協働ロボットと AI による作業連動システムの開発（第１報）

生駒晃大*、坂東直行**、渡辺博己*、安部貴大*

Development of cooperation system using collaborative robots with artificial intelligence（Ⅰ）

IKOMA Akihiro*, BANDO Naoyuki**, WATANABE Hiroki* and ABE Takahiro*

本研究では、DX の推進によるものづくりの高度化を実現するため、協働ロボットと AI 技術を組み合わせることで、人とロボットの協調作業や、製品の状態把握を支援するための技術開発に取り組んでいる。今年度は、ロボットアームによるピッキング動作を対象に、カメラで取得した画像から、対象物の位置や姿勢を推定するための深層学習による物体認識モデルについて検討した。実験の結果、Mask R-CNN によるセグメンテーションや、OnePose による 2D-3D マッチングにより、ピッキングに必要な物体の位置姿勢を推定できることを確認した。

１．はじめに

AI（人工知能）や IoT をはじめとするデジタル技術の進展により、様々な業界で DX（デジタル変革）の実現に向けた取り組みが行われている。製造業などのものづくりの現場では、DX による新たなビジネスモデルの創出や優位性の確立に向け、製造プロセスの自動化や生産システムの連携、製品データの分析などをより高度化するための技術が求められている。

生産工程の自動化技術では、従来の産業用ロボットに加え、人と同じ空間で作業可能な協働ロボットにも注目が集まっている。安全性や柔軟性に優れた協働ロボットの活用により、単純な繰り返し作業だけでなく、人の動きや生産ラインの変化に合わせた複雑な作業の自動化が期待されている。

そこで、本研究では、このような協働ロボットによる高度な自動化を実現するため、協働ロボットと AI 技術を組み合わせることで、人とロボットの協調作業や、対象物の状態把握を支援するための技術開発に取り組んでいる。今年度は、ロボットアームによるピッキング動作を対象に、深層学習（Deep Learning）を用いた対象物の位置姿勢の推定手法について検討したので報告する。

２．深層学習による物体位置姿勢の推定

ロボットアームによるピッキング動作を適切に行うためには、対象物が作業エリア上にどのような状態で配置されているのかを正確に把握する必要がある。そのため、ロボット周辺に取り付けたカメラや 3D スキャナで得られた画像や点群データを解析するための物体認識アルゴリズムが多数提案されている。

画像データを対象とした二次元的なアプローチとして、対象物のテンプレートや局所特徴量を用いたマッチング手法がある [1]。安価な単眼カメラのみでシステムを構築できる一方、マッチングに使用する特徴量の設計の良し悪しが認識精度に大きな影響を与える。また、点群データを対象とした三次元的なアプローチでは、対象物の 3D データを用いたモデルベースのマッチング手法がある [2]。対象物の三次元形状を認識できるため、ロボットによる把持位置の特定などに有効だが、高価なセンサや対象物の 3D モデルが必要となる。

そこで本研究では、ハンドクラフトな特徴量や 3D モデルを必要としない深層学習による物体認識モデルを使用したので、その手法について述べる。

２．１　セグメンテーションモデル

深層学習による画像データを入力とした物体認識モデルとして、画像中の物体位置を矩形で抽出する物体検出モデルや、画素単位で物体位置を特定するセグメンテーションモデルが存在する。セグメンテーションモデルは物体検出モデルと比較し、学習データの準備や学習時のパラメータ調整などに時間を要するが、より詳細な物体形状を抽出することが可能である。そのため、本研究では、物体検出モデルの一つである R-CNN[3]をセグメンテーションモデルに拡張した Mask R-CNN[4]を使用して物体位置の推定を行った。

Mask R-CNN によるセグメンテーションの流れを図 1 に示す。まず、畳み込みニューラルネットワーク（CNN）をベースとした特徴抽出器に画像を入力し、高次元の特徴マップを抽出する。次に、抽出した特徴マップから、検出対象となる物体が存在する領域を Region Proposal Network（RPN）により物体候補領域（RoI）として絞り込みを行う。最後に、特徴マップから切り出した各 RoI に対して、領域のサイズを揃えるためのリサイズ処理（RoIAlign）を適用後、識別処理により物体領域を示すマスク画像を生成する。

また、Mask R-CNN は、セグメンテーション用のマスク画像の生成だけでなく、物体クラスの分類や、物体検出のための矩形情報も推定するマルチタスクなモデルとなっている。そのため、モデルの学習時には、これら 3

