Vol.2018-MPS-117 No.6
2018/3/1

敵対的生成ネットワークを用いた インタラクション中の人間の振る舞いのモデル化

西村 優佑[1,a)]　中村 泰[1,b)]　石黒 浩[1,c)]

概要：近年, ロボットの社会進出が活発になるにつれて, 人間と親和性が高いロボットが求められている. 過去の研究から, ロボットの人間らしい動きがインタラクションに影響を与えることが示唆されている [6], [17], [19]. しかし, ロボットの人間らしい動作の生成を目的とした研究は少ない. そこで, 本研究では, より人間とインタラクションを行うロボットの実現を目的とし, GAN を用いてインタラクション中の人間の振る舞いをモデル化する手法を提案する.

キーワード：ヒューマンロボットインタラクション, Generative Adversarial Networks, 動作生成

Human behavior modeling during interaction using Generative Adversarial Networks

Yusuke Nishimura[1,a)]　Yutaka Nakamura[1,b)]　Hiroshi Ishiguro[1,c)]

Abstract: Nowadays, as robot's social advancement become active, more familiar robot is demanded. Previous studies suggested that robot make a influence to human interaction by moving like a human. However, there are few studies aiming at human like motion generation for robot. Therefore, in this research, we propose a method of huuman behavior modeling during interaction with the aim of achinving a robot that interacts more with humans.

Keywords: Human Robot Interaction, Generative Adversarial Networks, Motion Generation

1. はじめに

近年, ロボットの社会進出が活発になるにつれて, ロボットと人間がより日常的な場面でインタラクションをする機会が増えている. それに伴い, 人間と親和性が高いロボットが求められてきており, 人間とロボットの共生を目指した Human Robot Interaction（HRI）の研究が盛んに行われている [7], [9], [22]. 過去の研究から, ロボットの人間らしい動きがインタラクションに影響を与えることが示唆されている [6], [17], [19]. そこで, 本研究では, 人間との親和性の向上を目指して, 人間のインタラクション動作のモデル化に取り組む.

現在, ロボットの動作を記述する方法としてはルールベースが用いられることが多い [20], [21], [23]. インタラクション中の人間は対面している人間の仕草や周囲の音などに応じて様々な動作をする. そのため, 人間と自然なインタラクションを行うためのルールは視覚センサやマイクなどの入力とロボットの関節角などの出力に対して複雑な対応関係を持った煩雑なルールとなる. よって, ルールベース手法によるロボットのインタラクション動作の実装には多大な労力が必要になる. そこで, 機械学習手法を用いてロボットの動作のモデル化が行われているが [14], [16], [18], 様々な状況を認識しながら, 多様な動作を生成することができる動作モデルは未だにない. よって, 本研究では, この

[1] 大阪大学大学院基礎工学研究科
〒 560-8531 大阪府豊中市待兼山町 1-3
[a)] nishimura.yusuke@irl.sys.es.osaka-u.ac.jp
[b)] nakamura@is.sys.es.osaka-u.ac.jp
[c)] ishiguro@is.sys.es.osaka-u.ac.jp

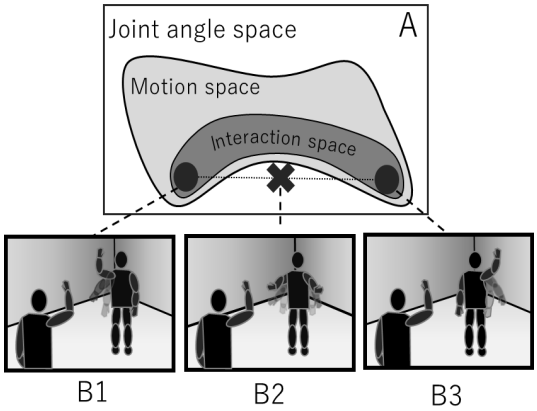

図 1 インタラクション中の動作
Fig. 1 Interactive motion

ようなロボットの動作モデルを人間の動作データから自動的に構築する手法の開発を目標とする.

