配置物体を考慮したAttention機構マルチモーダルネットワークによる衝突予測

○河合竹彦　平川翼　山下隆義　藤吉弘亘（中部大学）　杉浦孔明（慶応義塾大学）

1. はじめに

高齢者や障碍者の生活支援は人手不足が問題になっている．そのため，家庭で家事や介護などの生活支援を行う，Domestic Service Robot (DSR) の需要が高まっている [1]．DSR が生活支援を行う場合の基本的な動作として，物体を移動させる物体運搬タスクが挙げられる．物体運搬タスクは把持タスク，物体運搬タスク，物体配置タスクの３つに大別できる．DSR が行う物体配置タスクは，家庭環境で行われ，工業用ロボットと異なり，配置エリアに多様な物体が乱雑に配置され，障害物となる場合がある．このような状況で物体配置タスクを行うには，事前に有害な衝突が発生するかを予測し，衝突を防ぐ必要がある．この物体配置タスクにおける衝突予測を行う手法として PonNet [2] が提案されている．PonNet は乱雑に障害物が配置された配置エリアに，把持中の物体を配置する際の衝突を予測する．ネットワークの入力は，ロボットのヘッドカメラで撮影した RGB-D 画像と，把持物体の大きさ情報とカメラポーズを用いる．把持物体の大きさ情報を利用するのは，物体把持タスク時に物体認識を行うため，モデル情報が既知であることを前提としている．しかし，この前提では未知物体を把持した場合や，物体運搬中に把持物体の姿勢が変化した場合に対応できない．

そこで，本研究では，未知物体を含む物体配置タスクにおける衝突予測において，配置物体を考慮した Attentin 機構マルチモーダルネットワークを提案する．この問題設定のために，配置物体の姿勢変化を追加した衝突予測データセットをシミュレータで作成した．このデータセットには，ロボットのヘッドカメラで配置エリアを撮影した RGB-D 画像，ハンドカメラで配置物体を撮影した RGB 画像，物体配置後の衝突統計情報が含まれる．これらを Attention 機構を導入したマルチモーダルネットワークに入力し，有害な衝突を予測する．また，Attention 機構により衝突予測の判断根拠を説明可能にする．

2. 関連研究

物体配置タスクの衝突予測手法として，マルチモーダルな Attention 機構と Self-attention 機構を持つ手法 [2] がある．この手法は家庭環境のような乱雑に障害物が配置された状況での衝突予測のためのデータセット BILA-12k の作成と Attention 機構と Self-attention 機構を持つマルチモーダルネットワーク（PonNet）を提案している．PonNet は従来の平面検出による衝突予測手法 [6] に比べ高精度な衝突予測が可能である．BILA-12k データセットは SIGVerse [4] を用いて開発したシミュレータ（PonNet シミュレータ）を使用している．PonNet シミュレータは RoboCup@Home [3] の DSR 競技環境で使用された Unity ベースのシミュレータを基に開発している．RoboCup@Home の DSR 競技環境は,「World Robot Summit 2018 Partner Robot Challenge/Virtual Space」の大会で使用された SIGVerse [4] をベースとして使用しており，写実的でリアルな環境と物理演算が可能で，ROS アーキテクチャに沿ってロボットをエミュレートできる．このシミュレータを使用することで，実機ロボットで必要な様々なシーンの再現や教師ラベルの作成を，自動化できる．PonNet シミュレータでは，衝突予測のようなロボットや環境を破壊する可能性のある危険なタスクを再現でき，配置タスクを行う際は，物理演算により連鎖的な衝突や，配置エリアから転がり落ちる衝突などを再現可能である．ロボットは，Human Support Robot(HSR) を使用する．HSR は，DSR として開発された単腕ロボットであり，ヘッドに RGB-D カメラとハンドに RGB カメラを搭載している．BILA-12k データセットには，背景と照明が異なる 12 環境と 6 種類の家具が存在する環境において，障害物 28 種類をランダムに配置し，12,000 サンプルの衝突統計を収集している．データセットには，ロボットのヘッドカメラから撮影した RGB-D 画像と配置物体のサイズ情報，カメラポーズ，衝突統計情報が含まれる．

