\mathbb{E}[\|G'(x) - y'\|] + \beta\mathbb{E}[\|G''(x) - y''\|]$$

† 九州大学
Kyushu University
‡ NTT メディアインテリジェンス研究所
NTT Media Intelligence Laboratories

ⓒ2021 Information Processing Society of Japan

G はネットワークであり、x,y はそれぞれノイズが重畳した入力データと対になる正解データを表す。α, βは各要素のバランスを調整するためのパラメータとして用いる(α=0, β=0の時は関節回転角の平均二乗誤差となる)。

これにより、より関節回転角の変化特徴を生かした学習が可能になると考えられる。

4. データ作成およびネットワーク構築

4.1 データセット

モーションデータは汎用形式として BVH フォーマットが広く利用されており、このフォーマットで収録された公開データセットは複数存在するが、提供元それぞれで骨格構造が異なったり、関節回転角の回転軸の順番が異なるなど一貫性がない。そのため、本研究ではモーションキャプチャスーツを用いてモーションデータを複数取得し、これらを図 1 に示す骨格構造になるよう変換･出力した BVH ファイル(root ノードのみ 3 軸移動量と 3 軸回転角の計 6 パラメータを保持し、それ以外の関節ノードは 3 軸回転角のみをパラメータとして保持する形式)を用いる。ここで、本研究では関節回転角を対象としてノイズ除去を行うため 3 軸移動量のパラメータは除外した関節回転角そのままの値を正解データとし、これに対して疑似的にノイズを加えたデータを入力データとする。疑似的に重畳するノイズは、前章にて述べたようにマーカーレスモーションキャプチャにて重畳しやすいランダムノイズとスパイクノイズをそれぞれモデル化する。

4.1.1 ランダムノイズ

ランダムノイズは 2 次元姿勢推定[1, 2]における微小な推定誤差をモデル化したノイズであり、フレーム毎に微細な振幅として現れる。一般にフレーム内において被写体が大きく写っており、推定精度が高い場合はノイズが小さくなり、反対に小さく写っているなど、推定精度が低い場合はノイズが大きくなる傾向がある。

4.1.2 2 次元姿勢推定[1, 2]の結果を鑑みるとランダムノイズはガウスノイズでモデル化できるように考えられるが、正解データの関節回転角にガウスノイズを重畳させると、エラー! 参照元が見つかりません。に示すように骨格構造における root ノードから末端ノード(足先や手先など)に移動するに従って正解データの 3 次元座標との誤差が大きくなる。そこで、root ノードから末端ノードに移動するに従って、正解データの 3 次元座標との誤差を最小化しつつ、設定するガウスノイズの標準偏差を維持するように補正を加えたノイズを重畳させることで、ランダムノイズをモデル化する。スパイクノイズ

スパイクノイズは 2 次元姿勢推定[1, 2]において、対象の被写体が他の被写体と隣接した場合に生じやすい他者特徴点との接続誤りをモデル化したノイズであり、隣接し続けている間は長時間に渡って断続的に現れる。スパイクノイズは過去の推定フレームが正しいという仮定のもと、現フレームの 3 次元姿勢を予測することで推定誤誤差を軽減する手法[3]も提案されている。しかし、過去の推定フレームの推定精度に依存するため、本研究では過去の推定フレームを用いたノイズ軽減は行わない。

ノイズのモデル化には、骨格構造の中からランダムに関節を選択し、当該関節より先(root ノード方向とは反対側)の関節回転角を、データセット中における他の動作の関節回転角に置き換えることで代用する。なお、連続するフレーム長・同時に生じるスパイクノイズの最大関節数はパラメータとして与えられるよう設計する。

4.2 ネットワーク

DNN への入力データは、前節で述べた通り BVH ファイルのパラメータから root ノードの 3 軸移動量を除外したデータを時間方向に所定のフレーム数連結したテンソルを用いる。出力データも入力データと同サイズのテンソルとし、学習時には出力データと正解データの間の損失を 3 章に記す計算方法算出する。DNN のネットワークアーキテクチャは図 3 に示すように AutoEncoder を用い、同サイズのテンソルには Skip connection を設定する。