近年, 記録された多様なデータを生成する枠組みとして, ニューラルネットワークの生成モデルの 1 つである GAN（Generative Adversarial Networks）の研究が盛んである. 例えば, 文章からの画像を生成するモデル [10] などのように, 低次元の潜在変数から高次元の多様な疑似データを生成する研究が行われている. 本報告では, この GAN を用いてインタラクション動作モデルの作成に取り組む.

本研究の目的はモーションキャプチャなどで取得した人間の動作データを単に再現することではなく, ロボットが動作モデルによって人間と同じように多様なインタラクションを実現することである. このようなロボットの動作モデルを計測した人間の動作データから自動的に作り出せるようにすることで, 人間と共生するロボットの自然な動作の実現が期待できる. また, ロボットが人間らしい動きをすることで, ロボットの動作が人間にとって認識しやすくなり, 人間との親和性が高まる可能性がある [3], [9].

2. インタラクション動作のモデル化

インタラクション中の人間は同じようなシチュエーションでも異なる動作をする場合がある. 例えば「挨拶」を行う場合, 図 1 の B1 や B3 のように右手を挙げると左手を挙げる場合がある. しかし, この 2 つの典型例に対して, 関節角において平均を取った B2 のような動作は必ずしも自然な動作とはならない. すなわち, インタラクション中の動作においては “平均的な動作” が必ずしも図 1A の Intraction space に含まれないことが問題となる. そのため, 既存の動作生成手法では, この例のように複数の正解動作を扱うことが困難であるため, 本研究では GAN を用いたインタラクション動作のモデル化に取り組む.

2.1 ロボットのインタラクション動作生成

ロボットの動作生成を行うことを想定したシステムの概要図を図 2 に示す. 本研究では, ロボットと人間が対面している状況で, 視覚センサで得た情報から, それに対するロボットの自然な動作を生成することを問題として扱う. 提案システムの具体的な処理としては以下の通りである.

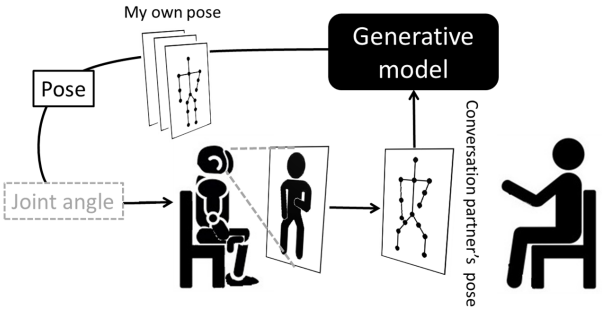

図 2 システム概要
Fig. 2 System diagram

(1) 視覚センサで得た対面者の画像から相手の特徴点座標を抽出する.
(2) 相手の特徴点座標から, それに対する適切な特徴点の座標データ群を動作モデルによって生成する.
(3) 生成された座標データをロボットの関節角に変換し, ロボットを制御する.

既存の骨格位置推定ライブラリ [2], [11], [13] によって人間の画像から特徴点座標は抽出可能である. そこで, 本報告では, この研究の第一歩として, (2) の実現を目的に GAN によるインタラクション動作モデルの作成に取り組んだ.

2.2 GAN を用いたインタラクション動作モデル

GAN は Generator と Discriminator と呼ばれる, 2 つのニューラルネットワークから構成される. Generator は出力が Discriminator に訓練データからサンプルされたデータ (以後リアルサンプルと呼ぶ) であると識別されるように学習する. Discriminator は入力されたデータがリアルサンプルか Generator が出力したデータ (以後フェイクサンプルと呼ぶ) かを正確に識別できるように学習していく [4]. GAN はこれら 2 つのネットワークを交互に学習させていくことで, リアルサンプルと見分けが付かないようなデータを出力する Generator を得る.

本報告では, GAN を用いて 2 つの動作モデルを実装した.

Motion GAN 学習の安定化を目的とした GAN の拡張モデルである WGAN-GP[*1] (Wasserstein GAN - Gradient Penalty)[5] を用いて, ガウス分布に従う潜在変数から約 1 秒程の 2 人の動作を生成する. GAN によって 2 名によるインタラクションの動作を生成できることを確かめる.