PonNet は，RGB-D 画像を入力とする 2 ストリームのマルチモーダル CNN をベースに Attention Branch Network (ABN) [5] の Attention Branch 構造を採用し，ネットワークの注視領域を可視化している．注視領域を可視化することでネットワークが入力画像のどこを見て衝突の有無を判断しているか視覚的に確認でき，解釈性の向上と精度向上が可能である．
また，ネットワークの入力には，RGB-D 画像とメタ情報として配置物体の大きさ情報とカメラポーズを含む 4 次元ベクトルを用いる．入力の RGB-D 画像は，ロボットのヘッドカメラで物体配置前の配置エリアを撮影している．このとき，配置物体の大きさ情報を入力可能であるのは，物体を把持した時点で物体検出等を行い，物体のモデル情報が既知である前提のもとに配置物体の情報を利用するためである．また，配置物体の大きさ情報は，3 次元ベクトル x, y, z を正規化して使用する．カメラポーズは，配置エリアの高さ情報であり，配置エリアが低い位置の場合は 0, 高い位置の場合は 1 とする．ベースネットワークには，ResNet-18 を用いている．RGB 画像は，中央 192×192 をクロップし，Depth 画像は法線マップへ変換したのち中央 192×192 をクロップをしている．RGB 画像と Depth 画像は，特徴抽出後に Self-attention 機構により重み付けし結合する．

