第 8 回横幹連合コンファレンス
2017.12.2-3 立命館大学 朱雀キャンパス

対象の状態変化を伴う道具の操りプロセスの記述・想起・再現

○島田 伸敬 （立命館大学） 松尾 直志 （立命館大学）

Representation, Recall, and Replay of Tool Operation with Target State Transition

∗N. Shimada (Ritsumeikan University), T. Matsuo (Ritsumeikan University)

Abstract– When a human manipulates a tool, the person not just grasps it but also may manipulates it with hands and fingers, so that it change its orientation and position, or deform the shape. Furthermore, according to the state transition of such object, the next operation is performed as a larger scale process by a series of such operations. We discuss what is necessary to and how to describe such manipulation process and reproduce the behavior by a robot.

Index terms– Object handling, Mimicking human behaviors, Process modeling

1 はじめに

モノの種類や機能を認識するロボットビジョンでは、近年非線形パターン識別器や深層機械学習を応用した成果が注目されているが、主に静的な物体の形状や見え方に着目した認識に主眼がおかれてきた [1]。しかし鎌倉ら [2] が示唆したように、人が手でモノを操作する指使いとそのモノの機能は関連しており、動きを伴う指使いとモノの形を突き合わせて初めてモノの機能を発現させる動的過程が理解できる。ロボットが人と同じ環境で道具類を使用するには、人がモノを操作する指使いをまねることが一番の近道である。

これは従来から教示として研究され、産業ロボットでは一早く実用化されたが、事前に人間にロボットを操作させ、その運動を再現するものであった。近年では、ロボットの構造が既知の場合に視覚・力覚フィードバックを用いて固形物や紐などの対象を操る制御手段の研究もがある [3]。数理的な制御モデリング手法は最適性の観点から、特定の状況下での把持など基本的な状態維持動作の実現を達成する。最近年では深層学習を取り込んだ強化学習手法の進展により、試行錯誤から自発的に行動を会得する研究も進んでいる。実世界の対象の状態を順番に変化させながら目的を達成することはまだハードルの高い課題である。

複雑な物体やシーンの操作は、ただの運動再現や状態維持ではなく、手順=プロセスとして記述される状態遷移を伴う一連の動作群である。たとえば箱を指をつかってこじ開ける、ひもを結ぶ、といった緻密で複雑な指使い (例えば Fig.1) が相当する。またある手順の一手を実行しても対象の物体やシーンが想定した状態変化を起こすとは限らず、改めて状況を観測した上で必要であれば手戻りして再度同じステップを繰り返す必要もある。この手順をロボットにモデリングさせ、物体形状やシーン状況の多少の変動にうまく適応しつつ模倣させらるか、途中の失敗からの回復も自動的に試みることができるか、が課題となる。

本稿では、物体／シーンを対象にした人の行動、とくに物体把持とシーン変遷を引き起こす動作を対象に、物体のパーツ形状と手指の把持姿勢の関係を観察から会得する研究、ならびに部屋内シーンを変化させる動作シーケンスのモデリングと想起についての著者らの研究成果を述べる。

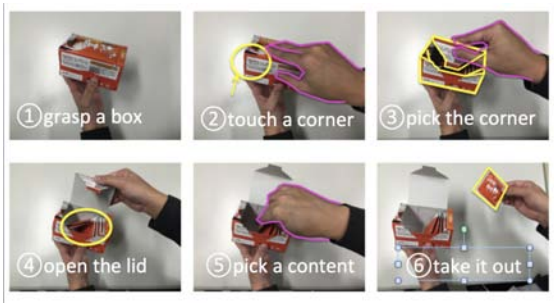

Fig. 1: picking out a teabag in a box

2 物体パーツの三次元形状観察に基づく把持パタンの想起

鎌倉 [2] によれば把持のパタンによって発揮する機能が異なるとされ、視覚情報から物体機能を推定する研究 [4] がある。本節では人が把持直前に物体のパーツ形状に合わせて手指形状を変えて把持を行う preshaping の模倣する機械学習の枠組を述べる。パーツの形状と把持パタンの関わりをモデリングすることで、物体形状をその機能タイプによってクラスタリングし推定することを目指す。

