初期ポーズ生成の改良とGCNの導入による ポーズシーケンス生成モデルの拡張

寺内 健人[1,a)]　柳井 啓司[1,b)]

概要：
近年, 画像生成は GAN の発展によって飛躍的な進歩を遂げていて, GAN はラベル, 画像, テキスト等の条件付けによってさまざまな画像変換タスクに応用されている. しかし, 動画生成では時間次元の情報が増えたことにより高度なモデリングが必要となるため, 未だ発展途上である. 動画生成では, ガイドを追加することでより高品質な動画生成に成功している. 人間の動作の動画生成の場合は 2 段階の方法もある. 本研究では, ポーズシーケンスの生成, ポーズシーケンスからの動画生成の 2 段階に分けた動画生成を考慮し, 特に 1 段階目のポーズシーケンスの生成に焦点を当てる. モデルに グラフ畳み込みネットワーク (Graph Convolutional Network, GCN) を組み込み, より明示的にポーズをモデル化することでポーズシーケンスの自然な生成をすることを目指す. 提案手法は，初期ポーズ生成の改良と GCN の導入によって, 従来手法よりも質の高いポーズシーケンスの生成が可能であることを実験により確認した.

1. はじめに

近年, 画像生成は Generative Adversarial Networks (GAN) や Variational Auto Encoder (VAE) の発展によって飛躍的な進歩を遂げている. GAN, VAE は, ラベル, 画像, 音声やテキスト等の条件付けによって様々な画像変換タスクに応用されている. しかし, 動画合成は時間次元が増えたことにより, 画像に比べてより高度なモデリングが必要とされるため, 未だ発展途上である.

動画生成では, 無条件なガイドなし動画生成, フローなどのガイドを用いた動画生成手法に分けられ, ガイドベースなものでは, ガイドを追加することでより高品質な動画生成に成功している. また, 生成する動画も, 人物動画, 雲の流れ等の自然風景の動画, 街並みの運転景色の動画など多岐にわたる.

人間の動作の動画生成の場合は 2 段階の方法もあり, ポーズシーケンスの生成, ポーズシーケンスから動画の生成に分けた生成をしている. また, 人体のポーズシーケンスの生成のみに焦点を当てた既存研究はいくつかあり, ポーズのリファレンスを接続することで長時間のポーズシーケンスを生成する研究, ポーズをフレームに分けて逐次生成する研究などが存在する. ポーズシーケンスの生成には動画生成のガイドだけではなく, ゲームのキャラクターを生成ポーズのアクションに応じて動かすなどの応用例があり, 多くの場面において有用である.

本研究では，既存の人体ポーズガイドの生成 [1] に比べて, グラフ畳み込みネットワーク (Graph Convolutional Network, GCN) [2] を用いてポーズをより明示的にモデル化することで少ない入力から自然なポーズシーケンス生成を可能にする. ポーズの操作がより簡単になることで, 動画合成タスクの幅がより広がる. したがって本研究の目的は 1 枚の人物画像を入力ラベルに応じたアクションで動かし, 動画とすることとする. これをポーズシーケンスの生成, ポーズシーケンスからの動画生成の 2 段階に分けて考える. 特に 1 段階目のポーズシーケンス生成に焦点を当て, 入力ラベルに応じたアクションでポーズシーケンスを生成することを主目的とする. 動作認識の手法にはポーズシーケンス表現はしばしば用いられており, ポーズシーケンス表現の解釈には GCN が用いられることが多い. しかし, ポーズシーケンスの生成に GCN を用いる手法は未だ少ない. この論文では, GCN を組み込み, ポーズの構造を考慮したモデルについて提案する.

[1] 電気通信大学 大学院情報理工学研究科 情報学専攻
[a)] terauchi-k@mm.inf.uec.ac.jp
[b)] yanai@mm.inf.uec.ac.jp

2. 関連研究

本章では, 本研究に関連する手法について説明する. 本研究で利用した手法については, 特に詳しく説明する.

2.1 画像生成

画像生成の手法は, VAE[3] や GAN[4] の発展によって近年大きく進歩した. GAN では, 尤もらしい画像を出力するジェネレータと入力された画像が本物かどうかを見分けるディスクリミネータを交互に敵対的に学習させることで本物に近い画像を出力することができる. VAE では, 変分下限を最大化することでオートエンコーダで画像を再構成しつつ潜在変数を正規分布にすることができる. 近年では, GAN の様々な改良により StyleGAN[5] や BigGAN[6] が高解像度の画像生成で著しい成功を上げているが, VAE の方式でも VQ-VAE2[7] など, GAN に匹敵する高精度の画像生成ができるモデルも提案されている. 動画生成では個々のフレームの品質が動画全体の品質に大きく寄与するため画像生成の手法が応用されることが多い.