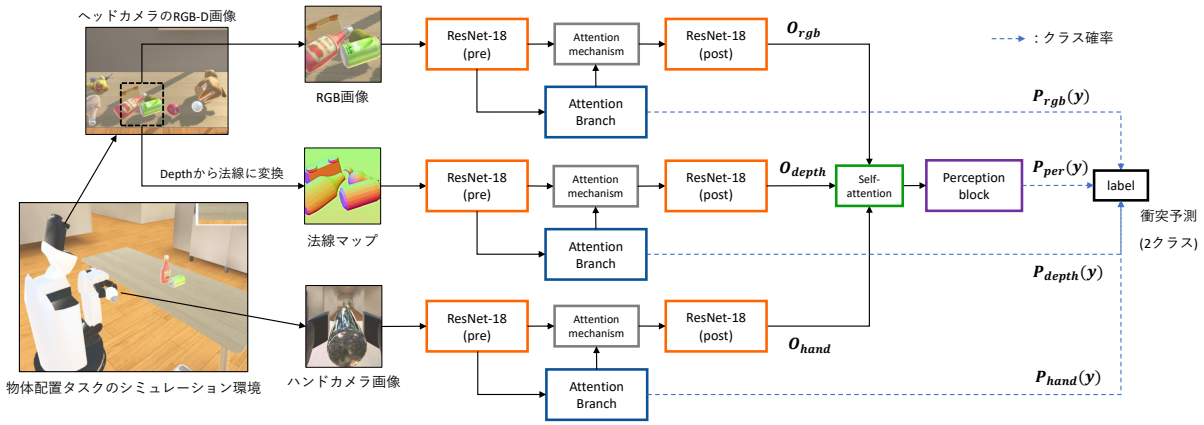

図 1: 提案手法の流れ

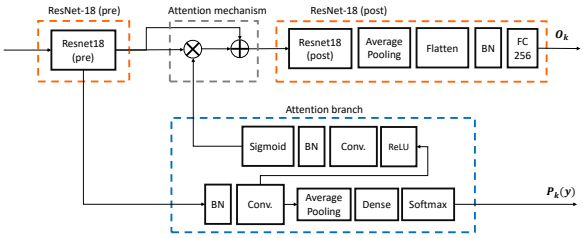

図 2: Feature Extractor の構造

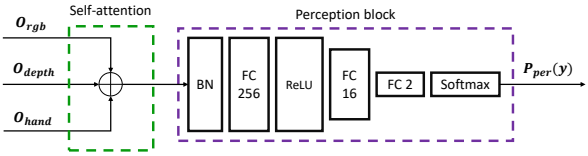

図 3: Classifier の構造

3. 提案手法

本研究では，ハンドカメラによる配置物体の考慮を追加したモデルフリー Attention 機構マルチモーダルネットワークを提案する．提案手法の流れを図 1 に示す．また，配置物体の姿勢変化を追加した物体配置タスクにおける衝突予測データセットを作成した．PonNet は，配置物体が既知でモデル化されていることを前提にしているため，未知物体でモデル情報がない物体に対応できない問題がある．この問題に対して，本手法はハンドカメラを追加したマルチモーダルネットワークを使用し，未知物体でモデル情報がない配置物体に対応するとともに，配置物体の姿勢を配置動作直前に認識し，衝突予測に役立てる．配置物体の姿勢を配置動作直前に認識する利点として，物体把持を行う際に重心がずれた把持を行った場合や，運搬時に把持物体の姿勢が変化した場合などに対応できることが挙げられる．また，従来手法では配置物体の大きさ情報を入力していたのに対し，画像を入力することで配置物体の形状をより正確に認識でき，形状による転がり易さなどを考慮できると考えられる．

3.1 ネットワーク構造

本手法のネットワークはヘッドカメラで撮影した RGB 画像，Depth 画像，ハンドカメラで撮影した RGB 画像を入力とする．本手法の Feature Extractor の構造を図 2 に示す．Feature Extractor は特徴抽出を行う ResNet 部分と Attention を学習する Attention branch 部分，それらを結合する Attention mechanisim で構成される．ここで ResNet-18(pre) は ResNet-18 の第 3 レイヤーまでの前半部分を示し，ResNet-18(post) は第 4 レイヤー部分である．また，インデックス $k \in \{rgb, depth, hand\}$ は入力を示す．図 2 の Attention branch 構造を用いることでネットワークが画像中のどこを見て衝突の有無を判断しているか可視化することができる．これによりネットワークの出力の解釈性が向上するともに視覚的説明が可能となる．ネットワークの学習ではまず，入力画像をベースネットワークの特徴抽出器に入力し，特徴マップを獲得する．得られた特徴マップを Attention Branch に入力することで Attention マップを獲得する．Attention マップは Attention mechanism により特徴マップと乗算を行い，特徴マップ上の Attention 領域を強調する．さらに，0 に近い値の乗算により失われた情報を補填するために Attention マップと特徴マップの加算を行う．また，Attention Branch のクラス予測確率 $\boldsymbol{P}_{rgb}(y)$，$\boldsymbol{P}_{depth}(y)$，$\boldsymbol{P}_{hand}(y)$ を用いて正解クラスとの予測誤差を算出し，誤差逆伝播により Attention branch の学習を行う．Attention 機構により強調された特徴マップを ResNet-18(post) に入力し 256 次元の特徴ベクトル $\boldsymbol{O}_{rgb}$，$\boldsymbol{O}_{depth}$ を得る．次に，Classifier の構造を図 3 に示す．Classifier では RGB の特徴ベクトル $\boldsymbol{O}_{rgb}$，Depth 特徴ベクトル $\boldsymbol{O}_{depth}$ とハンドカメラの特徴ベクトル $\boldsymbol{O}_{hand}$ を Self-Attentin 機構により重み付けを行う．重み付けは各特徴ベクトル $\boldsymbol{O}_k$ を学習可能な重み α_k と乗算する．α_k は式 (1) により求める．また，e_k は式 (2) により求める．

$$\alpha_k = \frac{\exp(e_k)}{\exp(e_{rgb}) + \exp(e_{depth}) + \exp(e_{hand})} \tag{1}$$

$$e_k = \boldsymbol{V}^{\mathrm{T}} \tanh(\boldsymbol{W}_k \boldsymbol{O}_k + \boldsymbol{b}_k) \tag{2}$$