2.1 手法の概要

Fig.2 に，持ち方を想起する学習モデルの作成手順を示す．本稿では持ち方を 30 次元の記述子（持ち方パラメータ）で表しており，持ち方と物体形状の組み合わせを把持パタンと呼ぶ．最終的には未知の物体画像を学習済みモデルに入力するとその物体に対応した持ち方パラメータが想起できるモデルを作成することを目標とする．

まず，Auto-encoder を用いて把持画像 (深度画像, 手のマスク画像，物体のマスク画像の 3 チャンネルから成る画像) から持ち方パラメータが写像される空間の学習を行う．持ち方パラメータの作成には学習済み Auto-encoder の Encoder 部を使い，持ち方パラメータから把持画像を復元する際に Decoder 部を利用する．次に Auto-encoder の学習結果である持ち方パラメータを教師とし，CNN を用いて物体のみ画像と持ち方パラメータの関係を学習する．本手法では入力画像中の物体領域の位置ずれの対策のため，画像の位置ずれに不変な記述子を生成する Shift Invariant Auto-encoder を使用する [9]。また同じ物体でも部分ごとに異なる持ち方がありうる。そこで 64px × 64px × 3 チャンネルの把持

画像から各チャンネルの 32px × 32px のパッチ画像を切り出して学習に用いた．

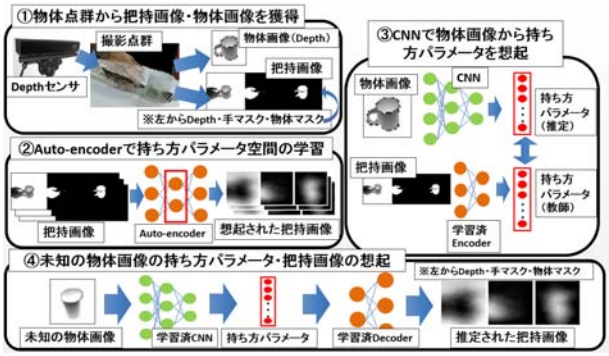

Fig. 2: Overview of the method

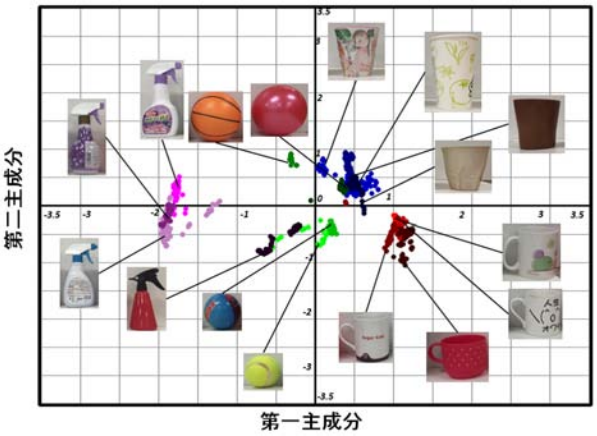

Fig. 4: Plots of Grasping descriptor for each object

2.2 把持パタン想起実験

使用した物体は，Fig.3 に示す通りマグカップ，(取っ手無し) コップ，ボール，スプレーの 4 カテゴリ、4 種類の計 16 物体である．それぞれの物体に対して，把持画像を約 100 枚ずつ作成し，把持パタンの想起モデルの学習を行った．

Fig. 3: 16 training objects

2.2.1 Auto-encoder による持ち方パラメータ空間の学習結果

まず，Shift Invariant Auto-encoder の学習により獲得した持ち方パラメータがカテゴリごとに持ち方パラメータ空間上で分かれているかを確認する為に，Fig.4 に持ち方パラメータ空間の第一主成分，第二主成分の分布を示した．

