Transformer モデルを用いたスケッチ指示による把持位置推定

○野田修平　平川翼　山下隆義　藤吉弘亘（中部大学）

生活支援ロボットの容易な操作手法としてスケッチ指示がある．指示画像と把持対象物体の画像からロボットの把持位置を推定する従来手法では，スケッチ位置のみの変更に対応できない場合がある．本研究では Transformer 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し，Decoder では，Encoder のシーン解析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．

1. はじめに

少子高齢化社会による労働力不足問題の解決策として，Domestic Service Robot（DSR）の社会実装が考えられる．DSR は人間とのコミュニケーションを通じて指示を理解し，適切に動作する必要がある．しかし，ロボットの社会実装には特別な装置やその専門知識が必要である．そこで，誰もが簡単かつ直感的にロボットを操作可能となるインターフェースが求められている．

スケッチによる指示は，専用の装置や知識を用いることなくロボットに指示する手法の一種である．スケッチ指示は，タブレットなどのデバイスにロボットが撮影した画像を表示し，画像に合わせた指示を描き込むことで，状況に応じた正確な指示を送ることができる．これを実現するには，画像を適切に処理し，スケッチを正確に認識する必要がある．

従来，画像認識には ResNet[1] などの CNN ベースモデルが用いられてきたが，近年では Transformer[2] が注目されており，その応用範囲は画像認識にも広がっている．Vision Transformer（ViT）[3] は，その代表的な例であり，画像のパッチを入力として処理することで，従来の CNN ベースモデルを超える性能を示している．

本研究では，Transformer モデルを用いたスケッチ指示による物体の把持位置推定手法を提案する．従来の CNN ベースモデルでは，深度画像とスケッチ画像を重ね合わせた画像を畳み込んで用いているため，それぞれの画像に対する関係性を正確に求めることができず，スケッチ位置のみの変更に対応できない場合があった．そこで，Encoder-Decoder 型の Transformer モデルを導入し，深度画像を Encoder に入力，スケッチ画像を Query として Decoder に入力することで，Cross-Attention を求め，それぞれの画像の位置関係を正確に算出する．

また，Transformer モデルの学習には適切なデータセットが重要であるため，Unity を用いてデータセットを作成した．従来の CNN ベースモデルでは深度画像とスケッチ画像を重ねて入力するが，提案手法ではそれぞれの画像を分けて入力するため，深度画像 1 枚に対して 10 種類のスケッチ画像を生成した．人間が描いたスケッチに近づけるため，スケッチ画像の作成には細線化と円の配置処理を行った．

実験では，CNN ベースモデルと比較することで有効性を検証する．評価は，正解ラベルとの誤差と把持成功率による定量的評価と，グリッパの推定結果の可視化と Attention の可視化による定性的評価を行う．

2. 関連研究

DSR を人間の生活空間で動作するための，専門的な知識がなくとも誰でも容易に操作を可能にする手法について述べる．代表的な手法として，自然言語指示とスケッチ指示がある．

2.1 自然言語指示によるロボット操作

近年，人間が自然言語を用いてロボットに指示を与え，ロボットの行動を生成する研究が進んでいる．これは自然言語処理技術とロボット制御技術を組み合わせたもので，直感的なコミュニケーション手段として期待されている．ロボットが人間の曖昧な指示を理解し，行動に変換することが求められるが，言語の複雑さや文脈による意味の変化までをロボットに理解させることは容易ではない．これらを克服するため，自然言語処理技術の重要性が増し，様々な研究が行われている．

Brohan らは，多様なタスクに対応可能なロボット学習モデルである Robotics Transformer (RT-1) を提案した [4]．自然言語処理分野では，多様性の高い大規模データセットと表現力の高いモデルの活用が進歩をもたらしているが，ロボット分野では同様の成果はまだ十分に得られていない．ロボット分野における大規模データセットは限られており，多様なロボットに対応できるモデルの作成が困難である．RT-1 は，ロボットのカメラからの画像と自然言語のタスク記述を入力とし，トークン化されたアクションを直接出力する Transformer 構造で構成されている．また，RT-1 は 17 ヶ月間に 13 台のロボットを用いて収集された 700 以上のタスクと 13 万エピソードからなる大規模データセットで学習されており，特定の下流タスクをゼロショットまたは小規模なデータセットで高いパフォーマンスを達成できる．

Korekata らは，自然言語による指示に従い日常生活で使用する物体を指定された目的地まで運搬する DSR の研究を行っている [5]．ロボット分野における既存のマルチモーダル方法は，ターゲットオブジェクト候補と目的地候補のすべての組み合わせに対する推論を必要とするため，計算複雑性の観点から実用的ではない．そこで，Korekata らは単一のモデルを用いてターゲットオブジェクトと目的地を個別に予測する Switching Head-Tail Funnel UNITER を提案した．これは，ターゲットオブジェクトと目的地を個別に予測することでタスクを解決するモデルである．モデルの性能に関しては，実環境に近いシミュレータ環境で作成したデータセットを用いて実験されているが，言語理解の精度の観点からベースライン方法を上回ることを示している．

