Vol.2016-CVIM-202 No.20
2016/5/12

3-D シーン観察に基づく 手と物体の関係性の学習と把持パタンの想起

川上 拓也[†2]　松尾 直志[†1]　小川 陽子[†2]　島田 伸敬[†1]

概要：本稿では日常的なシーンから物体の把持状態の推定と 3 次元的な想起を行う手法の提案をする近年，画像認識の分野では，画像から物体を認識し識別するという研究が盛んに行われている．物体というのは様々な機能を持っており，その機能は物体とそれを把持する人間の手の形と深い関係がある．しかし，手の動作と物体の関係から物体を識別するという課題は，手と物体が相互に隠蔽を行うため手の全体像の検出や姿勢の推定，領域分割などが一般的に困難である．そこで，日常的な物体把持シーンから持ち方の情報を抽出し，機械学習を用いて物体と持ち方の関係を学習させることで，物体からその物体の機能を発現させるような 3 次元的な持ち方を想起する手法を提案する．本手法では，RGB-D センサを用いて物体把持シーンの点群を撮影し，その点群から作成した把持画像から手と物体の共起性を表すような局所特徴を抽出して得たベクトルを持ち方パラメータとする．ここでいう把持画像というのはある物体を把持した状態の点群から得た，深度情報，手領域情報，物体領域情報の 3 チャンネルから成る画像の事である．その後，物体と持ち方パラメータの関係の学習を行い，学習に使用していない物体の持ち方をパラメトリックに記述する．また，物体から想起された持ち方パラメータを用いて把持画像の復元を行う．

Learning Hand-Object Interaction and Inference of Grasp Pattern Based on 3-D Scene Observation

TAKUYA KAWAKAMI[†2]　TADASHI MATSUO[†1]
YOKO OGAWA[†2]　NOBUTAKA SHIMADA[†1]

1. はじめに

1.1 研究の背景と目的

人間が把持を行う物体というのは様々な機能を持っている．また，人間は物体を把持する際に，その物体の機能に応じて手の形を変えて把持する[1]．本稿では機械学習を用いて，人間が物体にどういったアプローチを行うかという視点から物体の把持パタンを推定する．

視覚的情報から物体の機能を認識するという課題に取り組んでいる研究として，北橋らの報告[2]が挙げられるが，人間が物体を使用する際に移動を伴う物体に対象が限定されている．本稿では，物体を把持した際の物体と手の形状による機能の認識を行うため，使用の際に移動を伴わない物体も認識対象とする．

物体のみの画像から把持パタンが想起できれば，ロボットハンドの分野では，ロボットが物体を認識した際にその物体の形状から適切な把持パタンが推定できる．その把持パタンに合わせて指の関節角度や手首位置を設定できれば，物体を把持することができる．

室内監視システムの分野では，物体を認識した際にその物体の機能まで想起できるようになり，物体を機能ごとに分類や，その物体を用いる人間の行動推定ができるのではないかと考える．

†1 立命館大学院 情報理工学研究科
Ritsumeikan University Information Science
†2 立命館大学情報理工学部
Ritsumeikan University Graduate School of Information Science and Engineering

1.2 本稿の構成

図 1 に，持ち方を表すパラメータを想起する学習モデルを作成する手順を示す．本稿では，この学習モデルのことを『把持パタン想起モデル』と呼ぶ．最終的にはある物体の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．2 章で学習に使用する画像の作成手順，3 章で把持パタン想起に使用した学習モデルの説明を行う．

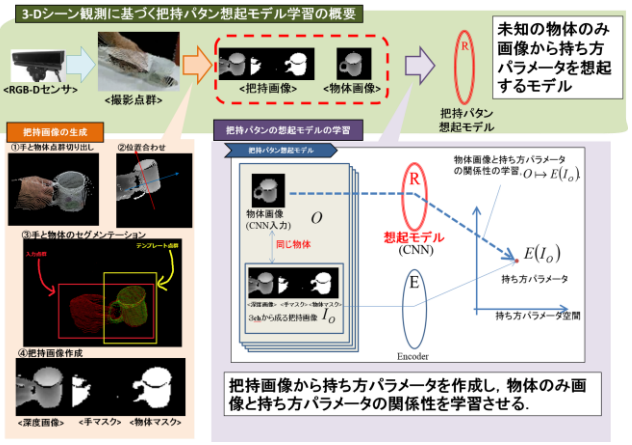

図 1　把持パタン想起モデル学習の概要

2. 3-D シーン観察に基づく把持画像の生成

2.1 RGB-D センサによる把持画像の撮影

川本らの室内ロギングシステム[3]を利用すると仮定して，机などに置いてある物体を把持し，持ち上げるというシーンから点群の時系列情報を撮影する．今回使用したセンサはマイクロソフト社の Kinect v2 センサである．センサ