＊ 情報技術部
＊＊ 機械部

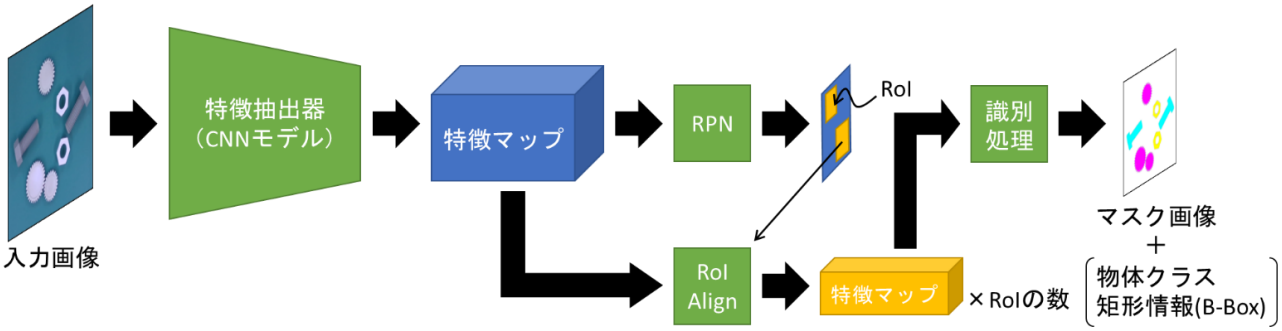

図１　Mask R-CNN によるセグメンテーション処理フロー

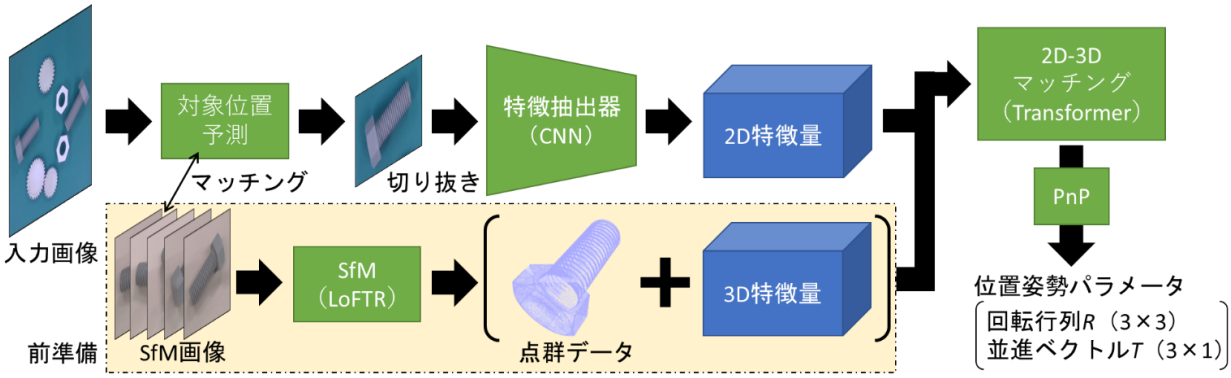

図２　OnePose による位置姿勢推定フロー

つの正解ラベルを付与したデータを準備し、各タスクの推論結果との誤差の合計を最小化するようなマルチタスク学習が行われる。

２．２　2D-3D マッチングモデル

前節のセグメンテーションモデルによる物体位置の推定は二次元平面上の位置情報となるため、ピッキングの際には、対象物との距離情報が別途必要となる。そのため、物体までの距離が既知な場合や、デプスカメラなどの画像と距離情報を同時に取得できるような状況での利用に限定されてしまう。そこで、画像データのみから物体の位置姿勢を推定することが可能な、2D-3D マッチングによる三次元の物体認識モデルについても検討した。

三次元の物体認識モデルの場合、対象物の 3D モデルをシミュレーション空間上で撮影したデータから 2D-3D の対応関係を学習するものが提案されている [5]。しかし、3D モデルが利用できない場合には適用が難しく、また、個々の物体ごとにモデルを学習する必要があるため、実利用を想定した場合の手間も大きい。そのため、本研究では、2D-3D マッチングのモデルとして、物体の 3D モデルが不要で、物体ごとの学習処理も必要ない OnePose[6][7]と呼ばれる手法を使用した。

