3D2-OS-37b-3

深層学習モデルによる動作指示に基づく衣服折り畳みタスク学習

Learning and generation of clothing folding task based on motion direction indication

鈴木彼方 *1*2 Kanata Suzuki　加瀬敬唯 *1*2 Kei Kase　尾形哲也 *1*2 Tetsuya Ogata

*1早稲田大学　理工学術院 Faculty of Science and Engineering, Waseda University

*2産業技術総合研究所　人工知能研究センター Artificial Intelligence Research Center, AIST

We describe a learning-based approach to co-work with humans by using neuro-dynamical models for learning multisensory information. Assuming that a robot works by receiving instructions from a human, the robot must acquire the meaning relationships of instructions that have characteristics such as ambiguity and context dependence, and switch or combine its motions to perform the required tasks. As a solution to this problem, we present a method with two phase deep learning model; a convolutional auto-encoder for conpressing highlly-dimensional camera image, and a hierarchical recurrent neural network (RNN) for integration of motions and image feture vectors. By embedding the short cyclic attracters to the network that consists of a direction phase and a behavior phase, our models self-organize the meaning relationships for sensory sequences and human instructions. To confirm the ability of our method, we demonstarated clothes-folding that consists of four short motions and three kinds of instructions that indicate the direction of each folding motion. The results showed that a robot can perform long seqencial task operation, and the proposed method self-organized its behaviors according to context.

1. 緒言

近年，少子高齢化対策や産業創出を目的として，人間とともに共同作業を行うこと目指したロボットが注目を集めている．共同作業において，ロボットは人間と作業内容を共有したり，人間から指示を受けてタスクを遂行する必要があるが，実現にはいくつかの課題が残る．実環境で求められるタスク動作は多岐にわたり，ロボットが取得するセンサ情報に合わせて適切に動作を選択，及び，生成する必要があるためである．人間はこれらの知覚情報を無意識化に処理することが可能であるが，従来のロボティクス分野で用いられていた作り込みによる手法 [Perez 15] を適用することは困難である．また，ロボットが人間と協調するためには，指示を受けてインタラクティブに動作を生成しなければならない．しかしながら，記号接地問題 [Harned 90] に表されるように，人間による記号的な指示と現実空間の意味関係は動的に変化し，通常，人間によるロボットへの指示はその動作の詳細まで具体的に明示されない．例えば，人間がロボットに物体を取るように指示を与えた場合，物体を把持する箇所はその視覚的な形状によって変わる．このような状況依存性を考慮した上で，ロボット自身が視覚情報と統合して判断，及び，動作を行う必要がある．

人間-ロボット間のインタラクションを目的として，近年では認知ロボティクスの分野で言語表現と行動の統合モデルが広く研究されている．中でも再帰的な結合を持つ神経回路モデルである Recurrent Neural Network (RNN) は，各時系列情報の意味関係を自己組織化的に獲得することが可能である [Ogata 07][Heinrich 14]．山田らは RNN を小型ロボットに適用し，言語-運動間の意味構造を反映したダイナミクスを獲得することで，ロボットとの即時的なインタラクションを実現した [Yamada 16]．しかしながら，視覚情報として物体の色重心座標を用いているため，物体形状の変化に対応することが困難であった．また，上記研究の多くは言語意味関係の獲得に注力しており，与えられたタスク動作における実環境下での汎化性を考慮していない．実際の共同作業下の物体形状は不定であり，視覚情報との統合が求められる．これに対し，我々はオンラインで画像情報を取り込むことにより，動作を受動的に切り替える手法に取り組んできたが [Yang 16]，言語的なインタラクションは行われておらず，限られたセンサ情報から一意に決められた動作を生成していた．

本研究では，実環境での共同作業タスクを想定し，階層型 RNN を多自由度ロボットに適用し，動作分岐を伴う衣類折り畳みタスクを行った．インタラクティブに動作を切り替えるため，オンラインで画像情報と 3 種類の動作指示を入力しながら動作生成を行う．また，外部状況により異なる動作を意味する入力に対し，適切なダイナミクスが獲得されたかを議論する．

連絡先: 鈴木彼方，早稲田大学理工学術院，東京都新宿区大久保 3-4-1，03-5286-2742，suzuki@idr.ias.sci.waseda.ac.jp

2. 統合学習モデル

本研究では前章で述べた RNN を含む 2 種類の深層学習器を組み合わせたモデルを用いて動作学習を行う (図 1)．深層学習器とは多層の階層構造をした神経回路モデルを指し，様々な高次元データを前処理なしに扱うことが可能である．