2.2 動画生成

動画生成手法は大きく分けてノイズからの無条件動画生成, フローやセグメンテーションをガイドに用いた合成に分けることができる. 無条件の動画生成では, 生成を前景と背景に分けた VGAN[8], 時間による潜在空間の移動について考慮した TGAN[9], 潜在空間を動きとコンテンツに分離した MoCoGAN[10], 多くのデータで学習することで高品質の動画生成を可能にした DVDGAN[11] などの手法がある. しかし, これらの手法は生成品質, 計算量に難があることが多い. 一方, ガイドベースの生成手法では生成を単純化し, 生成を高品質にすることができる. ポーズベースの手法 [1], フローベースの手法 [12], セグメンテーションベース [13] や 3D 情報をガイドに加えた手法 [14] などが存在し, ガイドベースの手法では, ガイドを編集することで生成を操作することができるなど, 応用できることが多い. したがって, 本研究ではポーズシーケンスの生成, ポーズシーケンスからの動画生成の 2 段階に分けて動画生成を考慮し, 特にポーズシーケンスの生成をより自然に行うことを目的とする.

2.2.1 Global Flow Local Attention

Global Flow Local Attention[12] では, ポーズシーケンスからの動画生成手法を提案している. 図 1 のように前フレームの情報, 次フレームのポーズからフローを推定し, 推定フローを用いてソース画像の特徴マップをワーピングすることで, 特徴マップを各フレームに対応した形に変形する. その後, ワーピングした特徴マップを局地的なアテンションを用いて埋め込むことで入力ポーズに対応した画像を生成する. 画像を各フレームで逐次生成することでポーズシーケンスと意味的に一致した動画を生成することができる. 本研究では, 生成したものをポーズシーケンスベースの動画生成手法を用いて動画を生成することができる.

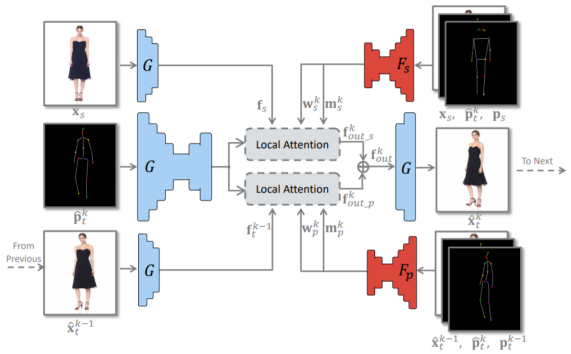


図 1: Global Flow Local Attention の概要.([12] より引用)

2.3 Graph Convolution Network

Graph Convolutional Network(GCN)[2] は, 格子状に関係性をもつ CNN とは異なり, 複雑な形状のグラフでの畳み込みを可能にする. 人体のポーズを GCN で畳み込む試みは行動認識 [15] , 2D からの 3D ポーズ推定 [16] で行われているが, ポーズ生成で行われている例は未だ少ないこの研究では, ポーズシーケンスの生成に GCN を適用することで時間と構造を共に捉え, より現実的なポーズシーケンスの生成を可能にすることを考える.

2.3.1 MS-G3D

MS-G3D[15] では, スケルトンベースの GCN を用いた行動認識のためのモデルを提案している. グラフの全ての頂点において, 近傍の情報を同じ距離の近傍ごとに同じカーネルで畳み込むマルチスケールな GCN を時間方向にもノードを繋げて畳み込む MS-G3D レイヤを考案し, 図 2 のように MS-G3D レイヤを重ねたパスと時間と構造を別に畳み込むパスの 2 つを組み合わせることで精度が向上することを示した. 本研究では, MS-G3D レイヤを用いて生成における時間と構造を同時に畳み込むことで時間的, 構造的に尤もらしいポーズシーケンスの生成をすることを考える. また, 事前学習モデルの最後の全結合層を取り除くことで分布間距離の測定における特徴抽出器としても用いる.

2.4 ポーズシーケンスの生成

ポーズシーケンスの生成はポーズガイドの動画生成の前

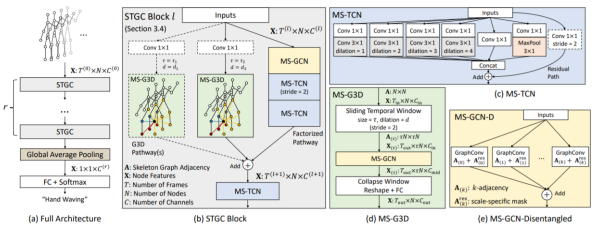


図 2: MS-G3D の概要.([15] より引用)

段階としての使用や, 3D モデルの動作に使うことができるなど, いくつかの有望な応用例がある. ポーズシーケンスの生成には CNN 等を用いて全体をまとめて生成する方法, RNN を用いて逐次生成する方法, また, いくつかのリファレンスを切り出し, その間を補間することで多様なモーション生成を可能にするリファレンスベースの方法 [17] などが存在する.

2.4.1 Cai らの研究

Cai らの研究 [1] では, 図 3 のように, ポーズシーケンスの生成, ポーズシーケンスからの動画生成の 2 段階に分けた動画生成について扱っている. ポーズシーケンスの生成では, 潜在変数の移動を学習するモデル, 潜在変数からポーズシーケンスを逐次生成するモデルを Discriminator を用いて敵対的ロスを用いて学習している. 最適化によりあるポーズに対応する潜在変数を探すことで, ポーズシーケンスの補間, 予測に使うこともできる. ポーズシーケンスの生成には, 時間の考慮は Discriminator 側に RNN があるのみであり, 生成器では潜在変数の移動のみで時間変化を考慮している. 本研究では, 全体生成モデル, 逐次生成モデルそれぞれで時間情報を考慮するためのモジュールを追加し, GRU で構造情報を同時に考慮する.

