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

RGB-D画像からの把持パタン想起に基づくハンドアームによる物体把持

矢野 将基[1,a)]　福原 宏弥[1]　松尾 直志[1]　島田 伸敬[1]

概要：物体が写った画像から人間がその物体を把持している様子を想起し、想起結果から把持位置・姿勢を獲得して物体把持を行う手法を提案する。物体のみが写った様子と、その物体を人間が把持している様子のペアを用いて、人間がその物体を把持する際の手と物体の関係性を学習しておく。そして、形状の似た未知の物体が写った画像が与えられた際に、人間がその物体のどの部分を、どのように把持するかを想起する。これにより、人間がその物体を把持する際、物体のどの部分に、どのような手の姿勢で触れるかが分かるため、その情報を基にロボットの手先の目標位置と向きを求め、ハンドアームロボットによって物体把持を行う。

キーワード：ハンドアームロボット, CNN(Convolutional Neural Network), 物体把持

Grasping an Object by a Hand-Arm Robot Based on Human Interaction Recalled from RGB-D Image

Masaki Yano[1,a)]　Hiroya Fukuhara[1]　Tadashi Matsuo[1]　Nobutaka Shimada[1]

Abstract: We propose a method that enables a robot to grasp on object based on how a human grasps it. By observing interactions by humans, we model the relationship between a shape of an object and how to grasp it. In advance, the model is trained with pairs of images before/after a human grasps an object. We can train the model without labeling interaction between an object and a hand. With this model, a robot can recall how a human grasps an object from an appearance of the object. The robot can grasp the object by moving its hand to a point of the object where a human touches for grasp it. By experiments for actual objects, we show availability of proposed technique.

Keywords: Hand-Arm Robot, CNN(Convolutional Neural Network), Grasping an Object

1. はじめに

近年、カメラやセンサの高性能、低価格化が進んでおり、高性能なセンサをロボットに搭載することが可能となっている。ロボットは一般家庭など、複雑な環境下にも活躍の幅を広げていくと考えられる。しかし、ロボットの動作一つ一つのソフトウェアを人間が開発するのは開発者への負担が大きくなるため困難である。そこで、人間が道具を操作するシーンをロボットが観察し、操作を自動的に学習することができれば、この問題の解決につながると考えられる。

本研究では、ある物体のみが写った画像と人間がその物体を把持した画像をセットで学習する。そして、未知の物体が写った画像が与えられた際、人間がその物体のどの部分を、どの方向から掴みにいくかを想起する。想起結果からハンドアームロボットの手先の目標位置と向きを求めて物体把持を行う。**図 1** に提案する手法の概要を示す。

本研究では、Rethink Robotics 社が開発している 7 自由度の双腕ロボット、Baxter を使用する。Baxter の手先には本研究室で製作したロボットハンドを装着し、Baxter 頭

[1] 立命館大学
Ritsumeikan University
[a)] yano@i.ci.ritsumei.ac.jp

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

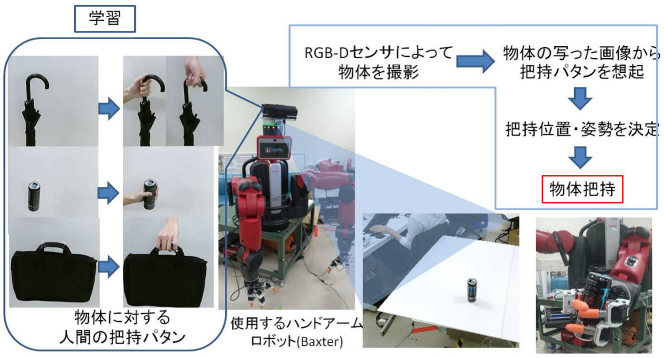

図 1　提案する手法の概要

部には RGB-D センサとして Kinect V2 を設置する。

2. RGB-D 画像からの把持パタンの想起

鎌倉 [1] によると、人間は物体の使用目的に応じて把持姿勢を変えているといい、その種類は限定されているという。これより、同じ使用目的の物体ならばほぼ同じ把持姿勢になると考えられる。

本研究における把持パタンを、「ある物体を把持する際、物体のどの部分を、どのように掴むか」と定義する。そして、物体のみが映った画像からその物体に対する把持パタンを推定することを把持パタンの想起とする。