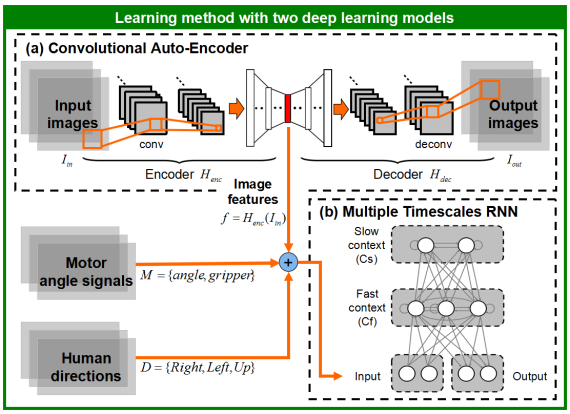

図 1: 砂時計型 NN(a) と階層型 RNN(b) からなる統合モデル．取得画像 I，運動情報 M，動作指示 D を統合的に学習する．

2.1 Convolutional Auto-Encoder

はじめに，ロボットより取得する実画像は高次元･大量のデータセットのため，砂時計型の深層学習器 [Hinton 06] に畳み込み層を組み込んだ Convolutional Auto-Encoder (CAE) を用いた表現学習を行う．CAE はエンコーダー層とデコーダー層から構成され，入出力が恒等写像 $I_{out} = H_{dec}(H_{enc}(I_{in}))$ となるように学習することで，中間層に入力画像 I_{in} の特徴をよく表す低次元の画像特徴量 $f = H_{enc}(I_{in})$ が獲得される．本研究では入力層近くに畳み込み層を導入することで，高解像度の画像から物体のエッジや形状を効率よく表現することが可能となる．また，画像特徴量から同様に復元するために，畳み込み処理を転置した逆畳み込み層を導入する．抽出された画像特徴量は視覚情報として，次節に示す運動，動作指示との統合学習に用いられる．

2.2 Multiple Timescale RNN

第一章で述べたように，ロボットは視覚情報からの受動的な動作変更，及び，動作指示に対する即時的な動作切り替えが求められる．本研究では視覚情報 f，運動情報 M，動作指示 D との統合学習器として，応答速度の異なる階層的なニューロン郡を持つ Multiple Timescales RNN (MTRNN)[Yamashita 08] を用いる．MTRNN は入出力層である IO 層と異なる発火速度 τ を持つコンテキスト層 (Cs/Cf 層) の 3 層からなり，それぞれ再帰的な入力を持つ．時刻 t における i 番目のニューロンの内部状態は以下の式で表される．

$$u_i(t) = \left(1 - \frac{1}{\tau_i}\right) u_i(t-1) + \frac{1}{\tau_i}\left[\sum_{j \in N} w_{ij} x_j(t-1)\right] \quad (1)$$

ここで，w は各ニューロン間の結合重み，N は結合元のニューロン数を表し，モデルは BPTT 法 [Werbos 90] を用いてパラメータの最適化が行われる．上記に示した各ニューロンの発火速度の違いから，MTRNN は人間による動作指示と物体の形状遷移を含むタスクプロセスを階層的に自己組織化する．

2.3 動作指示に基づいたダイナミクスの埋め込み

ロボットはタスク途中に人間の指示に基づいて即時的な動作の切り替え，及び，組み合わせを行う．学習する教示データは人間による入力が行われる動作指示フェーズと，動作生成フェーズにより構成され，ロボットはタスク中にこれらのフェーズを交互に行う．本研究ではこれらのシーケンスをアトラクタ上に構成し，即時的に反応するニューロン郡 (Cf 層) にダイナミクスとして埋め込むことで，人間の指示に対応する動作生成を可能とする．加えて，非即時的に反応するニューロン郡 (Cs 層) に表現されるタスクプロセスと統合することにより，物体に対する柔軟な動作対応と視覚-指示間の意味表現を可能とする．

3. 実験

3.1 実験設定

本手法の有効性を検証するため，産業用ヒューマノイドロボット Nextage を用いたロボット実験を行う．ロボットは各腕 6 自由度と Gripper 開閉 1 自由度を持ち，計 14 自由度の多自由度系となっている．また，ロボットの頭部カメラから取得する RGB カメラ画像は 112x112x3[pixel] の計 37632 次元を，CAE により 10 次元の画像特徴量に圧縮する．従って，次節で述べる人間による動作指示 3 種類と合わせて，MTRNN の入力は合計 27 次元となる．

