骨格データを用いた身体部位の領域推定による動画生成システムの設計

M2022SE002 本多 優斗
指導教員：沢田 篤史

1 はじめに

現在，様々な手法で動画の自動生成システムの導入や研究が行われている．そのなかでも機械学習を用いた動画生成システムの研究が注目されている．

機械学習を用いた動画生成システムの課題として手足の関節がねじれるなど物理的にあり得ない挙動を生成してしまうことが挙げられる．また，機械学習を用いたシステムはブラックボックス性により原因となる箇所の特定が難しく保守が困難なことも挙げられる．これらの原因として機械学習システムの処理の詳細を人間が完全に把握することができないことがある．これにより，仮に物理的に整合しない動画が生成されるような不具合が生じたとしても，修正すべき箇所の特定や修正の実施が困難となっている．

既存の機械学習を用いた動画生成システムでは，実際の動画や物体のベクトルなどのデータをもとに動画を生成する [1][2][3]．しかし，ベクトルの情報だけでは関節などの位置関係が明確でないので動画生成時に不整合を起こしてしまう可能性がある．

本研究の目的は物理的に整合した動画を生成することができるシステムの実現することである．また，その際に問題となる保守性を担保できるようにすることも目的とする．この目的を達成するための技術課題として次の 2 つを設定する．

1. 骨格データと身体の部位データを用いた動画生成システムのアーキテクチャ設計．
2. 提案したアーキテクチャにより保守性が担保されることの確認．

1. に対して骨格データを用いた動画生成システムにおいて適切に処理を分割しモジュール化するアプローチを行う．そのためにレイヤードスタイル [4] とパイプアンドフィルタースタイル [4] を用いて動画生成システムのアーキテクチャを設計する．レイヤードスタイルによりピクセルデータを扱う層，各フレームにおける骨格の位置情報を処理する層，位置情報をもとに姿勢の情報を処理する層に階層を分割して処理することで，各層における役割を定義することができる．また，レイヤードスタイルの各層は独立しているので変更を加えた際の他の層への影響が少ない．パイプアンドフィルタースタイルを用いることで各層内のフィルタを処理ごとにモジュール化し独立させることができる．これにより層内の処理の変更が他の処理に与える影響を少なくすることができる．処理を各モジュールごとに分割することにより各モジュールから得られる出力を確認するのが容易なので修正箇所の特定も容易となることが期待できる．

2. に対してはシステムの出力から不具合が起きている箇所の特定をすることやアーキテクチャ内のモジュールを別のモジュールに変更してもシステムが機能するか評価する．システム保守性が担保されていることを示すためにシステムの修正箇所を容易に特定できることやシステム内のモジュールを別の同じ処理を行うことができるモジュールに変更できるか示す必要がある．

本研究では，骨格データを用いた動画生成システムの実装を行う際に機械学習システムの処理を層ごとに分け各処理をモジュール化することで保守性の担保された動画生成システムのアーキテクチャを設計した．システムの実装は Python を用いて行った．

実装したシステムを用いて行った動画生成実験では，十分な品質の動画を生成することはできなかったが，提案するアーキテクチャに基づいて実装したことにより，原因となる箇所の特定を行うことができた．

動画生成システムの保守性の向上は定性的であるが示すことができた．本研究では，修正箇所の特定や別の同じ処理を行うモジュールへの変更可能性について考察した．

2 既存研究とその問題点

Tulyakov ら [1] 研究では，ガウス分布で求めた次のフレームでの動体の位置情報をもとに計算したベクトルを入力データとして動画生成を行う MoCoGAN を提案した．MoCoGAN は，背景と動体を分けて生成することができることから，それぞれに特化した生成が可能である．これにより動きが滑らかに表現されている動画の生成を行うことができる．Yan ら [2] の研究では通常の人間動画だけでなくその動画から抽出した骨格動画を学習させ動画生成を行う．骨格動画を学習させたことにより人物の生成が従来のものよりも骨格を捉えた生成を行うことができ，動画中の人物の動きの滑らかさの向上にもつながっている．Yamamoto ら [3] の研究では，動画の次のフレームを生成するための時系列データに，オプティカルフローにより計算されたデータを利用している．オプティカルフローデータを学習させたことにより動体を捉えた動画の生成が可能となり，フレーム間の動きの滑らかさの向上につながっている．