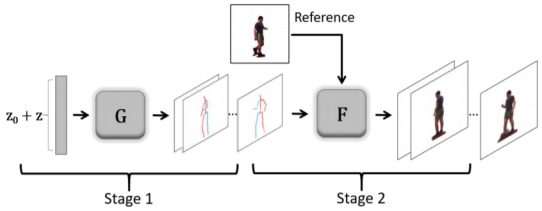


図 3: Cai らの研究の概要.([1] より引用)

2.4.2 Action2Motion

Action2Motion[18] はポーズシーケンスの生成のみに焦点を当てた研究であり, アクションラベルのみの条件付けからポーズシーケンスを生成する. この際, VAE の事前分布は通常正規分布に従うが, 時間が変化するに従い事前分布は変化すると仮定し, 前の時間ステップの情報から事前分布を推測している. 実際には図 4 のように前フレームのエンコーダ, 次フレームのエンコーダをそれぞれ用意し, 前フレームのエンコーダの出力の分布を次フレームのエンコーダの出力の分布と近づける Prior Loss を導入し, デコーダで再構成し, MSE Loss と合わせて学習する. さらに, エンコーダ, デコーダで追加で時間情報を考慮するために, GRU[19] を用いて時間経過に伴う情報を捉えている. テスト時に前フレームのエンコーダは次フレームのエンコーダと似た出力になるため, 前フレームのポーズから次フレームのポーズを逐次生成することができる. ポーズシーケンスの表現は身体の中心座標とそれぞれの関節の角度の表現を使用している. また, ポーズシーケンスの生成のみに焦点を当てたデータセットである HumanAct12 を紹介している. このデータセットは以前のデータセットのモーションキャプチャで得られたポーズシーケンスアノテーションと異なり, 時間的に滑らかでノイズが少ない. 本研究の逐次生成モデルでは Action2Motion を基に 1 フレーム目の生成を見直し, デコーダに GCN を導入することでより自然なポーズシーケンスの生成をする.

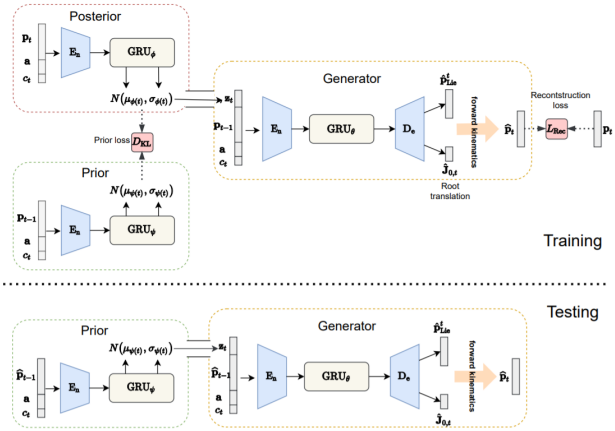


図 4: Action2Motion の概要.([18] より引用)

3. 提案手法

本研究ではカテゴリ条件付きポーズシーケンス生成を目的とした手法を提案する. Action2Motion[18] を基にした 1 フレームずつ逐次生成するモデルを提案し, GRU, GCN を用いて時間と構造をそれぞれ考慮する. 学習方法, ポーズ表現についても Action2Motion を基に学習する.

3.1 提案モデル

逐次生成モデルでは, Action2Motion に (1)1 フレーム目の考慮の見直し, (2) デコーダに GCN を追加, の 2 つの変更を加えることで生成品質の改善を目指す. 基本的な構成は Action2Motion のものと同一に前フレームエンコーダ

E_pr, 次フレームエンコーダ E_po, デコーダ D の構成のモデルであり, 1 フレーム目を考慮する初期フレームエンコーダ, 初期条件デコーダを追加し, デコーダに GCN を組み込むことでより表現力のあるモデルを学習することを目的とする.

3.1.1 1 フレーム目の考慮の見直し

Action2Motion では, 最初のフレームの生成時, 前フレームの情報を 0 として扱っているため, 初期フレームの多様性が少なくなる. したがって, 最初のフレームの特別な扱いが必要と考え, 最初のポーズを VAE のように正規分布になるよう初期フレームエンコーダでエンコードしノイズからの多様な生成を可能にする. また, 最初のフレーム生成時, デコーダの条件として潜在空間から初期フレーム条件デコーダでデコードした情報を与える. アーキテクチャ全体は図 5 のようになる.