[*1] GAN は学習途中で損失関数の勾配が発散する場合があるため, 学習の不安定さが問題となっている. この問題を理論的に解析し, 安定な学習法を提案した WGAN[1] が存在する. WGAN-GP はこの WGAN を改良したモデルである.

Conditional Motion GAN　ラベルなどの条件から, 条件付きの生成を行う Conditional GAN[8] を用いて, 片方の人間のデータからもう一方の人間の特徴点座標を生成する. GAN によって, 対面者の計測したデータを条件とした動作生成を行えることを確かめる.

2.3 データセット

本研究では, ニューヨーク州立大学が公開している SBU-Kinect-Interaction Dataset[15] を利用した. 図 3 のように, このデータセットには 2 名によるインタラクションを Kinect で撮影して得た RGB 画像, Depth 画像, 人間の特徴点座標が含まれている. 各インタラクションは合計 21 組のペアによって, アクション側とリアクション側に分かれて交互に行われ, 総撮影時間は約 10 分である. インタラクションの内容は Approaching, Departing, Kicking, Pushing, Hugging, Shaking hands, Exchanging と Puncing の計 8 種類である.

本動作モデルの学習には, データセットに含まれている人間の特徴点の 2 次元座標のみを使用し, 20 組分のデータを訓練データとして使用した. また, 残り 1 組の 8 種類のインタラクションデータを Conditional Motion GAN のテストデータとして評価に用いた.

3. 2 名によるインタラクション動作の生成

Motion GAN におけるネットワーク概要を図 4 に示す. 本モデルでは, 連続する 16 フレーム分のスケルトンの 2 次元座標を訓練データとする. 1 人当たりのスケルトンは 15 点の特徴点から構成されているため, 各時刻 t における 2 人分の特徴ベクトルは $\mathbf{x_t} \in R^{60}$（t=1,...,16）となる. よって, リアルサンプル $\mathbf{X}$ は $\mathbf{X}=[\mathbf{x_1},...,\mathbf{x_{16}}] \in R^{16\times60}$ となる.

Generator はガウス分布に従う潜在変数ベクトル $\mathbf{z} \in R^{100}$ を入力として, フェイクサンプル $\hat{\mathbf{X}}=[\hat{\mathbf{x}}_\mathbf{1},...,\hat{\mathbf{x}}_\mathbf{16}] \in R^{16\times60}$ を出力する. Generator はこのフェイクサンプルを Discriminator [*2] へ入力し, その出力の期待値を最小化するようにネットワークの重みを更新していく. Generator には, 4 層の転置畳み込み層を使用し, 合計 6 層のニューラルネットワークを用いた. Generator の損失関数を以下に示す.

$$L_G = -E\left[D(\hat{\mathbf{X}})\right]$$

Generator の活性化関数は出力層に tanh を使用し, それ以外は Leaky ReLU を使用した.

Discriminator は $\mathbf{X}$ もしくは $\hat{\mathbf{X}}$ を入力として, 入力のリアルサンプルらしさを示す値 (以後評価値と呼ぶ) を出力する. WGAN では, Discriminator にリアルサンプルとフェイクサンプルをそれぞれ入力し, 出力された評価値の期待値の差を最大化するようにネットワークの重みを更新していく. WGAN の損失関数を以下に示す [1].

$$L_{wgan} = E\left[D(\boldsymbol{X})\right] - E\left[D(\hat{\boldsymbol{X}})\right]$$

WGAN-GP では, この損失関数に一様分布に従う重み係数 $\epsilon \sim U(0, 1)$ を用いて以下のペナルティ項を加える [5].

$$\tilde{x} = \epsilon D(\boldsymbol{X}) + (1-\epsilon)D(\hat{\boldsymbol{X}})$$
$$L_{gp} = E\left[\left(\left\|\nabla D(\tilde{\boldsymbol{X}})\right\|_2 - 1\right)^2\right]$$