このとき，$\boldsymbol{V}, \boldsymbol{W}_k$ は学習可能な重み行列であり，$\boldsymbol{b}_k$ は学習可能なバイアスベクトルである．Self-attention に

図 4: 配置物体の姿勢 (左:姿勢 1, 右:姿勢 2)

表 1: データセットの衝突有りラベル内訳 [%]

O-O	T-O	A-O	O-D	Any	P1	P2
3.15	21.11	14.41	10.46	30.33	23.19	76.06

より重み付けを行った 256 次元ベクトルを Perception block に入力し，最終的な予測 $\boldsymbol{P}_{per}(y)$ を出力する．

3.2 データセット作成

本研究では，モデルフリーかつ配置物体の姿勢を考慮した配置タスクの衝突予測データセットを作成した．データセットの作成は PonNet シミュレータを用い，背景と照明が異なる 12 環境に障害物 28 種類，家具 6 種類を配置し，1 環境につき 1,000 サンプルを取得した合計 12,000 サンプルを 1 セットとした．これを，必要なサンプル数が得られるまで繰り返し，デーセットを構築する．データセット作成手順を以下に示す．まず背景環境と家具を順番に選ぶ．次に配置エリアへ障害物をランダムに 1 から 3 個選択し，配置エリア上に落下させることでランダムに障害物を配置する．このときの配置エリアの状態の RGB-D 画像を HSR のヘッドカメラを用いて撮影する．

配置物体の姿勢を図 4 に示す．次に HSR が姿勢 1 または姿勢 2 の状態の配置物体を把持する．このときの配置物体の姿勢は交互に登場し，配置物体の状態の RGB 画像を HSR のハンドカメラを用いて撮影する．最後に把持物体を配置エリア中央に配置し，衝突情報を記録する．以上を繰り返すことで，データセットを作成する．また，衝突有りの判定は障害物同士の衝突，ロボットアームと障害物の衝突，配置物体と障害物の衝突，障害物と環境の衝突において閾値 0.1m/s 以上の衝突とする．

実際に作成したデータセットの衝突有りラベルの内訳を表 1 に示す．O-O は配置エリア内の障害物同士の衝突を表す．T-O は配置物体と障害物の衝突を表す．A-O は障害物とロボットアームの衝突を表す．O-D は障害物と配置エリアおよびその周辺環境への衝突を表す．Any は以上の条件の衝突いずれかが一つ以上発生した場合を表す．P1 は Any の条件で配置物体が姿勢 1 だった場合であり，P2 は姿勢 2 であった場合である．姿勢 2 の衝突有りラベルは姿勢 1 の 3.2 倍程度あり，姿勢 2 は衝突が発生しやすいことがわかる．例として，図 5 のように，障害物の配置が同じ場合でも姿勢 1 で衝突が起きないが，姿勢 2 では衝突が起きる場面が挙げられる．このように配置エリアの状態だけでなく，配置物体の姿勢も考慮することが重要である．

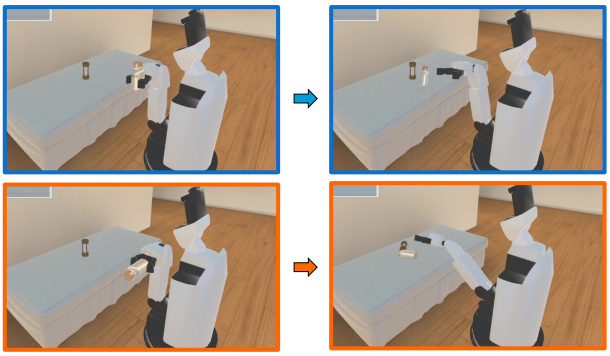

図 5: 配置物体の姿勢によって衝突が発生する場面
(上:姿勢 1) 衝突なし
(下:姿勢 2) 配置物体と障害物が衝突

表 2: 精度比較

手法	モデル情報	ハンド画像	正解率 [%]
Ponnet[2]	✓	-	84.49 ±0.22
Ponnet	-	-	84.25 ±0.30
提案手法	-	✓	87.10 ±0.33