既存研究の問題点として次の 2 つが挙げられる．

1. 手足の関節がねじれるなど物理的にあり得ない挙動を生成してしまう．
2. 基本的に唯一の機械学習モジュールからなるシステムであり，そのブラックボックス性によりシステムの修正箇所の特定や再利用が困難．

1. について既存の動画生成システムでは人間の身体部位の可動域を考慮したデータでの学習を行っておらずその特徴を学習することができない．それにより結果とし

て手足のねじれなど物理的に整合していない生成が行われてしまったと考えられる．

2. について既存の機械学習システムは 1 つの機械学習モジュールに処理を依存していることが挙げられる．1 つの機械学習モジュールに処理を依存した場合，機械学習にはブラックボックス性が存在するので処理内容の詳細を把握することが難しく，原因箇所の特定が困難である．また，1 つの機械学習モジュールに処理を依存した場合，再利用時に変更の余地が無く再利用性に優れていない．

3 目的と技術課題

本研究の目的は物理的に整合した動画を生成することができるシステムの実現とその際に問題となる保守性を担保できるようにすることである．本研究では，物理的に整合した動画を生成するには人間の骨格データを入力する必要があると考えた．仮に，骨格データと動画データを 1 つの機械学習モジュールに処理させるような構造をとる場合，唯一の機械学習モジュールに入力されるデータの種類が増え，期待される処理内容も複雑化する．一方で，機械学習モジュールのもつブラックボックス性により不具合の特定がますます困難になる．結果として，変更が必要な箇所をどのように変更すれば不具合を解消することができるが分からないなど保守性に問題を抱えたシステムが実装されてしまうことがある．

本研究の技術課題は次の 2 点である．

1. 骨格データと身体の部位データを用いた動画生成システムのアーキテクチャ設計．
2. 提案したアーキテクチャにより保守性が担保されることの確認．

1. では，保守性を担保するために機械学習のブラックボックス性を低減することのできるアーキテクチャを定義する必要がある．機械学習システムの保守性を担保するためには機械学習のブラックボックス性が与える影響を少なくするためのアーキテクチャ設計を行う必要がある．

2. では，設計したアーキテクチャに基づいて骨格データを用いて動画を生成するシステムを設計実装する必要がある．その故で，システムが物理的に整合した妥当な動画を生成することができるか，また，仮に不具合が生じた場合にその原因特定がしやすいかを評価する必要がある．これらを通じて提案方法の妥当性や有用性を示す必要がある．

4 アプローチ

4.1 技術課題へのアプローチ

技術課題を解決するために，次のようなアプローチを行う．

1. に対して骨格データを用いた動画生成システムにおいて適切に処理を分割しモジュール化するアプローチを行う．そのために，レイヤードスタイルとパイプアンドフィルタースタイルを用いることで各層ごとに役割を分けフィルタごとにモジュール化することで保守性を向上させる．

2. に対しては，骨格データとそれに対応する動画データから物理的に整合した動画生成ができることを実験により確かめる．また，仮にシステムの出力から不具合が起きていることが判明した場合，その原因箇所の特定をすることが容易に行えることを定性的に示す．さらに，アーキテクチャ内のモジュールを別のモジュールに変更してもシステムが機能することについても定性的に評価する．

4.2 システム全体のアーキテクチャ

動画生成を行うために必要な処理は動画生成 GAN に入力するデータの前処理と前処理したデータを入力し動画を生成を行う処理である．この処理は，人間が映像の構成内容を理解する構造と似ている．人間の脳が映像を処理するとき映像から動きや形から何を表しているのかを推測することにより映像の構成を理解している．

