Convolutional Neural Networkによる 把持位置に基づいたマルチクラス物体認識

長谷川 昂 宏^{*1} 山 内 悠 嗣^{*1} 山 下 隆 義^{*1} 藤 吉 弘 亘^{*1}
秋 月 秀 一^{*2} 橋 本 学^{*2} 堂 前 幸 康^{*3} 川 西 亮 輔^{*3}

Multi-Class Object Recognition Based on Grasping Position by Convolutional Neural Network

Takahiro Hasegawa^{*1}, Yuji Yamauchi^{*1}, Takayoshi Yamashita^{*1}, Hironobu Fujiyoshi^{*1},
Shuichi Akizuki^{*2}, Manabu Hashimoto^{*2}, Yukiyasu Domae^{*3} and Ryosuke Kawanishi^{*3}

Automatization for the picking and placing of a variety of objects stored on shelves is a challenging problem for robotic picking systems in distribution warehouses. Here, object recognition using image processing is especially effective at picking and placing a variety of objects. In this study, we propose an efficient method of object recognition based on object grasping position for picking robots. We use a convolutional neural network (CNN) that can achieve highly accurate object recognition. In typical CNN methods for object recognition, objects are recognized by using an image containing picking targets from which object regions suitable for grasping can be detected. However, these methods increase the computational cost because a large number of weight filters are convoluted with the whole image. The proposed method detects all graspable positions from an image as a first step. In the next step, it classifies an optimal grasping position by feeding an image of the local region at the grasping point to the CNN. By recognizing the grasping positions of the objects first, the computational cost is reduced because of the fewer convolutions of the CNN. Experimental results confirmed that the method can achieve highly accurate object recognition while decreasing the computational cost.

Key Words: Object Recognition, Convolutional Neural Network, Picking

1. は じ め に

物流倉庫のピッキングロボットにおいて、棚 (shelf) に収納された多種多様な物体の中から指定された物体をピック&ブレースする技術は重要な課題である。これまでの産業用ロボットにおけるピッキング作業は、ピッキング対象物体の種類や形状を少数に限定したり、剛体の物体を対象にすることで実現されてきた。例えば、工業部品等のピンピッキングの問題においては、高度な物体認識技術を必要とせず、テンプレートマッチング等のシンプルなアルゴリズムが用いられる [1] [2]。また、同一物体のばら積みピッキングのような問題においては、距離画像や三次元点群を用いた物体の把持位置検出法のみが用いられる [3] [4]。

しかし、物流倉庫のピッキング作業のように、様々な種類、形状が混在するピッキング問題 [5] においては高精度な物体認識が必要となる。このような問題では、単純に把持位置を検出するだけでなく、把持する物体がどのような物体かを特定しなければならない。多種多様な物体を高精度に認識することができる機械学習法として Convolutional Neural Network (CNN) が提案されている [6]~[10]。CNN は、与えられた画像があらかじめ定義したカテゴリの中で、どのカテゴリに属するかを求める画像識別を得意とする。そのため、画像中に複数の物体が存在する場合においては、あらかじめ画像から複数の物体候補を検出する。検出した物体候補領域をそれぞれ CNN へ入力して画像識別することで、どのような物体がどこにあるかを求めることができる [11] [12]。よって従来の CNN により、物体のピック&ブレースを実現するには、Fig. 1 (a) に示すように **Step 1**: 画像全体からの物体候補検出, **Step 2**: 検出した候補領域を CNN で識別, **Step 3**: 識別した物体領域から把持可能な位置を検出, の3段階の処理が必要となる。Step 1 の物体候補検出では、物体らしい矩形領域のみ検出して物体のクラスは識別しない。そして、Step 1 で検出したすべての候補領域画像を CNN

原稿受付 2017 年 4 月 26 日

^{*1}中部大学

^{*2}中京大学

^{*3}三菱電機株式会社

^{*1}Chubu University

^{*2}Chukyo University

^{*3}Mitsubishi Electric Corporation

■ 本論文は有用性 (要素分野) で評価されました。

Fig. 1 Workflow of object recognition for pick-and-place unit

へ入力して物体を識別する。最後に、Step 2 で識別した領域に対して把持位置を検出する。

ここで、物体の把持位置は画像中の複数の物体上に検出されるため、把持位置自体を物体候補と見なすことができる。よって、Step 1 の物体候補領域を把持位置領域に置き換えて効率的なピッキングシステムを実現することができる。把持位置は、物体上の把持しやすい領域を検出するように設計されている。そのため、様々な物体の中で共通するような形状を捉えた把持位置が検出され、識別時に物体の形状情報が失われる。しかし、把持位置周辺領域の物体のテクスチャやカラー情報は得られるため、学習過程で特徴量を獲得する CNN を用いることで、物体を識別する。本研究では CNN による把持位置に基づいた物体認識手法を提案する。提案手法は、Fig. 1 (b) に示すように **Step 1**: 画像全体から把持可能位置を検出、**Step 2**: 検出した把持位置領域を CNN で識別、の 2 段階の処理となる。Step 1 で画像全体から把持可能位置を検出することで、従来のように、物体候補をあらかじめ検出する必要がないため、効率的な物体認識を実現できる。

