Heterogeneous Learning によるオブジェクトネスと物体把持位置の検出

長谷川 昂宏† Xuanyi Sheing† 荒木 諒介† 山内 悠嗣† 山下 隆義† 藤吉 弘亘†
†中部大学
E-mail: hf@cs.chubu.ac.jp

Abstract

本研究では，Deep Convolutional Neural Network を用いたピッキングロボットのための物体把持位置検出法を提案する．従来，学習アプローチによる物体把持位置検出法として，2 段階の Deep Neural Network を用いた手法が提案されている．1 段目のネットワークにより物体から各方向毎にラスタスキャンして複数の把持位置候補を検出する．2 段目のネットワークでは，複数の把持位置候補から 1 つに絞り込むことで物体の把持位置を検出する．しかし，物体の把持位置を検出するために 2 つの Deep Neural Network を複数回ラスタスキャンして使用するため非効率という問題がある．そこで，本研究では畳み込み層を用いた Deep Convolutional Neural Network により画像中の物体特徴を自動的に捉え，1 度のラスタスキャンで最適な把持位置を効率的に検出する．さらに提案手法では，Heterogeneous Learning として全結合層の出力ユニットにオブジェクトネスユニットと把持座標点ユニットを割り当てる．これにより，入力画像の物体らしさの識別と把持座標点の推定を同時に解くことができる．評価実験により，提案手法は従来の把持位置検出法と同等以上の精度で効率的に把持位置を検出することを確認した．

1 はじめに

産業用ロボットや生活支援ロボットにおいて必要とされているタスクはロボットが対象物体 (工業部品や日用品) を正確に把持することである．ロボットから物体を把持するには，対象物体を撮影した画像から自動的に物体の最適な把持位置を検出する必要がある．これはロボットシステムにおいて重要な前処理であり，基本的なタスクとなるため可能な限り計算コストを抑えて効率化しなければならない．

ピッキングロボットを対象とした物体把持位置検出法として，これまでに多くの手法が提案されている [1, 2, 3, 4, 5, 6]．これらの手法には機械学習ベースの手法とテンプレートベースの手法に分けられる．機械学習ベースの代表的な手法として 2 段階の Deep Neural Network (DNN) を用いた把持位置検出法 [7] がある．これは，4 点の把持位置を結んだ矩形領域でロボットの把持位置を検出する．まず，1 段目の DNN で様々な把持矩形領域で画像をラスタスキャンする．そして，1 つの物体に対して複数の把持位置候補を検出する．その後，2 段目の DNN により複数の把持位置候補から最適な把持位置を検出する．この手法では，1 つの物体の把持位置を検出するために 2 段階の DNN で処理しなければならない．また，様々な把持矩形を考慮するため，矩形のサイズと方向を変化させ，8000 回以上のラスタスキャンが必要となる．そのため，2 段階の DNN を用いた把持位置検出法は計算コストが高く非効率である．

一方，テンプレートベースの代表的な手法として Fast Graspability Evaluation[8] が挙げられる．Fast Graspability Evaluation は機械学習を使わずにハンドモデルと物体領域の 2 値パターン画像の単純な畳み込みにより把持位置検出を実現している．これは，ロボットハンドが物体に接触する領域と衝突する領域の 2 種類を 2 値パターン画像のテンプレートとしてあらかじめ保持しておく．そして，対象物体を抽出した 2 値画像に対して接触領域と衝突領域のテンプレートを畳み込むことで把持可能性 (Graspability) を算出する．Fast Graspability Evaluation もまたロボットハンドモデルの様々な方向やハンド開き幅を考慮したテンプレートを作成し，それら全てを物体領域に畳み込む必要がある．

本研究では Deep Convolutional Neural Network (DCNN) を用いた物体の把持位置検出法を提案する．DCNN は畳み込み層と全結合層から構成されるネットワークモデルである．畳み込み層を用いることにより，入力画像から自動的に物体の画像特徴を捉えることができるため，物体の最適な把持位置を 1 度のラスタスキャンで推定することができる．また，全結合層の出力には物体らしさ (オブジェクトネス) を識別するユニットと把持座標点を推定するユニットを割り当てる．このように出力ユニットに複数の異なるタスクを割り当てた学習方法は Heterogeneous Learning と呼ばれ，1 つの DCNN で複数の異なるタスクを同時に解くことが可能となる．提案手法ではラスタスキャンによる探索を 1 度のみ行い，オブジェクトネスユニットの出力値を用いて最適な把持位置を推定する．そのため，非常