本研究では, WGAN-GP の損失関数に加えて L_2 ノルムを足したものを Discriminator の損失関数とした. Discriminator 内の重みの数を N, i 番目の重みを $w_i (i = 1, \ldots, N)$ とすると, L_2 ノルムは以下の通りに表される.

$$L_2 = \sum_{i=1}^{N} w_i^2$$

Discriminator には, 4 層の 3 次元の畳み込み層を使用し, 合計 5 層のニューラルネットワークを用いた. Discriminator の損失関数を以下に示す. α と β は各項に対する重み付け係数で, 本研究ではそれぞれ 1.0 と 0.0001 とした. Discriminator の活性化関数はすべて Leaky ReLU を使用した.

$$L_D = L_{wgan} + \alpha L_{gp} + \beta L_2$$

Generator と Discriminator の構造は Video GAN [12] を参考にした. ハイパーパラメータは WGAN-GP の論文と同じ値を使用し, 学習は勾配法の一種である Adam (Adaptive moment estimation) を用いた. 訓練データは [-0.5,0.5] の範囲で正規化した.

3.1 比較対象

Motion GAN を評価するために, 生成モデルの 1 つである PPCA (Probabilistic Principal Component Analysis) との比較を行った. 潜在空間として累積寄与率が 99%を超えた 74 次元分の主成分を用いた.

3.2 結果

図 5 と図 6 に PPCA 及び Motion GAN による生成結果を示す. また, 図 7 に 胴体からの相対座標系における手と足の速度分布を, 表 1 に各速度分布の尖度を示す.

図 5 のように, PPCA からは動きが遅い動作が多く生成された. それに対して, 図 6 に示す通り, GAN からは訓練データのインタラクションに近い多様なインタラクショ

表 1　速度分布の尖度
Table 1　Kurotosis of velocity distribution

	Trace	GAN	PPCA
Hand	24.30	8.46	0.87
Foot	18.01	12.08	0.63

[*2] WGAN-GP では, Discriminator を Critic と呼称するが, 本論文では Discriminator と呼ぶ.

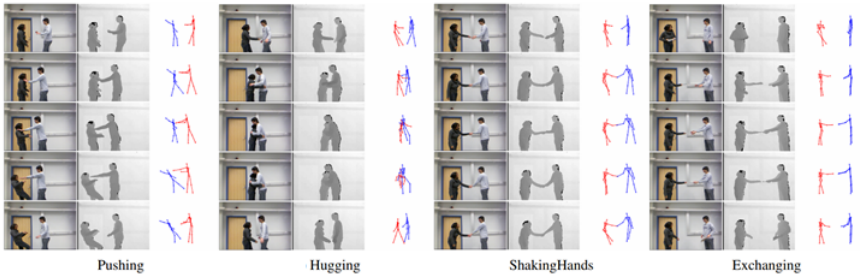

図 3　SBU-Kinect-Interaction Dataset[15]