の位置や角度は固定したままで，センサからの距離 90cm の位置に物体を置き撮影を行った．撮影の手順は図 2 の『撮影手順』に示す通り，初期フレームには物体のみが映っていると仮定し，その後物体を把持して持ち上げるといった手順である．

2.2 ICP による位置合わせ

撮影画像から把持画像を作成する手順を図 2 に示す．①に示す通り，撮影した点群に 3 次元トリミングと平面除去を行い，その点群から手と物体の点群のみを抽出する．次に，その点群に対し Iterative Closest Point(ICP)アルゴリズムを用いて初期フレームの物体点群と重なるように位置合わせを行う．

ICP アルゴリズムというのは，②に示す通り，ある空間上の点群 A を同一空間上の点群 B とできるだけ重なるような変換を行う変換行列を求めるアルゴリズムである．本手法では，初期フレーム以外の点群に対し，一つ前のフレームの点群と重なるような変換行列を ICP で求め，その行列を使い，全点群を初期フレームの物体位置に重なるような位置合わせを行う．その後，③に示す通り，位置合わせをした点群から初期フレームの点群と重なっている点を物体点群とし，それ以外の点を手の点群とする．その点群を用いて，④に示す通り，深度画像，手のマスク画像，物体のマスク画像の 64×64×3ch からなる把持画像を作成する．学習時にはこの画像の中心から 32×32 を切り出して使用している．深度画像は，センサからの距離 85cm～105cm を 0~255 の値にスケール変換しており，点が投影されていないピクセルに関しては値を 0 としている．各マスク画像は 0 と 255 の二値画像としている．

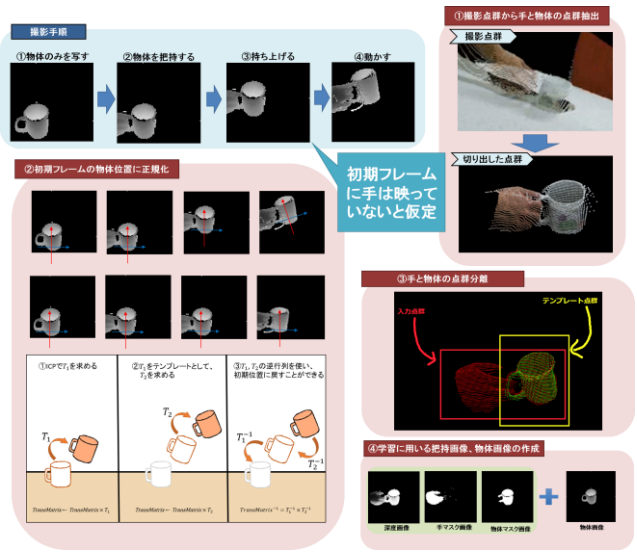

図 2　学習に用いる把持画像，物体画像の作成

3. 機械学習を用いた把持パタンの想起

物体のみ画像から，把持パタンを表すパラメータである持ち方パラメータの想起を行う．把持パタンを想起するモデルは，松尾らの AutoEncoder(AE)と Convolutional Neural Network(CNN)を組み合わせたモデルを使用する[4].

3.1 AutoEncoder による持ち方パラメータ空間の学習

Auto Encoder を用いて把持画像から 30 次元の持ち方パラメータが写像される空間を学習する．AE は，教師なしのニューラルネットで今回使用するモデルは中間層のユニット数が入出力層より少ないボトルネック型のネットワークである．このようなネットワークには入力を低次元化する機能があり，入力画像をより抽象度の高い情報にすることができる．これにより，物体の細かい形状を無視し，汎化性能が上がることを期待する．

この AE の入力層から中間層までを Encoder 部，中間層から出力層までを Decoder 部と呼び，本稿では持ち方パラメータの作成に Encoder 部を使い，持ち方パラメータから把持画像の復元に Decoder 部を利用している．

AE のモデル構造は，encoder 部が，

1. 畳み込み(32×32×3→24×24×16×3)
2. Tanh
3. プーリング(24×24×3×16→12×12×3×16)
4. Tanh
5. Reshape(一次元配列へ直す)
6. 線形結合(6912→1500)
7. Tanh
8. 線形結合(1500→150)
9. Tanh
10. 線形結合(150→30)
11. Tanh

Decoder 部が，

1. 線形結合(30→150)
2. 線形結合(150→1500)
3. 線形結合(1500→3072)

となっている．

学習に使用する画像は画素を-1～1 に正規化している．また，プーリングには Max プーリングを採用している．

3.2 CNN による物体画像からの持ち方パラメータの想起

次に学習済み AE の学習結果である持ち方パラメータを教師とし，CNN で物体のみ画像との関係を学習させる．CNN は関係性を学習するためのニューラルネットで，主に画像を入力とした場合に用いる．図 4 に示す通り，入力は 32×32 の物体画像で，出力層は教師の次元数と同じ 30 ノードとする．