また, Action2Motion では, 最初のフレームの生成時, 前フレームの情報を全て 0 として扱っている. その結果, エンコーダの出力はカテゴリにのみ依存し, 最初のフレームの多様性は再パラメタ化のノイズのみで保たれるため, ほとんど一様なポーズ出力となる. したがって, 最初のフレームのみ特別な扱いが必要になると考え, 最初のポーズを VAE 形式で潜在空間が正規分布となるよう初期フレームエンコーダ E_f でエンコードし, ノイズからの多様な生成を可能にする. また, 最初のフレームの生成時, デコーダの条件として潜在空間から初期フレーム条件デコーダ D_f でデコードした情報を与える. また, GRU の初期値を学習可能パラメータとすることで柔軟な学習ができるようにする. アーキテクチャ全体は図 5 のようになる.

ポーズシーケンスの長さ T, ジョイントの数 J, データセットのポーズシーケンスを $P = \{p_1, p_2, ..., p_T\}$, i 番目のポーズ表現を $p_i \in \Re^{J\times 3}$ と表し, ポーズシーケンス $\hat{P} = \{\hat{p_1}, \hat{p_2}, ..., \hat{p_T}\}$ を生成する. i 番目のポーズの条件ベクトル c_i はアクションカテゴリのワンホットベクトル α, 相対時間のスカラー値 t_i のセットの (α, t_i) とする. 初期フレームエンコーダの出力 $(\mu_f, \sigma_f^2) = E_f(P, c_1)$ から再パラメタ化で初期潜在変数 z_f を生成, 初期フレーム条件デコーダの出力 $\dot{p_0}$ を $\dot{p_0} = D_f(z_f)$ で生成, それらを用いて初期フレーム $\hat{p_1}$ は $\hat{p_1} = D(z_f, \dot{p_0}, c_1)$ の式で生成する. 2 フレーム目以降は Action2Motion 同様 $(\mu, \sigma^2) = E(p_{i-1}, c)$ の式で μ, σ^2 を求め, 再パラメタして z を生成, $\hat{p_i} = D(z, \hat{p_i}, c_i)$ の式でポーズ $\hat{p_i}$ を生成する.

3.1.2 デコーダ

デコーダのレイヤ構成は図 5 のようになり, デコーダは Action2Motion 同様に GRU[19] を 2 層重ねる. その後,

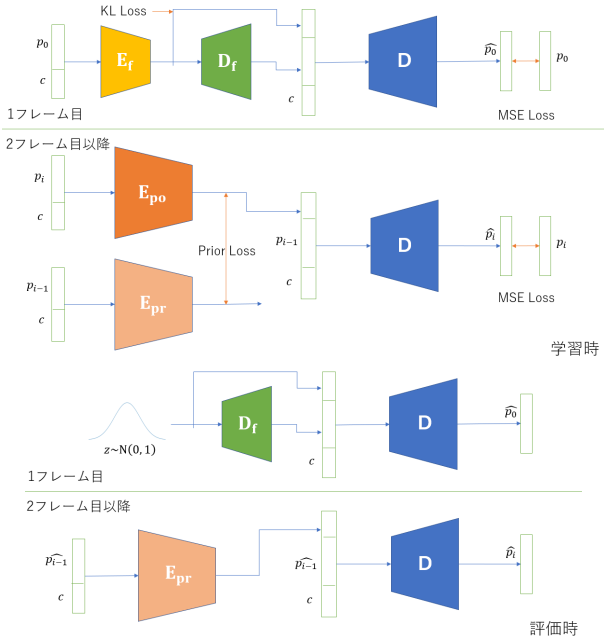


図 5: 逐次生成モデルの概要.

Linear 2 層でポーズ表現に埋め込み, GCN に接続することで構造表現を学習する. GCN は 3 層重ねることで複雑な表現を獲得する. GCN は MS-G3D の GitHub 実装の一部である MS-GCN を利用する. 最後に, ジョイントごとに別の重みで全結合層を通すことでポーズ出力とする. ジョイントごとに全結合層を通すことで, ジョイントごとの意味的な一貫性を保つ.

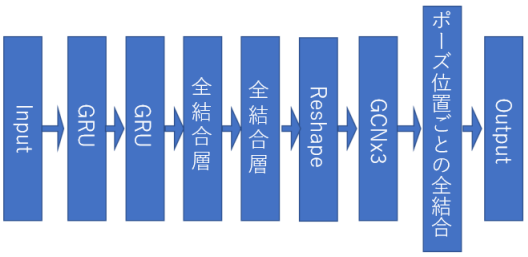


図 6: デコーダのレイヤ構成.

3.2 損失関数

目的関数は Action2Motion 同様の再構成損失, 前フレーム損失に加え, 初期フレームエンコーダの出力を正規分布に近付けるために KL 損失を加える.

$$L_f = -\frac{1}{2}\sum_{j=1}^{dim(z_f)}(1 + \log\sigma_j^2 - \mu_j^2 - \sigma_j^2)$$

KL 損失を加えることで初期フレームをノイズから生成することを可能にする.

3.3 ポーズ表現

学習するポーズシーケンスの表現は Action2Motion の表現の同様を用い, 身体の中心ジョイントの 3 次元直交座標系の座標と各ボーンの角度表現で構成する. 角度表現はリー代数表現であり, 3 つのパラメータで回転を表す. 3D 座標に変換する際は, 身体の中心のジョイントから接続するボーンを順に表現の角度に傾け, ボーンの長さだけ移動した位置を求めることで隣のジョイントの座標を求める. この時, ボーンの長さは手動で調整する. ボーンの長さをポーズ表現に含めないことで体格によらない動きの安定した生成が可能となる.

