F-023

深層学習モデルによる自己と他者の共有身体イメージの獲得
A Deep Neural Network Model for Development of Shared Body Image of Self and Other

野口 渉† 飯塚 博幸†‡ 山本 雅人†‡
Wataru Noguchi Hiroyuki Iizuka Masahito Yamamoto

1. はじめに

他者の動作の模倣においては観察した他者の行動が自己の運動へ変換されるが，そのためには自己と他者の身体姿勢・動作の対応関係を認識する必要がある．しかし，自己と他者の身体の対応関係は感覚情報から直接得られず，学習・経験を通して獲得されるものだと考えられる．

深層学習の分野においては，学習を通して自己と他者の身体の対応関係を獲得可能なモデルが提案されている．例えば，人の同じ動作を同時に撮影した一人称と三人称視点の動画において，同時刻の一人称・三人称視点の画像フレームが特徴空間上の近い点にマッピングされるように深層学習モデルを学習させることで，異なる視点間での身体の対応を獲得可能なモデルが提案されている [1]．また，著者らは，自己と他者の情報を同一のモジュールを用いて処理することにより，視覚の予測学習を通して自己と他者で共有された空間位置の表現が獲得されることを示している [2]．ただし，[2] においては身体的な動作ではなく，空間の移動に関する動作に限っており，身体の姿勢の対応は扱っていない．

本研究では，[2] と同様に自己と他者の情報を共有のモジュールによって処理するモデルを用い，自己と他者の身体姿勢の対応を獲得するモデルを構築する．とくに，身体の 3D モデルの予測学習をとおして，自己と他者で共有な身体イメージを獲得可能なモデルを提案する．

2. ロボットシミュレーション

自己と他者の身体イメージの獲得をシミュレーションするために，自己と他者 2 体のヒト型ロボットが存在する仮想環境を構築する．図 1 (a) に構築した仮想環境を示す．ロボットは左右の肩と肘の関節にそれぞれ 3 軸回転の自由度，合計で 12 自由度をもち，自身の関節角度と動作に伴う関節角速度，また，頭部前面に備えたカメラによる視覚を観測する．とくに視覚画像により正面に位置する他者の姿を観測可能である（図 1 (b)）．また，ロボットは，これらの感覚情報に加え，3 次元の仮想空間上におけるロボットの形状と姿勢を表す 3D モデルを認識可能である．

2.1 3D モデルの認識

3D モデルは Signed Distance Function (SDF) により表現される．SDF は 3D 形状を 3 次元空間座標からモデル表面までの符号付き距離によって間接的に表現する関数であり，SDF を近似することで効率よく連続な 3D 形状を表現する深層学習モデルが提案されている [3].

†北海道大学 人間知・脳・AI 研究教育センター Center for Human Nature, Artificial Intelligence, and Neuroscience, Hokkaido University
‡北海道大学 大学院情報科学研究院 Faculty of Information Science and Technology, Hokkaido University

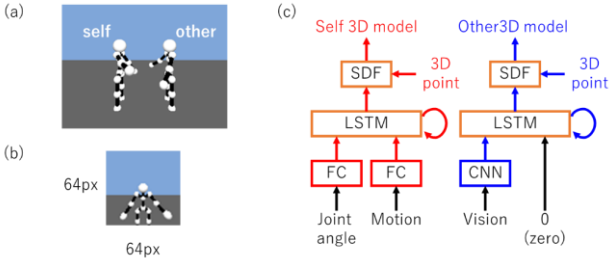

図 1 (a) 自己ロボットと他者ロボット．(b) 自己ロボットの視覚．(c) 重ね合わせネットワーク

3. 重ね合わせネットワークモデル

自己と他者の共有身体イメージを獲得可能なモデルとして重ね合わせネットワーク を構築する（図 1 (c)）．重ね合わせネットワークは自己と他者それぞれの情報を処理する self 側ネットワーク（図 1 (c) 左）と other 側ネットワーク（図 1 (c) 右）に分けられる． Self 側では，現時刻における自己ロボットの関節角度と関節角速度を入力として受け取り，次時刻における自己ロボットの 3D モデル（SDF）を予測する．Other 側では，自己ロボットの視覚を入力として受け取り他者ロボットの 3D モデル（SDF）を予測する．

3.1 共有モジュール

重ね合わせネットワークは self 側と other 側ネットワークにおいて自己と他者の情報を共有のモジュールにより処理する [2]．上述の通り，self 側，other 側ネットワークで入力される感覚情報が異なるが，それらの感覚情報はそれぞれ別々のモジュールによりエンコードされたのち，共通の LSTM (Long Short-Term Memory) モジュールに入力され，self 側，other 側でそれぞれ別々に平行に処理される．とくに，self 側の関節角度と other 側の視覚情報が LSTM の同じ受容ニューロンに入力され，self 側の運動感覚に対応する受容ニューロンについては，other 側ではゼロベクトルが入力される．LSTM の出力は後述の SDF モジュールに入力され，3D モデルの予測出力が生成されるが，SDF モジュールも self 側と other 側で共通である．このように自己と他者の情報を共通のモジュールを用いて処理するメカニズムのもと，感覚情報の予測学習を行うことによって自己と他者で共有の内部表現が獲得され得ることが示されている [2].