OnePose による位置姿勢推定の手順を図 2 に示す。OnePose は 3D モデルを使用しない代わりに、対象物を複数方向から撮影した画像データから Structure from Motion（SfM）により 3D の点群データを事前に作成する必要がある。この際、テクスチャの少ない物体を対象とした SfM では、画像間の特徴点マッチングが不安定になり、十分な点群を生成できない問題がある。そのため、深層学習による画像間マッチングの手法である LoFTR[8]を用いることで、特徴点の検出に依存しない SfM を実現し、点群生成の精度を向上している。

OnePose では、SfM による点群データの生成後、学習処理を行わず、そのまま推論処理を実行することが可能である。推論処理では、まず、入力画像全体から対象物の大まかな位置を予測する。この処理は、SfM で使用した画像群と入力画像との LoFTR によるマッチングにより、一致度の高い部分を切り抜くことで実現している。次に、切り抜いた画像に対して、CNN モデルによる特徴量の抽出を行い、点群データの生成過程で得られた特徴量との 2D-3D マッチングを行う。マッチングには Attention 機構を用いた Transformer ベースのモデルを使用し、特徴量同士の相互の依存関係を広く捉えたマッチングを実現している。最後に、マッチングの出力結果として、画像上の二次元座標に対応付けられた三次元点群座標とのペアが得られるため、得られた対応点を RANSAC によるノイズ点除去を含む PnP アルゴリズムにより処理することで、物体の回転と位置の情報を表す位置姿勢パラメータを推定する。

３．実験

３．１　Mask R-CNN による位置推定

一つ目の実験として、Mask R-CNN を用いたセグメンテーションによる位置推定の検証を行った。本実験では、図 3 に示すような歯車形状のワークを使用し、歯数が 20、25、30 枚の 3 種類と、直径が 30、35、40mm の 3 種類の組み合わせで、合計 9 種類のワークを使用した。なお、

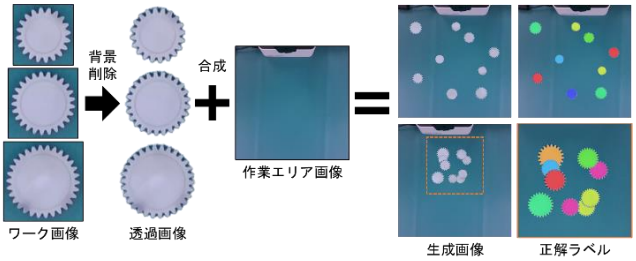

図３　セグメンテーション用データセットの構築手順

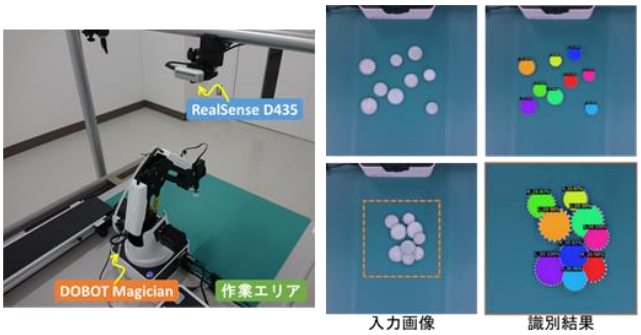

図４　実験環境とセグメンテーション結果

歯車の厚みは全て 7mm となっている。

また、本実験では、推定されたワークの位置情報をもとにロボットアームによるピッキング動作についても検証を行った。

３．１．１　データセットの構築とモデルの学習

セグメンテーションモデルの学習に必要なデータセットの構築を簡易化するため、本実験では、図 3 に示す手順で生成したデータで学習を行った。まず、作業エリアを写した画像と、各ワークを個別に撮影し、ワーク部分のみを切り出した透過画像を準備する。次に、作業エリアの画像を背景に、切り抜いたワーク画像をランダムな位置に貼り合わせる。これにより、合成位置をもとにラベル付けを自動で行いながら、様々な配置パターンを持った学習データを作成した。

Mask R-CNN の学習には、生成した画像 300 枚を使用し、画像サイズは 1080×1080pixel とした。また、特徴抽出器となる CNN モデルには、ImageNet データセットで事前学習済みの ResNet50 を使用した。NVIDIA RTX A6000 を搭載したマシンによる学習完了までの時間は約 40 分であった。