Fig.4 より，同じ物体は持ち方パラメータ空間上でも概ね近い位置にプロットされていることが分かる．また別物体であっても，同じカテゴリであれば近い位置に集まっている．スプレーカテゴリの物体は他の物体と形状が大きく異なるため，Fig.4 の空間上でも他のカテゴリのサンプルと離れた位置にプロットされている．

2.2.2 CNN による把持画像の想起結果

次に，Fig.5 に想起された持ち方パラメータから Decoder を用いて作成した復元画像を示す．

手形状は取っ手の部分を把持しているような手マスク画像が想起された．コップは胴の部分を包むような手形状が想起され，ボールは全体を覆うような手形状が想起されている。スプレーに関してもレバーを引く指の形が再現されている。

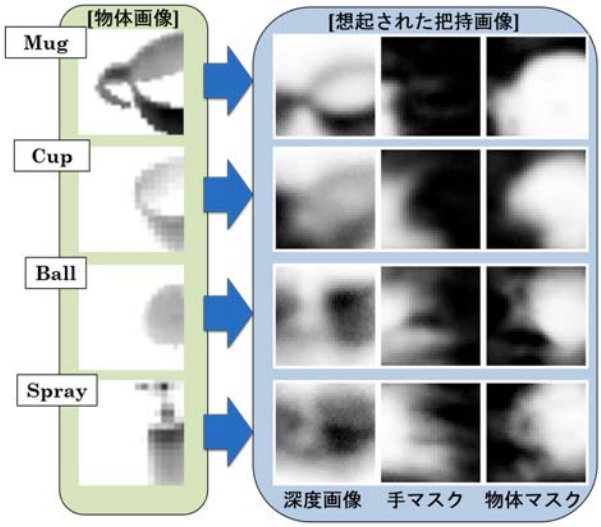

Fig. 5: Recalled grasping images for 3-D shapes of object parts

2.2.3 マグカップの取っ手の有無による持ち方の変化

最後に，同物体であってもパーツ毎に異なる持ち方が想起されるのか，一つの物体画像から異なる二カ所をパッチ画像として切り出し，それぞれ想起を試みた．Fig.6 にその結果を示す．今回はマグカップの取っ手を含めたパッチ画像と取っ手を隠したパッチ画像からそれぞれ把持画像の想起を行った．取っ手のあるパッチ画像に関しては取っ手を握るような手形状が想起され，取っ手を隠したパッチ画像に関しては底の部分を支えるような持ち方が想起された．また，マグカップの取っ手を隠したパッチ画像とコップの似たような部分のパッチ画像を比較したところ，概ね同じような手形状が想起された．この結果から，学習モデルがマグカップの重要なパーツである取っ手を認識し，そのパーツに適切な手形状を想起していることが確認された．

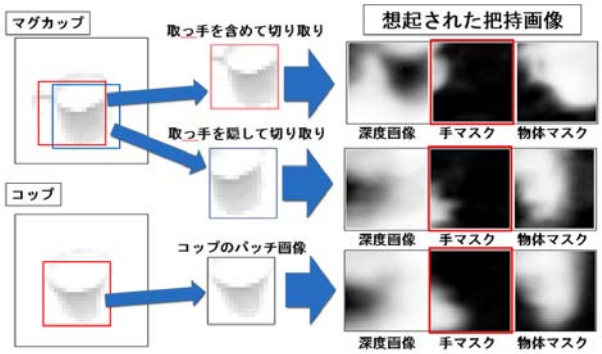

Fig. 6: Variations of grasping type for cups with/without handle

3 シーン変遷の変遷をトリガーにした動作プロセスの想起

室内等において家具や道具類をその機能に応じて使いこなす技能をロボットに与える手段として、Learning from Demonstartion (LfD)[6] が提案されている。人の使用例を画像として提示することで道具使いの行動を教示するのが骨子だが[7],[8]、道具や背後のシーンが行動によって変化する場合、その変化・変遷が予期したものであるかどうかを確認しながら、場合によって手戻りをしてやり直すような課題は難易度が高い。本節では、人の行動とそれによるシーンの状態（3-D 形状）の変化の連鎖を動的なプロセスとして学習し、現在のシーンの状態に応じて、ゴールとして想定されるシーン状態に到達するために必要な動作列を生成する枠組みについて述べる。