4. 実験

説明したモデルを用いて実際に学習を行い, その結果を報告する. カテゴリに対応したポーズシーケンスの生成が出来ていることを確認するため, 実際に結果を画像として描画し, 定性的に評価する. また, 比較するベースラインは Action2Motion[18] とし, 分布間距離の測定を用いて定量的な評価を行い, 学習するデータセットは HumanAct12[18] を用いる. Adam オプティマイザを利用して 6000epoch 学習を繰り消す.

4.1 データセット

データセットは Action2Motion で提示されている HumanAct12 を用いる. HumanAct12 は 9 から 403 フレームの長さ, 表 1 の 12 個の粗いカテゴリ, 34 個の詳細カテゴリの 1191 個のポーズシーケンスが含まれている動作認識, 生成用のデータセットである. 各ポーズシーケンスは 24 個のジョイントから構成されており, 各ジョイントは図 7 の通りになっている. 生成に使うデータはランダムにシーケンスを選び, そのシーケンスからランダムに 32 フレーム切り出す. ポーズシーケンスが 32 フレームに満たない場合, 32 フレームになるまで最後のフレームでパディングする. 実際のポーズシーケンスは図 8 のようなものが含まれる. カテゴリ条件付けは 12 個の粗いカテゴリを利用する.

表 1: HumanAct12 に含まれるカテゴリ.

粗いカテゴリ	詳細カテゴリ
準備体操	手首の準備運動
	胸筋の準備運動
	肘の準備運動
	右腕で身体を傾ける
	左腕で身体を傾ける
	右に反る
	左に反る
歩く	歩く
走る	走る
跳ねる	垂直に跳ねる
	手を挙げて跳ねる
飲む	右手のボトルで飲む
	左手のボトルで飲む
	右手のコップで飲む
	左手のコップで飲む
	両手で飲む
ダンベルを上げる	右手でダンベルを上げる
	左手でダンベルを上げる
	両手でダンベルを上げる
	ダンベルを頭の上まで上げる
	ダンベルを両手で持って膝を曲げる
座る	座る
食べる	右手で食べる
	左手で食べる
	パイ, ハンバーガーを食べる
ハンドルをきる	ハンドルをきる
電話をする	電話を取り, かけ, 元に戻す
	左手で電話する
ボクシングする	左, 右で殴る
	右, 左で殴る
	左アッパー
	右アッパー
投げる	右手で投げる
	両手で投げる

4.2 定性評価

生成結果の例は食べるアクションは図 9, 走るアクションは図 10 のようになった. 3 次元のジョイントの位置を 3 次元空間上に描画し, キーポイントの繋がりを直線で表すことでポーズとしている. 図は左上の画像を 1 フレーム目として 4 フレーム毎の結果を表示している. 学習したモデルを用いてそれぞれ異なるノイズベクトルから生成されている.

提案モデルでは, データセットに似た動きができており, 歩きやジャンプのような複雑な動きでさえうまく生成できている. しかし, 最初の数フレームは不自然なポーズの遷移をしていることがわかる. また, Action2Motion と比較すると定性的には生成品質の違いはわからなかった.

4.3 潜在空間の補間

モデルがデータセットの記憶をしていないことを示すために, 潜在空間の補間を行う. 2 つの潜在変数の中間の潜在変数が 2 つのポーズシーケンスの中間のポーズシーケンスが生成されていることを示すことができれば, 学習したモデルは潜在空間内で連続的な表現が獲得できていると言える. 乱数から生成した 2 つの潜在変数とその中間の潜在変

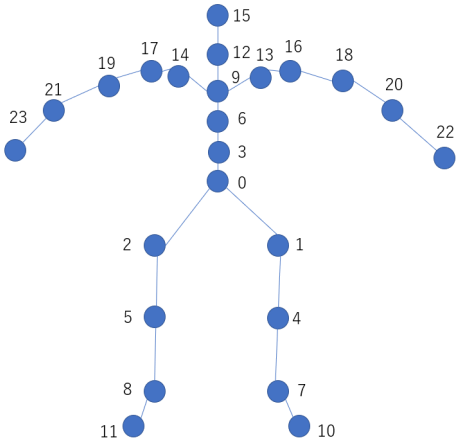
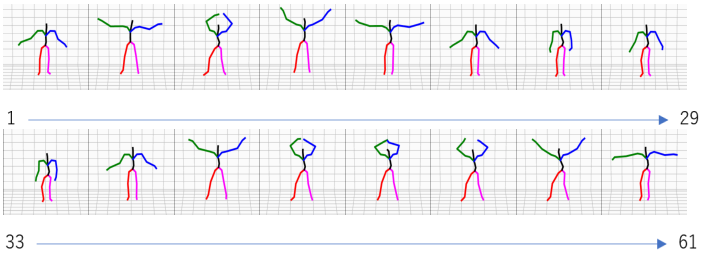


図 7: データセットのジョイントの位置.