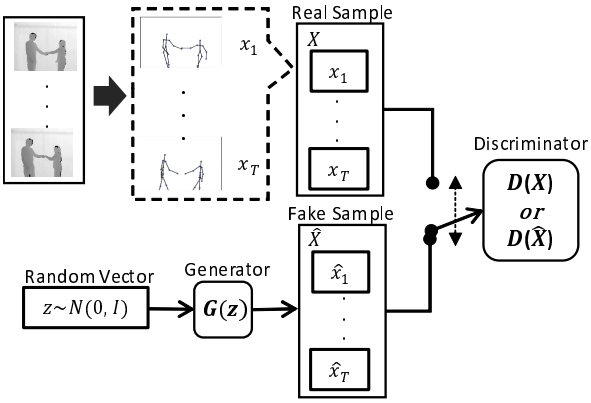

図 4　Motion GAN
Fig. 4　Motion GAN

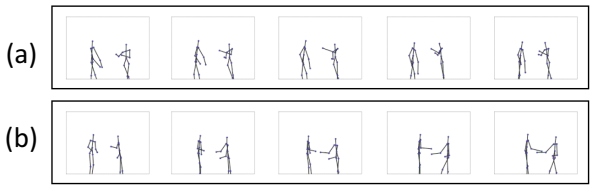

図 5　PPCA 生成結果
Fig. 5　Generation result of PPCA

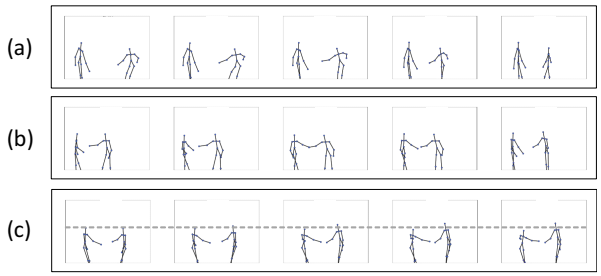

図 6　GAN 生成結果
Fig. 6　Generation result of GAN

ン動作が生成された．しかし，図 6(c) ように，本モデルによって生成される動作の中には，途中で身長が変化しているデータも生成されているため，この部分については今後改善していく必要がある．

図 7 から，訓練データの速度分布は原点でピークを持っていることが分かる．それに対して，PPCA は少しずれた位置でピークを持っており，GAN は PPCA よりも訓練

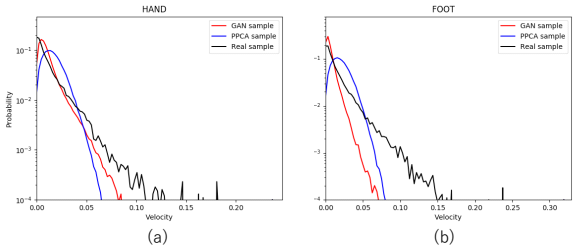

図 7　手と足の速度分布
Fig. 7　Hand and Foot of velocity destribution

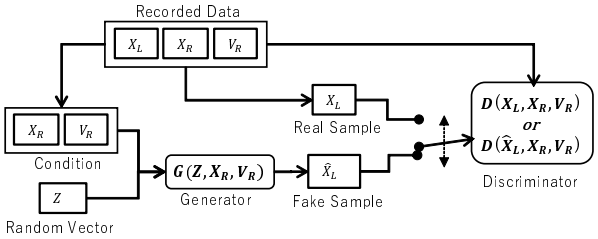

図 8　Conditional Motion GAN
Fig. 8　Conditional Motion GAN

データに近い位置でピークを持っていることが分かる．また，表 1 より，統計量の 1 つである尖度において，PPCA よりも GAN の方がより訓練データに近い速度分布になっていることが分かる．

これらのことから，PPCA よりも GAN の方がより訓練データの速度分布に近い多様なインタラクション動作を生成できていると考える．

4.　対面相手の振る舞いに応じたインタラクション動作の生成

Conditional Motion GAN を図 11 に示す．本モデルでは，右側のスケルトン座標 $\mathbf{x_R} \in R^{30}$ と各特徴点座標の速度 $\mathbf{v_R} \in R^{30}$ を対面者のデータとして，左側のスケルトン座標を $\mathbf{x_L} \in R^{30}$ 自エージェントの座標として考える．本モデルの目的はこの対面者の動作データに応じた，自然な動作を生成することである．時刻 t でのスケルトン座標及び速度をそれぞれ $\mathbf{x_L}(t)$, $\mathbf{x_R}(t)$, $\mathbf{v_R} \in R^{30}$ として，右側のスケルトン座標における速度ベクトル $\mathbf{v_R}(t)$ の計算式を以下に

示す.

$$\mathbf{v_R(t)} = \mathbf{x_R(t)} - \mathbf{x_R(t-1)}$$

そして, T 時刻分の各データをそれぞれ $\mathbf{X_L}$,$\mathbf{X_R}$,$\mathbf{V_R}$$\in R^{T\times 30}$ とする. 本モデルでは, 連続する 4 フレーム分のデータ $\mathbf{X_L}$=[$\mathbf{x_L}$(1),$\mathbf{x_L}$(2),$\mathbf{x_L}$(3),$\mathbf{x_L}$(4)]$\in R^{4\times 30}$ をリアルサンプルとして本モデルの学習を行った.