に効率的な把持位置推定が可能となる．

1.1 関連研究

本章では，従来の把持位置検出法について述べる．把持位置検出法は，把持位置を教師データとして学習する機械学習ベースの手法とロボットのハンドモデルや物体の 3 次元モデルをテンプレートとして使用するテンプレートベースの手法に分類できる．以下に機械学習ベースの手法とテンプレートベースの手法について説明する．

機械学習ベースの手法

機械学習をベースとする把持位置検出の手法は，学習用画像データセットにあらかじめ最適な把持位置を教師データとして与えることで物体の把持位置を学習する [9, 10]．Jiang らは把持位置を矩形領域で表現することでロボットハンドの回転に加え，ハンドの開き幅も機械学習により推定した [11]．そして，2 段階の把持位置検出を構築することでより良い把持位置の検出を実現した．Lenz らは Jiang らの 2 段階把持位置検出に Deep learning を導入した，2 段階の DNN による把持位置検出を提案にした [7]．Deep learning[12] は幅広いタスクで高い性能を達成しているため，様々な研究に用いられている [13, 14]．2 段階の DNN を用いた把持位置検出手法は，1 段目に Small Neural Network を使用して 1 つの物体に対して複数の把持位置候補を検出する．2 段目では，1 段目のネットワークよりもユニット数を多くした Larger Neural Network を用いて把持位置候補を 1 つに絞り込むことにより把持位置を検出する．図 1 に Lenz らの 2 段階の DNN を用いた把持位置検出の例を示す．ユニット数の少ない Small Neural

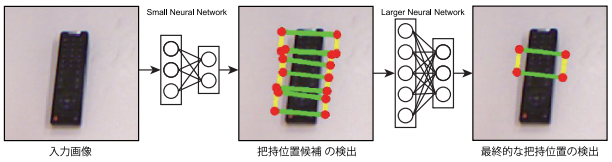


図 1　2 段階の DNN による把持位置検出．

Network を用いて画像全体から複数の把持位置候補を検出し，把持位置候補から Larger Neural Netwaork で把持位置を絞り込むことで計算コストを抑える工夫がされている．しかし，様々な矩形の方向やサイズを考慮するため，8000 回以上のラスタスキャンが必要となる．

テンプレートベースの手法

テンプレートベースの手法では物体の把持位置を検出するために，対象物体の 3 次元モデルやロボットハンドのモデルをテンプレートとして保持する．画像から得られた対象物体領域に対してテンプレートを当てはめることで，最適な把持位置を検出することができる．対象物体の 3 次元モデルをテンプレートとする場合，テンプレートを用いてポイントクラウドで表現された入力シーンに存在する物体の姿勢推定を行う．そして，推定した姿勢を基に最適な把持位置を決定する．対象物体の姿勢推定をするには円や円柱などの単純なモデルを用いて大まかな姿勢を近似する方法 [1, 2] や 2 点のオリエンテーションペアを用いて姿勢を推定する方法 [3, 4] がある．また，Iterative Closest Point (ICP) を用いて高精度に姿勢を推定する方法も提案されている [5, 6]．物体の 3 次元モデルを用いた手法は剛体の物体に対して非常に効果的である．

テンプレートベースの手法には，ロボットのハンドモデルをテンプレートする手法がある．ハンドモデルをテンプレートとすることで，物体の 3 次元モデルを必要とせず非剛体の物体の形状変化やオクルージョンが発生しても最適な把持位置を推定することができる．ロボットハンドモデルをテンプレートとした把持位置検出法として Fast Graspability Evaluation [8] が提案されている．Fast Graspability Evaluation はロボットハンドが物体に接触する領域と衝突する領域の 2 種類の 2 値パターンをテンプレートとして保持する．そして距離画像に対してセグメンテーションすることで，2 値化した物体領域を抽出する．物体領域に対して接触領域と衝突領域をそれぞれ畳み込んだ結果から Graspability マップを生成する．このとき，様々な回転角と開き幅のハンドモデルのテンプレートを畳み込み，Graspability マップをそれぞれ生成する．そして，Graspability マップがピークとなる位置とハンドパラメータを物体の把持位置とする．図 2 に Fast Graspability Evaluation の処理の流れを示す．Fast Graspability Evaluation は 3 次元