4. 評価実験

本章では，提案手法の有効性を検証するための評価実験を行う．評価実験では物体配置タスクにおける衝突予測を行い，各手法との精度比較を行う．

4.1 実験条件

本実験には 3.2 節で作成したデータセットを使用する．学習に 42,100 枚，検証に 5,300 枚，評価に 5,300 枚を使用する．ベースモデルには ResNet-18 を使用する．学習回数は 10 Epoch，最適化関数に Adam を使用し，初期学習率を 10^{-5} とする．また，学習結果には，各学習 epoch において，最も検証精度が高いモデルの評価結果を用い，それぞれ 5 回学習した結果の平均と分散を用いる．

4.2 実験結果

各モデルの正解率の比較を表 2 に示す．このとき，モデル情報は，配置物体の大きさである 3 次元ベクトル x, y, z と配置場所の高低の 1 次元ベクトルを表す．また，ハンド画像はハンド画像の使用の有無である．PonNet の配置物体のモデル情報有りと無しを比較した場合，正解率が同程度であることがわかる．これは，モデル情報が物体の大きさ情報のみであり，衝突予測に重要な把持物体の現在の姿勢を考慮できないためだと考えられる．よって，物体把持後も把持物体の姿勢変化が考えられる条件下では事前のモデル情報は衝突予測に貢献しないといえる．次に提案手法はハンドカメラの画像を使用することで精度が向上している．このことからハンドカメラの画像を利用することで配置動作の直

