

画像生成ネットワークの逆伝播に基づく 繰り返し更新による物体姿勢推定を用いた物体把持

○荒木諒介 真野航輔 大西剛史 (中部大) 平野正徳 ((株) CKD)
平川翼 山下隆義 藤吉弘亘 (中部大)

1. はじめに

ロボットが物体を把持するためには、カメラで撮影した RGB 画像から直接把持位置を推定して世界座標系に変換する方法や、何らかの方法で物体の姿勢を推定し、そこから把持可能な点を紐付ける方法がある。店舗で商品を陳列するロボットは、物体を把持するだけでなく、バーコードや賞味期限が表示されている面を認識する必要があるため、把持対象物体の姿勢を推定する手法が適している。コスト削減の観点では、高精度な距離センサを導入することが難しいため、安価な距離センサや単眼カメラのみで物体の姿勢を推定することが望ましい。

物体の姿勢推定タスクの目標は、カメラ座標を原点とした物体の 3 次元回転を表す行列 R (または四元数 q) と、並進移動量 t を求めることである。本研究では、入力された姿勢パラメタの物体画像を生成するネットワークを構築し、ネットワークの逆伝播に基づく繰り返し処理による物体の 6D 姿勢推定を提案する。はじめに、中間表現ネットワークを用いて、姿勢推定対象物体の画像をセグメンテーションラベルなどの中間表現に置き換える。次に、画像生成ネットワークを用いて、ランダムな姿勢パラメタから中間表現画像を生成する。この 2 つの画像の誤差を画像生成ネットワークに逆伝播し、姿勢パラメタの更新量を求める。この処理を粗い解像度の画像で繰り返し、ある程度の精度が得られたら高解像度の画像で繰り返し、正確な姿勢パラメタを推定する。これにより、単眼カメラの画像のみを用いた姿勢推定を実現する。

2. 従来手法

深層学習の登場以前から、センサから取得した点群と 3D CAD モデルを位置合わせするモデルベースの手法が多く提案されている。モデルベースの手法は、2 つの情報を繰り返し位置合わせすることで、正確な姿勢を推定できる。Iterative Closest Point (ICP) [2] は、最もよく利用される手法のひとつであり、多くの派生アルゴリズムが研究されている。ただし、取得した点群が欠損してしまう場合やノイズが多く含まれる場合は対応点探索ができず、位置合わせができない。そのため、点群の取得が難しい半透明物体や、鏡面反射する物体、黒い物体には適応が困難である。

畳み込みニューラルネットワーク (CNN) の登場以来、深層学習を用いたアピランスベースの姿勢推定手法が増えている。そのほとんどが、多視点から撮影した画像を学習して姿勢推定を行う手法である。Render for CNNs[7] は、CNN を用いて単一の画像に対して物体クラスとカメラ視点を直接推定する手法である。SSD-

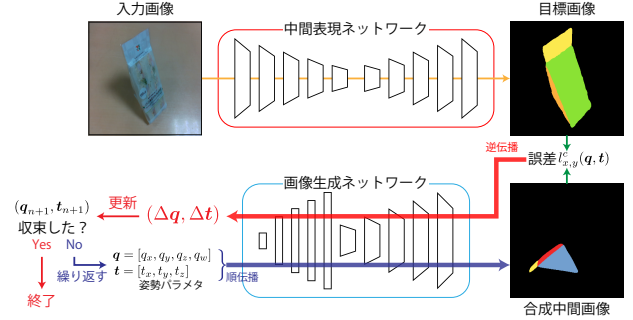


図1 提案手法による姿勢推定の流れ。

6D[4] や Single Shot 6D[8] は、物体検出手法を 6D 姿勢推定用に拡張した手法である。PVNet[5] は、特徴点抽出と Perspective-n-Point の 2 段階で姿勢を推定する。アピランスベースの手法は点群を用いないため、モデルベースの手法では位置合わせが困難な物体に対して頑健である。しかし、物体のテクスチャ、照明や光源位置の変化が姿勢推定に悪影響を与える。そこで、BB8[6] や SSD-6D では、得られた姿勢パラメタを初期値として ICP の処理を行い、精度を向上している。