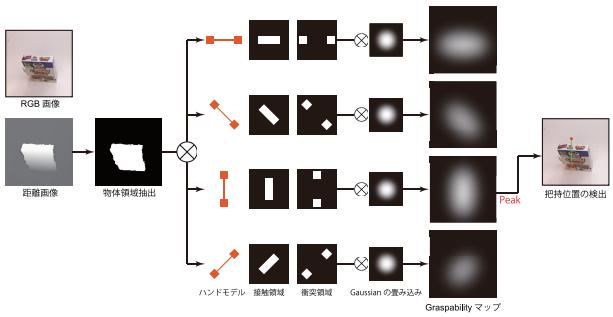


図 2　Fast Graspability Evaluation による把持位置検出．

物体モデルの姿勢推定が必要なく，高い再現性で把持位置を検出することができる．しかし，Fast Graspability Evaluation も様々な方向のハンドモデルを画像に畳み込まなければならない．

1.2 提案手法の概要

提案手法の特徴を以下に示す．

- Deep Convolutional Neural Network による把持位置検出

従来法は 2 段階の DNN を用いて把持位置を検出するため非効率である．提案手法は畳み込み層を用いた DCNN を用いることで入力画像の特徴を自動的に捉えて最適な把持位置を検出する．また，1 度のみの画像のラスタスキャンで物体把持位置を検出できるため効率的である．

- Heterogeneous Learning によるオブジェクトネスと把持位置の検出

提案手法では全結合層の出力ユニットに物体らしさを表すオブジェクトネスユニットと把持位置を推定する把持座標点ユニットを割り当てる．オブジェクトネスユニットを割り当てることにより，入力画像の物体らしさと把持位置を同時に求めることができる．オブジェクトネスユニットの出力値を用いて物体らしい位置の把持点を検出することができる．

2 提案手法

本研究では，Deep Convolutional Neural Network を用いた Heterogeneous Learning によるオブジェクトネスと把持位置検出を提案する．以下に提案手法の詳細を述べる．

2.1 回帰型 Deep Convolutional Neural Network

DCNN は畳み込み層とプーリング層を階層的に構成し，それら 2 つの層から得られた特徴マップを全結合層に入力する．提案手法で用いる DCNN は畳み込み層を 4 層，全結合層を 2 層とする．畳み込み層は，フィルタサイズ $n \times n$ の重みフィルタを畳み込み，そのレスポンス値 v を活性化関数 $f(v)$ に通す．その後，$f(v)$ を特徴マップとして格納する．各層の畳み込みフィルタは M 個使用し，それぞれのフィルタで特徴マップを生成する．活性化関数にはシグモイド関数，Rectified Linear Unit (ReLU)，Maxout が一般的に用いられる．本研究では活性化関数に ReLU を用いる．ReLU は式 (1) に示すように v が負の値となった場合に 0 を返し，正の値となった場合に v の値をそのまま返す関数である．

$$f(v) = \max(0, v) \tag{1}$$

活性化関数に ReLU を用いることにより，v が大きな値となった場合でも勾配を得ることができる．プーリング層では特徴マップを縮小させる処理を行う．プーリングには，Max Pooling や Average Pooling，L_p Pooling がある．本研究では Max Pooling を用いて特徴マップを縮小させる．Max Pooling は様々なプーリングの手法において性能が良いとされおり，あらかじめ決定した領域における最大値により間引きを行うことで特徴マップを縮小させる．このように，入力画像に畳み込みとプーリングを階層的に行うことで，画像の特徴を獲得する．畳み込みとプーリングにより獲得された特徴マップを 1 次元に変換して全結合層の入力とする．全結合層では式 (2) に示すように重み付きの全結合を計算する．

$$h_i(\boldsymbol{v}) = f\left(\sum_{j=1}^{N} w_{ij} v_j + b_i\right) \tag{2}$$

全結合層においても畳み込み層と同様に活性関数 $f(\cdot)$ を適用して出力値 $h_i(\boldsymbol{v})$ を獲得する．回帰型 DCNN では，全結合層の出力ユニットに回帰で求めたい x 座標と y 座標を割り当てる．本研究では，2 点の把持位置と 4 点の把持位置の座標を回帰で求める．よって，2 点の把持位置を求める場合は出力ユニットに 4 個の回帰ユニットが割り当てられ，4 点の把持位置を求める場合は 8 個の回帰ユニットが割り当てられる．