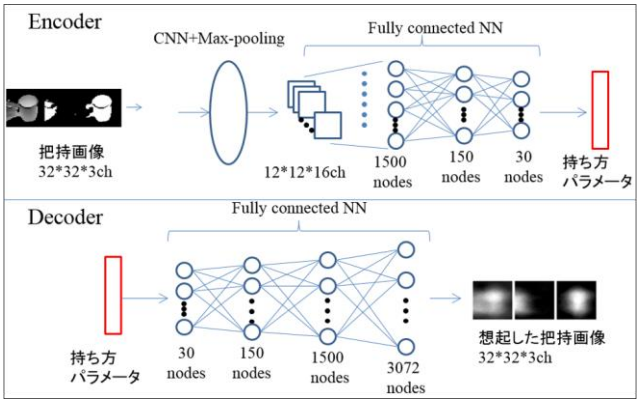

図 3　AutoEncoder モデル構造

モデル構造は,

1. 畳み込み(フィルタ：5×5×16 枚)
2. Tanh
3. プーリング
4. 正規化
5. 畳み込み
6. Tanh
7. プーリング
8. 正規化
9. Reshape
10. 線形結合(6400→128)
11. Tanh
12. 線形結合(128→30)
13. Tanh

となっている.

また，プーリングには L2 プーリングを採用している.

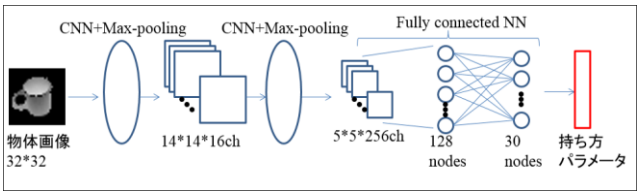

図 4　CNN モデル構造

3.3　把持画像の想起

本稿では, この AE と CNN を組み合わせたモデルを把持パタン想起モデルと呼び，このモデルを用いて把持パタンを想起する．また，想起された持ち方パラメータを AE の Decoder に入力することにより，把持画像の想起も行う.

4.　把持パタンの想起結果

4.1　把持パタン想起モデルの学習

使用した物体は，図 5 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 カテゴリの物体である．カテゴリごとに 1 種類の物体を用意し，把持画像をそれぞれ 4 0 枚作成し，把持パタンの想起モデルの学習を行った．まず, 2 章で作成した把持画像を用いて Auto Encoder の学習を行う．学習済みの Encoder を用いて把持画像から持ち方パラメータを抽出し, CNN の学習の教師として使用する．最後に CNN で教師である持ち方パラメータと物体のみ画像の関係性を学習させる.

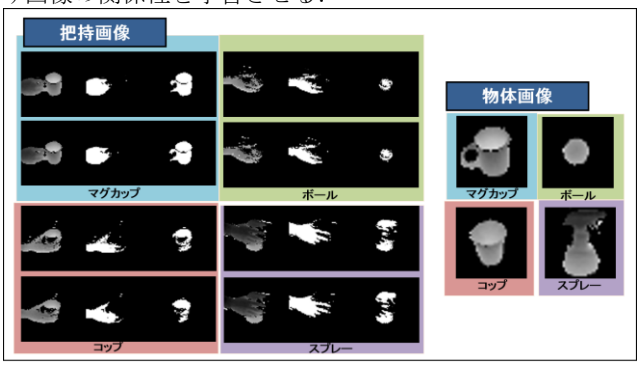

図 5　学習に使用した画像

4.2　把持パタン想起モデルの学習結果

学習済の把持パタン想起モデルの想起結果の分布を図 6 に示す．入力物体のカテゴリごとに色分けしており，持ち方パラメータの第一主成分，第二主成分の二軸でグラフ化している．また，グラフから，同じカテゴリの物体は同じような位置に分布されていることが分かる.