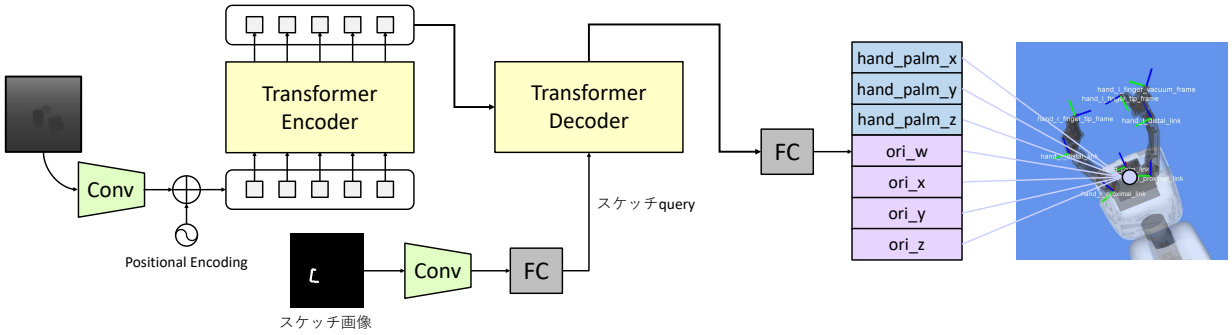


図 1　提案手法のモデル構造

2.2　スケッチ指示によるロボット操作

スケッチ指示は自然言語指示よりも視覚的であり，ロボットに対して直感的なコミュニケーションが可能である．人間は物体や動作の概念をタブレットなどのデバイス上で描いて示すことで，複雑な指示を正確に伝えることが可能である．自然言語指示と比べて直感的かつ視覚的に情報を伝えることが可能であるが，ロボットがスケッチを正確に理解し，それに応じて行動する必要がある．

Lin らは，グラフベースのスケッチを用いて，スケッチで描かれた物体に関連する把持形状を生成する方法で把持方向学習を行った [6]．グラフベースのスケッチを用いることで，個人差によるスケッチの曖昧さやばらつきを効率的に低減することが可能である．また把持方向学習に関しては，多クラス分類タスクとして扱い，5 次元の把持パラメータを回帰する手法を用いた．スケッチの抽象さや曖昧さを考慮に入れ，スケッチの構造を組み込むことでグラフの利点を活用し，表現能力を強化している．実際の人のスケッチやロボット実機を用いた実験でも有用性が示されている．

岩永らは，ピックアンドプレースタスクにおいて，スケッチ指示を用いた把持位置姿勢推定を CNN ベースモデルを用いて行った．[7][8]．データセットには，0 から 255 の範囲で正規化した深度画像と，把持対象物体に対する Human support robot(HSR) のグリッパを模したコの字型の手書き線をチャンネル方向に重ね合わせ，指定の領域で切り取った画像を用いた．テストデータと人の描いたスケッチによる評価が行われたが，人の書いたスケッチに対しては，正確な推論が困難であった．

3.　提案手法

生活支援ロボットの実用化には，簡単かつ直感的な操作手法の実現が必要である．そこで，本研究では岩永らの CNN ベースモデルのスケッチ指示手法に対して，Transformer モデルを導入することで，正確かつ汎化性能の高い把持位置推定を実現する．

CNN ベースモデルを用いた従来手法では，深度画像とスケッチ画像をチャンネル方向に重ね合わせた 2 チャンネルの画像を入力として用いる．しかし，深度画像とスケッチ画像を重ね合わせた画像を畳み込んで用いており，早い段階で 2 種類の特徴量が混合するため，それぞれの画像に対する関係性を正確に求めることができず，一部のスケッチパターンにおいて正確に推論できない場合があった．また本手法では，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が，物体を把持する位置は 1 か所とは限らない．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．

これを解決するために，本研究では Transformer ベースモデルを導入し，Encoder に深度画像を入力することでシーンの解析を行い，Decoder で Encoder で解析したシーンの結果とスケッチ画像間の Cross-Attention を求めることで把持位置を予測する．図 1にモデル構造を示す．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．最後に Decoder からの出力を FC 層を用いて把持位置姿勢 7 次元の値を出力する．これにより，早い段階で特徴量が混合することなく，それぞれの画像の関係性を適切に求めることが可能である．また，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に対しても，深度画像の特徴を Encoder，深度画像とスケッチ画像の関係性を Decoder で捉えるため，従来手法に比べて柔軟な対応が可能である．