2.2 Heterogeneous Learning によるオブジェクトネスと把持位置の学習

Heterogeneous Learning は複数のタスクを単一の DCNN で扱うための学習法である．本研究では，Heterogeneous Learning を用いることで，単一の DCNN で回帰タスクである把持位置と識別タスクであるオブジェクトネスの学習を行う．Heterogeneous Learning では，複数のタスクを解くために全結合層の出力ユニットに各タスクを割り当てる．本研究では，全結合層の出力ユニットに把持位置を推定する把持座標点ユニット (回帰タスク) と入力画像の物体らしさを表すオブジェクトネスユニット (識別タスク) を割り当てる．これにより，入力画像中の物体らしい領域における把持位置を出力する．図 3 に本研究で使用する 2 点の把持位置を検出する DCNN の構造を示す．DCNN の各層の詳細な構成は表 1 に示す．

DCNN の学習では畳み込みフィルタの重みと全結合層の結合重みおよびバイアスを決定する．DCNN では学習で求める重みが膨大な数となるため，最適な重みを決定するために誤差逆伝搬法を用いる．誤差逆伝搬法では，初期値として重みに乱数を与えて教師信号との誤差が小さくなるように繰り返し重みを更新する．把持座標は回帰推定するため，誤差関数として式 (3) のような二乗誤差関数 E_m を用いる．

$$E_m = ||\boldsymbol{T}_r - \boldsymbol{O}_r||_2^2 + (T_c - O_c)^2 \tag{3}$$

把持座標の教師信号 $\boldsymbol{T}_r$，把持座標ユニットの出力値 $\boldsymbol{O}_r$，オブジェクトネスの教師信号 T_c，オブジェクトネスユニットの出力値 O_c から誤差を求める．

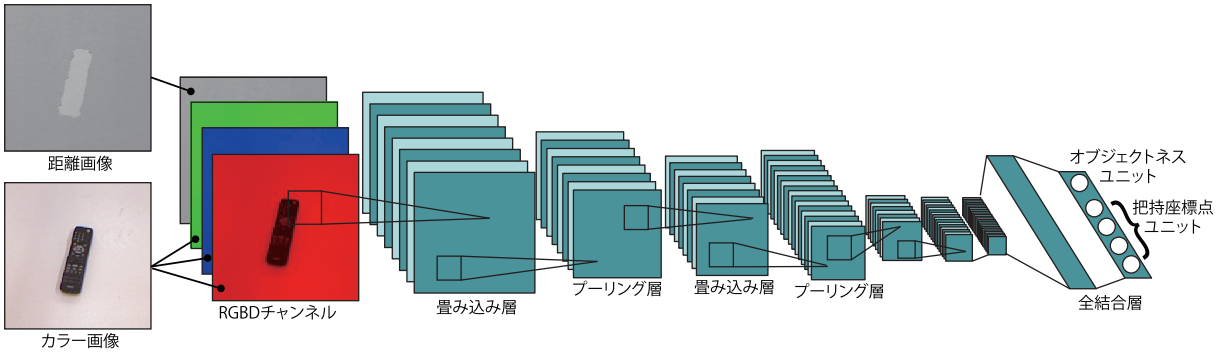


図 3　提案手法の DCNN の構造.

表 1　提案手法で使用する DCNN の詳細.

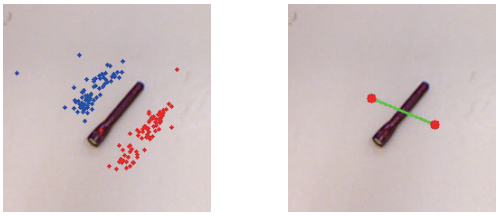
Layer	詳細
畳み込み層 1 層目	畳み込みフィルタ：9 × 9 活性化関数：ReLU プーリング：2 × 2
畳み込み層 2 層目	畳み込みフィルタ：9 × 9 活性化関数：ReLU プーリング：2 × 2
畳み込み層 3 層目	畳み込みフィルタ：7 × 7 活性化関数：ReLU プーリング：2 × 2
畳み込み層 4 層目	畳み込みフィルタ：5 × 5 活性化関数：ReLU プーリング：なし
全結合層 1 層目	ユニット数：2000
全結合層 2 層目	ユニット数：5 or 9