(a) 座るアクション



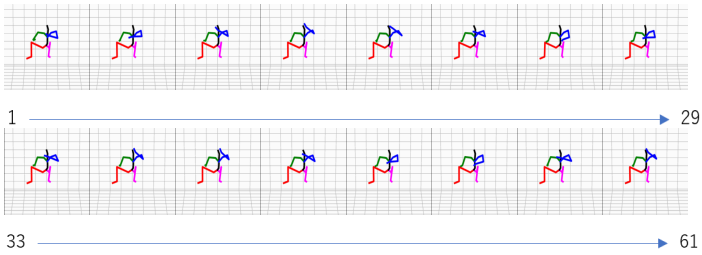
(b) 跳ねるアクション

(c) 投げるアクション

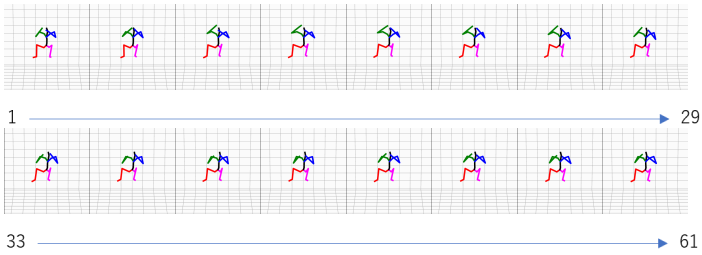
図 8: データセットのポーズシーケンスの例.

数から学習したモデルを用いてポーズシーケンスをそれぞれ生成する. この実験では初期フレームの生成に与える潜在変数についてのみ考え, 他のフレームのエンコーダでは乱数による振動を抑えるために分散を 0 とする. 同じアクション内と別アクション間の 2 つの設定で補間し, 別アクション間の補間ではアクションラベルも同時に補間する.

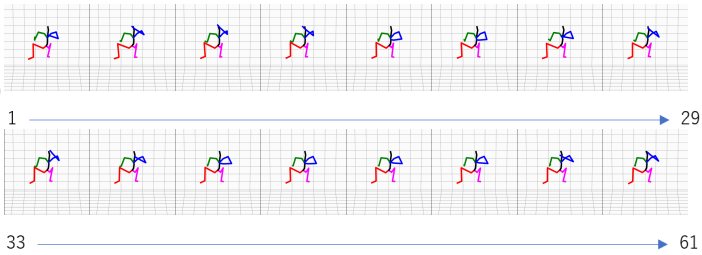
同じアクション内での補間結果を図 11, 別アクション間



(a) データセット



(b) Action2Motion



(c) 提案モデル

図 9: 食べるアクションの生成結果.

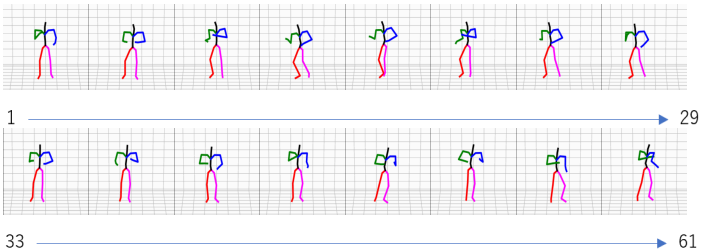
の補間結果を図 12 に示す. 図は左の画像を 1 フレーム目として 8 フレーム毎の結果を表示している. 上下のポーズシーケンスを生成する潜在変数を補間することで中央のポーズシーケンスを生成している. 上の補間は準備運動のアクション, 下の補間はジャンプと座るアクションの補間である. 同じアクション, 別のアクションの補間それぞれにおいて, 2 つの潜在変数の中間から生成されたポーズシーケンスは 2 つの潜在変数から生成されたポーズシーケンスの両方の特徴を引き継いでおり, 補間ができていると言える.

5. 定量評価

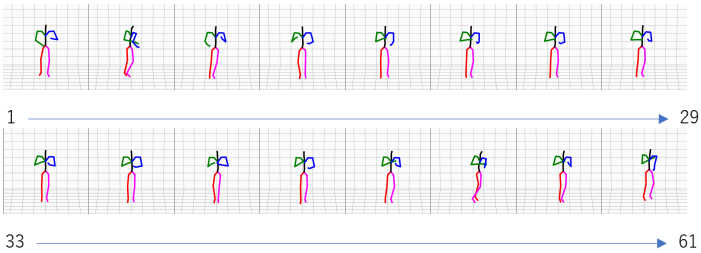
5.1 分布間距離の測定

FID[20], FVD[21] に触発され, リアルデータと生成データの間の分布間距離を測定することで生成したデータがデータセットに近い分布での生成ができているかを測定する尺度とする. 動作認識モデル MS-G3D の学習済みモデルでリアルデータと生成データの特徴ベクトルを抽出後, 特徴ベクトル間の分布間距離を FID 同様に測定する. MS-G3D は NTU-RGB-D120[22] で学習済みのモデルを用い, 最後の全結合層の直前の出力を特徴ベクトルとする.