本研究のシステムは人間の脳が映像処理を行う構造を模倣させるためにレイヤードスタイルとパイプアンドフィルタースタイルを組み合わせて設計をした．レイヤードスタイルを用いることにより，ピクセルデータを扱う信号レベルの処理を行う層，骨格の位置情報を扱うベクトルレベルの処理を行う層，位置情報をもとに姿勢の情報を処理する人物レベルの処理を行う層に分割した．レイヤードスタイルにより定義された各層は独立しており，変更を加えた際に他の層への影響が少ないという利点がある．パイプアンドフィルタースタイルにより各層の処理をフィルタごとに分割しモジュール化することで各モジュールから得られた出力を確認することができ修正箇所の特定を容易に行うことができる．また分割したモジュールは独立しており変更を加える際に他のモジュールに与える影響を小さくすることができるという利点がある．

本研究のシステムは図 1 に示した通り人物処理層, ベクトル処理層，信号処理層の 3 つの層により構成されている．この 3 つの層により人間の脳が映像を処理する構造を定義した．

人物処理層は，画像内で検出された人物の姿勢や向きの人物レベルのデータを扱う層である．この層により画像内の人物の情報が解析され，各部位の状況などを補正することができる．この層は人間の脳が画像内の物体の形を捉えた際に処理する工程を模倣している．主な処理として LSTM を用いた骨格データの修正が挙げられる．

ベクトル処理層は，人間の骨格座標や身体の領域データから得られる RGB データのベクトルレベルの人物の情報を処理することができる．この層は人間の脳が数値として人間を捉えた際に処理する工程を模倣している．主な処理として AlphaPose[5] を用いた骨格推定やセマンティックセグメンテーションを用いた色の情報の抽出が行われている．

信号処理層は，既存の動画生成や人物処理層とベクトル処理層で変換された信号レベルのデータを処理する層である．この層では人間の脳が姿勢や座標のデータといった情報から映像内の構成要素を処理する工程を模倣している．この層の処理には既存の動画生成モジュールを使用した動画生成や生成された動画を処理して得られたデータを入力することで動画生成を行うモジュールが存在している．

各層を超えたデータの受け渡しは各層の中間に存在するモジュールに入力することで実現している．例として修正した姿勢から骨格の位置情報を推定する人物骨格推定モジュールが挙げられる．

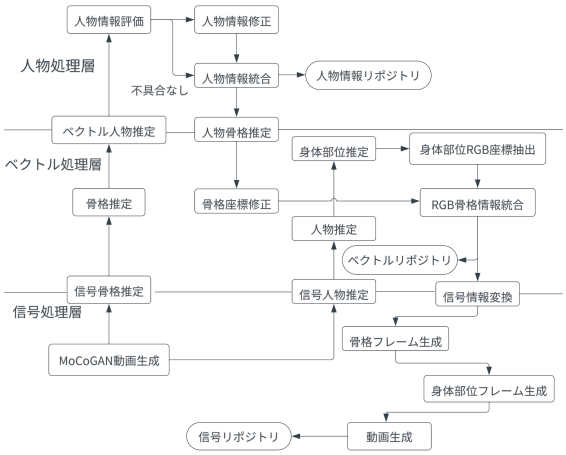

図 1　動画生成アーキテクチャ

セマンティックセグメンテーションを各フレームに適用することでフレーム内のどこの領域に人物が存在していたのかがわかる．これにより求められた領域内の情報を抽出すれば人間の色の情報を獲得することができる．

5　動画生成システムの実験

本研究のシステムは実行時に骨格データとピクセルデータの 2 つのデータを入力して動画生成を行うことができるようにシステムのアーキテクチャ設計を行った．しかし，2 つのデータを入力する場合に多くの GPU のメモリ容量を使用しなければならず，現在の実装環境ではメモリ不足で動作させることができなかった．

代替案として図 2 に示すように、骨格データのみを入力データとして実装し、動画生成を行った．このアーキテクチャでは学習時には設計したアーキテクチャと同様にセマンティックセグメンテーションで得られたピクセル値を動画生成モジュールに学習データとして入力するが，実行時にはこの処理を行わずに骨格データのみを処理し入力する．このアーキテクチャに基づいて動画生成を行なったが期待していた結果は得られなかった．