学習用の画像データセットは Cornell 大学の研究グループから公開されている Cornell Grasping Dataset[1]を使用する．Cornell Grasping Dataset は 280 種類の日用品アイテムを撮影した画像が 870 枚用意されている．各画像には RGB 画像，距離画像，4 点把持位置の教師信号が含まれている．提案手法はラスタスキャンベースの手法であるため，学習ではラスタスキャンのウィンドウ (250 × 250 画素) を切り出した RGB 画像と距離画像を学習データとして与える．よって，DCNN には RGB 画像の 3 チャンネルと距離画像の 1 チャンネルを用いた 4 チャンネルの RGB-D 画像を入力として与える．4 点把持位置を検出する場合，Cornell Grasping Dataset に含まれている把持座標の教師信号を用いる．2 点把持位置を検出する場合，Cornell Grasping Dataset に 2 点の教師信号が含まれていないため，Fast Graspability Evaluation [8] により検出した把持座標データを教師信号とする．オブジェクトネスユニットの教師信号は学習画像に対象物体が含まれている場合に 1 を付与し，対象物体が含まれていない背景画像には 0 を付与する．背景画像の把持座標の教師信号は，2 点把持位置の場合，左把持座標に (0, 0) を与え，右把持座標に (249, 0) を与える．4 点把持位置の場合は左上把持座標に (0, 0)，右上把持座標に (249, 0)，左下把持座標に (0, 249)，右下把持座標に (249, 249) を与える．

[1]http://pr.cs.cornell.edu/grasping/rect_data/data.php

2.3 オブジェクトネスを用いた把持位置検出

提案手法ではオブジェクトネスユニットを用いることにより，物体の最適な把持位置を検出する．ラスタスキャンしたウィンドウから DCNN により把持位置を検出した場合，各ウィンドウ毎で把持位置が検出される．そのため，図 4(a) に示すように 1 枚の画像から多数の把持位置が検出される．そこで，検出された全ての把持位置を用いて Parzen window により把持位置の統合処理を行う (図 4(b))．Parzen window は図 5(a) に示すように把持点 $\boldsymbol{x}$ に対して任意の関数 $\delta(\boldsymbol{x})$ で投票することで統合を行う．任意の位置 $\boldsymbol{x}$ において Parzen window の投票値 $P(\boldsymbol{x})$ は式 (4) のようになる．



(a)DCNN により検出された把持位置
(青：左把持位置，赤：右把持位置)

(b)Parzen Window による
把持位置の統合結果

図 4　提案手法により検出された把持位置の統合.

$$P(\boldsymbol{x}) = \frac{1}{G}\sum_{t=1}^{G}\delta(\boldsymbol{x}-\boldsymbol{x}_t) \tag{4}$$

本研究では，関数 $\delta(\cdot)$ に一般的に用いられるガウス関数を用いる．しかし，Parzen window をそのまま適用した場合，物体の周辺以外に検出された把持位置によって投票結果が曖昧になる (図 6(a), (b))．そこで，提案手法では式 (5) に示すように，関数 $\delta(\cdot)$ にオブジェクトネスユニットの出力値 O_c を重みとして掛けることで，

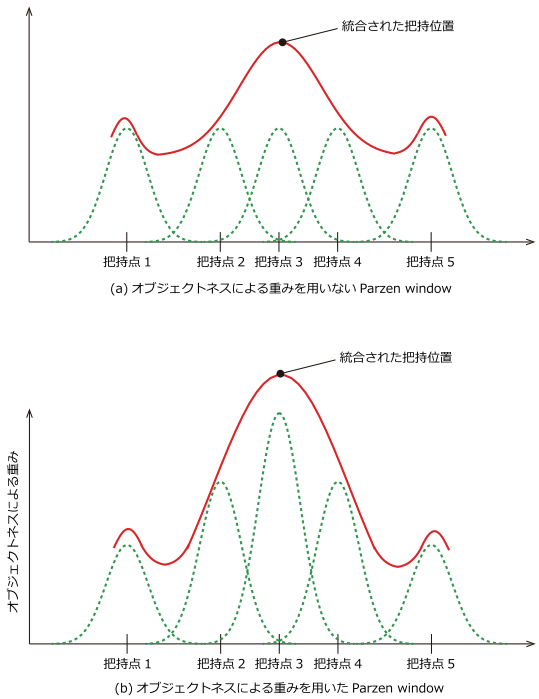


図 5　Parzen window による把持位置の統合.