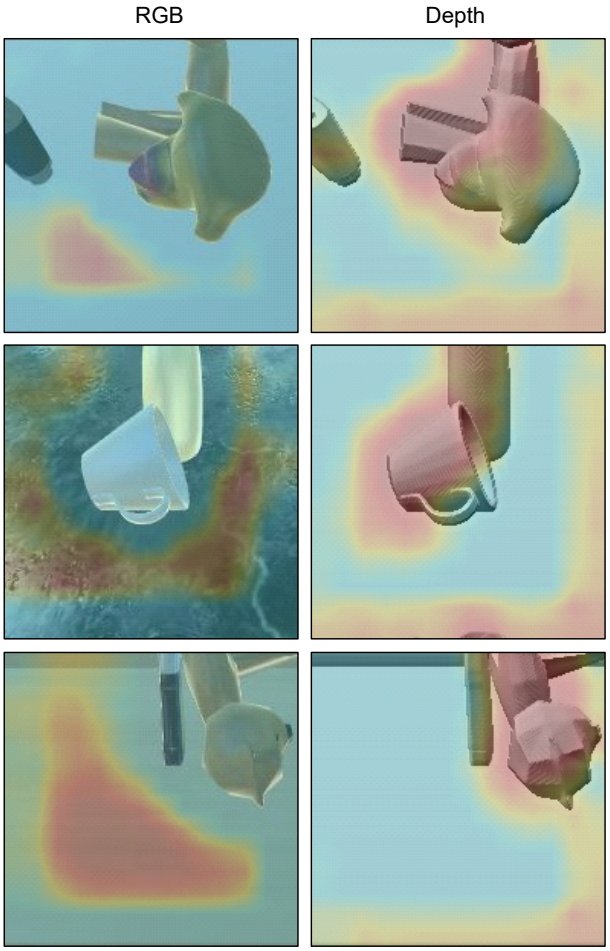

図 6: Attention map の可視化結果

前に配置物体の姿勢や大きさを考慮しているといえる．例えば，図 5 のような配置エリアの障害物配置が同じであるが，配置物体の姿勢が違う場合でもハンドカメラによって姿勢情報が得られ，適切な予測が可能になる．次に，正解ラベルが衝突有りかつ予測結果が衝突有りの正解例の Attention map の可視化結果を図 6 に示す．図 6 の RGB 画像では，障害物が存在しない場所に注視領域が発生しており，平面を注視していることがわかる．次に，Depth 画像では障害物に注視領域が発生しており，障害物を注視していることがわかる．このことから，RGB と Depth で異なる特徴を学習し，お互いを補う性質が精度向上に貢献していると考えられる．次に，ハンド画像の Attention の可視化結果を図 7 に示す．把持物体を注視していることがわかる．これによって，物体の種類や大きさと衝突の発生しやすさの関係を学習していると考えられる．また，特定の物体のみに Attention が発生する傾向あった．これは，姿勢 1 の状態や大きさが小さい物体など，衝突の危険性が低い物体とそうでない物体を区別していると考えられる．

5. おわりに

本研究では，ハンドカメラによる配置物体の考慮を追加したモデルフリー Attention 機構マルチモーダルネットワークを提案し，配置物体の姿勢変化を追加した物体配置タスクの衝突予測データセットを作成した．作成したデータセットで学習した結果，ハンドカメラを追加したことで衝突予測の精度が向上した．今後は配置物体の姿勢の種類を増やし，様々な条件での衝突予測向上を目指す．

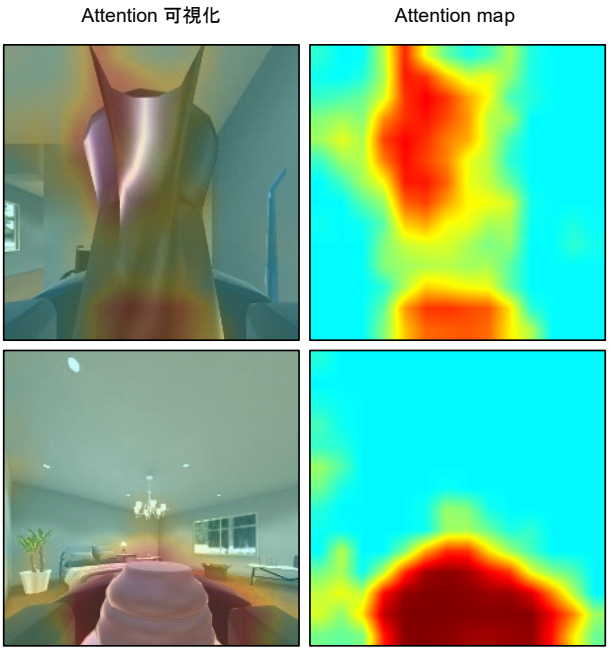

図 7: ハンドカメラ画像の Attention map

参 考 文 献

[1] 寺田 耕志. *et al.*: “生活支援ロボット Human Support Robot（HSR）における ROS 活用”，日本ロボット学会誌，vol.35，no. 4，pp. 280-283，2017.

[2] A. Magassouba. *et al.*: “Predicting and attending to damaging collisions for placing everyday objects in photo-realistic simulations”，Advanced Robotics. 2021. `https://doi.org/10.1080/01691864.2021.1913446`

[3] L. Iocchi. *et al.*: “RoboCup@Home: Analysis and results of evolving competitions for domestic and service robots”，Artificial Intelligence，vol. 229，pp. 258-281，2015. `https://www.sciencedirect.com/science/article/pii/S0004370215001174`

[4] T. Inamura. *et al.*: “Development of RoboCup@Home Simulation towards Long-term Large Scale HRI”，RoboCup 2013: Robot World Cup XVII,pp.672–680,2014.

[5] H. Fukui. *et al*: “Attention Branch Network: Learning of Attention Mechanism for Visual Explanation”，IEEE Conference on Computer Vision and Pattern Recognition, pp. 10705-10714, 2019.

[6] Wang, C.,and Guo, X.,: “Plane-Based Optimization of Geometry and Texture for RGB-D Reconstruction of Indoor Scenes”，IEEE 3DV, pp. 533–541, 2018.