本研究で扱うタスクは、shelf に収納されている物体の名前とピッキング対象物体名が記述された JSON ファイルをロボットシステム側で受け取り、ピッキングのための物体認識を実行する。これは現在、Amazon.com の物流倉庫において実用化されている shelf 運搬用の自律移動ロボットによる物流システムを想定したタスクである [13]。Amazon.com での物流システムでは、Kiva とよばれる自律移動ロボットが、商品を収納した shelf をピッキング作業場所へと運搬する。このとき、運搬動作による shelf の振動により商品の配置が変化する恐れがあるため、商品の配置情報は記録されないが、shelf に収納されている商品名は記録される。このような物流システムの場合、ピッキング時の物体認識処理が実行される前に shelf 内の物体を知ることができる。そこで、物体認識を行う CNN の出力層において、shelf 内の物体情報を考慮した制約付き softmax 関数を用いて物体認

識を高精度化する。制約付き softmax 関数は shelf に存在する物体に対応する CNN の出力ユニットの値のみを softmax 関数へ入力する。制約付き softmax 関数を用いることにより、shelf に存在しない物体との誤認識を減らすことができ、信頼度の高いクラス確率を出力することができる。

2. ピッキングロボットのための物体認識

ピッキングロボットにおける物体認識は、画像中にどのような物体がどこにあるかを特定したうえで把持位置を検出しなければならない。そのため、従来の物体検出法を利用したピッキングロボットのための物体認識は以下の 3 段階の処理が必要となる。

Step 1: 画像全体からの物体候補検出

Step 2: 検出した候補領域を CNN で識別

Step 3: 識別した物体領域から把持可能な位置を検出

ここでは、Step 1 と Step 2 が物体検出の処理に相当し、Step 3 がピッキングのための把持位置検出の処理になる。物体検出は機械学習を用いた手法が主流となっており、handcrafted feature とよばれる研究者の知見に基づいて設計された画像特徴量 [14]～[16] と統計的学習手法 [17]～[19] の組み合わせによる方法が実現されている。このような手法では、特徴量抽出における処理時間が比較的速いため、様々なスケール・アスペクト比の検出ウィンドウを画像内でラスタスキャンすることで対象物体を検出する。

さらに、特徴量抽出も一つの機械学習のフレームワークで獲得する Convolutional Neural Network [6] が提案され、高い認識精度が得られることから非常に注目されている。CNN は高い認識精度が得られる一方で、様々な重みフィルタを繰り返し画像に畳み込むため計算コストが高い、handcrafted feature を用いた物体検出のように、ラスタスキャンを行うと数万～数十万単位の検出ウィンドウ画像を CNN で識別することになるため、膨大な計算時間が必要となる。

そこで、Region proposal とよばれる物体候補領域をあらかじめ検出しておくことで、CNN での識別処理コストの増加を抑えるアプローチが用いられる。Region proposal として検出される領域は、物体が存在する領域として信頼度の高い領域のみを検出する。そのため、Region proposal に基づく CNN の物体検出は、ラスタスキャンで画像を網羅的に CNN で識別するよりも、はるかに計算コストが抑えられるため有力な物体検出法である。Region proposal 検出法と CNN を組み合わせた物体検出法として、Regions with Convolutional Neural Network (R-CNN) [11] が提案されている。Region proposal を検出する手法には、類似するセグメンテーションを繰り返し結合する Selective search [20] や画像の勾配やエッジに着目した BING [21]、Edge boxes [22] 等が提案されている。R-CNN では、Region proposal 検出法として Selective search を採用している。

R-CNN により、画像中の物体位置検出と識別を高精度に求めることができるが、それでも物体候補領域を繰り返し CNN で識別する計算コストや、Selective search 等による Region proposal 検出自体の計算コストが高い理由で、実用的な処理時

間で物体を認識することができない。これらの問題を解決するために、Fast R-CNN [12] と Faster R-CNN [23] が提案されている。Fast R-CNN では、CNN の処理時間に関して最もボトルネックである重みフィルタの畳み込み処理を画像全体に 1 度だけ適用する。そして、Selective search により入力画像から検出した Region proposal の矩形を、畳み込み処理で得られた特徴マップ上に射影する。特徴マップ上に射影した各矩形領域を比較的計算コストの低い全結合ニューラルネットワークにより識別することで、物体検出処理を高速化している。Faster R-CNN では、Region proposal 検出も CNN のフレームワークで学習させ、より良い物体候補を検出し、Fast R-CNN と同様のアプローチを組み合わせることで R-CNN や Fast R-CNN よりも高精度かつ高速な物体検出を実現している。Fast/Faster R-CNN においても、Step {1, 2, 3} で構成される 3 段階の物体認識に当てはまる。