3.2 タスクデザイン

第一章で述べた動作分岐や組み合わせを伴う動作として，本研究では，衣服の折り畳みタスクを行う．衣服は子供用の半袖シャツを扱い，ロボットの目の前の卓上の可動域内に設置する．衣服の折り畳みタスクは 4 回の折り畳み動作の組み合わせからなり，ここでは明示的にそれぞれを Stage 1-4 とする．はじめにシャツの両袖を 2 回折り畳み (Stage 1, Stage 2)，次に縦方向に二つ折りにし (Stage 3)，最後に横方向に二つ折りにする (Stage 4)．各 Stage 間に，次に折り畳む方向に対応する指示 3 種類”Right”, ”Left”, ”Up”が入力される．従って，タスクの組み合わせは 4 通り存在する (図 2)．このうち動作指示は折り畳む方向のみを表すため，タスクプロセスに伴う物体形状の遷移に関する情報は与えられない．例えば，Pattern 1 中の Stage 1 と Stage 4 において与えられる動作指示は同じにも関わらず，求められる折り畳み動作は異なる．ロボットはコンテキストに埋め込まれた画像情報を考慮したうえで，適切な動作を生成する必要がある．また，実験では教示データとして Pattern 1-3 を，テストデータとして未知の組み合わせである Pattern 4 を扱う．タスクを遂行するためには，動作指示から折り畳み動作を切り替え，適切に組み合わせねばならない．

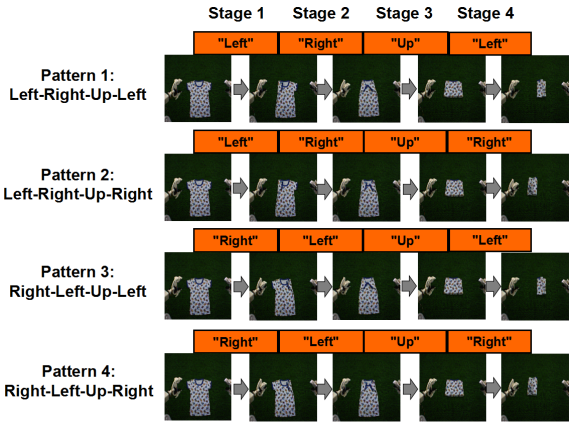

図 2: 折り畳みタスクにおける，4 回の折り畳み動作の組み合わせ．画像はロボットから見たタオルの形状を示す．

3.3 結果と考察

はじめに，学習済モデルのパフォーマンステストを行った．ロボットは画像情報と動作指示をオンラインで受け取り，MTRNN の次時間の出力として予測される関節角度情報を両腕の入力に用いる．未知の組み合わせ (Pattern 4) における動作生成結果を図 3 に示す．図は各 Stage ごとに破線で区切られており，Stage 1 と Stage 4 では同じ動作指示”Right”が入力されているが，ロボットは視覚情報に基づいて適切な動作が生成したことが確認された．図 3 の動作は，物体の位置をロボットの届く範囲で変更しながら試行した 6 回のうち，最も悪かった結果を示しており，各関節角度のステップごとの平均二乗誤差は 0.00331 であった．これは図中の正しい関節角度の値 (実線) とほとんど外れておらず，十分にタスク遂行可能な値である．

次に，動作指示に基づいた衣服折り畳みタスクがいかに内部表現されているかを解析するため，コンテキスト層 (Cf，Cs) の内部状態を主成分分析により可視化した結果を図 4 に示す．図上部にしめす Cf 層の内部状態には，動作指示ごとの 3 つのクラスタが現れており，図下部にしめす Cs 層の内部状態には，Stage ごとの 4 つのクラスタが現れている．ここで Stage 1 と Stage 4 は同じ指示，異なる折り畳み動作であることから，コ

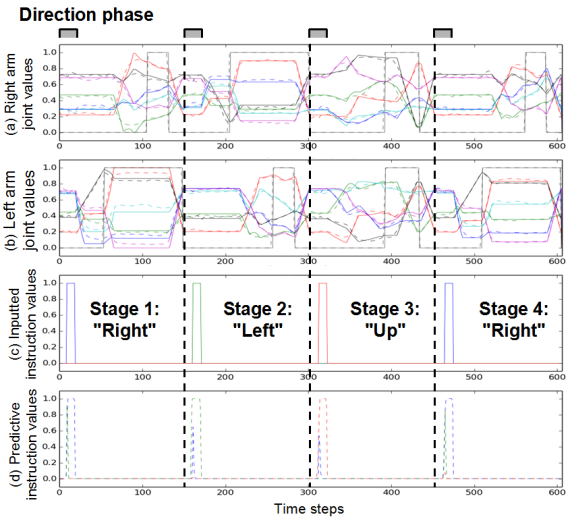

図 3: 未知の組み合わせ (pattern 4) における動作生成結果. (a) 右腕の出力, (b) 左腕の出力, (c) 入力された動作指示, (d) モデルにより予測された動作指示, をそれぞれ表す. 実線は正しい値を示しており, 破線はモデルによる予測を表す.