GAN を用いない方法での実装も行った．こちらは人物が映る動画フレームを生成することができた．しかし，手の指などを正確に生成することができなかった．

6　考察

6.1　保守性についての考察

本研究では，骨格データを入力することで動画生成を行うシステムの実装を行なった．実験の結果，十分な品質を持つ人物動画の生成をすることはできなかった．ただし，本研究のアーキテクチャ設計により出力結果から図 3 の赤枠の箇所である骨格座標修正モジュール，動画生成モジュールに不具合があることを容易に特定することができた．

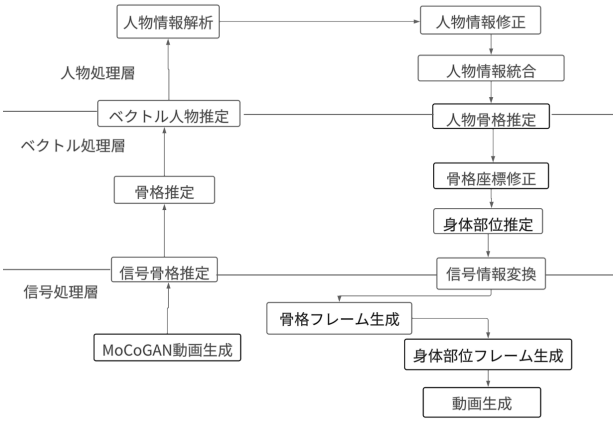

図 2　実装したシステムのアーキテクチャ

特定が容易だった理由としてレイヤードスタイルによって処理を役割ごとに各層に割り当て，パイプアンドフィルタースタイルによってフィルタごとに出力結果を確認することができたことが挙げられる．各層に役割を持たせることにより出力結果からどの層から得られたデータなのかを容易に把握することができた．また，フィルターごとに処理をモジュール化することによりモジュールの出力結果を確認することで容易に問題となっているモジュールを特定することができた．

既存の機械学習を用いた動画生成システムでは 1 つの機械学習モデルに処理を依存している．このような場合本研究で実装したシステムのように処理ごとの出力結果を確認することができないので修正箇所を容易に特定することができない．

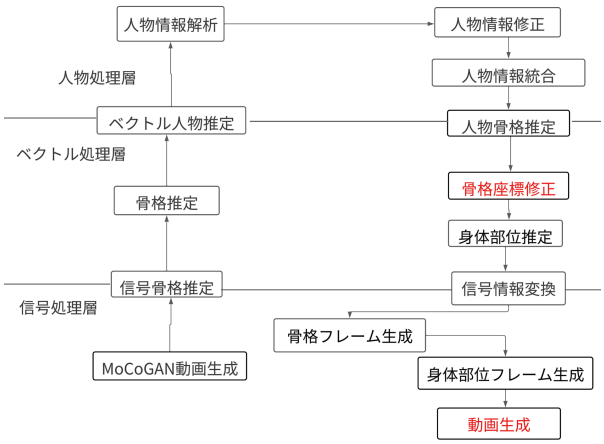

図 3　修正箇所

6.2　モジュールの変更可能性についての考察

本研究で提案したアーキテクチャの変更可能性を確認するために利用した既存の機械学習モジュールと同様の動作を行う別の機械学習モジュールに変更しても機能するか考察した．

既存の動画生成システムでは 1 つのモジュールに処理を依存しているので変更することは難しいが本研究のシ

ステムはモジュールに処理を分割しているので処理ごとのモジュールの変更は可能であると考えられる．

本研究ではシステムに入力する動画データを得るために MoCoGAN を使用したが他の動画生成システムで生成した動画でも人物動画であれば変更は可能だと考えられる．その理由として，本研究で用いている骨格推定などの各モジュールには入力動画の解像度の制限がないことが挙げられる．

本研究では骨格推定に AlpaPose を使用した．このモジュールの変更が可能である条件として，入出力が同様の骨格推定モジュールの存在を挙げれば良いと考えた．AlphaPose は 25 個の骨格座標を推定するモジュールであり OpenPose も同様に 25 個の骨格座標を推定するモジュールである．AlphaPose から OpenPose に変更することは 25 個の骨格の位置座標データを AlphaPose と同じ形式で出力させることのできるモジュールを実装すれば可能である．よって別の骨格推定モジュールへの変更は容易であると言える．