R-CNN や Fast R-CNN における Region proposal 検出は、物体にプリントされている文字や模様、ロゴマーク等のテキストを捉えて Region proposal を過剰に検出することがある。そのため、ニューラルネットワークによる識別回数が増加してしまい処理時間が遅くなる。また、Faster R-CNN では、Region proposal 検出、物体矩形回帰、物体クラス識別の三つのタスクを一つの CNN で処理するため 18 層の大規模なネットワークを必要とする。さらに、Region proposal に基づく手法は検出した物体矩形内に複数の物体が存在すると、Step 3 において対象物体とは異なる物体から把持位置を検出することがあり、誤った物体を把持する可能性がある。

そこで、本研究では物体の把持位置に基づいた効率的な物体認識手法を提案し、コンパクトかつシンプルなネットワーク構造での物体認識を実現させる。また、物体の把持位置は画像中の複数の物体上に検出されることを考えると、検出された把持位置領域は物体候補と見なすことができる。よって、Step 1 において把持位置を検出し、検出した把持位置を物体候補とすることで、以下のような 2 段階構造の効率的な物体認識が可能である。

Step 1 : 画像全体から把持可能位置を検出

Step 2 : 検出した把持位置領域を CNN で識別

Step 1 では把持位置検出が物体候補の役割も担うため、Faster R-CNN のように Region proposal を含めた CNN を学習する必要はなく、全 8 層のコンパクトなネットワークで識別問題を解くことができる。提案手法では、検出された把持位置の周辺領域を CNN へ入力し、把持位置が検出された物体を識別することで、そのままピッキング動作に移ることができるため、ピッキングロボットのための効率的な物体認識が可能である。

3. 提案手法

本研究では、Convolutional Neural Network による把持位置に基づくマルチクラス物体認識を提案する。提案手法では、まず画像中のすべての物体から把持可能な位置を検出する。前段処理に把持位置検出を行うため、様々な物体の把持位置を高速に検出できる Fast Graspability Evaluation [4] を使用する。Fast Graspability Evaluation はロボットの二次元ハンドモデルをテンプレートとして保持し、距離画像から得られた物体セグメンテーションに対して二次元ハンドモデルが最適に当てはまる位置を把持位置として検出する。Fast Graspability Evaluation で検出した把持位置に対して CNN により物体認識を行う。

3.1 把持位置に基づく Convolutional Neural Network

提案手法は、画像から検出された把持位置に対して CNN により把持位置が検出された物体を認識する。提案手法の CNN は Fig. 2 に示すように畳み込み層と全結合層で構成され、畳み込み層は画像の特徴抽出、全結合層は特徴量の識別の役割を担う。畳み込み層は、重みフィルタを画像または特徴マップに畳み込み、その応答値を活性化関数に通して特徴マップとする。その後、プーリングにより特徴マップを縮小させる。活性化関数には Rectified Linear Unit (ReLU) [24]、プーリングには Max pooling を用いる。また、ミニバッチごとに特徴マップの分布が平均 0、分散 1 となるように正規化する Batch Normalization [25] を導入する。提案手法の CNN 構成の詳細を Table 1 に示す。提案手法の把持位置に基づいた CNN は、検出された把持位置の周辺領域を固定サイズで切り取った RGB 画像が入力となる。CNN の入力画像サイズは 227×227 ピクセルとする。全結合層の出力ユニットはクラス数すなわち認識対象の物

Fig. 2 Our CNN architecture

Table 1 Details of our CNN architecture

Layer	Details
1st layer (convolution)	Filter size : 11×11
	Stride : 4
	Padding : 0
	Normalization : Batch normalization
	Activation function : ReLU
2nd layer (convolution)	Pooling : Max pooling
	Filter size : 5×5
	Stride : 1
	Padding : 2
	Normalization : Batch normalization
3rd layer (convolution)	Activation function : ReLU
	Pooling : Max pooling
	Filter size : 3×3
	Stride : 1
	Padding : 1
4th layer (convolution)	Normalization : -
	Activation function : ReLU
	Pooling : -
	Filter size : 3×3
	Stride : 1
5th layer (convolution)	Padding : 1
	Normalization : -
	Activation function : ReLU
	Pooling : Max pooling
6th layer (fully connected)	Filter size : 3×3
	Stride : 1
7th layer (fully connected)	Padding : 1
	Normalization : -
8th layer (fully connected)	Activation function : ReLU
	Dropout (training) : $\bigcirc$
	Activation function : ReLU
	Dropout (training) : $\bigcirc$
	Activation function : Softmax
	Dropout (training) : -