3.2 SDF モジュール

SDF モジュールは LSTM の出力に加え，3 次元座標の値を入力として受け取り，入力座標と 3D モデル表面との符号付き距離を出力する．任意の 3 次元空間座標に対応する符号付き距離を出力するように学習を行うことで，3D モデルを表現する SDF を近似することができ [3]，ロボットの形状と姿勢を表現する身体イメージの獲得が可能である

と考えられる．とくに，この SDF モジュールを self 側，other 側ネットワークで共有することによって，身体イメージを自己と他者で共有することを可能とする．

4. 自己と他者の共有身体イメージの獲得

提案した重ね合わせネットワークを自己と他者の 3D モデルの予測学習により訓練し，自己と他者で共有な身体イメージの獲得をシミュレーションする．

4.1 3D モデル予測学習

自己ロボットが両腕を動かしながら感覚情報と 3D モデルを観測する時系列データにおいて重ね合わせネットワークに 3D モデルの予測学習を行わせる．ただし，他者ロボットは一試行の時系列中では動作を行わず同じ姿勢を保ち，異なる試行において異なる姿勢をとる．

学習データとして 100 ステップからなる時系列を 1000 個作成した．また，1 ステップにおいて SDF を学習するための 3 次元点と符号付き距離のペアは 10000 個である． Self 側の関節角度エンコーダ，関節角速度エンコーダ，SDF モジュールとして全結合ニューラルネットワーク，other 側の視覚エンコーダとして畳み込みニューラルネットワークを用いる．学習は確率的勾配降下法により SDF の予測誤差を最小化することで行う．

また，本来自己と他者の 3D 空間中の位置や回転は異なり，自己と他者の身体の対応を認識するためには，位置・回転の違いも認識する必要があるが，本論文においては，自他の空間中での位置と回転の違いは既知であるという条件のもと学習を行う．具体的には，self 側，other 側ネットワークにおいて 3D 座標点はそれぞれ自己，他者ロボット中心の座標系に変換した上で SDF モジュールへ入力する．

4.2 自己身体イメージ の獲得

まず，self 側ネットワークにより自己ロボットの 3D モデル予測学習を行った．図 2 (a) に学習後のモデルが出力した 3D モデルの例を示す．出力された SDF を 3D モデル表面の法線方向を示す法線マップとして 2 次元画像上に投影しており，実際の自己ロボットの姿勢と 3D 形状を正しく予測できていることがわかる．つまり，モデルは予測学習により，内部に自己ロボットの身体イメージを構築したといえる．

4.3 自己身体イメージを用いた他者 3D モデルの予測

つぎに，前節で獲得された自己ロボットの身体イメージを用いて，他者ロボットの 3D モデルの予測学習を行った．ここで，other 側ネットワークにおいて self 側と共有の LSTM と SDF デコーダのパラメータは固定して学習を行っており，視覚情報を他者の 3D モデルを予測するための姿勢情報にエンコードする学習を行うことになる．

Self 側の学習の場合と同様に，学習後に出力された 3D モデルを可視化したところ（図 2 (b)），他者ロボットの姿勢に対応する 3D モデルを出力していることが確認された．また， LSTM の内部状態を PCA により次元削減を行うことで可視化した（図 3）．自己・他者ロボットに全く同じ動作をさせた場合において，self 側と other 側の内部状態をそれぞれ示しているが，self 側と other 側でほぼ同様な時間変化を示しており，自己ロボットと他者ロボットの 3D モデルを出力するために同じ内部表現を用いていることがわかる．また， self 側と other 側ネットワークでは入力される感覚情報（関節角度と視覚）が異なるにもかかわらず，関節角度と視覚を同様な内部表現にエンコードするように学習されたことも示している．これらの結果から，重ね合わせネットワークが共有モジュール上に自己と他者で共有の身体イメージを獲得，つまり，自己と他者の身体の対応を獲得したと考えられる．

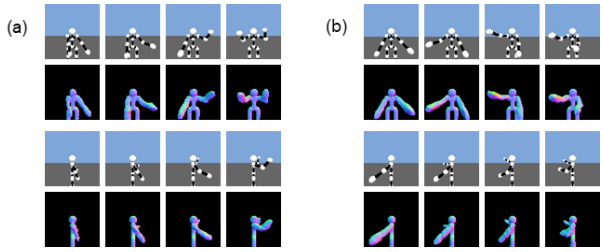

図 2　学習後のモデルの出力した 3D モデルの法線マップによる可視化．(a) 自己ロボット．(b) 他者ロボット．ロボットを正面・側面から映した場合を図示．

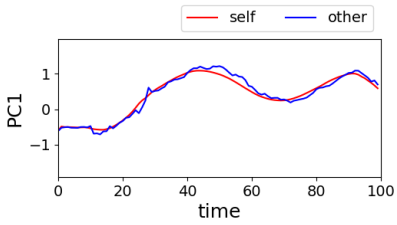

図 3　共有モジュール(LSTM)の内部状態

5. おわりに

本論文では，自己と他者の情報を共有のモジュールで処理する重ね合わせネットワークモデルにおいて，自己と他者の身体の 3D モデルの予測学習を行うことで，共有モジュール上に自己と他者の共有身体イメージが獲得されることを示した．今後は，自己と他者の空間中での位置・回転の違い，すなわち，視点の違いが未知である状況においても，自己と他者の身体の対応を獲得可能なモデルの構築を試みる．

謝辞

本研究の一部は JSPS 科研費 JP20K19880，および，北海道大学情報基盤センター人工知能対応先進的計算機システム共同研究の助成を受けたものである．

参考文献

[1] Sermanet, P., Lynch, C., Hsu, J., and Levine, S.: Time-Contrastive Networks: Self-Supervised Learning From Multi-View Observation: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 14-15 (2017).

[2] Noguchi, W., Iizuka, H., Taguchi, S. and Yamamoto, M.: Spatial Representation of Self and Other by Superposition Neural Network Model: Proceedings of the Artificial Life Conference 2019, pp. 531-532 (2019).

[3] Park, J. J., Florence, P., Straub, J., Newcombe, R., and Lovegrove, S.: Deepsdf: Learning Continuous Signed Distance Functions for Shape Representation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 165-174 (2019).