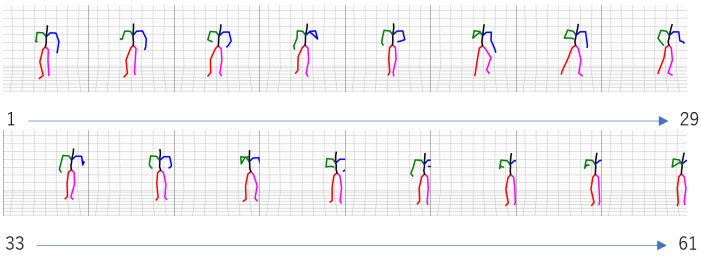
事前学習したデータセットと学習に使うデータセット



(a) データセット



(b) Action2Motion



(c) 提案モデル

図 10: 走るアクションの生成結果.

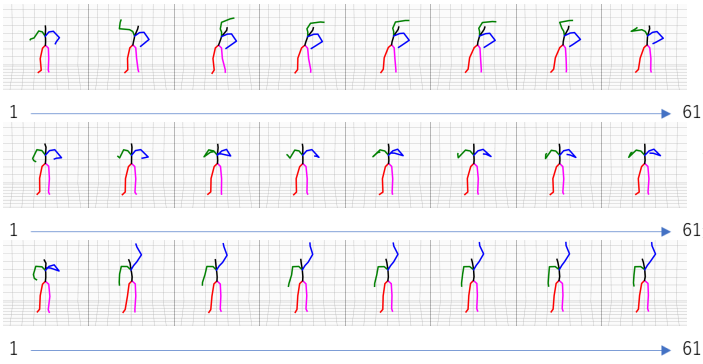


図 11: 準備運動アクションでの補間.

間のポーズ表現には差異があり, 直接適用することが出来ない. したがって, ポーズを NTU-RGB-D120 の形式に変換する. 対応するジョイントがあるものはそのジョイントに合わせ, ないものは周辺のジョイントの位置から補間する. 測定するポーズの特徴ベクトルの集合を $V_f = \{v_f1, v_f2, ..., v_fn\}$, データセットのポーズの特徴ベクトルの集合を $V_r = \{v_r1, v_r2, ..., v_rn\}$, 測定する生成ポーズシーケンスの特徴ベクトルの平均を μ_f, 分散共分散行列 Σ_f, データセットのポーズシーケンスの特徴ベクトルの平均を μ_r, 分散共分散行列 Σ_r として,

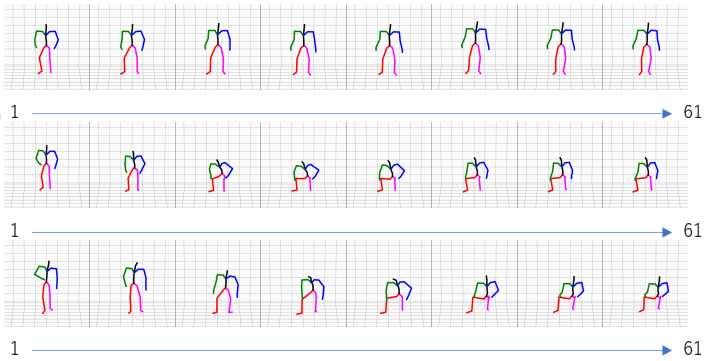


図 12: ジャンプと座るアクションの間の補間.

$$d^2 = |\mu_r - \mu_f|^2 + Tr(\Sigma_r + \Sigma_f - 2(\Sigma_r\Sigma_f)^{1/2})$$

の式で分布間距離 d^2 を測定する. 1191 個のポーズシーケンスをリアルデータ, 生成データからそれぞれサンプリングして測定する.

測定結果は表 2 のようになった. 既存研究と比較し, 提案モデルが上回り, データセットの動きをよりしっかり捉えられていることがわかる.

表 2: 分布間距離の測定結果.

手法	分布間距離
提案モデル	10.22
Action2Motion	14.26

5.2 アブレーション研究

本研究では, Action2Motion に (1)1 フレーム目の考慮の見直し, (2) デコーダに GCN を追加, の 2 つの変更を加えることで生成品質の改善を図った. この 2 要素がそれぞれ生成品質に与えている影響を調べるための実験を行う. (1) のみの変更を加えた提案モデル (1), (2) のみの変更を加えた提案モデル (2), 両方の変更を加えた提案モデル (1, 2), さらにベースモデルの Action2Motion の分布間距離を測定する. また, 1 フレーム目の考慮を見直したことで, 1 フレーム目の多様性が上がっていることが想定される. それを示すため, 1 フレーム目のみのポーズの多様性を測定する. 多様性 D は以下の式で求める.

$$D = \frac{1}{JN^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{J}(p_{i,1,k} - p_{j,1,k})^2$$

結果は表 3 のようになり, 提案モデル (1,2) は提案モデル (1) と比較して大きく分布間距離が小さくなっており, データセットに近い分布ができている. GCN の追加によって, より効率的なポーズの考慮ができることがわかる. また, 提案モデル (1, 2) は提案モデル (2) と比較し, 分布間距離

では大きな差はないが, 初期フレームの多様性は大きく上がっている. 提案モデル (2) では, GCN を導入したことでベース手法の Action2Motion に比べて多様性がなくなっているが, 提案モデル (1, 2) では初期フレームの見直しによってより多様な出力ができることが示せた.