体数+背景クラスとなり、最も応答値の高いユニットが認識クラスとなる。把持位置に基づいた CNN により、検出された把持位置のみに認識処理を実行すれば良いため計算コストを抑えることができる。また、最初に把持可能な位置をすべて検出するため、CNN で認識した後にそのままピッキングが可能であるため効率的である。

3.2 学習画像

CNN の学習画像は、あらゆる向きで配置された単一物体の画像を使用する。学習画像の撮影は三菱電機社製の産業用ロボット MELFA に搭載されている MELFA-3D Vision を用いる。MELFA-3D Vision はカメラとプロジェクタで構成され、アクティブステレオ法により RGB 画像と距離画像を撮影する。画像は、単一物体を様々な姿勢で配置して撮影する。撮影した距離画像から Fast Graspability Evaluation [4] により把持位置を検出し、検出された把持位置の 227×227 ピクセルの範囲を RGB 画像から切り取って学習用パッチ画像とする。CNN の物体認識は、様々な姿勢や環境下においても認識精度を保つために Data Augmentation を行う。学習画像生成の流れを **Fig. 3** に示す。提案手法では、Data Augmentation として回転、平行移動、照明変化、ノイズ付加をランダムで画像に施して学習用

Fig. 3 Training patch image generation **Fig. 4** Examples of training patch image

パッチ画像を増加させる。Data Augmentation により生成した学習用パッチ画像の例を **Fig. 4** に示す。

3.3 制約付き softmax 関数

CNN によるクラス識別では、softmax 関数によりクラス確率を算出する (**Fig. 5 (a)**)。CNN の出力ユニットの値を h 、クラス数を C とすると、クラス確率は、式 (1) の softmax 関数 $P_r(\cdot)$ により算出される。

$$P_r(h_i) = \frac{\exp(h_i)}{\sum_{j=1}^C \exp(h_j)} \quad (1)$$

ここで、 i は CNN の出力ユニットの値 h に対するインデックスである。本研究では、shelf に存在する物体に対応する出力ユニットのみを用いて計算する制約付き softmax 関数によりクラス確率を計算する (**Fig. 5 (b)**)。例えば shelf に {1, 3, 4} 番めの出力ユニットに対応する物体が収納されているとき、制約付き softmax 関数 $P_c(\cdot)$ は式 (2) のように定義できる。

$$P_c(h_i) = \frac{\exp(h_i)}{\sum_{j=\{1,3,4\}} \exp(h_j)} \quad (2)$$

Fig. 5 Softmax functions

制約付き softmax 関数は shelf に収納されている複数の物体が既知である場合において有効な関数であり, shelf に存在しない物体との誤認識を減らすことができる. 制約付き softmax 関数 $P_c(\cdot)$ による物体識別は inference で使用し, 学習時では通常の softmax 関数 $P_r(\cdot)$ を用いて CNN を学習させる.

4. 評価実験

提案手法の有効性を確認するために評価実験を行う. 本実験では, R-CNN [11], Faster R-CNN [23] と提案手法である把持位置に基づく CNN の認識精度と処理時間を比較する. R-CNN の Region proposal 検出には Selective search [20] を使用する.

4.1 データセット

本実験で使用するデータセットは国際物流ロボットコンペティションである Amazon Picking Challenge (APC) [5] で使用された商品を使用する. 2015 年に開催された APC 2015 では全 25 種類の物体, 2016 年に開催された APC 2016 では全 39 種類の物体が使用された. Fig. 6 に APC 2015 の認識対象物体, Fig. 7 に APC 2016 の認識対象物体を示す. APC 2015 のデータセットは [26] で公開されているデータを使用する. CNN の学習には shelf の中に単一の物体が配置されている画像のみを用いる. APC 2015 の学習画像は 750 枚, APC 2016 の学習

Fig. 6 Objects of Amazon Picking Challenge 2015

Fig. 7 Objects of Amazon Picking Challenge 2016

画像は 1709 枚である. 評価用画像は shelf の中に複数の物体が配置されている画像を使用する. APC 2015 の評価画像は 594 枚, APC 2016 の評価画像は 200 枚である.

