Transformer モデルを用いた 2D 手書き指示による把持位置推定

ER20066 野田修平　　指導教授：藤吉弘亘

1. はじめに

生活支援ロボットの実現には，誰でも簡単に操作可能なインターフェースが重要である．岩永らは，人の手書きのスケッチ画像からロボットの把持位置を CNN で推定する手法を提案している [1]．シーンの深度画像とスケッチ画像を同時に入力して畳み込み処理を行うため，スケッチの位置のみを変更した場合に対応できない場合がある．そこで，本研究では Transformer[2] を用いた Encoder-decoder モデルによる把持位置推定を提案する．Encoder では，深度画像からシーンの中の各物体を解析する．Decoder では，Encoder のシーン解析を結果とスケッチ画像との関係性を Cross-Attention により求めることで，物体とスケッチの位置関係を正確に算出する．

2. スケッチ画像を用いた把持位置推定

岩永らは，簡単かつ直観的な指示が可能なスケッチ画像を用いた把持位置推定を提案している [1]．岩永らのモデル構造は ResNet18 をベースにしており，入力はシーンの深度画像と物体に対して書かれたスケッチ画像をチャンネル方向に重ね合わせた 2 チャンネル画像である．出力は，グリッパの手首，左指先，右指先の 3 点の 3 次元座標である．本手法は，深度画像とスケッチ画像のペアをあらかじめ用意する必要がある．しかし，物体を把持する位置は 1ヶ所とは限らない．そのため，学習していない深度画像とスケッチ画像のペアを入力した場合，位置関係を正確に理解できない問題がある．

3. 提案手法

本研究では，Transformer をベースとする把持位置推定モデルを提案する．提案手法のモデル構造を図 1 に示す．まずシーンの深度画像を 3 層の畳み込み層で処理する．そして，得られた特徴マップに Positional Encoding を加えたものを Encoder に入力する．Encoder では物体や床の位置関係を特徴量として抽出する．スケッチ画像は 3 層の畳み込み層で処理した後，全結合層でベクトルに変換して Decoder に入力する．出力は，グリッパの手首の位置の 3 次元座標と姿勢のクォータニオンである．Decoder では，Encoder で得た特徴量とスケッチの入力データとの Cross-Attention を求めることで，正確かつ汎化性能の高い把持位置推定を実現する．

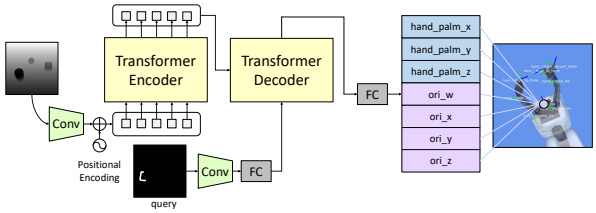

図 1: スケッチ画像を入力とする Transformer モデル

4. 評価実験

従来手法 [1] と提案手法の比較により，有効性を検証する．

4.1. 実験条件

学習の最適化手法には Adam，エポック数を 200，バッチサイズを 128，学習率を 0.0001 として学習を行う．損失関数は位置の 3 次元ユークリッド距離の誤差と姿勢の平均二乗誤差の和を用いる．

4.2. データセットの作成

本実験では Unity 環境で作成したデータを用いる．把持対象物体の深度画像は，3 種類の物体をランダムな位置に出現させ，オブジェクトの位置に応じてカメラの角度を変えて撮影した画像である．スケッチ画像はグリッパを模したコの字型オブジェクトを把持対象物体の周辺に配置し，スケッチ線を中心に円を配置した画像とする．1 枚の深度画像に対して異なる 10 通りのスケッチを用意する．データ数は それぞれ訓練データ 104,000 枚，検証データ 13,000 枚，テストデータ 13,000 枚である．

4.3. 実験結果

表 1 に従来手法と提案手法の位置と姿勢の推定誤差の平均値と最小値，最大値を示す．表 1 から平均値は同程度であるが，提案手法の最大値が大きく低下していることが確認できる．このことから従来手法では対応できなかった場合も対応できるようになったと考えられる．この例を図 2, 図 3 に示す．

表 1: 誤差の比較

	平均値	最大値	最小値
従来手法	0.060	**2.168**	0.007
提案手法	0.061	**0.619**	0.004

図 2 は従来手法による推論結果の例，図 3 は提案手法による推論結果である．左 2 枚と右 2 枚はそれぞれ同一のシーンに異なるスケッチを与えた時の結果である．スケッチの正解姿勢をオレンジ色，出力結果を青色のオブジェクトで表す．左 2 枚のシーンでは，従来手法は把持できない誤った位置と姿勢を予測しているが，提案手法は把持可能な位置を予測できた．以上により，提案手法は様々なスケッチ入力に対応可能であるといえる．

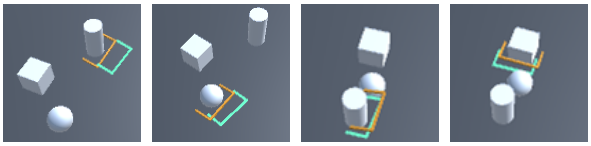

図 2: 従来手法による推論結果

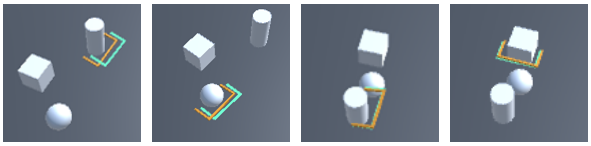

図 3: 提案手法による推論結果

Attention の可視化結果の例を図 4 に示す．Encoder では，シーン全体を注視している．一方で，Decoder ではスケッチした指示周辺の物体を注視していることが確認できる．

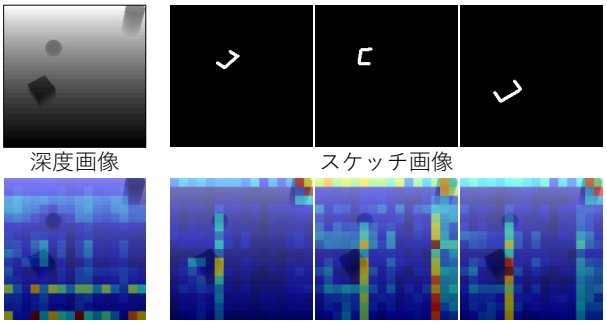

図 4: Attention の可視化結果

5. おわりに

本研究では，物体とスケッチの位置関係を正確に求めるため，Encoder-decoder 型の Transformer モデルを用いて把持位置推定を行い，汎化性能向上の効果を確認した．今後は推定した把持位置を用いたロボット動作の生成を行う予定である．

参考文献

[1] 岩永優香 等，“2D 手書き指示でロボットに人の意図を伝えるインタフェースの開発と評価～深層学習を用いた把持位置姿勢指示手法の検討～”，日本ロボット学会学術講演会，2022.

[2] A.Vaswani, *et al.*,“ *Attention is all you need* ”, *NearIPS, 2017.*