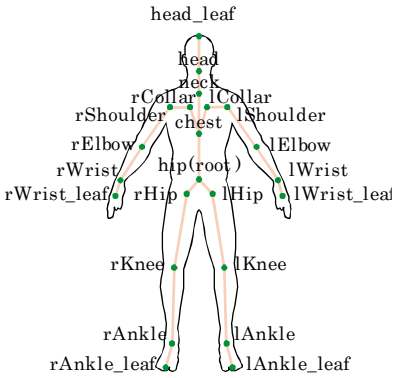

図 1　骨格構造

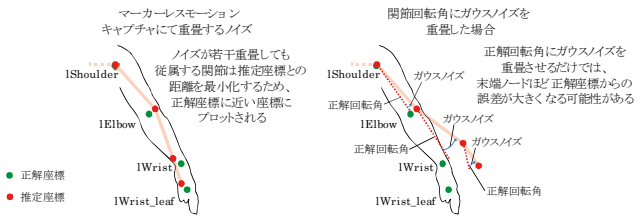

図 2　ガウスノイズ重畳時の誤差

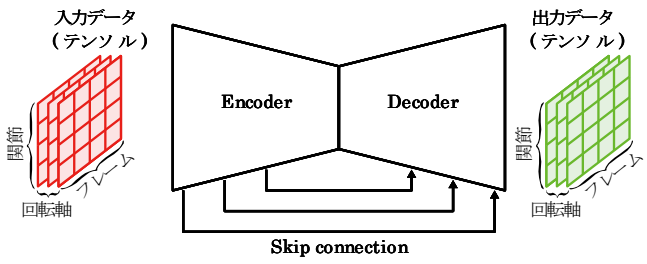

図 3　DNN のネットワークアーキテクチャ

5. 実験・考察

5.1 データセット

学習データは 6 名の被写体を対象に、日常動作やストレッチ・スポーツ等の動作を演じた BVH ファイル(30fps)を対象に、時間方向のフレーム数を 16 に設定したテンソルを学習用(Train)に 1,000,000 サンプル、検証用(Validation)に 500,000 サンプル生成した。

ランダムノイズの最大標準偏差には 10 を設定し、サンプル毎にランダムに設定した。スパイクノイズのフレーム長は標準偏差 16 の正規分布で決定し、同時に生じるスパイクノイズの出現数は最大 3 として出現箇所と共にランダムに決定した。

5.2 ネットワーク・損失関数

AutoEncoder には時間軸方向にのみダウンサンプリングする(関節数のダウンサンプリングは行わない)ようにカスタマイズした ResNet50 を用い、学習率 1.0e-3・バッチサイズ 512 に設定した Adam [6]にて学習を行った。

損失関数にて計算対象とする微分値は関節回転角に対して 1 次微分値(角速度に相当)と 2 次微分(角加速度に相当)を計算し、それぞれ正解データとの誤差を計算したうえで、回転角の誤差との平均値を損失として計算した。なお、関節回転角に対して時間微分を行う際、回転角の時間差分に対して時間間隔Δt(本データセットでは 30fps のため 1/30)で除算されるため、強いランダムノイズやスパイクノイズが重畳したデータには、角速度・角加速度に極めて大きな値が検出され、損失が収束しなくなる傾向が確認された。そのため、本実験においては関節回転角のスケールと一致させるため、$\alpha = 1/900, \beta = 1/810{,}000$ を用いて学習した。

5.3 実験結果

前節に記載の微分損失を用いた損失関数を用いて学習させた場合(提案手法)と、関節回転角の平均二乗誤差を損失関数として用いて学習させた場合(従来手法)を各 8 回学習させた際の Validation 損失の平均推移を図 4 に示す。なお、提案手法は学習過程においては微分損失を含めて損失計算が行われるが、図 4 に示す損失は関節回転角の平均二乗誤差のみのデータを表す。

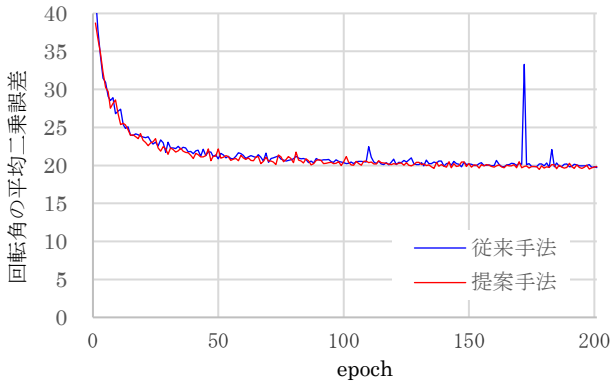

図 4 学習過程における損失値の推移

視認できるほどの優位差は確認できなかったが、先頭 100epoch(従来手法において後半に数回異常値が検出されたため)の各損失に対して積分したところ、平均的に約 1%損失を削減できる効果が確認できた。なお、前節に記載の通り本施行では収束のため微分損失の影響を意図的に下げて実験を行ったが、最適なα, βを用いることで効果をさらに強調できると考えられる。