物体らしい把持位置の重みを高くし，物体らしくない把持位置の重みを低くして投票する (図 5(b))．図 6(c) に左把持位置の重み付け投票結果，図 6(d) に右把持値位置の重み付け投票結果を示す．オブジェクトネスユニットにより重み付けした投票をすることで，より正確な把持位置を検出することが可能となる．

$$\delta(\boldsymbol{x}) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}}{2\sigma^2}\right) \cdot O_c \tag{5}$$

3　評価実験

提案手法の有効性を確かめるために評価実験を行う．評価実験では，Cornell Grasping Dataset から学習に使用していない画像を使用する．比較手法は 2 段階の DNN による把持位置検出法 [7] を用いる．また，2 点把持位置を検出する場合と 4 点把持位置を検出する場合に分けて評価を行う．

3.1　2 点把持位置検出

2 点把持位置検出では検出した把持座標点と教師信号の把持座標点とのユークリッド距離を用いて精度を評価する．式 (6) の条件を満たした場合に把持位置の検出成功，それ以外を検出失敗として検出率を比較する．

$$\frac{E_l + E_r}{E_t} \leq T \tag{6}$$

ここで，E_l は検出した左把持座標と教師信号の左把持座標とのユークリッド距離，E_r は検出した右把持座標と教師信号とのユークリッド距離である．E_t は教師信号の左把持座標と右把持座標間のユークリッド距離である．図 7 にしきい値 T を変化させたときの 2 点把持

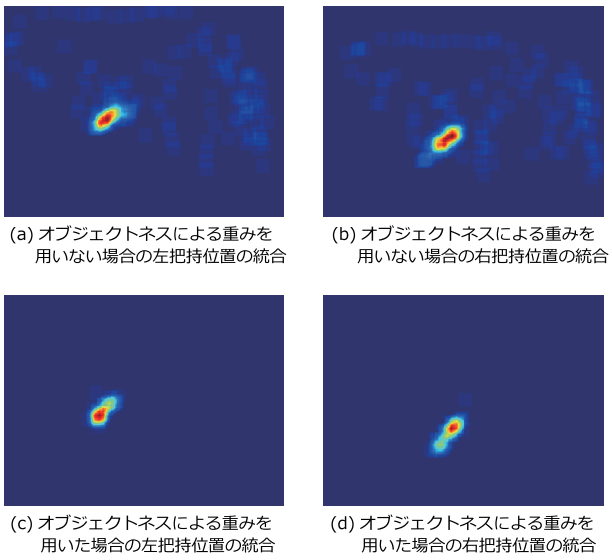


図 6　オブジェクトネスユニットを用いた投票マップ.

位置の検出率 [%] を示す．赤色で示す線は提案手法，青色で示す線は従来法である 2 段階の DNN による手法を示す．図 7 より，提案手法は従来法と比べ，検出率

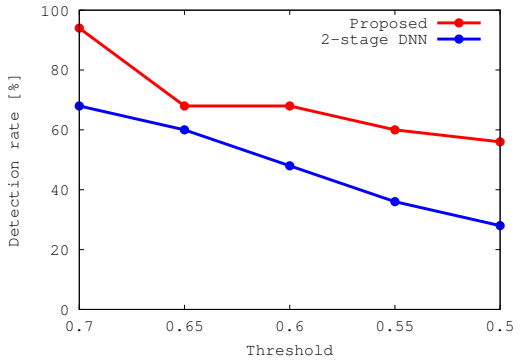
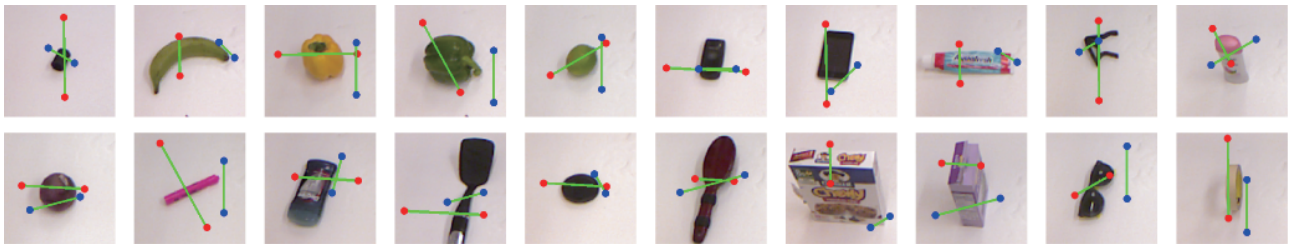


図 7　2 点の把持位置検出の精度.