ンテキストでクラスタが別となっている. このことからモデル内部において, 切り替えを行うための即時的な反応を示す表現と, 視覚情報を含む統合的なタスクプロセスを示す表現が獲得されていることがわかる. モデルの出力はこれら両コンテキストを統合したものであることから, 人間の指示-視覚情報と統合して状況に即した動作を生成することが可能である.

4. 結言

本研究では, 人間-ロボット間の共同作業を目的として, 人間による動作指示に基づくロボットのオンライン動作生成について述べた. 階層型 RNN を実装した多自由度の工業用ヒューマノイドロボットに画像特徴量, 運動情報, 動作方向を表す 3 種類の動作指示を学習させたところ, 複数の動作からなる衣服折り畳み動作の生成が可能となった. その際, 動作指示-視覚情報を統合して適切なダイナミクスの生成や切り替え, 及び, 意味関係の表現が確認された. 今後の展望として, 語彙の追加, 別物体や別動作へのモデルの適用を行う.

謝辞

本研究は, 文科省科研費基盤研究 (A)(No.15H01710), NEDO の支援の下で行われた. ここに謝意を表します.

参考文献

[Harned 90] S. Harned, "The symbol grounding problem," Physica D, Vol. 42, pp. 335–346, 1990.

[Heinrich 14] S. Heinrich, and S. Wermter, "Interactive Language Understanding with Multiple Timescale Recurrent Neural Networks," 24th International Conference on Artificial Neural Networks, LNCS. vol. 8681, pp. 193–200, 2014.

[Hinton 06] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, vol. 313, pp. 504–507, 2006.

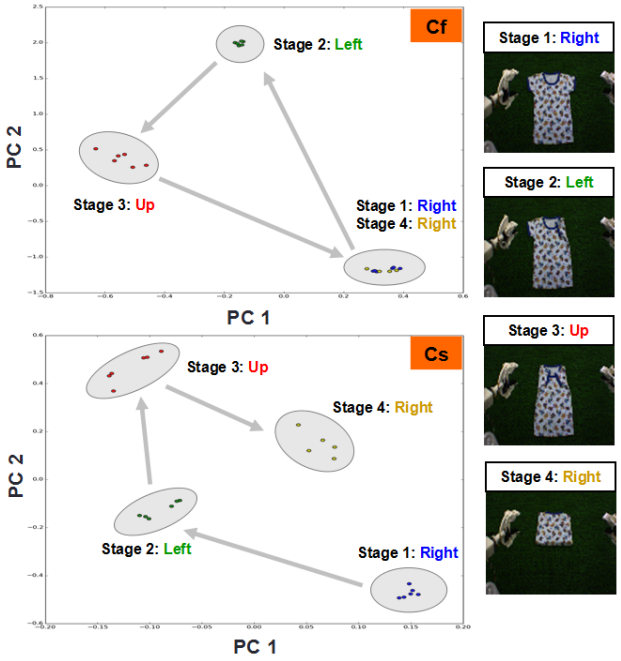

図 4: 各折り畳み動作中の MTRNN のコンテキスト層の内部状態の平均 (PC1-PC2). 図中の右部は動作指示入力時における, ロボットから見た物体状態.

[Ogata 07] T. Ogata, M. Murase, J. Tani, K. Komatani, and H. Gokuno, "Two-way Translation of Compound Sentences and Arm Motions by Recurrent Neural Networks," IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1858–1863, 2007.

[Perez 15] C. Perez-D'Arpino and J. A. Shah, "Fast target prediction of human reaching motion for cooperative human-robot manipulation tasks using time series classification," in 2015 IEEE International Conference on Robotics and Automation, pp. 6175–6182, 2015.

[Werbos 90] P. J. Werbos, "Backpropagation through time: what it does and how to do it," Proceedings of the IEEE, vol. 78, no. 10, pp. 1550–1560, 1990.

[Yamada 16] T. Yamada, S. Murata, H. Arie, and T. Ogata, "Dynamical Integration of Language and Behavior in a Recurrent Neural Network for Human-Robot Interaction," Frontiers in Neurorobotics, vol. 10, Article 5, pp. 1–17, 2016.

[Yamashita 08] Y. Yamashita and J. Tani, "Emergence of functional hierarchy in a multiple timescale neural network model: a humanoid robot experiment", PLoS Computational Biology, 4, 11, pp.e000220-1–e1000220-18, 2008.

[Yang 16] P.C. Yang, K. Sasaki, K. Suzuki, K. Kase, S. Sugano, and T. Ogata, "Repeatable Folding Task by Humanoid Robot Worker using Deep Learning," IEEE Robotics and Automation Letters, vol. 2, no. 2, pp. 397–403, 2016.