4.　データセットの作成

Transformer モデルにおいて，どのようなデータセットを用いて学習するかは非常に重要である．適切なデータセットを用いて学習するために Unity を用いて，データセットを作成した．CNN ベースモデルでは，深度画像とスケッチ画像を重ねてモデルに入力しているため，1 枚の深度画像に対して 1 枚のスケッチ画像が必要である．しかし，提案手法ではそれぞれの画像を分けてモデルに入力するため，1 枚の深度画像に対して複数枚のスケッチ画像を用いることが可能であり，本研究では深度画像 1 枚に対して 10 種類のスケッチ画像を生成した．実験に用いる画像は深度画像と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り，そ

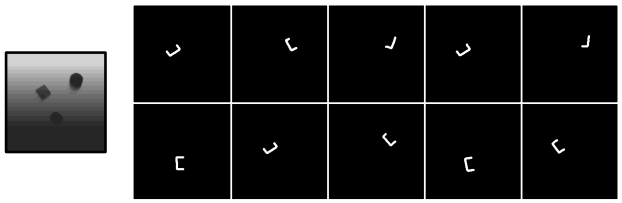


図 2　深度画像 1 枚に対する 10 種類のスケッチ画像

図 3　線の太さを一定にする処理の工程

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

れぞれ 224×224 ピクセルの画像である．2 つの画像の組み合わせの例を図 2に示す．画像の枚数は訓練データが 104,000 枚，検証データが 13,000 枚，テストデータが 13,000 枚である．

スケッチ画像は，従来手法に用いられたデータ同様に，物体を把持するようにコの字型 3D オブジェクトを配置して撮影するが，今回のデータセットの作成においては，その後に細線化と円の配置を行う処理を追加した．処理の工程を図 3に示す．Unity 上のスケッチはオブジェクトであるため，奥行が存在し，撮影の際にカメラから遠いほど細くなってしまう．そのため，線の太さを一定にする処理を行う．まず，撮影したスケッチ画像に対して細線化処理を行う．その後，細線化した線を中心に半径 2 ピクセルの円を配置する．この工程を行うことで線の太さが一定になり，角が丸みを帯びるため，より人間が描いたスケッチに近づけることが可能である．

5.　評価実験

本章では，提案手法の有効性を検証するために評価実験を行う．作成したデータセットを用いて CNN ベースモデルと Transformer ベースモデルで学習し，結果を比較する．CNN ベースモデルには ResNet18 を用いる．学習条件は，バッチサイズは 128，学習率は 0.001，エポック数は 200 である．また，位置の誤差の算出には 3 次元ユークリッド距離，姿勢の誤差の算出には平均二乗誤差を用いて，その和を損失関数とする．

5.1　定量的評価

推論時の真値との誤差と把持成功率による定量的評価を行う．把持位置は推論結果と真値との 3 次元ユークリッド距離，姿勢は平均 2 乗誤差を用いて評価する．推論時の各誤差を表 1に示す．平均値は僅かに従来手法の精度が提案手法を上回っている．一方，最大値は従来手法に比べて提案手法の値が小さくなった．このことから，物体やスケッチの配置次第では従来手法では対応できない場合があるが，提案手法では対応が可能であることが確認できた．従来手法では入力画像をチャンネル方向に重ねて畳み込む形式であり，深度画像とスケッチ画像の関係性を正確に理解できていないが，提案手法では，Encoder で深度画像の情報を的確に解析し，Decoder でスケッチ画像との関係性を算出することで，様々な場合に対応できるようになったためと考えられる．

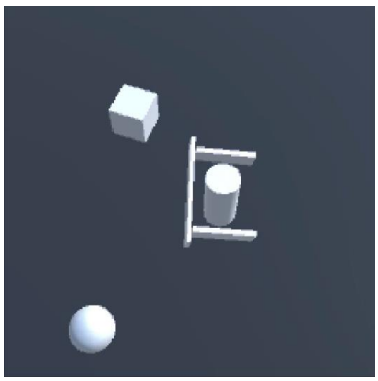


図 4　把持実験環境

表 2　把持成功率

	成功回数 [回]	失敗回数 [回]	成功率 [%]
従来手法	10,310	2,690	79.3
提案手法	10,618	2,382	81.7

また，データセット作成時に用いた Unity 環境を使用して，モデルの出力結果に応じた位置姿勢でグリッパを配置し，把持対象物体も同時に配置して把持実験を行う．実験環境を図 4に示す．グリッパを模したコの字型のオブジェクトの先端部分を，物体を挟むように移動させ，物体を把持できた場合を成功とした．テストデータの 13,000 通りの配置で検証し，成功回数と失敗回数を記録した．実験結果を表 2に示す．提案手法では，従来手法に比べて把持率が 2.4% 上昇した．このことからも一部の従来手法では対応できない物体やスケッチの配置に対しても，対応が可能であることが確認できた．