松尾ら [2] はある物体の典型的な把持パタンを学習することで、未知の物体に対する把持パタンを想起する手法を提案している。この手法では、CAE(Convolutional Auto-Encoder)[4] を用いて、ある物体を把持している際の、手と物体の相互作用を 30 次元の持ち方パラメータで表現している。また、Decoder を用いて持ち方パラメータから物体を把持している様子のテクスチャ画像、手領域のマスク画像、物体領域のマスク画像を復元する。物体のみの画像と、その物体を人間が把持している画像の 2 枚の組をセットにして CAE と Decoder の学習を行う。学習には CNN(Convolutional Neural Network)[3] を用いる。

把持パタン想起では、入力された濃淡画像の各位置ごとに 32 × 32 のウィンドウ (パッチ) を生成する。Encoder を用いてパッチごとの持ち方パラメータと手と物体間での相互作用確率の想起を行う (**図 2**)。そして、Decoder を用いて持ち方パラメータから物体を把持している様子のテクスチャ画像、手領域のマスク画像、物体領域のマスク画像を復元する (**図 3**)。

本研究ではこの手法を用いて物体が写った画像からその物体に対する把持パタンの想起を行い、想起結果のうち手領域マスク画像と物体領域マスク画像を使用して手先の目標位置と物体の把持位置を推定し、物体把持を行う。

