第 28 回日本バーチャルリアリティ学会大会論文集（2023 年 9 月）

人間の３Ｄモーションデータの誇張表現生成ネットワークモデルの提案

Proposal of Network Model for Emphasis of Human 3D Motion

飯田航平 1)，清水創太 1)
Kohei IIDA, Sota SHIMIZU

1) 芝浦工業大学大学院電気電子情報工学専攻 （〒108-8548 東京都江東区豊洲 3-7-5, ma22012@shibaura-it.ac.jp）

概要：本研究では，元の人間の動きを誇張表現した VR 用 3D アバター動作を生成するネットワークモデルを提案する．ここでの誇張表現とは，アニメーションなどでしばしばみられる，一見物理的にはおかしいにもかかわらず，逆にわかりやすい動作を表現することを意味する．誇張表現を適用することで，人間にとって逆に違和感のない動作生成を目指す．

キーワード：アニメーション，機械学習，モーションキャプチャ，誇張表現

1. はじめに

現在，VR などの仮想空間でアバターを動かす方法として主に専用の機器を用いる手法と可視光カメラを用いた手法が報告されている．専用の機器を用いた手法には，ビーコン式 VR 機器，近赤外光カメラ，Kinect のような ToF 方式のものなどが上げられる．一般的に，高額ではあるが精度が良く違和感のないモーションキャプチャを行うことができる．一方，可視光カメラを用いた手法には，カメラからの可視光映像に OpenPose 等を適用し，姿勢推定する手法が上げられる[1]-[3]．使用するカメラの台数に違いがある場合があるが，カメラのみで安価にモーションキャプチャを実現していることが特長である．しかし，映像からの姿勢推定に学習器を用いていることもあり，実時間処理や精度の面で改善の余地が残されている．

本研究では，特に単眼カメラを用い，精度の問題に対して「違和感の軽減」という別の指標を導入することにより，独自の 3D モーションキャプチャデータの取得を目指す．

また，WIRED.jp では，プレイステーションのスタジオで働くゲーム効果音職人的技術者の仕事を紹介している．「ゴッド・オブ・ウォー」などのゲームでは，実物通りに音を録音して効果音が作られているのではなく，様々なアイテムを活用して効果音を再現している．しかし，ゲームをプレイしているときに違和感を意識せず，むしろ臨場感を感じ，よりゲームに没入できるようになっている．

さらに，「モーションキャプチャデータを用いた加速度制御手法によるメンタルモーション生成」という研究では，物理法則に則ったものが観客にとっては不自然に移ることがあることを紹介しており，アニメーションの中で用いられる誇張表現が違和感を打ち消すのに役立つと述べている[4]．当該研究では，物理法則に則っていないにもかかわらず，人間がリアルに感じる動きのことを「メンタルモーション」と呼んでいる．

こうした背景から，一見物理的にはおかしいにも関わらず，逆にわかりやすい誇張表現を適用することで逆に人間にとって違和感の少ない動作生成を行うネットワークモデルを提案する．

2. 関連・先行研究

単眼カメラを用いた姿勢推定の研究は数多くされている[1]-[3]．OpenPose は定点可視光カメラ 1 台のみで姿勢推定を行う手法である．このようなカメラを用いた姿勢推定では主に動画像から関節点どうしの関係を深層学習により推定している．そのため，必ずしも違和感のない姿勢推定を実現しているとは言えない．本研究では，上述の研究で得られた 3D モーションデータを加工することで違和感のないモーションに変更して生成することを目指す．

また，モーションデータの制御という点では，キャラクターのスケルトンを制御するキャラクターコントロールと呼ばれる研究がある[5][6]．AI4Animation: Deep Learning for Character Control という一連の研究では，キャラクターアニメーションのための包括的なフレームワークをまとめている．その中でも，DeepPhase は教師なしで非構造化されたモーションデータから周期的特徴を学習することで連続的で滑らかな動きを生成することが出来る[5]．

さらに，ある基本的なモーション(例えば，「歩く」，「走る」など)に，独自のスタイル(「おじいさん」，「楽しく」)を合成するモーションスタイル変換という研究が報告されている[7]．「Unpaired Motion Style Transfer from Video to Animation」では，モーションスタイル変換のためのデータ駆動型フレームワークを紹介している[7]．この研究では，動画からのスタイル抽出を実現しており，ラベルのないモーションデータのスタイル変換が可能である．誇張表現も大局的な定義で見るとスタイル変換の一種であり，上述の研究におけるスタイル合成とみなすことが出来る．

違和感の軽減という指標から，モーションキャプチャのリアルな動きをセルアニメーションの誇張表現を含めた動きに変換する研究も報告されている[4][8]．「モーションキャプチャデータを用いた加速度制御手法によるメンタルモーション生成」では，物理法則に則っていないが，人間がリアルに感じる動き「メンタルモーション(誇張表現)」が違和感を軽減させることに役立つと述べている[4]．