表 3: アブレーション研究の結果.

手法	分布間距離	多様性
Action2Motion	14.26	0.02083
提案モデル (1)	13.35	0.02219
提案モデル (2)	10.21	0.00917
提案モデル (1, 2)	10.22	0.02143

6. おわりに

本研究では, 初期フレームの考慮をし, GCN を組み込むことで構造情報を明示的にとらえたモデルを提案した. 定性的に, 提案モデルは複雑で多様な動きの生成ができることがわかった. また, 潜在変数の補間により, モデルは潜在空間の連続表現が学習できていることがわかった. 分布間距離の測定の実験の結果, 提案モデルの有効性が示された.

今後の展望として, 評価指標が 1 つしかなく, 十分とは言えないため, 第三者による主観評価を含む広範な評価も必要である. また, ポーズシーケンスから動画を生成するモデルについて検討し, 実際に生成ポーズから動画を生成したい.

参考文献

[1] Cai, H., Bai, C., Tai, Y. and Tang, C.: Deep video generation, prediction and completion of human action sequences, *Proc. of European Conference on Computer Vision*, pp. 366–382 (2018).

[2] Niepert, M., Ahmed, M. and Kutzkov, K.: Learning convolutional neural networks for graphs, *Proc. of International Conference on Machine Learning*, pp. 2014–2023 (2016).

[3] Kingma, D. and Welling, M.: Auto-encoding variational bayes, *Proc. of International Conference on Learning Representations* (2014).

[4] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y.: Generative adversarial nets, *Advances in Neural Information Processing Systems*, pp. 2672–2680 (2014).

[5] Karras, T., Laine, S. and Aila, T.: A style-based generator architecture for generative adversarial networks, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 4401–4410 (2019).

[6] Brock, A., Donahue, J. and Simonyan, K.: Large Scale GAN Training for High Fidelity Natural Image Synthesis, *Proc. of International Conference on Learning Representations*, (online), available from ⟨https://openreview.net/forum?id=B1xsqj09Fm⟩ (2019).

[7] Razavi, A., van den Oord, A. and Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2, *Advances in Neural Information Processing Systems*, pp. 14866–14876 (2019).

[8] Vondrick, C., Pirsiavash, H. and Torralba, A.: Generating Videos with Scene Dynamics, *Advances in Neural Information Processing Systems*, Vol. 29, pp. 613–621 (2016).

[9] Saito, M., Matsumoto, E. and Saito, S.: Temporal generative adversarial nets with singular value clipping, *Proc. of IEEE International Conference on Computer Vision*, pp. 2830–2839 (2017).

[10] Tulyakov, S., Liu, M., Yang, X. and Kautz, J.: MoCoGAN: Decomposing Motion and Content for Video Generation, *Proc. of IEEE Computer Vision and Pattern Recognition* (2018).

[11] Clark, A., Donahue, J. and Simonyan, K.: Adversarial video generation on complex datasets, *arXiv preprint arXiv:1907.06571* (2019).

[12] Ren, Y., Li, G., Liu, S. and Li, T. H.: Deep Spatial Transformation for Pose-Guided Person Image Generation and Animation, *IEEE Transactions on Image Processing* (2020).

[13] Wang, T., Liu, M., Zhu, J., Liu, G., Tao, A., Kautz, K. and Catanzaro, B.: Video-to-Video Synthesis, *Advances in Neural Information Processing Systems* (2018).

[14] Mallya, A., Wang, T.-C., Sapra, K. and Liu, M.-Y.: World-Consistent Video-to-Video Synthesis, *Proc. of European Conference on Computer Vision* (2020).

[15] Liu, Z., Zhang, H., Chen, Z., Wang, Z. and Ouyang, W.: Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 143–152 (2020).

[16] Zhao, L., Peng, X., Tian, Y., Kapadia, M. and Metaxas, D. N.: Semantic Graph Convolutional Networks for 3D Human Pose Regression, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 3420–3430 (2019).

[17] Xu, J., Xu, H., Ni, B., Yang, X., Wang, X. and Darrell, T.: Hierarchical Style-based Networks for Motion Synthesis, *Proc. of European Conference on Computer Vision* (2020).

[18] Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M. and Cheng, L.: Action2Motion: Conditioned Generation of 3D Human Motions, *Proc. of ACM International Conference Multimedia* (2020).

[19] Chung, J., Gulcehre, C., Cho, K. and Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling, *Advances in Neural Information Processing Systems* (2014).

[20] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. and Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium, *Advances in Neural Information Processing Systems*, pp. 6626–6637 (2017).

[21] Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M. and Gelly, S.: Towards accurate generative models of video: A new metric & challenges, *arXiv preprint arXiv:1812.01717* (2018).

[22] Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L. and Kot, A.: NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 42, No. 10, pp. 2684–2701 (2020).