3. 想起のための手と物体の関係性の学習と学習結果の評価

学習に使用する画像は、**図 4** に示した通りの持ち方で、

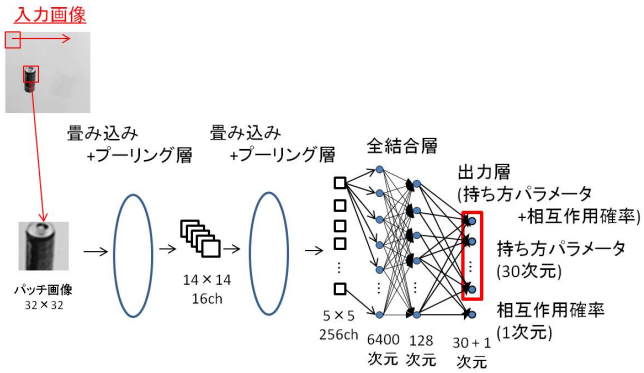

図 2　Encoder を用いたパッチごとの持ち方パラメータと相互作用確率の想起

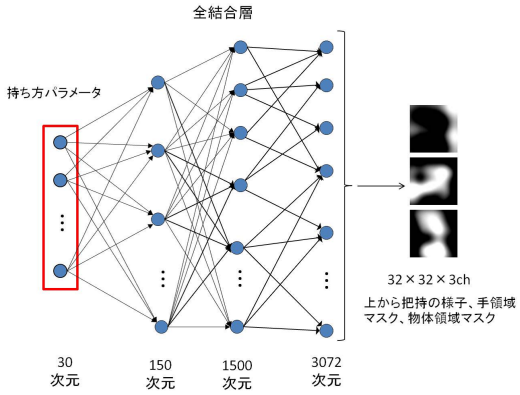

図 3　Decoder を用いた持ち方パラメータからの 3 次元画像の復元

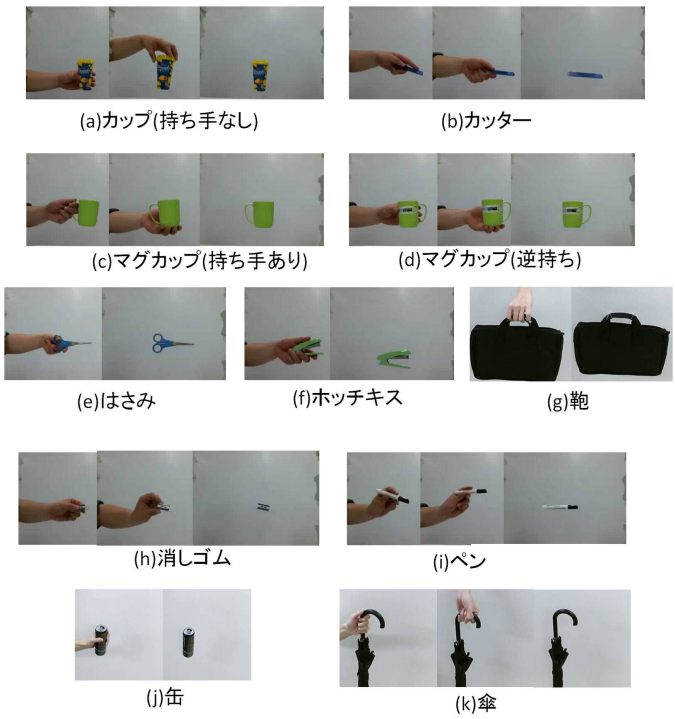

図 4　学習に使用する物体と把持パタン

持ち方の数は計 18 通りである。各持ち方に対する画像の枚数は 120 枚であり、120 × 18=2160 枚の画像を使用して学習を行う。

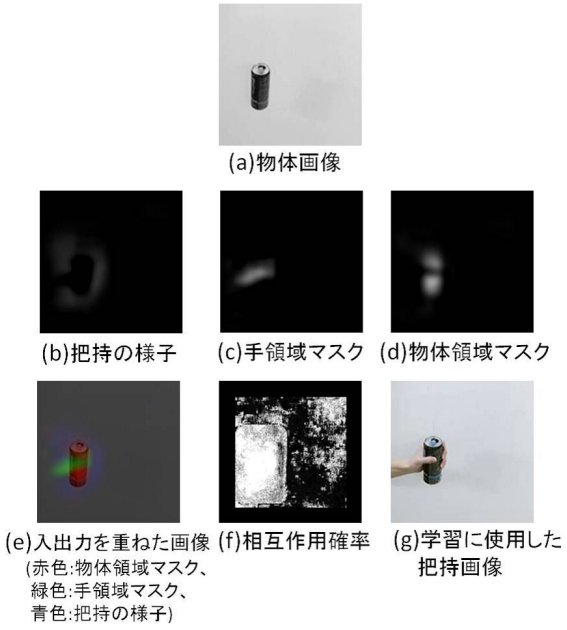

図 5 学習済み物体に対する把持パタン想起結果

図 5 は学習に使用した缶 (図 4 の (j)) について把持パタンの想起を行った結果である。ここで、図 5(b) はパッチごとの想起結果のうち把持の様子のテクスチャ、図 5(c) は手領域のマスク、図 5(d) は物体領域のマスクを、相互作用確率を重みとして足し合わせた結果の画像である。また、図 5(e) の入出力を重ね合わせた画像は、青色部分は把持の様子を表すテクスチャ、緑色部分は手領域のマスク、赤色部分は物体領域のマスクを表している。図 5(f) の相互作用確率は、入力画像中の各位置におけるウィンドウにおける相互作用の確率を、ウィンドウの中心にあたる位置座標にプロットした画像である。例えば、入力画像中の $(1, 1)$ から $(32, 32)$ の領域のパッチに対する相互作用確率は、確率マップの $(17, 17)$ にプロットしている。ここで、確率マップの左、上側 15pixel と右・下側 16pixel の部分を中心とするパッチを作成することはできないため、その部分の相互作用確率を 0.0 で埋めている。これは、入力画像と確率マップの画像サイズを合わせるためである。入出力重ね合わせ画像 (図 5(e)) と学習に使用した把持画像 (図 5(g)) を比較すると、把持画像における缶の領域は赤色、手領域は緑色、その周辺は青色に塗り分けられている。しかし、手首位置より左の部分は緑色に塗られていない。その理由として、今回は 32×32 のウィンドウ毎に想起を行っており、手首付近のウィンドウに缶が入らなかったためであると考えられる。現に、図 5(f) に示した相互作用の確率マップでは缶周辺部分以外は確率が低くなっていることが確認できる。このことから、学習済み画像について概ね期待通りの結果が得られていると考えられる。

図 6 に図 5 と同じ種類の、未学習の物体に対する把持パタン想起結果を示す。図 6(e) では、物体の領域が赤色、物体の左側に緑色、周辺が青色にくっきりと塗り分けられて