4.2 物体認識における精度

R-CNN, Faster R-CNN と提案手法である把持位置に基づく CNN の認識精度を比較する. 各手法において制約付き softmax を使用して認識結果を出力する. 制約付き softmax 関数は, shelf に収納されている物体クラスのみでクラス確率を計算し, その中での最大値を推定クラスとする. APC 2015 データセットの各物体の認識率を Fig. 8 に示す. APC 2015 データセットでは, shelf の中に 2 個の物体が配置される場合と 3~4 個の物体が配置される場合にかけて評価する. グラフの横軸は物体番号であり, Fig. 6 に示す番号に対応している. グラフの最終列は全物体の平均認識率である.

shelf 中の物体数が 2 のとき, 提案手法は R-CNN と同等の

Fig. 8 Recognition rate of APC 2015 dataset

平均認識率が得られていることが確認できる (Fig. 8(a)). shelf 中の物体数が3~4 の場合で、提案手法は R-CNN よりも平均認識率が 9.2% 向上した (Fig. 8(b)).

APC 2016 データセットの各物体の認識率を Fig. 9 に示す. APC 2016 データセットでは, shelf の中に 2~3 個の物体が配置される場合, 4~5 個の物体が配置される場合, 6~10 個の物体が配置される場合にわけて評価する. グラフの横軸の物体番号は, Fig. 7 に示す番号に対応している. 提案手法は R-CNN と同等以上の平均認識率であり, 物体数が 6~10 個の場合においては, Faster R-CNN の平均認識率を 4.3% 上回る結果が得られた.

以上の結果から, 提案手法は R-CNN と同等以上の認識精度で効率的なピッキングロボットシステムに適用できると言える.

本実験では Faster R-CNN の精度が最も高いが, これは Region proposal の検出も CNN によって獲得するため, 識別に適した物体候補領域が検出できていると考えられる. しかし, Faster R-CNN は Region proposal 検出, 物体矩形回帰, 物体クラス分類をすべて CNN により処理するため, 提案手法の CNN と比較して大規模なネットワークを必要とする.

提案手法は APC 2016 データセットにおいて Faster R-CNN

と同等の精度でありながら, R-CNN や Faster R-CNN よりも短い処理時間で効率的に物体を認識することができる. 各手法の処理時間については, 4.4 節で比較する.

最後に, 2016 年 7 月にドイツ・ライプツィヒにて開催された Amazon Picking Challenge 2016 [27] において, 我々の参加チーム (Team C²M) の物体認識結果を Fig. 10 に示す. 競技は shelf から指定された物体を取り出し, tote とよばれる箱に物体を移動させる “pick task” と tote から shelf へ物体を収納する “stow task” の 2 種類のタスクが設定されている. Fig. 10 は pick task において提案手法の物体認識が試行された 6 シーンの認識結果である. pick task では, shelf に収納されている物体名が記述された JSON ファイルが使用可能であるため, 制約付き softmax 関数をそのまま適用させることができる. pick task における認識率は 50/54 (= 正解認識結果 / 検出した把持位置) で 92.5% であり, 実問題においても高精度な物体認識を実現した. stow task では, tote から取り出した物体を仮置き台へ移動させ, 仮置き台上の物体を提案手法により識別した. stow task においては, JSON ファイルを使用せず, ロボットに搭載された力覚センサにより tote から把持した物体の重量を測定する. 物体の重量から, 把持物体を 4 種類の候補に絞り込

Fig. 9 Recognition rate of APC 2016 dataset

むことで、仮置き台での認識処理は制約付き softmax 関数を使用することが可能となり、stow task においても高精度な認識を実現した。

4.3 把持位置検出における精度

ここでは、物体認識後の把持位置検出の正解率を比較する。4.2 節において、R-CNN や Faster R-CNN は検出した物体矩

形に対する認識率を比較した。R-CNN や Faster R-CNN 等の Region proposal に基づく手法は、認識物体の位置を矩形領域として検出した後、検出した矩形領域内から把持位置を検出する必要がある。R-CNN や Faster R-CNN は以下の二つの要因で誤った物体を把持することがある。

(1) 物体矩形の誤認識による誤った物体の把持。

Fig. 10 Recognition results for our team in Amazon Picking Challenge 2016

(2) 物体矩形内の異なる物体の写り込みによる把持位置の誤検出。一つめの要因については、4.2 節の実験で示す認識精度の割合で誤った物体を把持してしまう。さらに、R-CNN や Faster R-CNN では二つめの要因が発生する。検出した物体矩形のクラスが正解であっても、Fig. 11 に示すように shelf の中に多数の物

体が密集して配置されている場合、物体矩形内の異なる把持位置を検出する場合がある。一方、提案手法は把持位置ごとに物体認識を行うため、誤った物体を把持する要因は誤認識した場合のみである。R-CNN や Faster R-CNN により検出した物体矩形から Fast Graspability Evaluation [4] を用いて把持位置

Table 2 Accuracy rate of grasping position within bounding box [%]