5.2　定性的評価

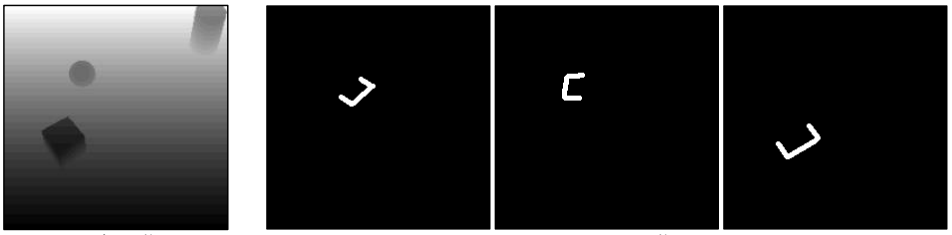
Attention の可視化とグリッパの位置と姿勢の推定結果の可視化による定性的評価を行う．図 5にグリッパの推定結果を示す．左 2 枚と右 2 枚はそれぞれ同一のシーンに異なるスケッチを与えた時の結果である．スケッチの正解位置姿勢をオレンジ色，出力結果を青色のオブジェクトで表す．左 2 枚のシーンでは，従来手法は把持出来ない誤った位置と姿勢を予測しているが，提案手法は把持可能な位置を予測できた．以上により，提案手法は様々なスケッチ入力に対応可能であるといえる．

提案手法で獲得した Attention を図 6に示す. Encoder 側では画像下部に強い Attention，画像全体に弱い Attention が確認できる．Decoder 側ではスケッチに対応した Attention が確認できる．しかし，後方の円柱周辺などの適さない場所にも Attention が確認できる．この結果から，スケッチ画像の特徴そのものは捉えられているが，Encder 側の Attantion の影響や線形層の影響で

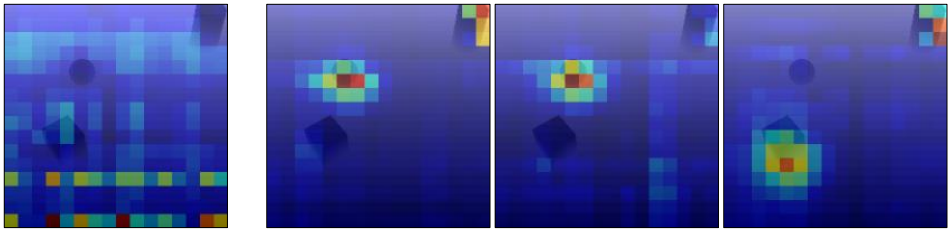
CNNベースモデルによる推論結果

Transformerモデルによる推論結果

図 5　グリッパの推定結果の可視化



深度画像　　スケッチ画像



Encoder　　Decoder

図 6　深度画像と各スケッチに対する Attention

不適切な Attantion も発生していると考えられる.

6.　おわりに

本研究では,物体とスケッチのそれぞれの位置関係を正確に求めるための Encoder-decoder 型の Transformer モデルによる把持位置推定手法を提案した. Transformer モデルを用いることで，従来手法の CNN ベースモデルに比べて汎化性能の向上を実現した. また, Attention の可視化を行い, Encoder がオブジェクト全体を, Decoder がスケッチの特徴を捉えていることを確認した．今後は推定した把持位置を用いたロボット実機による把持実験を行う予定である.

参　考　文　献

[1] A. Brohan, et al：“Deep Residual Learning for Image Recognition”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

[2] A. Vaswani, et al：“Attention is All you Need”, Advances in Neural Information Processing Systems, vol. 30, pp. 5998–6008, 2017.

[3] A. Dosovitskiy, et al：“An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”, arXiv preprint arXiv:2010.11929, 2021.

[4] A. Brohan, et al：“RT-1: ROBOTICS TRANSFORMER FOR REAL-WORLD CONTROL AT SCALE”, arXiv:2212.06817, 2022.

[5] R. Korekata, et al：“Switching Head–Tail Funnel UNITER for Dual Referring Expression Comprehension with Fetch-and-Carry Tasks”, IEEE/RSJ International Conference on Intelligent Robots and Systems, 2023.

[6] H. Lin, et al：“I Know What You Draw: Learning Grasp Detection Conditioned on a Few Freehand Sketches”, IEEE International Conference on Robotics and Automation, 2022.

[7] 岩永ら：“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.

[8] 岩永ら：“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～物拾いタスク指示手法に関するパイロットスタディ”, 日本ロボット学会学術講演会, 2023.