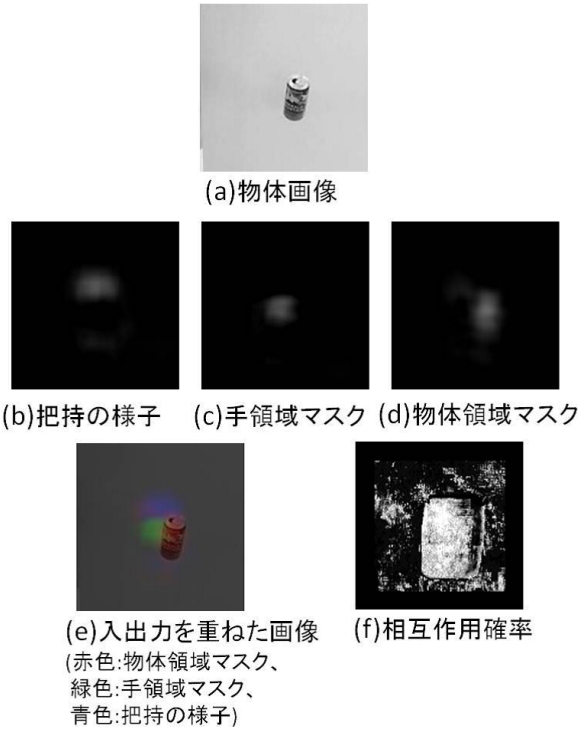

図 6 未学習物体に対する把持パタン想起結果

いる。これは学習に使用した把持画像と同様、缶の左側から中心付近を把持することを示している。また、図 6(f) に示す相互作用の確率マップでは、物体周辺の確率が高く、離れた点は低くなっている。これらのことから、学習済み物体と同種類の未学習画像についても期待通りの結果が得られていると考えられる。

4. 想起結果に基づく把持動作の実装

4.1 RGB-D 画像からの把持パタン想起

Kinect V2 によって取得したカラー画像には背景も含まれているが、今回は手先の可動範囲外の領域については考慮する必要がない。そこで、Kinect V2 によって取得した深度情報を使用し、想起領域の抽出を行う。本研究では距離が 1.5m 以内の領域に対してラベリング処理を行い、領域数が最大となる領域に対して想起を行う。

想起結果からロボット座標系における手先の目標位置と物体の中心位置を取得する。まず、想起結果画像のうち手先領域マスク画像と物体領域マスク画像に対して、しきい値を基に 2 値化処理を行う。そして、2 値化画像に対してラベリング処理を行い、領域数が最大となるラベルの重心を手先の目標位置・物体の中心位置の座標とする。

次に、求めた座標を画像座標系からカメラ座標系に変換し、さらにロボット座標系に変換する。画像座標系からカメラ座標系への変換には本研究室内で製作した Kinect 座標系変換サーバを使用する。座標系変換サーバでは Kinect SDK に用意されている CoordinateMapper クラスのメンバ関数である、MapColorFrameToCameraSpace 関数を使

用して座標系の変換を行っている。カメラ座標系からロボット座標系への変換は、キャリブレーションによって変換行列を作成し、この変換行列によって行う。カメラ座標系のある点 (X_{ci}, Y_{ci}, Z_{ci}) と、それに対応するロボット座標系の点 (X_{ri}, Y_{ri}, Z_{ri}) の組を 10 組程度を取得し、疑似逆行列を計算することで変換行列を求めることができる。

4.2 想起結果に基づく手先移動と把持

物体把持のために、ロボットアーム・ロボットハンドの各関節の目標関節角度を計算する。まず、ロボットアームであるが、逆運動学を計算して手先位置・姿勢から関節角度を計算するためのモジュールが Baxter の製作会社によって用意されているため、これを使用する。

手先の目標位置は 4.1 で述べた手先の目標位置をそのまま使用する。手先姿勢は**図 7** に示す通り、手先の目標位置から物体中心位置に向かうベクトルと手先の法線ベクトルが一致する姿勢とする。

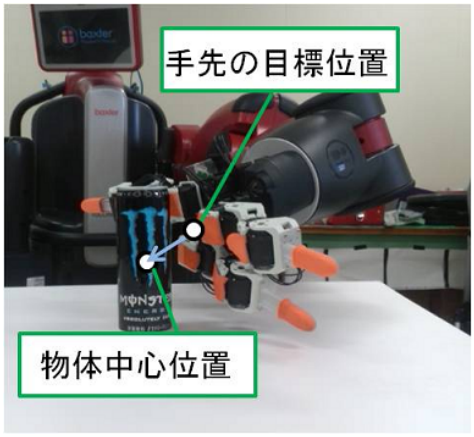

図 7 手先の目標姿勢

ロボットハンドの姿勢は、缶を把持した状態 (**図 8**) の各関節の関節角度をあらかじめ記録しておく。そして、手先が目標に移動した後にその角度を再現することで握る動作を行う。