Generator はガウス分布に従う潜在変数ベクトル $\mathbf{Z}\in R^{24}$ に加えて条件として $\mathbf{X_R}$, $\mathbf{V_R}$ を入力し, $\hat{\mathbf{X}}_\mathbf{L}$ を出力する. $\mathbf{Z}$ を時刻 t に対応する潜在変数ベクトルを $\mathbf{z}(t)\in R^6$ (t=1,2,3,4) とする. 本モデルでは, $\mathbf{z}(t)$,$\mathbf{x_R}(t)$,$\mathbf{v_R}(t)\in R^{30}$ (t=1,2,3,4) を 1 フレーム分のデータとして, Generator は各フレームのデータそれぞれに対して, 同じ重みを持つようにした. Generator には, 6 層のニューラルネットワークを使用した. 活性化関数は出力層に tanh を使用し, それ以外は ReLU を使用した. Generator の損失関数を以下に示す.

$$L_G = -E[\log(1 - D(\hat{\mathbf{X}}_\mathbf{L}, \mathbf{X_R}, \mathbf{V_R}))]$$

Discriminator は $\mathbf{X_L}$,$\mathbf{X_R}$, $\mathbf{V_R}$ もしくは $\hat{\mathbf{X}}_\mathbf{L}$, $\mathbf{X_R}$, $\mathbf{V_R}$ を入力として, 入力がリアルサンプルである確率を出力する. 本モデルでは, ネットワークの性能向上を目的として, 識別器を各フレームに対して識別を行う D_1 と 4 フレーム分のデータに対して識別を行う D_2 の 2 つのネットワークを用いた. D_1 は $\mathbf{x_L}$(t) , $\mathbf{x_R}$(t) , $\mathbf{v_R}$(t) もしくは $\hat{\mathbf{x}}_\mathbf{L}$(t) , $\mathbf{x_R}$(t) , $\mathbf{v_R}$(t) を入力として, D_2 は $\mathbf{X_L}$, $\mathbf{X_R}$,$\mathbf{V_R}$ もしくは $\hat{\mathbf{X}}_\mathbf{L}$, $\mathbf{X_R}$, $\mathbf{V_R}$ を入力として, 入力がリアルサンプルである確率を出力する. 各ネットワークはリアルサンプルを入力した時に 1 を, フェイクサンプルを入力した時に 0 を出力するように学習していく. D_1 には, 3 層の 2 次元の畳み込み層を使用し, 合計 5 層のニューラルネットワークを用いた. D_2 には, 4 層の 3 次元畳み込み層を使用し, 合計 5 層のニューラルネットワークを用いた. Discriminator の損失関数を以下に示す. 活性化関数は出力層にシグモイド関数を使用し, それ以外は Leaky ReLU を使用した.

$$\begin{aligned}
L_{d1_{real}}(t) &= E\left[\log(D_1(\mathbf{x_L(t)}, \mathbf{x_R(t)}, \mathbf{v_R(t)}))\right] \\
L_{d1_{fake}}(t) &= -E\left[\log(1 - D_1(\hat{\mathbf{x}}_\mathbf{L}\mathbf{(t)}, \mathbf{x_R(t)}, \mathbf{v_R(t)}))\right] \\
L_{D_1}(t) &= L_{d1_{real}}(t) + L_{d1_{fake}}(t) \\
L_{d2_{real}} &= E\left[\log(D_2(\mathbf{X_L}, \mathbf{X_R}, \mathbf{V_R}))\right] \\
L_{d2_{fake}} &= -E\left[\log(1 - D_2(\hat{\mathbf{X}}_\mathbf{L}, \mathbf{X_R}, \mathbf{V_R}))\right] \\
L_{D_2} &= L_{d2_{real}} + L_{d2_{fake}} \\
L_D &= \sum_t L_{D_1}(t) + L_{D_2}
\end{aligned}$$

D_1 の構造は DCGAN を参考にし, Generator と D_2 の構造は Video GAN を参考にした. また, ハイパパラメータや学習アルゴリズムは DCGAN の論文と同じものを使用し, 訓練データは [-0.5,0.5] の範囲で正規化した.