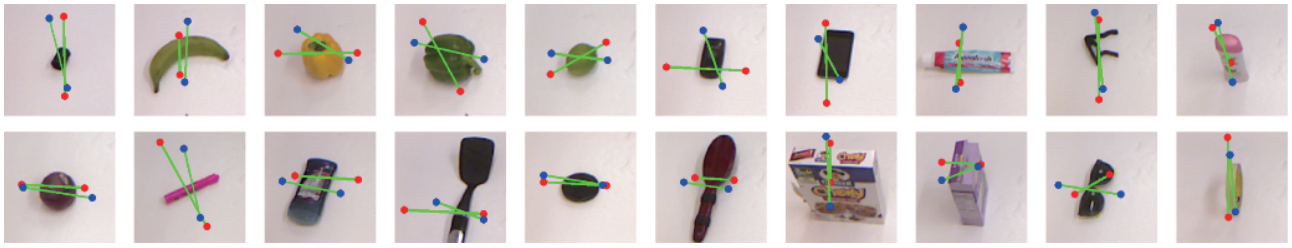
が平均で 21%向上した．これは，DCNN により画像の特徴を自動で獲得することで最適な把持位置を検出できているためである．また，把持位置の統合処理においてオブジェクトネスユニットの出力値で重み付けを行うことで，より良い把持位置を検出できていると考えられる．図 8 に提案手法と従来法による 2 点の把持位置の検出結果を示す．図中の赤色の点は正解座標点，青色の点は各手法により検出した結果である．

3.2　4 点把持位置検出

4 点把持位置検出では検出した 4 点把持位置を結んだ矩形領域と教師信号の矩形領域の重なり率を用いて精度を評価する．式 (7) の条件を満たした場合に把持位置の検出成功，それ以外を検出失敗として検出率を比較



(a)2 段階 DNN(従来法) による 2 点の把持位置検出例



(b) 提案手法による 2 点の把持位置検出例

図 8　2 点の把持位置の検出結果.

する.

$$\frac{R_d \cap R_t}{R_d \cup R_t} \geq T \tag{7}$$

ここで，R_d は検出した 4 点の把持位置を結んだ矩形領域，R_t は教師信号の 4 点把持位置を結んだ矩形領域である．図 9 にしきい値 T を変化させたときの 4 点把持位置の検出率 [%] を示す．赤色で示す線は提案手法，青色で示す線は従来法を示す．図 9 より，提案手法は従

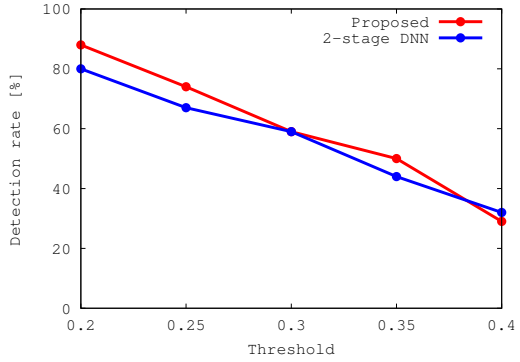


図 9　4 点の把持位置検出の精度.

来法と比べ，検出率が平均で 3.6%向上した．提案手法は 1 度のラスタスキャンで従来法と同等以上の性能で把持位置を検出することが可能であるため，DCNN と Heterogeneous Learning を用いた把持位置検出は有効であると考えられる．図 10 に提案手法と従来法による 4 点の把持位置の検出結果を示す．図中の赤色の点は正解座標点，青色の点は各手法により検出した結果である.

4　おわりに

本研究では，Heterogeneous Learning によるオブジェクトネスと物体把持位置検出法を提案した．把持位置検出に DCNN を用いることで，画像中の物体の特徴を自動的に捉え，1 度のラスタスキャンで効率的に把持位置を検出することが確認できた．また，Heterogeneous Learning を用いることで，画像の物体らしさをオブジェクトネスユニットで出力し，オブジェクトネスユニットの出力値を用いて，複数の把持位置を統合することで，高精度な把持位置検出が可能となった．今後の課題として，Heterogeneous Learning を用いた物体の認識と把持位置の同時推定を検討する.

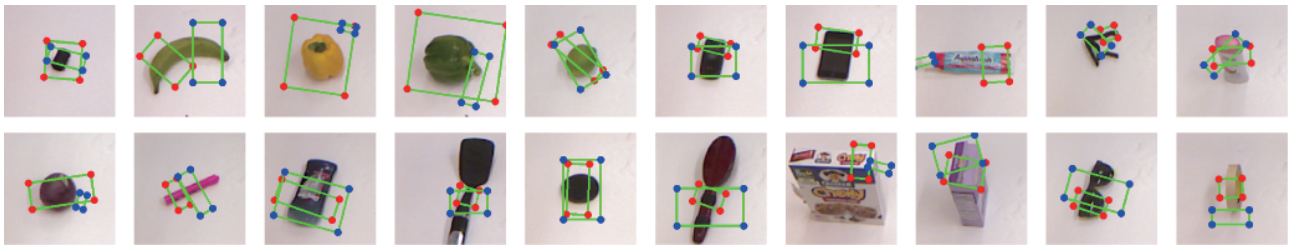
参考文献

[1] K. Harada, K. Nagata, T. Tsuji, N. Yamanobe, A. Nakamura, and Y. Kawai, “Probabilistic approach for object bin picking approximated by cylinders”, International Conference on Robotics and Automation, pp.3742–3747, 2013.