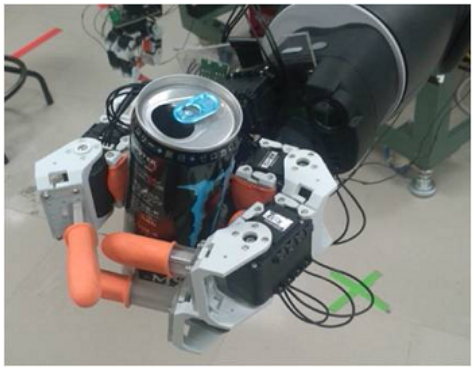

図 8 缶を把持した状態

5. 物体把持実験

これまでに述べた手法により、実際にロボットが物体を把持できるかどうかを検証する。

5.1 実験環境

今回は、**図 9** に示す缶とカバンに対して把持実験を行う。これらはいずれも学習に使用したものである。

(a) 実験に使用する缶　(b) 実験に使用するカバン

図 9 実験に使用する物体

また、実験に使用する物体は**図 10** に示す通り、いずれも高さ 70cm の台上、70cm 先の位置に立てて配置した。

図 10 実験環境

5.2 缶を把持する場合の実験結果

5.2.1 RGB-D 画像からの把持パタン想起

図 11 はロボット頭部に設置した Kinect V2 によって取得したカラー画像である。Kinect V2 によって取得した深度情報を使用し、図 11 のうち、距離が 1.5m 以内の領域を抽出した画像を生成し、その画像に対して把持パタンの想起を行う。

サーバへの入出力画像を重ね合わせた結果を**図 12** に示す。赤色が物体領域、緑色が手領域を表すマスクである。先述の通りウィンドウごとの想起結果を確率に基づいて足し合わせているため、色が濃い部分ほどそれぞれの領域である確率が高いことを示している。缶の上部を中心に物体領域、その左に手領域を表すマスクが示されている。

想起結果の手領域と物体領域マスク画像に関して、しきい値を基に 2 値化処理を行う。2 値化画像を**図 13** に示す。

図 11　Kinect V2 により取得したカラー画像

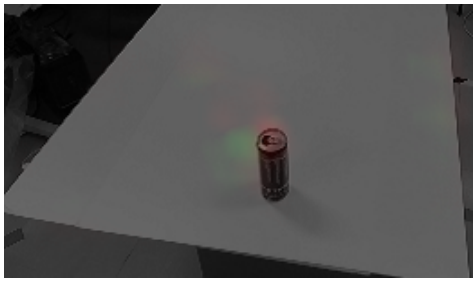

図 12　入力画像と想起結果を重ね合わせた様子; 赤色:物体領域マスク、緑色:手領域マスク

本実験では手領域マスクのしきい値を 0.10、物体領域マスクのしきい値を 0.15 に設定した。2 値化処理によって得られた領域に対してラベリング処理を行い、領域数最大ラベルの重心位置を求めて手先の目標位置・物体中心位置とした。図 13 より、缶の上部に対して物体領域、その左側に示されている。そのため、缶の上部を左側から掴みに行くことで把持できると考えられる。

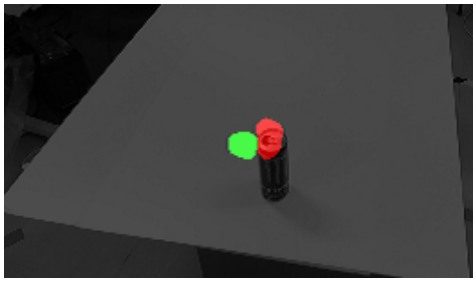

図 13　想起結果のうち、物体領域・手領域に関してしきい値を超えた領域; 赤色:物体領域マスク、緑色:手領域マスク

5.2.2　想起結果に基づく手先移動と把持

図 14 に把持動作開始時からの動作の様子を示す。図 13 で示した通り、缶の上部を把持している。今回の実験では、開始位置から目標位置に移動するまでが約 9 秒、目標位置に移動してから把持動作を行うまでが約 1.3 秒であった。13 秒後の結果が示す通り、缶を倒すことなく把持することができた。

5.3　カバンを把持する場合の実験結果

5.3.1　RGB-D 画像からの把持パタン想起

図 15 は Kinect V2 によって取得したカラー画像である。カバンについても同様に、図 15 のうち、距離が 1.5m 以内の領域を抽出した画像を生成し、把持パタンの想起を行う。