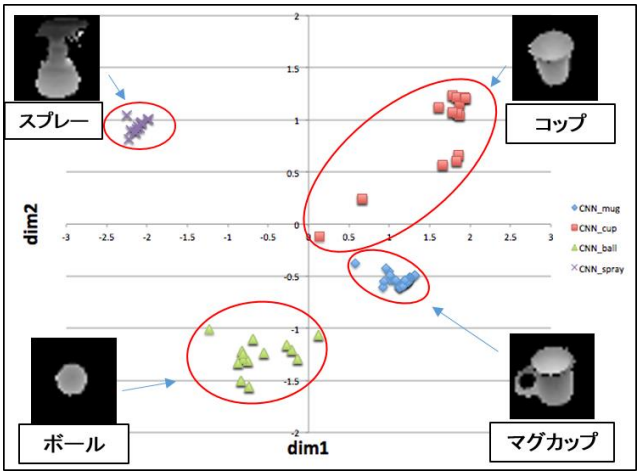

図 6　既知の画像を用いた把持パタン想起結果

4.3　未知の物体画像を用いた把持パタンの想起

学習済みの把持パタン想起モデルに学習に用いていない物体画像を入力し，その出力結果をグラフ化した．図 7 に想起した持ち方パラメータの分布を示す．図 6 と同様に入力物体のカテゴリごとに色分けしており，持ち方パラメータの第一主成分, 第二主成分の二軸でグラフ化している．コップとボールに関しては，既知の画像を用いた想起結果よりもまとまりがあるように見える．しかし，持ち方パラメータの分布を見ると，既知の画像を用いた際の分布と同じような傾向にあることが分かる.

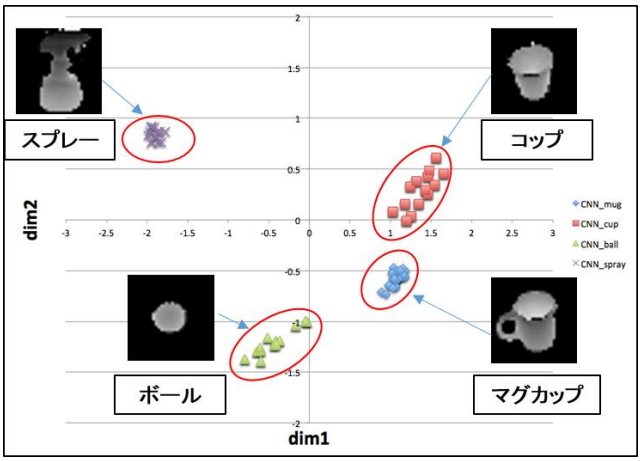

図 7　未知の画像を用いた把持パタン想起結果

4.4 把持画像の復元

前節で想起した持ち方パラメータを用いて把持画像の想起を行う．学習済み AE の Decoder 部に想起した持ち方パラメータを入力し，把持画像の復元を行った．図 7 にその結果を示す．図の左に把持パタン想起モデルに入力した物体画像，中央に想起した持ち方パラメータから復元した把持画像，右に同一カテゴリの物体把持画像を配置している．想起画像の手マスク画像を見ると，カテゴリごとに異なった手の形をしている事が分かる．手の位置も実際の把持画像の手マスクと同じようなパターンで復元されている．ただ，32×32×3 の画像を 30 次元まで圧縮しているため，全体的にぼやけているように見える．

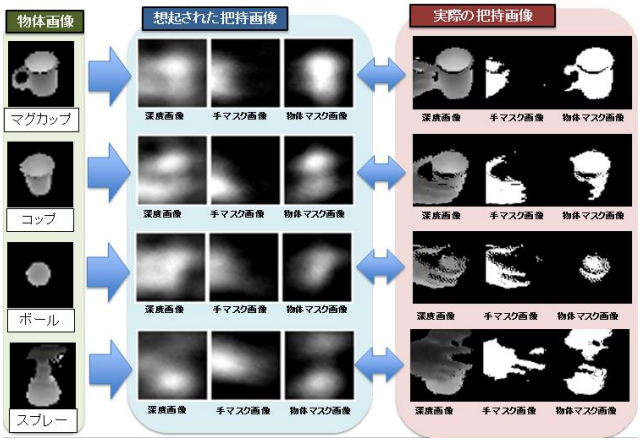

図 7　把持画像の復元

5. 今後の課題

今後は物体をどのような方向から撮影しても同じような持ち方パラメータを抽出するために，様々な角度の物体画像を学習に組み込む必要がある．

その上で，学習に使う画像を日常的なシーンから自動で収集し，未知のカテゴリの物体に対する持ち方の想起を行う予定である．

6. 参考文献

[1]鎌倉，“手の形 手の動き”，医歯薬出版株式会社，1989.
[2]北橋ほか，“動作と物体の統合的認識とそのモデル化”，情報処理学会研究報告．CVIM， 88(2005-CVIM-150), pp.109-116，2005.
[3]川本ほか，“階層型イベント検知に基づく人と物の関わりのロギングシステム”，第 18 回画像の認識・理解シンポジウム，SS5-37，2015.
[4]Matsuo et.al，“Extraction of Descripter of Hand-Object interaction”, 第 18 回画像の認識･理解シンポジウム, OS1-4, 2015.

7. 謝辞

本研究の一部は文部科学省私立大学戦略的研究基盤形成支援事業（平成 25 年～平成 28 年, S1311039 ）により実施しました．本研究は JSPS 科研費 24500224, 15H02764 の助成を受けたものです．