Dataset	# of objects	Proposed grasping position based CNN	R-CNN with selective search	Faster R-CNN
APC 2015	2	82.88	83.31	90.36
APC 2015	3-4	74.71	55.92	68.66
APC 2016	2-3	92.72	79.97	91.51
APC 2016	4-5	92.69	56.85	82.49
APC 2016	6-10	88.89	55.61	74.54

Table 3 Breakdown of object recognition time [ms]

	Proposed grasping position based CNN		R-CNN with selective search		Faster R-CNN	
	CPU	GPU	CPU	GPU	CPU	GPU
Selective search	-		1,662.3		-	
Detection of grasping positions	579.6		570.1		559.4	
Inference processing of CNN	245.1	8.9	622.8	26.8	10,551.9	69.2
Total	824.7	588.5	2,855.2	2,259.2	11,111.3	628.6

Fig. 11 Examples of grasping position within detected bounding box by Faster R-CNN

候補を検出し、把持位置候補ごとに正解率を求めると **Table 2** となる。提案手法は検出した把持位置から物体を識別する手法であるため、4.2 節で示した認識率と同じ精度となる。また、すべての手法において shelf 中の物体のみを検出対象とする制約付き softmax を適用する。APC 2015 データセットの shelf 内の物体数が 2 の場合は、物体同士の重なりが少ないため Faster R-CNN の正解率が高い結果となる。しかし、shelf 中の物体数が増えると R-CNN や Faster R-CNN は複数の物体を一つの矩形で検出するケースが多くなり、把持位置の正解率が大幅に低下する。提案手法は、shelf 中の物体数が 3 以上のすべてのデータセットにおいて最も良い精度であることが確認できる。

4.4 処理時間

APC 2016 データセットのうち、物体数が 2~3, 4~5, 6~10 個の画像をそれぞれ 10 枚ずつ用意し、合計 30 枚の画像の平均処理時間を比較する。処理時間は CPU : Intel Core i7-7700 3.6~4.2 [GHz], GPU : NVIDIA GeForce GTX 1080 Ti を用いて計測する。提案手法の平均把持位置数は 5, R-CNN の平均物体候補数は 14 である。全手法において、把持位置検出は Fast Graspability Evaluation を用いる。**Table 3** に認識に必

要な処理時間の内訳を示す。Table 3 より、提案手法の CPU での処理時間は R-CNN と比較して約 3.4 倍、Faster R-CNN と比較して約 13.4 倍高速であることが確認できる。R-CNN は、Selective search による Region proposal 検出の処理時間の割合が非常に大きい。また、Selective search は物体にプリントされている文字やロゴマーク等のテキストに反応して過剰に物体候補を検出するため、CNN の実行回数が多くなり処理時間が遅くなる。Faster R-CNN は、Region proposal 検出、物体矩形回帰、物体クラス分類の三つのタスクを一つの CNN で処理するため、提案手法の CNN と比較して大規模なネットワークが必要となる。そのため、CNN の実行に多くの処理時間が必要となる。GPU を用いた場合においても、提案手法の処理時間が最も短いため、提案手法は高速かつ効率的にマルチクラス物体認識が可能であると言える。

5. おわりに

本論文では、Convolutional Neural Network による把持位置に基づいたマルチクラス物体認識を提案した。把持位置に基づく CNN による物体認識は R-CNN と同等以上の認識精度が得られ、Faster R-CNN よりも高速な処理が可能である。入力画像から把持位置を検出し、把持位置周辺情報のみを用いて物体を認識することで効率的なピッキングロボットシステムへの応用が期待できる。評価実験では、APC 2016 データセットで 88% 以上の平均認識率を達成した一方で、APC 2015 データセットの一部の物体で提案手法の精度低下が確認された。このような、提案手法にとって不向きな物体や状況における認識精度の向上は改善すべき点であり、今後の課題である。

また、本論文では Amazon Picking Challenge で使用された 25~39 種類の物体を対象としたが、実際の物流倉庫を想定した大規模な種類の物体において、提案手法の高精度化が望まれる。このような問題に対処するには、物体識別と把持位置検出を単一の CNN の枠組みで実現する。物体識別と把持位置検出を同時に解くような CNN を構築することで、把持しやすさと識別に有効な把持位置を同時に学習できると考えており、より高精度な物体識別が期待できる。さらに、Fast/Faster R-CNN のように物体領域ごとに識別を行う 2 段階のアルゴリズムではなく、入力画像からダイレクトに複数の把持可能位置と物体クラスを

推定するような CNN を構築することで、より効率的なピッキングシステムを実現させる。

謝 辞 この成果は、国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の委託業務の結果得られたものです。また、本研究のロボットシステム構築に携わった三菱電機株式会社 先端技術総合研究所 関真規人氏、ご協力いただいた皆様に心より感謝いたします。