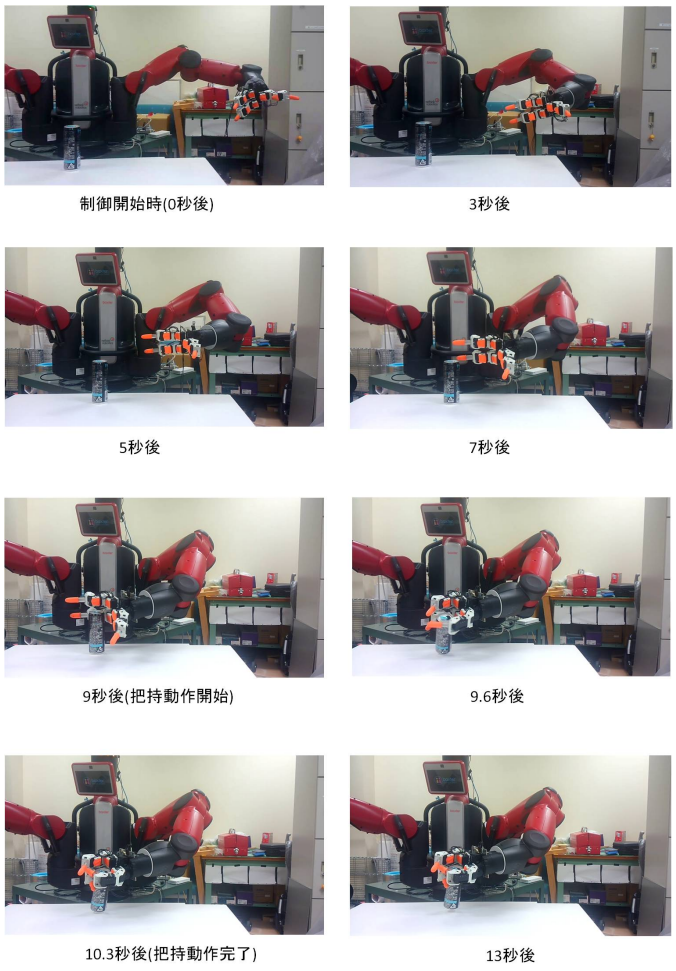

図 14　缶を把持する様子

図 15　Kinect V2 により取得したカラー画像

サーバへの入出力画像を重ね合わせた結果を図 16 に示す。物体領域はカバンの左端上部と中央から右端上部、手領域は持ち手の上部にマスクが強く示されている。

想起結果の手領域と物体領域マスク画像に関して、しきい値を基に 2 値化処理を行う。2 値化画像を図 17 に示す。本実験では手領域マスクのしきい値を 0.10、物体領域マスクのしきい値を 0.15 に設定した。図 17 より、物体領域はカバンの左端上部と中央から右端上部、手領域は持ち手の上部に現れた。物体位置は領域数が大きい方の重心位置と

図 16　入力画像と想起結果を重ね合わせた様子; 赤色:物体領域マスク、緑色:手領域マスク

するため、中央から右端上部の物体領域の重心位置を物体の中心としている。

図 17　想起結果のうち、物体領域・手領域に関してしきい値を超えた領域; 赤色:物体領域マスク、緑色:手領域マスク

5.3.2　想起結果に基づく手先移動と把持

図 18 に把持動作開始時からの様子を示す。今回の実験では、開始位置から目標位置に移動するまでが約 7.6 秒、目標位置に移動してから把持動作を行うまでが約 2.2 秒であった。また、カバンは把持後も机上に接着していたため、把持できているかどうかを確認するために腕を上げるための制御コマンドを手動で送信し、確認を行った。確認のための制御コマンドは制御開始から約 19 秒後に送信した。図 18 が示す通り、持ち上げた後も、カバンを落とすことなく把持することができた。