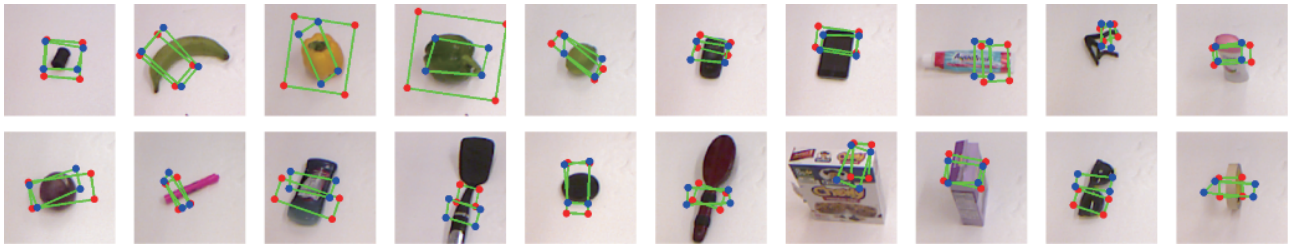
[2] M. Nieuwenhuisen, D. Droeschel, D. Holz, J. Stuckler, A. Berner, J. Li, R. Klein, and S. Behnke, “Mobile bin picking with an anthropomorphic service robot”, International Conference on Robotics and Automation, pp.2327–2334, 2013.

[3] B. Drost, M. Ulrich, N. Navab, and S. Ilic, “Model globally, match locally: Efficient and robust 3D object recognition”, , 2010.

[4] C. Choi, Y. Taguchi, O. Tuzel, M. Y. Liu, and S. Ramalingam, “Voting-based pose estimation



(a)2 段階 DNN(従来法) による 4 点の把持位置検出例



(b) 提案手法による 4 点の把持位置検出例

図 10　4 点の把持位置の検出結果.

for robotic assembly using a 3D sensor", International Conference on Robotics and Automation, pp.1724–1731, 2012.

[5] P. J. Besl, and N. D. McKay, "Method for registration of 3-D shapes", Robotics-DL tentative, pp.586–606, 1992.

[6] D. Chetverikov, D. Svirko, D. Stepanov, and P. Krsek, "The trimmed iterative closest point algorithm", International Conference on Pattern Recognition, vol.3, pp.545–548, 2002.

[7] I. Lenz, H. Lee, and A. Saxena, "Deep Learning for Detecting Robotic Grasps", International Journal of Robotics Research, vol.34, no.4-5, pp.705–724, 2015.

[8] Y. Domae, H. Okuda, Y. Taguchi, K. Sumi, and T. Hirai, "Fast graspability evaluation on single depth maps for bin picking with general grippers", International Conference on Robotics and Automation, pp.1997–2004, 2014.

[9] A. Saxena, J. Driemeyer, and A. Y. Ng, "Robotic grasping of novel objects using vision", International Journal of Robotics Research, vol.27, no.2, pp.157–173, 2008.

[10] J. Glover, D. Rus, and N. Roy, "Probabilistic models of object geometry for grasp planning", Science and Systems IV, pp.278–285, 2008.

[11] Y. Jiang, S. Moseson, and A. Saxena, "Efficient grasping from rgbd images: Learning using a new rectangle representation", International Conference on Robotics and Automation, pp.3304–3311, 2011.

[12] Y. Bengio, "Learning deep architectures for AI", Foundations and trends in Machine Learning, vol.2, no.1, pp.1–127, 2009.

[13] Q. V. Le, "Building high-level features using large scale unsupervised learning", International Conference on Acoustics, Speech and Signal Processing, pp.8595–8598, 2013.

[14] K. Sohn, D. Y. Jung, H. Lee, and A. O. Hero III, "Efficient learning of sparse, distributed, convolutional feature representations for object recognition", International Conference on Computer Vision, pp.2643–2650, 2011.