3. 提案手法

本論文では、画像生成ネットワークの逆伝播に基づく繰り返し処理による、物体の 6D 姿勢推定手法を提案する。提案手法による姿勢推定の流れを図1に示す。

はじめに、任意の物体検出手法を用いて姿勢推定対象の物体を検出する。予め学習した中間表現ネットワークに検出された物体画像を入力して、中間表現画像に変換する (図1のオレンジ色矢印)。ここで得られた画像を目標画像と呼ぶ。次に、ランダムに決められた初期姿勢パラメタ (q, t) を、予め学習した画像生成ネットワークに入力し、その姿勢パラメタの中間表現画像を生成する (図中の青色矢印)。これを合成中間画像と呼ぶ。次に、目標画像と合成中間画像の誤差を求め (図中の緑色矢印)、画像生成ネットワークに逆伝播する (図中の赤色矢印)。これにより、初期姿勢パラメタを把持対象物体の姿勢パラメタに近づける更新量を求める。更新された姿勢パラメタで再度順伝播、逆伝播を行う。これを繰り返して更新量が収束した場合、より高解像度な合成中間画像を用いて再度繰り返し処理を行う。この処理を coarse-to-fine による繰り返し処理と呼び、最終的には正確な姿勢パラメタを推定できる。

3.1 中間表現ネットワーク

中間表現ネットワークは、入力された RGB 画像を中間表現画像へ変換する。中間表現を用いることで、従

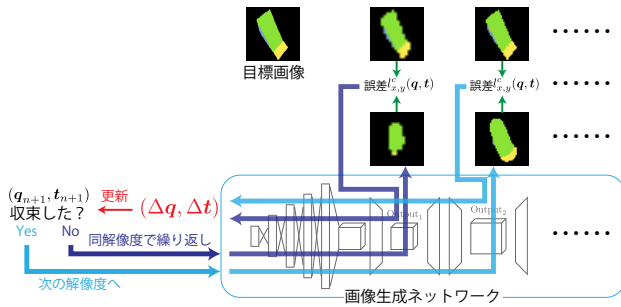


図4 Coarse-to-Fine による繰り返し処理の流れ。

80, 160 × 160, 320 × 320} の誤差であり、クラスごとの二乗誤差で計算する。

3.3.2 教師データと学習方法

教師画像は、中間表現に変換した 3D モデルをランダムな姿勢パラメタで回転および平行移動し、黒色の背景に透視投影した画像である。姿勢パラメタを入力データとして、教師画像を生成するように誤差を最適化する。入力データと教師画像のペアは無限に生成できるが、ここでは 400 万ペアを学習する。最適化関数はモメンタム SGD を用いて、学習率を 0.0001、モメンタムを 0.9 に設定する。

3.3.3 Coarse-to-Fine による繰り返し処理

生成画像の解像度は、位置合わせの精度と推定速度のトレード・オフの関係にある。高速かつ高精度な推定を実現するために、図 4 に示すような coarse-to-fine による繰り返し処理を行う。はじめに、粗い解像度の画像で位置合わせを行い、回転と並進パラメタ両方の更新量が閾値を下回るまで更新し続ける。ただし、片方のパラメタのみ更新量が閾値を下回った場合、そのパラメタを固定してもう一方のパラメタを更新し続ける。次に、先程の処理で得られたパラメタを初期値として、より高解像度の画像で同じように繰り返す。こ

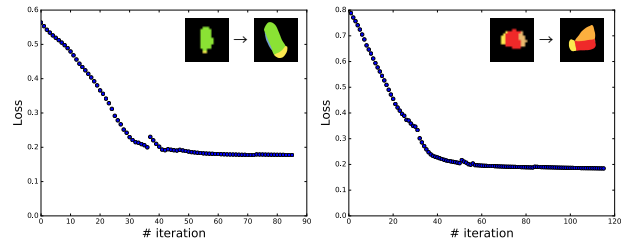


図6 繰り返し処理による誤差の推移。

れを複数解像度に渡って繰り返すことで、単一解像度で姿勢推定するよりも高精度になり、速度の低下も抑制できる。閾値は、画像を数ピクセルだけ変化させたときの勾配を基に決定する。

4. 評価実験

本手法は、コンビニエンスストアで働くロボットが商品の姿勢を推定し、商品棚に陳列するタスクを想定している。そこで、コンビニエンスストアで販売されている 4 つの商品（サンドイッチ、おにぎり、カップスープ、弁当）を用いて物体姿勢推定用のデータセットを作成し、評価に用いる。データセットの画像は、Intel RealSense D435i を用いて撮影されている。評価指標として、各軸に対する回転誤差と並進誤差の平均と標準偏差を用いる。回転誤差は、姿勢パラメタのクォータニオンをオイラー角に変換し、各回転軸に対する正解角度と推定されたオイラー角のなす角で表す。並進誤差は、各軸の並進移動量の差である。