３．１．２　ロボットアームによるピッキング

学習したセグメンテーションモデルを用いたピッキング動作の検証のため、図 4（左）に示す環境で実験を行った。センシングデバイスとしてデプスカメラ（RealSense D435）を作業エリア上部に設置することで、セグメンテーション用の画像と、ピッキングに必要なワークまでの距離情報を同時に取得する。ピッキング用のロボットには 4 軸の小型ロボットアーム（Dobot Magician）を使用し、エンドエフェクタには吸着グリッパを装着している。なお、カメラとロボットの位置関係は事前にキャリブレーションを行っている。

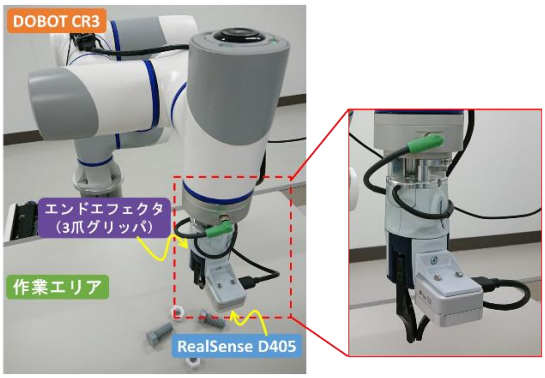

図５　OnePose による位置姿勢推定の実験環境

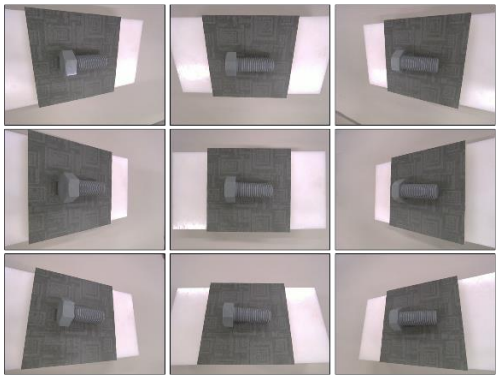

図６　ロボットを動かして撮影した SfM 用の画像例

実際に作業エリアにワークを配置してセグメンテーションを行った結果が図 4（右）である。合成画像のみで学習したモデルであっても、正しくワークの種類と位置が識別できていることが確認できる。

ピッキングの際は、デプスカメラで取得した距離情報をもとに、カメラに近い位置のワークから順にピッキングを行う。セグメンテーションで抽出した領域の重心位置を把持位置とすることで、積み上げられたワークに対しても問題なくピッキングが行えることを確認した。

３．２　OnePose による位置姿勢推定

二つ目の実験として、OnePose を用いた位置姿勢推定の検証を行った。実験環境として、図 5 に示すように、6 軸のロボットアーム（Dobot CR3）と小型のデプスカメラ（RealSense D405）を使用した。デプスカメラはエンドエフェクタ（ARH350A）の先端に固定されており、ロボットの動きに合わせ、把持対象物の詳細な状態をより近い距離から撮影するような実験環境となっている。また、実験には、M16 サイズの樹脂製のボルトとナットの 2 種類のワークを使用した。

３．２．１　SfM による点群データの構築

まず、OnePose による推論処理に必要な点群データを SfM により構築する。SfM の実行には対象物を囲うように撮影した複数の画像が必要となる。そこで、本実験では、カメラがアーム先端に固定されていることを利用し、ロボットを動作させてカメラ位置を調整することで、SfM 用の画像を自動的に収集できるような制御を組み込んだ。実際にボルトを撮影した画像の例を図 6 に示す。

収集した画像を用いた SfM によるボルトとナットの点群データの生成結果を図 7 に示す。ナットの点群については、形状を概ね捉えることができている。一方、ボルトについては、頭部と先端に点群が集中し、ねじ部についてはデータの欠落が大きい結果となった。これは、画像間でのねじ部のマッチングが失敗していることが原因であり、OnePose で使用する学習済みの LoFTR では、ねじ部の特徴を捉えられなかったためと考えられる。なお、SfM の構築には、どちらのワークも 162 枚の画像を使用し、点群生成と 3D 特徴量の抽出に要した処理時間は、3.1.1 項と同一の環境で約 5 分であった。

３．２．２　物体位置姿勢の推定

構築した点群と 3D 特徴量を使用した OnePose によるワークの位置姿勢推定の結果を図 8 に示す。図 8（左）の画像を入力とし、推論処理により出力された位置姿勢パラメータをもとに 3D のバウンディングボックスを描画した結果が図 8（右）となっている。