参 考 文 献

- [1] D.M. Tsai, C.T. Lin and J.F. Chen: "The Evaluation of Normalized Cross Correlations for Defect Detection," *Pattern Recognition Letters*, vol.24, no.15, pp.2525–2535, 2003.
- [2] H.G. Barrow, J.M. Tenenbaum, R.C. Bolles and H.C. Wolf: "Parametric Correspondence and Chamfer Matching: Two New Techniques for Image Matching," *International Joint Conference on Artificial Intelligence*, pp.659–663, 1977.
- [3] M. Nieuwenhuisen, D. Droeschel, D. Holz, J. Stückler, 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.
- [4] 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.
- [5] Amazon Robotics Challenge, <https://www.amazonrobotics.com/#/roboticschallenge>, 2017.
- [6] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner: "Gradient-based learning applied to document recognition," *Proc. of the IEEE*, vol.86, no.11, pp.2278–2324, 1998.
- [7] A. Krizhevsky, I. Sutskever and G.E. Hinton: "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, pp.1097–1105, 2012.
- [8] K. Simonyan and A. Zisserman: "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich: "Going deeper with convolutions," *Conference on Computer Vision and Pattern Recognition*, pp.1–9, 2015.
- [10] K. He, X. Zhang, S. Ren and J. Sun: "Deep residual learning for image recognition," *Conference on Computer Vision and Pattern Recognition*, pp.770–778, 2016.
- [11] R. Girshick, J. Donahue, T. Darrell and J. Malik: "Rich feature hierarchies for accurate object detection and semantic segmentation," *Conference on Computer Vision and Pattern Recognition*, pp.580–587, 2014.
- [12] R. Girshick: "Fast r-cnn," *International Conference on Computer Vision*, pp.1440–1448, 2015.
- [13] R. D'Andrea: "Guest editorial: A revolution in the warehouse: A retrospective on kiva systems and the grand challenges ahead," *Automation Science and Engineering*, vol.9, no.4, pp.638–639, 2012.
- [14] P. Viola and M. Jones: "Rapid object detection using a boosted cascade of simple features," *Conference on Computer Vision and Pattern Recognition*, vol.1, pp.I–I, 2001.
- [15] D.G. Lowe: "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol.60, no.2, pp.91–110, 2004.
- [16] N. Dalal and B. Triggs: "Histograms of oriented gradients for human detection," *Conference on Computer Vision and Pattern Recognition*, vol.1, pp.886–893, 2005.
- [17] Y. Freund and R.E. Schapire: "A decision-theoretic generalization of on-line learning and an application to boosting," *European conference on computational learning theory*, pp.23–37, 1995.
- [18] B.E. Boser, I.M. Guyon and V.N. Vapnik: "A training algorithm for optimal margin classifiers," *Proc. of the fifth annual workshop on Computational learning theory*, pp.144–152, 1992.
- [19] L. Breiman: "Random forests," *Machine learning*, vol.45, no.1, pp.5–32, 2001.
- [20] J.R. Uijlings, K.E. Van De Sande, T. Gevers and A.W. Smeulders: "Selective search for object recognition," *International Journal of Computer Vision*, vol.104, no.2, pp.154–171, 2013.
- [21] M.M. Cheng, Z. Zhang, W.Y. Lin and P. Torr: "BING: Binarized normed gradients for objectness estimation at 300fps," *Conference on Computer Vision and Pattern Recognition*, pp.3286–3293, 2014.
- [22] C.L. Zitnick and P. Dollár: "Edge boxes: Locating object proposals from edges," *European Conference on Computer Vision*, pp.391–405, 2014.
- [23] S. Ren, K. He, R. Girshick and J. Sun: "Faster R-CNN: Towards real-time object detection with region proposal networks," *Advances in Neural Information Proc. Systems*, pp.91–99, 2015.
- [24] X. Glorot, A. Bordes and Y. Bengio: "Deep sparse rectifier neural networks," *International Conference on Artificial Intelligence and Statistics*, pp.315–323, 2011.
- [25] S. Ioffe and C. Szegedy: "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *International Conference on Machine Learning*, pp.448–456, 2015.
- [26] Amazon Picking Challenge RGB-D + PointCloud Dataset 2015, http://mprg.jp/research/apc_dataset_2015_, 2015.
- [27] 藤吉弘亘, 松元叔一, 岡田慧: "「特別講演」 Amazon Picking Challenge 2016 の参加レポート", *パターン認識・メディア理解研究会*, vol.116, no.461, pp.123–129, 2017.