3.1 LSTM による動作とシーン変化の共起性のモデリングと想起

現時点における 3-D シーンの状態（深度画像）と人体の 3-D 姿勢スケルトン、および目標到達状況における 3-D シーンと人体姿勢を入力として指定すると、目標到達のために次にとるべき人体姿勢およびそれによって予期される変化した 3 − D シーンの状態が出力されるリカレント型ニューラルネットを構築する。ここでは LSTM Long Short-Term Memory(LSTM)[5] モデルを用いた。

3.1.1 Sparse Auto-Encoder によるシーン特徴

3-D シーンの状態として深度画像そのものを用いるとその高次元表現のためにモデル学習が困難になるため、予め想定されるシーン画像のセット（ここでは時系列としての画像の近接性を考えない）に対し Sparse Auto-Encoder[9] を用いて低次元表現を得る。Fig.7 に Auto-Encoder のパラメータ学習に用いた 7 種類のシーン状態カテゴリに相当する深度画像例を示す。

Fig. 8 に今回用いた Auto-Encoder の構造を示す。入力は 32x32 のサイズに縮小したシーン深度画像であり、エンコードされた 50 次元の中間層をシーン特徴記述子として用いる。

3.1.2 LSTM による人の行動とシーン変化のプロセスモデリング

Fig. 9 に LSTM の学習に用いた訓練シーケンスの例を示す。ここではホワイトボードに文字を書く、椅子を引いて座る、座って箱を持ち上げるなどシーンの状態に依存して生起し状態の変化を引き起こす 6 種類の行動クラスについて複数のシーケンスを用意した。

Fig. 7: Examples of training scene images for learning Sparse Auto-Encoder

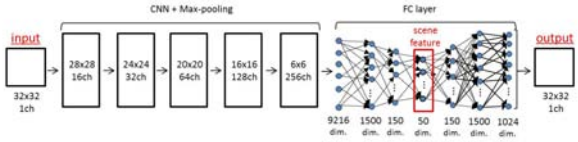

Fig. 8: Network structure of the Auto-Encoder

Fig. 10 に今回構築した動作とシーンの変遷シーケンスを出力する LSTM モデルのネットワーク構成を示す。入力特徴はスパース Auto-Encoder で圧縮された 3-D 深度シーン画像の 50 次元の低次元特徴と 75 次元の人体のスケルトン特徴 (Kinect から取得できる 25 関節の三次元位置）からなる。それぞれ現時刻と目標となるゴール状態におけるシーン特徴とスケルトン特徴のペアを入力し、モデルは将来ゴール状態になるための次フレームにおける人体およびシーン特徴を出力する。

人の動作によって必ずしも学習に用いた理想的なシーン状態変化がおこるとは限らず、状況によっては人体特徴とシーン特徴のゴールまで進捗度合いに差異が出る。そこで、人体特徴とシーン特徴独立に現時点とゴール状態のフレーム差を推定する 2 層全結合回帰モデルを別途学習しておく。まずこのモデルにより人体特徴とシーン特徴ごとにゴール状態の何フレーム手前の状態にあるかを推定し、それを各特徴といっしょに LSTM に入力して次フレームの予測を出力する。次フレーム特徴の予測を行う LSTM モデルは 600 ノード 2 層の LSTM ユニットから構成される。