本研究ではセマンティックセグメンテーションを行う際に DeepLabV3[6] を利用した．このモジュールを変更する場合，DeepLabV3 同様に人間に対応することができるモジュールである必要がある．一般的にセマンティックセグメンテーションを行なった場合出力されるデータとして推定した範囲を塗りつぶした画像データが出力される．本研究で実装した色の情報を抽出するモジュールはその塗りつぶされた範囲にある色の情報を抽出することができるモジュールであるので，DeepLabV3 と同様に塗りつぶされた色の範囲を出力できるようにモジュールを追加すれば変更することが可能である．

本研究の実験から修正箇所の特定を容易に行うことができ，システムの保守性が既存の動画生成システムよりも優れていることを示すことが出来た．特定した原因となる箇所を修正することで人物動画生成を行うことができるシステムを実現することができると考えた．また，レイヤードスタイルとパイプアンドフィルタースタイルによる人間の脳の処理構造を模倣したアーキテクチャ設計は他の機械学習システムにも応用できる可能性がある．

7　おわりに

近年，様々な方法を用いて動画生成システムは研究されている．その中でも機械学習を用いた手法が盛んに研究されており，学習させるデータの種類や機械学習モデルのアーキテクチャ設計などが主な研究内容として存在している．

既存の研究ではベクトルなどの数値や動画のフレームを学習させることで動画生成を行なっている．しかし，ベクトルなどの数値を学習させるだけでは人物の身体を生成した際の不整合を完全に除去することはできない．

既存の研究の課題として手足の関節に不整合が生じ物理的にあり得ない人物の挙動を生成してしまうことが挙げられる．また，機械学習を用いたシステムはブラックボックス性によりモジュールの再利用も困難なことが挙げられる．

本研究の目的は物理的に整合した動画を生成することができるシステムの実現することである．また，その際に問題となる保守性を担保できるようにすることも目的とした．この目的を達成するための技術課題として次の 2 つを設定した．

1. 骨格データと身体の部位データを用いた動画生成システムのアーキテクチャ設計．
2. 提案したアーキテクチャにより保守性が担保されることの確認．

2 つの技術課題に対して次のようなアプローチを行った．

1．に対して骨格データを用いた動画生成システムにおいて適切に処理を分割しモジュール化するアプローチを行った．2. に対してシステムの出力から不具合が起きている箇所の特定をすることやアーキテクチャ内のモジュールを別のモジュールに変更してもシステムが機能するか評価した．

本研究は，骨格データを用いた動画生成システムの実装時のブラックボックス性を軽減し保守性を担保することができるようにアーキテクチャを設計した．提案したシステムを Python を用いて実装したが人物動画の生成を行うことが出来なかったが原因となる箇所を特定することが容易にできた．

今後の課題として特定した原因となる箇所を修正することで動画生成を行うことができるか確認する必要がある．

参考文献

[1] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, Jan Kautz, ”MoCoGAN: Decomposing Motion and Content for Video Generation” , IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , pp.1526-1535, 2018

[2] Yichao Yan, Jingwei Xu, Bingbing Ni, Xiaokang Yang, ”Skeleton-aided Articulated Motion Generation” , the 25th ACM international conference on Multimedia, pp. 199-207, 2017.10

[3] Katsunori Ohnishi, Shohei Yamamoto, Yoshitaka Ushiku, Tatsuya Harada, ” Hierarchical Video Generation From Orthogonal Information: Optical Flow and Texture” , AAAI Conference on Artificial Intelligence, Vol. 32 No.1, pp. 2387-2394 , 2018.04.26

[4] Mary Shaw and David Garlan, ”Software Architecture: Perspectives on an Emerging Discipline”, Pentice Hall, 1996.

[5] Hao-Shu Fang, Jiefeng Li Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu,Yong-Lu Li, Cewu Lu, ”AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time” , IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 45, pp. 7157-7173, 2023.06.01

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, Hartwig Adam, ”Rethinking Atrous Convolution for Semantic Image Segmentation” , arXiv,1706.05587v3, 2017.12.5