学習データに含まないモーションに対してデータセット作成時と同様の手法で疑似的にノイズを重畳して連続するモーションデータを生成し、このモーションデータに対して提案手法・従来手法それぞれを適用した際の、ある関節回転角におけるノイズ除去結果(正解値との回転角差分)を図 5 に示す。

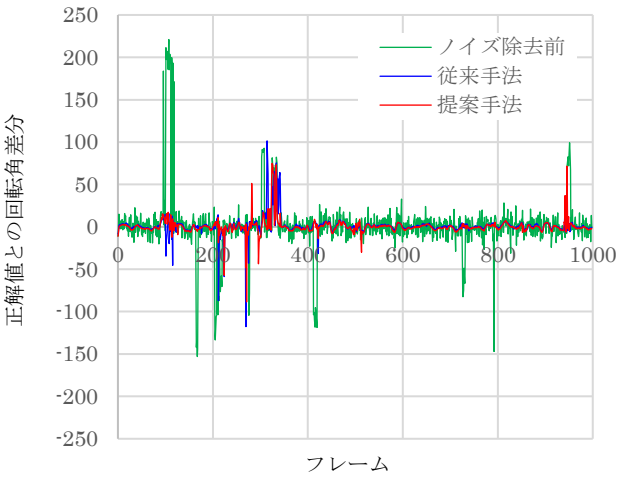

図 5 ノイズの除去結果

図 5 において、ノイズ除去前の信号では、±10°前後のノイズがランダムノイズを表しており、稀に現れる強い差分信号がスパイクノイズを表す。ノイズの除去効果としてはランダムノイズ・スパイクノイズ共に従来手法と同様に除去できている様子が確認できる。スパイクノイズに対しては、短いフレームで生じるものであれば除去されている様子が確認できたが、長いフレームに対して断続的に重畳する場合には、除去しきれない傾向が確認された。これら除去しきれない長いフレームに対して断続的に重畳するスパイクノイズに対しては、入力データとして 16 に設定したフレーム数を増加させたり、損失計算に例えば GAN などを用いて人間らしい動作か否かの評価値を新規に組込むことで、さらに軽減できると考えられる。

6. 結論・課題

本研究では、マーカーレスモーションキャプチャデータに重畳するノイズ除去を目的とし、関節回転角が時系列方向に滑らかに連続して変化するという特徴を反映させるため、時系列方向の微分損失を含めて計算する損失関数を設計した。この結果、従来手法(関節回転角の平均二乗誤差)と

ⓒ2021 Information Processing Society of Japan

同様のノイズ除去効果を発揮しつつ、特に学習の早期 epoch にて約 1%の損失削減効果が確認された。

今後の課題として、以下のものが考えられる。

・パラメータα, βの最適化
・損失関数の設計変更

本実験では関節回転角のスケールと一致させるパラメータα, βを用いたが、収束を阻害しない範囲で最適化を行うことで更なる損失削減ができる可能性がある。また、GAN 等の導入により「人間らしさ」を評価値として損失関数に組込むことで、ランダムノイズ・スパイクノイズをさらに削減可能性がある。

参考文献

[1] Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR. (2017)
[2] Papandreou, G., Zhu, T., Chen, L.C., Gidaris, S., Tompson, J., Murphy, K.: Personlab: person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In: ECCV. (2018)
[3] Takuya O., Yosuke I., Kazuki Y., Wataru T., Yoshihiko N.: Video Motion Capture from the Part Confidence Maps of Multi-Camera Images by Spatiotemporal Filtering Using the Human Skeletal Model. In: IROS. (2018)
[4] Daniel H.: Robust solving of optical motion capture data by denoising. ACM Trans. Graph, 37(4), (2018).
[5] Utkarsh M., G. Roshan L., Siddhartha C., Parag C.: A deep recurrent framework for cleaning motion capture data. arXiv, (2017).
[6] Diederik P. K., Jimmy B.: Adam: A method for stochastic optimization. In International Conference on Learning Representations (ICLR), 2015.

ⓒ2021 Information Processing Society of Japan