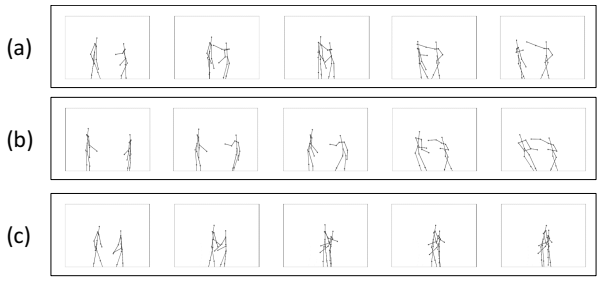

図 9 テストデータ
Fig. 9 Test data

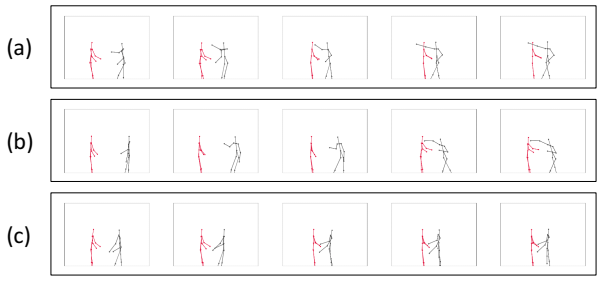

図 10 回帰モデル出力結果
Fig. 10 Generation result of regression model

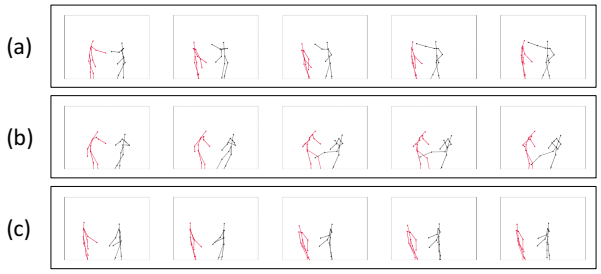

図 11 Conditional Motion GAN 生成結果
Fig. 11 Generation result of Conditional Motion GAN

4.1 比較対象

本モデルを評価するために, 線形回帰モデルとの比較を行った. $\mathbf{X_R}$ と $\mathbf{V_R}$ を説明変数, $\hat{\mathbf{X}}_\mathbf{L}$ の各座標を目的変数として, GAN と同じ訓練データを用いて重回帰分析を行い, GAN と同じテストデータに対する出力結果を比較する.

4.2 結果

比較対象となるテストデータを図 9 に, 回帰モデルの出力結果を図 10 に, Conditional Motion GAN の生成結果を図 11 に示す. インタラクションは計 8 種類あるが, リアクション側の動作生成結果の中で, 特徴的な結果のみを示す.

図 10 に示す通り, 線形回帰モデルはほとんど静止状態に近い姿勢が出力されており, 特に図 10(a) では, 相手の大きな座標変化に対応できず, 相手の手先が頭部を貫いていることが分かる.

それに対して, 図 11(a) と (b) に示す通り, GAN は元のインタラクションらしい動作が生成されていることが分かる. しかし, 図 11(c) のようにテストデータでは, Hugging

を行っているのに対して, GAN は対面者を避けるような動作が生成されている. Hugging の動作を扱えていないことが分かる. また, 本モデルは同じ条件 $\mathbf{x_A}$, $\mathbf{v}$ に対しても, $\mathbf{z}$ の値に応じて, 生成されるスケルトン $\mathbf{x_B}$ の大きさや動作が変化することが理想である. しかし, 本モデルでは, $\mathbf{z}$ を変化させても, スケルトンの大きさのみが変化し, 動作はあまり変化しない. そのため, 本モデルは mode collapse[*3] を起こしている可能性があり, この部分についても今後改善が必要であると考える.

5. 結論

本研究では, 人間の動作データを GAN に学習させることによって, PPCA よりも多様な動作生成を行えることを確かめ, 対面者の動作に応じた動作生成を行える可能性を示した.