4.1 実験結果

ICP と提案手法の性能比較を行う。正解姿勢とその中間表現画像を 51 ペア用意し、各軸に対する回転誤差と並進誤差の平均と標準偏差を比較する。比較結果は表 1 に示す。4 つのアイテムについて、提案手法による姿勢推定が高い精度を達成した。また、提案手法の

表 1 データセットでの精度比較。c2f は Coarse-to-fine による繰り返し処理を示す。括弧内は標準偏差である。

対象物体	手法	解像度 [pix]	R の誤差 [deg]			t の誤差 [mm]		
			R_x	R_y	R_z	t_x	t_y	t_z
サンドイッチ	ICP		41.4 (38.0)	26.5 (18.2)	94.2 (51.9)	7.2 (15.3)	5.3 (7.1)	3.4 (1.9)
		320	22.9 (22.3)	12.9 (17.7)	16.4 (34.7)	4.0 (4.1)	3.0 (2.4)	11.5 (11.0)
		c2f	25.1 (20.9)	13.7 (17.4)	17.4 (31.8)	3.4 (3.7)	3.0 (2.4)	12.7 (9.8)
おにぎり	ICP		82.4 (43.8)	44.4 (28.3)	85.3 (56.2)	2.6 (1.6)	2.8 (1.5)	3.9 (2.8)
		ours	54.0 (35.0)	33.3 (28.1)	62.7 (64.2)	2.5 (1.6)	2.8 (1.6)	5.9 (5.9)
		c2f	54.9 (51.0)	53.4 (38.1)	84.6 (56.5)	5.3 (10.3)	5.2 (6.6)	6.2 (6.4)
カップスープ	ICP		54.9 (51.0)	53.4 (38.1)	84.6 (56.5)	5.3 (10.3)	5.2 (6.6)	6.2 (6.4)
		ours	25.8 (39.0)	73.6 (36.3)	20.9 (50.2)	2.3 (2.2)	3.1 (3.3)	9.3 (17.4)
		c2f	25.8 (39.0)	73.6 (36.3)	20.9 (50.2)	2.3 (2.2)	3.1 (3.3)	9.3 (17.4)
弁当	ICP		52.1 (41.6)	46.1 (30.7)	82.1 (55.8)	2.6 (1.5)	2.4 (1.3)	4.0 (3.0)
		ours	53.2 (48.2)	46.7 (31.6)	80.4 (63.7)	6.5 (5.5)	6.1 (5.2)	21.6 (14.4)
		c2f	53.2 (48.2)	46.7 (31.6)	80.4 (63.7)	6.5 (5.5)	6.1 (5.2)	21.6 (14.4)

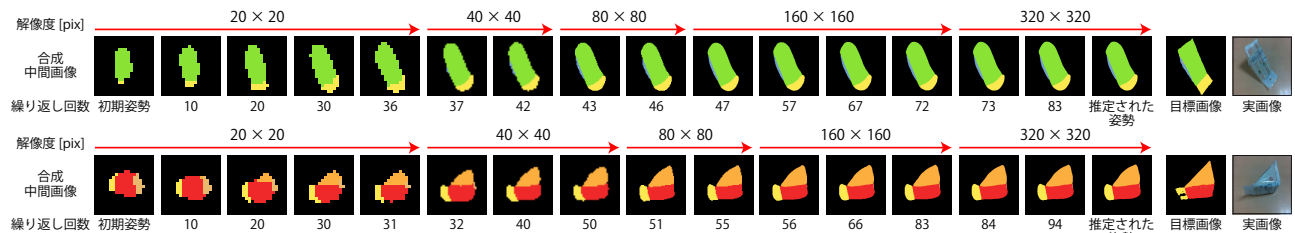


図5 繰り返し処理による姿勢推定の様子。

表 2 把持実験の結果.

入力画像									
Segmentation 結果									
姿勢推定結果									
把持成功	初回	✓	✓	✓	✓	×	×	×	×
	持替	-	-	-	✓				
陳列成功		✓	✓	✓	✓				

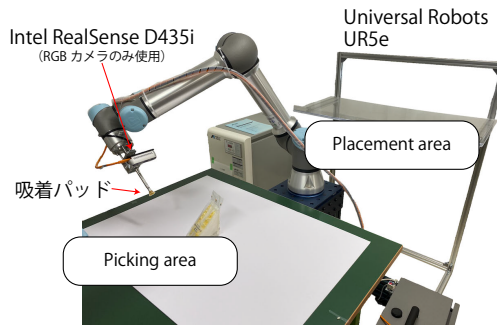


図 7 ロボットシステムの全体像.