学習時の損失関数を式 1 に示す。

$$
\begin{aligned}
E(t) \quad = \quad & w_1 \left| s_{t+1} - y_{t+1}^{(0,1,\ldots,t)} \right|^2 + \frac{w_2}{N-2} \sum_{t=2}^{N-1} \left| s_{t+i} - y_{t+i}^{(0,1,\ldots,t)} \right|^2 \\
+ \quad & w_3 \left| s_{t+N} - y_{t+N}^{(0,1,\ldots,t)} \right|^2 + w_4 \left| b_{t+1} - z_{t+1}^{(0,1,\ldots,t)} \right|^2 \\
+ \quad & \frac{w_5}{N-2} \sum_{t=2}^{N-1} \left| b_{t+i} - z_{t+i}^{(0,1,\ldots,t)} \right|^2 + w_6 \left| b_{t+N} - z_{t+N}^{(0,1,\ldots,t)} \right|^2 \\
+ \quad & \frac{w_7}{N-1} \sum_{t=0}^{N-1} \left| z_{t+i}^{(0,1,\ldots,t)} - z_{t+i+1}^{(0,1,\ldots,t)} \right|^2 \qquad (1)
\end{aligned}
$$

Fig. 9: Examples of training actions for learning LSTM model

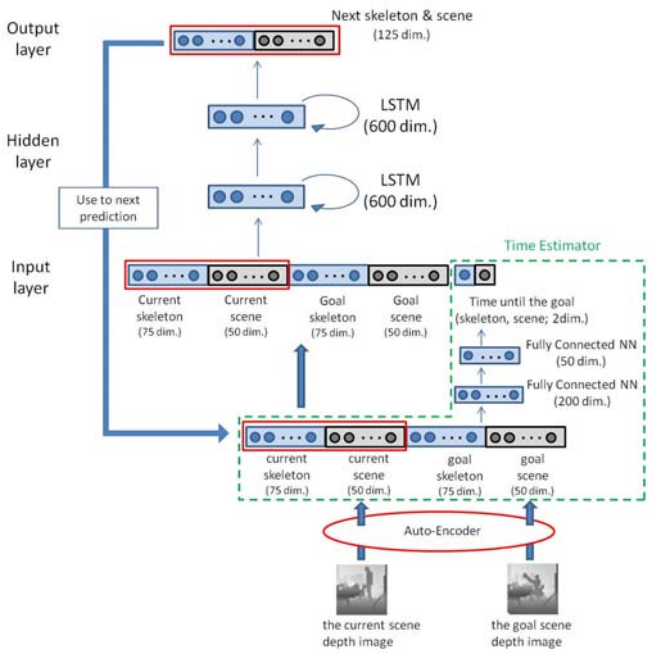

Fig. 10: Network structure of human action and scene change estimator

$w_1, \cdots, w_7$ はそれぞれの項の重み係数、$y_{t+i}^{(0,1,\ldots,t)}$, $s_{t+i}^{(0,1,\ldots,t)}$、$z_{t+i}^{(0,1,\ldots,t)}$、$b_{t+i}^{(0,1,\ldots,t)}$ はそれぞれ、$t+i$ における人体スケルトン特徴と低次元シーン特徴の推定値と実測値である。

3.1.3 動作とシーン変化の想起

Fig. 11 に訓練されたモデルによる動作とシーン変化の想起結果を示す。初期入力として最初のシーン特徴と人物スケルトン特徴、ならびに目標となるゴール状況におけるシーンと人物スケルトン特徴をモデルに与える。出力された次フレームにおけるシーン特徴と人物特徴を再帰的に入力し、次々に動作とシーン変化を想起させる。

Fig. 11 shows the result of recalling by the model.