長谷川昂宏 (Takahiro Hasegawa)

2015 年中部大学大学院博士前期課程修了。同年より同大学大学院博士後期課程に在籍。コンピュータビジョン、パターン認識・理解の研究に従事。

山内悠嗣 (Yuji Yamauchi)

2012 年中部大学大学院博士後期課程修了。2012 年中部大学大学院博士研究員、2014 年中部大学助手。2010 年独立行政法人日本学術振興会特別研究員 DC2。コンピュータビジョン、パターン認識の研究に従事。(日本ロボット学会正会員)

山下隆義 (Takayoshi Yamashita)

2002 年奈良先端科学技術大学院大学博士前期課程修了, 2002 年オムロン株式会社入社, 2011 年中部大学大学院博士後期課程修了 (社会人ドクター), 2014 年中部大学講師, 2017 年中部大学准教授, 人の理解に向けた動画像処理, パターン認識・機械学習の研究に従事. 画像センシングシンポジウム高木賞 (2009 年), 電子情報通信学会 情報・システムソサイエティ論文賞 (2013 年), 電子情報通信学会 PRMU 研究会研究奨励賞 (2013 年) 受賞.

秋月秀一 (Shuichi Akizuki)

2016 年中京大学大学院博士後期課程修了. 2016 年度独立行政法人日本学術振興会特別研究員 (DC2,PD). 2017 年 4 月より慶應義塾大学理工学部電子工学科助教, 中京大学特別研究員 (兼任). 博士 (情報科学). 三次元物体認識技術の研究に従事. 2012 年度画像センシングシンポジウム優秀学術賞, 2015 年度 MIRU インタラクティブ発表賞等受賞. 電子情報通信学会, 精密工学会, IEEE 各会員.

堂前幸康 (Yukiyasu Doma)

2006 年北海道大学大学院情報科学研究科博士前期課程修了. 博士 (情報科学). 2008 年より三菱電機株式会社先端技術総合研究所研究員. 2014 年より同主席研究員. 2015 年より産業技術総合研究所知能情報システム部門特定専門研究員, 現在に至る. ロボットビジョン, マニピュレーションの研究開発に従事. 米国 R&D100 賞, 論文賞 (JRM 2012, 日本ロボット学会誌 2016, FA 財団 2017) など受賞. IEEE, 電子情報通信学会, 情報処理学会, 精密工学会各会員. (日本ロボット学会正会員)

藤吉弘亘 (Hironobu Fujiyoshi)

1997 年中部大学大学院博士後期課程修了, 1997 年米カーネギーメロン大学ロボット工学研究所 Post-doctoral Fellow, 2000 年中部大学工学部情報工学科講師, 2004 年中部大学准教授, 2005 年米カーネギーメロン大学ロボット工学研究所客員研究員 (～2006 年), 2010 年中部大学教授, 2014 年名古屋大学客員教授. 計算機視覚, 動画像処理, パターン認識・理解の研究に従事. ロボカップ研究賞 (2005 年), 情報処理学会論文誌 CVIM 優秀論文賞 (2009 年), 情報処理学会山下記念研究賞 (2009 年), 画像センシングシンポジウム優秀学術賞 (2010, 2013, 2014 年), 電子情報通信学会情報・システムソサイエティ論文賞 (2013 年) ほか. (日本ロボット学会正会員)

橋本 学 (Manabu Hashimoto)

1987 年大阪大学大学院工学研究科前期課程修了. 同年三菱電機 (株) 入社. 生産技術研究所, 先端技術総合研究所にてロボットビジョン, ヒューマン認識技術などの研究開発に従事. 2008 年中京大学情報理工学部機械情報工学科教授. 2013 年より改組により同大学工学部機械システム工学科教授. 博士 (工学). 1998 年度日本ロボット学会実用化技術賞, 2012 年度画像センシングシンポジウム優秀学術賞, 2015 年度 MIRU インタラクティブ発表賞, 精密工学会画像応用技術専門委員会小田原賞, 2017 および 2018 年度 IWAIT Best Paper Award 等受賞. 電子情報通信学会, IEEE 各会員. (日本ロボット学会正会員)

川西亮輔 (Ryosuke Kawanishi)

2007 年 3 月静岡大学工学部機械工学科卒業, 2009 年 3 月静岡大学機械工学専攻博士前期課程修了, 2012 年 3 月静岡大学情報科学専攻博士後期課程修了, 現在三菱電機株式会社先端技術総合研究所に勤務. 2008 年 8 月 Third Asia International Symposium on Mechatronics にて Best Student Paper Award, 2009 年 10 月 IEEE Robotics and Automation Society Japan Chapter Young Award, 2014 年 11 月 R&D 100 Awards, 2014 年 12 月 FA 財団論文賞を受賞. ロボットビジョン, ロボットの知能化に関する研究に従事. 精密工学会, 博士 (工学). (日本ロボット学会正会員)