標準偏差も ICP と比較して小さいことから、提案手法は安定して姿勢を推定できる。

Coarse-to-fine による繰り返し処理の有効性を確かめるために、サンドイッチを用いて単一の解像度のみで姿勢推定を行った。精度の比較は表 1 内に示す。推定速度は、 320×320 の場合は 19.5 秒、Coarse-to-fine の場合は 9.4 秒である。Coarse-to-fine による繰り返し処理は、単一の解像度のみで姿勢推定を行った場合と同等の性能を得ながらも、推定速度を向上した。

繰り返し処理による姿勢推定の様子を図 5 に、誤差の推移を図 6 に示す。初めは低解像度で姿勢推定を行い、徐々に解像度を上げて目標姿勢に近づいていることがわかる。

4.2 アームロボットを用いた把持実験

提案手法で得られた姿勢推定結果から適切な把持位置を求められることを確認するために、アームロボットを用いた把持実験を行う。ロボットシステムは図 7 に示すように、UR5e を用いる。ビジョンセンサとして Intel RealSense D435i を搭載するが、距離情報は用いない。実験は、ランダムな姿勢で配置されているコンビニ商品（サンドイッチ）を、正しい姿勢で棚に陳列するタスクを行う。本手法は、物体検出を用いて物体のおおよその位置を検出することが前提であるため、Picking area には単一の物体を配置する。物体の姿勢を推定し、見えている領域が最も広い面の法線方向にアプローチして吸着把持を行う。このとき、物体の正面を把持できない場合は正しい姿勢で陳列できないため、物体を持ち直して正面を把持する。

表 2 に、把持実験の結果を示す。表の上部では、8 パターン用意した姿勢に対して、入力画像、セグメンテーション結果と姿勢推定結果の画像を示す。表の下部で

は、上記の姿勢推定結果を用いた場合の把持成功可否、持ち替えの成功可否、陳列の成功可否を示す。左側 4 つの結果は、陳列が成功したケースである。3 つ目のサンドイッチは種類が異なるが、セグメンテーションを用いることでテクスチャの差異を吸収し姿勢推定と陳列に成功した。4 つ目は物体の正面を把持できないため、背面を把持して一度持ち直すことにより陳列に成功した。しかし、姿勢によっては推定結果が大きく異なるケースや、推定結果がおおよそ正しくても把持できないケースが確認された。

5. おわりに

本研究では、画像生成ネットワークの逆伝播に基づく繰り返し更新による物体の姿勢推定手法、およびそれを用いた物体把持手法を提案した。今後は、LINEMOD などの姿勢推定データセットで評価して、高精度化を目指す。

謝 辞 本研究は、JSPS 科研費 JP20J15720 の助成を受けたものである。

参 考 文 献

- [1] Badrinarayanan, V. *et al.*: “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation”, T-PAMI, Vol. 39, No. 12, pp. 2481–2495, 2017.
- [2] Besl, P. J. and McKay, N. D.: “A Method for Registration of 3-D Shapes”, T-PAMI, Vol. 14, No. 2, pp. 239–256, 1992.
- [3] Karras, T. *et al.*: “Progressive Growing of GANs for Improved Quality, Stability, and Variation”, ICLR, 2018.
- [4] Kehl, W. *et al.*: “SSD-6D: Making Rgb-Based 3D Detection and 6D Pose Estimation Great Again”, ICCV, pp. 1521–1529, 2017.
- [5] Peng, S. *et al.*: “PVNet: Pixel-wise Voting Network for 6DoF Pose Estimation”, CVPR, pp. 4561–4570, 2019.
- [6] Rad, M. and Lepetit, V.: “BB8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects without Using Depth”, ICCV, pp. 3828–3836, 2017.
- [7] Su, H. *et al.*: “Render for CNN: Viewpoint Estimation in Images Using CNNs Trained with Rendered 3D Model Views”, ICCV, pp. 2686–2694, 2015.
- [8] Tekin, B. *et al.*: “Real-Time Seamless Single Shot 6D Object Pose Prediction”, CVPR, pp. 292–301, 2018.