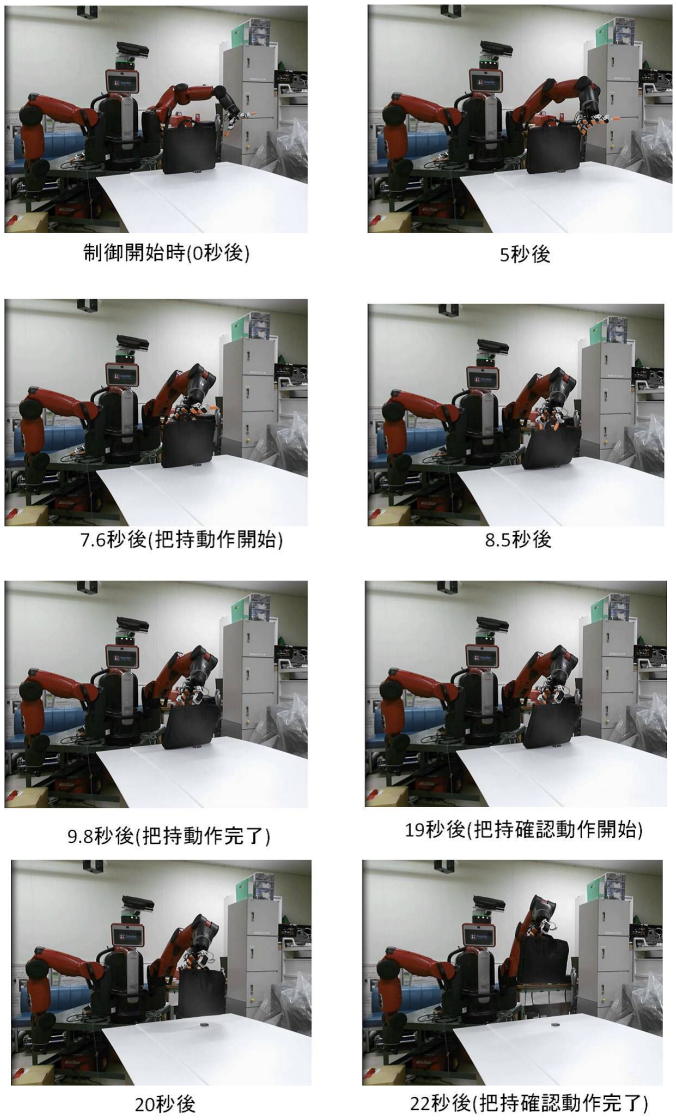

図 18　カバンを把持する様子

6.　おわりに

本研究では RGB-D 画像から把持パタン想起を行い、想起情報を用いて物体把持を行う手法を提案した。そして缶について把持実験を行い、手法の有用性を検証した。実験では、缶の下部において物体・手領域マスクが期待通りに示されなかった。原因としては机上に置いた物体の見え方と学習時の見え方が異なっていたことが挙げられる。そのため、異なった見え方の学習画像を追加して学習を行うことで解決できると考えられる。

今後の課題として、3 次元把持パタン想起への拡張が挙げられる。人間の 3 次元把持パタンを想起するために川上ら [5] の手法を用いることを検討している。現在、ロボットハンドによる握り方はあらかじめ作成した姿勢を再現しているため、どの物体に対しても同じ握り方で把持している。把持パタン想起によって人間の 3 次元手形状を想起することができるようになると、物体ごとに適した把持姿勢を獲得できるようになり、物体ごとに握り方を指示する必要がなくなる。

また、CNN を用いて人間の 3 次元手形状の深度情報とロボットハンドの関節角度の対応関係を学習し、3 次元手形状の深度情報からロボットハンドの関節角度を獲得できるようにする。これにより、3 次元手形状の想起結果から対応するロボットハンドの姿勢を生成できるようになる。

謝辞　本研究は JSPS 科研費 24500224, 15H02764 の助成を受けたものです。

参考文献

[1] Noriko Kamakura: ”手のかたち手のうごき”, 医歯薬出版株式会社,1989.
[2] Tadashi Matsuo, Nobutaka Shimada: ”Construction of Latent Descriptor Space of Hand-Object Interaction”,The 22nd Joint Workshop on Frontiers of Computer Vision (FCV2016), pp. 117-122, 2016.

[3] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner: ”Gradient-based learning applied to document recognition.” Proceedings of the IEEE 86.11 (1998): pp. 2278-2324.
[4] J. Masci, U. Meier, D. Cirean, and J. Schmidhuber: ”Stacked convolutional auto-encoders for hierarchical feature extraction.”, Artificial Neural Networks and Machine LearningICANN 2011. Springer Berlin Heidelberg, 2011. pp. 52-59.
[5] 川上 拓也, 松尾 直志, 小川 陽子, 島田 伸敬: ”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” , コンピュータビジョンとイメージメディア研究会 (CVIM), 2016 (発表予定)