また, 本報告の動作モデルでは, 姿勢データしか用いていない. しかし, 実際の人間は周囲の音や相手の姿勢, 対話の内容に応じて動作は変化すると考えられる. そのため, 今後は音声情報なども入力含めた動作モデリングに取り組むと共に, 実ロボットの動作生成に取り組む.

参考文献

[1] Martin Arjovsky, Soumith Chintala, and Lon Bottou. Wasserstein gan, 2017.
[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.
[3] Rui Fang, Malcolm Doering, and Joyce Y. Chai. Embodied collaborative referring expression generation in situated human-robot interaction. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, pp. 271–278, New York, NY, USA, 2015. ACM.
[4] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.
[5] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans, 2017.
[6] M. Huber, M. Rickert, A. Knoll, T. Brandt, and S. Glasauer. Human-robot interaction in handing-over tasks. In *RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication*, pp. 107–112, Aug 2008.
[7] Takamasa Iio, Masahiro Shiomi, Kazuhiko Shinozawa, Katsunori Shimohara, Mitsunori Miki, and Norihiro Hagita. Lexical entrainment in human robot interaction. *International Journal of Social Robotics*, Vol. 7, No. 2, pp. 253–263, Apr 2015.
[8] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets, 2014.
[9] A. H. Qureshi, Y. Nakamura, Y. Yoshikawa, and H. Ishiguro. Robot gains social intelligence through multimodal deep reinforcement learning. In *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, pp. 745–751, Nov 2016.
[10] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis, 2016.
[11] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017.
[12] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics, 2016.
[13] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016.
[14] Xiaolin Wei, Jianyuan Min, and Jinxiang Chai. Physically valid statistical models for human motion generation. *ACM Transactions on Graphics (TOG)*, Vol. 30, No. 3, p. 19, 2011.
[15] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L. Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*. IEEE, 2012.
[16] 哲也稲邑, 仁彦中村, 巌樹戸嶋, 英明江崎. ミメシス理論に基づく見まね学習とシンボル創発の統合モデル. 日本ロボット学会誌, Vol. 22, No. 2, pp. 256–263, mar 2004.
[17] 垣尾政之, 宮下敬宏, 光永法明, 石黒浩, 萩田紀博. 倒立振子移動機構を持つ人型ロボットの反応動作の違いが人に与える印象の変化. 日本ロボット学会誌, Vol. 28, No. 9, pp. 1110–1119, 2010.
[18] 高野渉, 山根克, 杉原知道, 山本江, 中村仁彦. 身体的記号化モデルに基づく人間とヒューマノイドロボットのコミュニケーション理論. 日本ロボット学会誌, Vol. 28, No. 6, pp. 735–745, 2010.
[19] 崇行神田, 倫太今井, 哲雄小野, 浩石黒. 人－ロボット相互作用における身体動作の数値解析. 情報処理学会論文誌, Vol. 44, No. 11, pp. 2699–2709, nov 2003.
[20] 神田崇行, 鎌島正幸, 今井倫太, 小野哲雄, 坂本大介, 石黒浩, 安西祐一郎. 人間型対話ロボットのための協調的身体動作の利用. 日本ロボット学会誌, Vol. 23, No. 7, pp. 898–909, 2005.
[21] 前田陽一郎. ファジィルールを用いた基本心理ベクトルに基づく自律移動ロボットの感情生成実験. 日本ファジィ学会誌, Vol. 12, No. 6, pp. 816–825, 2000.
[22] 美紀渡辺, 浩平小川, 浩石黒. ミナミちゃん：販売を通じたアンドロイドの実社会への応用と検証. 情報処理学会論文誌, Vol. 57, No. 4, pp. 1251–1261, apr 2016.
[23] 光永法明, 宮下善太, 宮下敬宏, 石黒浩, 萩田紀博. コミュニケーションロボット robovie-iv の開発とオフィス環境での日常対話. 日本ロボット学会誌, Vol. 25, No. 6, pp. 822–833, 2007.

[*3] 訓練データの一部の最頻値のみを学習してしまい, 学習の結果, Generator がその最頻値周辺のデータのみを出力してしまうことを mode collaps という.