本研究では，モーションスタイル変換のフレームワークをベースラインに，誇張表現をスタイルと捉え，メンタルモーション生成ネットワークを AI により構築する．

3. 誇張表現生成ネットワーク

一連の 3D モーション M を 2D アニメーション A のスタイルを取り入れて誇張表現を生成することが，我々の目的である．そのため，2D アニメーションの動きをスタイルとして扱える特徴を抽出することが重要である．本研究では Aberman らの研究を参考に，2D アニメーションの誇張表現 A をスタイルとして抽出し，モーション M に 2D アニメーション A のスタイルを合成して誇張表現を実現する．図 1 に本研究で提案するネットワークモデルを示す．ここでは，アニメーションから誇張表現を抽出するエンコーダ E_a と，モーション M からスタイルを除外した一般的動作を抽出するエンコーダ E_m からスタートする．デコーダ F では，Huang らが提案した画像のスタイル合成などに使われている Adain 層を用いて，E_m に誇張表現 E_a を合成する[9]．これらのエンコーダ，デコーダモデルによって生成されたモーションを誇張表現識別器 D に入力して，真偽を識別して学習を進める．本研究では，各種誇張表現をラベルとするデータセットを作成するのが難しいため，敵対的生成ネットワークを用いて，教師なし学習を行う．

3.1 エンコーダ

エンコーダでは一次元の時間畳み込み層を適用して特徴を抽出する．このとき，元の一連の 3D モーションを M，2D アニメーションを A とし各エンコーダへの入力とする．モーション M は関節の回転運動である単位四元数で表現する．

$$M \ni R^{p\times 4j}$$

2D アニメーションビデオ A は関節点の位置で表現する．

$$A \ni R^{p\times 3j}$$

このとき，p はポーズクリップ数，j は関節数である．

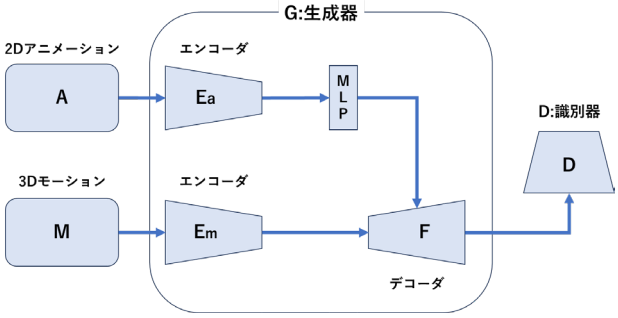

図 1　提案する動作の誇張表現生成ネットワーク

3.2 デコーダ F

適応的インスタンス正規化(Adain)を行う，いくつかの残差ブロックと時間解像度を上げるストライドを有する畳み込み層構造である．Adain は特徴の活性化にアフィン変換を用いて，誇張表現動作を生成する．

3.3 識別器 D

識別器 D では，入力された 3D モーションがデコーダから出力されたものか元の 2D アニメーション A かを識別することで，学習が進むにつれて誇張表現を違和感なく組み込んだモーションの生成を実現する．

4. おわりに

本研究では，モーションデータと 2 次元アニメーションデータに対して，時間畳み込みを行うことで特徴を抽出し，この 2 つを Adain 層によって合成する生成器 G と，生成器 G の出力を識別する識別器 D を含む誇張表現動作生成ネットワークを提案した．今後は 2D アニメーションからの誇張表現ためのより良い特徴量の抽出手法の考案・実装を目指す．次に，誇張表現動作を生成する際に，速度を考慮した手法の考案・実装を目指す．また，提案ネットワークから実際に生成されたモーションの解析及びアンケート調査を行い，違和感についての評価を行いたい．

謝　辞

本研究の一部は科研費 No. 21K03983 に基づいて実施されている．本研究の遂行に当たり，技術面のみならず多くのご助言，ご協力頂いた芝浦工業大学デザイン工学科人間支援知能ロボティクス研究室の皆さまに心より御礼申し上げます．

参考文献

[1] Z. Cao, T. Simon, S. Wei, Y. Sheikh, Realtime multi-person 2D pose estimation using part affinity fields, Proc. of CVPR (2017).

[2] A. Toshev and C. Szegedy, Deeppose: Human pose estimation via deep neural networks, Proc. of CVPR (2014).

[3] X. Peng, Z. Tang, F. Yang, R. S. Feris, and D. Metaxas, Jointly optimize data augmentation and network training:

Adversarial data augmentation in human pose estimation, Proc. of CVPR (2018).

[4] 今間俊博, 近藤邦雄, 栗山仁, 古家嘉之, モーションキャプチャデータを用いた加速度制御手法によるメンタルモーション生成, 日本図学会図学研究, 第 39 巻, 第 2 号, 通巻 108 号, pp. 3-10 (2005).

[5] S. Deng, M. Luo, C. Ai, Y. Zhang, B. Liu, L. Huang, Z. Jiang, Q. Zhang, L. Gu, S. Lin, Synergistic Doping and Intercalation: Realizing Deep Phase Modulation on MoS2 Arrays for High-Efficiency Hydrogen Evolution Reaction, Angew. Chem., Int. Ed., 58, pp. 16289– 16296 (2019)

[6] S. Starke, Y. Zhao, F. Zinno, and T. Komura, Neural animation layering for synthesizing martial arts movements, ACM TOG, 40(4), pp.1–16 (2021).

[7] K. Aberman, Y. Weng, D. Lischinski, D. Cohen-Or, and B. Chen, Unpaired motion style transfer from video to animation, ACM TOG, 39 (4), pp. 64:1-64:12 (2020).

[8] 今間俊博, 齋藤隆文, 阿部翔悟, アニメーションにおける動きの種類分析と誇張表現の適応手法, 日本図学研究, 第 47 巻, 第 2-3 号, pp.13-23 (2013).

[9] X. Huang and S. Belongie, Arbitrary style transfer in realtime with adaptive instance normalization, Proc. of ICCV, pp. 1510–1519 (2017).