描画されたバウンディングボックスを確認すると、図 8（上段）の結果では、ワークの位置姿勢ともに正しく推定できていることが確認できる。一方、図 8（中段・下段）の結果では、ワークの位置は概ね特定できているが、ワークの姿勢にズレが生じる結果となった。また、画像を入力後、物体位置の予測処理でワークの大まかな位置を正しく切り出せないものも多くみられ、全体的な推定精度としては不安定な結果となった。

これらの結果より、OnePose による位置姿勢推定では、モデル自体の学習を行わずに推論処理を実行できる利点がある一方、SfM での密な点群の構築や、位置姿勢推定前のワーク位置の切り出しが課題となることが判明した。

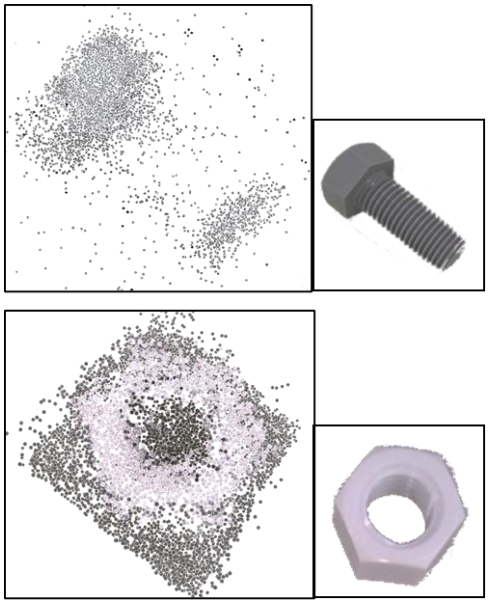

図７　3D 点群データの生成結果
（上：ボルト、下：ナット）

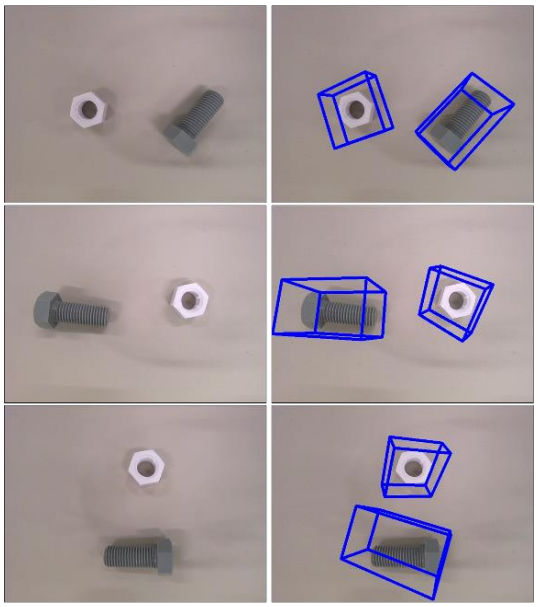

図８　OnePose による位置姿勢推定結果
（左：入力画像、右：推定結果を描画）

４．まとめ

本研究では、ロボットアームによるピッキング動作を行うための AI による物体認識モデルに関して、二次元での位置を把握するためのセグメンテーションモデルと、三次元での位置姿勢を把握するための 2D-3D マッチングモデルについて検討を行った。

セグメンテーションモデルでは、合成画像で学習した Mask R-CNN を使用し、推定した位置情報からロボットアームによるピッキングが行えることを確認した。また、2D-3D マッチングモデルでは、物体の 3D モデルや学習処理が必要ない OnePose による位置姿勢推定の結果や、より精度を高めるための課題についても確認できた。

今後は、より製造工程に近い環境下での物体把握や、人との協調作業の実現に向けた技術開発を行っていく。

【謝　　辞】

本研究の一部は、一般財団法人越山科学技術振興財団の研究助成金により実施させて頂きました。

【参考文献】

1) 橋本学, 精密工学会誌, Vol.87, No.8, pp.666-670, 2021
2) S. Hinterstoisser, et al., Proc. of ECCV, pp.834-848, 2016
3) R. Girshick, et al., Proc. of CVPR, pp.580-587, 2014
4) K. He, et al., Proc. of ICCV, pp.2961-2969, 2017
5) 秋月秀一, 映像情報メディア学会誌, Vol.73, No.2, pp.210-213, 2019
6) J. Sun, et al., Proc. of CVPR, pp.6825-6834, 2022
7) X. He, et al., arXiv:2301.07673, 2023
8) J. Sun, et al., Proc. of CVPR, pp.8922-8931, 2021