上段は「座る (sitting)」に対する想起結果で、濃淡の背景は想起されたシーンの深度画像を表す。初期状態として椅子の後ろに人物がたった状態、ゴール状態として椅子に腰掛けた状態を与えたところ、まずはじめに椅子を引く動作が出力されてシーン画像内で椅子がテーブルから引かれ、その後に椅子に座るシーケンスが想起された。

下段は「書く (drawing)」に対する想起結果で、「座る」のときと同じ初期状態を与え、ゴール状態としてホワイトボードの前に立つ状態を指定するとその前に移動して腕を動かす動作が想起された。

「座る」に対する想起においては、まず最初に椅子を引く行動が出力されるが、そのときもしシーンが期待（想起）どおりに椅子が引けた状態にならなかった場合、椅子を引く動作を繰り返す必要がある。現状のモデルではこの繰り返す動作が不完全であり、椅子が引けないシーンをわざと入力しつづけた場合には人物の動作が 1 フレーム進むごとにじわじわとゴールの姿勢に近づいてしまう結果がみられた。これは損失関数において現在の入力シーン特徴と想起された次フレームのシーン特徴の差異が大きくなることのペナルティを課すように変更すべきことを示唆していると考えている。

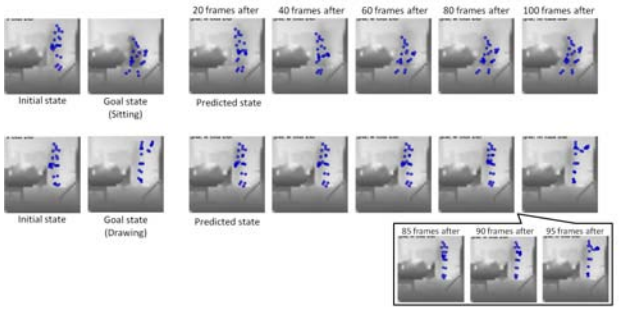

Fig. 11: The result of the human actions and the scene change recalling; Blue marks are the person's skeleton

4 まとめ

本稿では，物体の形状やシーンの状態と人間の行動の間の共起性をモデリングして、物体機能に応じた把持の仕方や対象の状態変化を逐次的に引き起こす手順動作の生成を獲得する枠組みの提案を行った。物体やシーン状態をトリガーにして人と同じような動作をロボットが模倣することができれば、全く白紙状態から強化学習等でスキルを獲得しなくても人と行動環境をともにし人を観察することによって自動的に自ら物の使い方を会得するロボットを構成するステップになると考える。現時点ではやや不完全な記述モデルの提示に留まるが、想起された手の把持状態画像をもとにハンドロボットに実際に preshaping を引き起こさせる実験を現在行っており、その成果については改めて報告をしたい。

参考文献

1) Zhang, et al., "Object detection via structural feature selection and shape model", IEEE Trans. on Image Processing, vol. 22, no. 12, pp.4984–4995 (2013)
2) 鎌倉, " 手のかたち 手の動き ", 医歯薬出版 (1989)
3) 山川ら,"ロボットハンドの構造・運動を考慮した操りスキルの統合に基づく結び目の生成計画 ", 日本ロボット学会誌, Vol. 31, No. 3, pp.283–291 (2013)
4) Tadahiro Kitahashi et.al, " Cooperative Recognition of Human Movements and Objects and Its Modeling ", Information Processing Society of Japan. CVIM, pp.109-116 (2005)
5) Hochreiter, Sepp, Jürgen Schmidhuber. "Long short-term memory", Neural computation 9.8 pp. 1735-1780 (1997)
6) Atkeson, Christopher G., and Stefan Schaal. "Robot learning from demonstration", ICML. Vol. 97. pp. 12-20 (1997)
7) Lee, Kyuhwa, et al. "A syntactic approach to robot imitation learning using probabilistic activity grammars.", Robotics and Autonomous Systems 61.12, pp. 1323-1334 (2013)
8) Sermanet, Pierre, et al. "Time-Contrastive Networks: Self-Supervised Learning from Multi-View Observation.", arXiv preprint arXiv:1704.06888 (2017).
9) T. Matsuo, N. Shimada, "Construction of Latent Descriptor Space of Hand-Object Interaction",The 22nd Joint Workshop on Frontiers of Computer Vision (FCV2016), pp. 117-122 (2016)