

1

脳の情報処理モデルに基づく 3次元点群とRGB画像を用いた マルチモーダルな物体認識手法の 実装および評価

Osaka University
Haruhito Ando

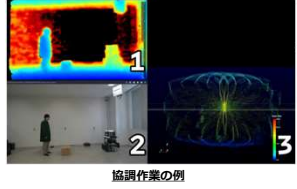
NS 研究会 24/02/29

1

2

研究背景

- Beyond 5G / 6G の通信技術の発達[1]
 - 多様なセンサーの利活用[2,3,4]
 - RGBカメラ, LiDARセンサー, 赤外線センサー
 - 通信技術を活用した新たなユースケースへの注目
- デジタルツインを活用したロボット制御
 - 人とロボットが協調作業を行う物流倉庫
 - 安全性や効率性のための確率的なロボット制御
 - 確率的な認識技術が必要



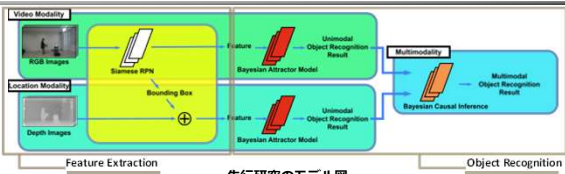
協調作業の例

[1] S. V. Kulkarni et al., "Beyond 5G: A Survey on 6G Research," *IEEE Access*, vol. 10, pp. 37 988-10 000, 2022.
[2] S. Wang, S. Liu, C. Fan, and F. Tang, "Survey of the Applications of Intelligent Sensors," *IEEE Access*, vol. 10, pp. 37 988-10 000, 2022.
[3] S. Wang, S. Liu, C. Fan, and F. Tang, "Survey of the Applications of Intelligent Sensors," *IEEE Access*, vol. 10, pp. 37 988-10 000, 2022.
[4] S. Wang, S. Liu, C. Fan, and F. Tang, "Survey of the Applications of Intelligent Sensors," *IEEE Access*, vol. 10, pp. 37 988-10 000, 2022.

2

3

先行研究概要



先行研究のモデル図

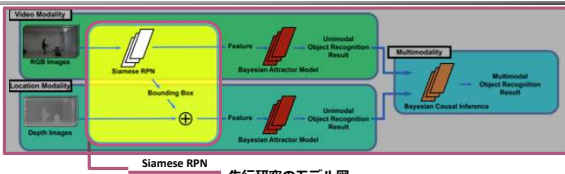
- 脳の情報処理モデルを活用したマルチモーダルな物体認識手法[5]
 - 映像と深度画像による観測を繰り返し観測対象に関する事後確率密度（確信度）を算出
 - 確信度を用いた不確実性を考慮した物体認識によりリスクを評価可能
- モダリティ間の依存関係
 - 特徴量抽出部分を共有しているため後段でのマルチモーダル統合の前に認識結果間に依存関係が発生

[5] S. Wang, S. Liu, C. Fan, and F. Tang, "Survey of the Applications of Intelligent Sensors," *IEEE Access*, vol. 10, pp. 37 988-10 000, 2022.

3

4

Siamese RPNを用いた特徴量抽出



先行研究のモデル図

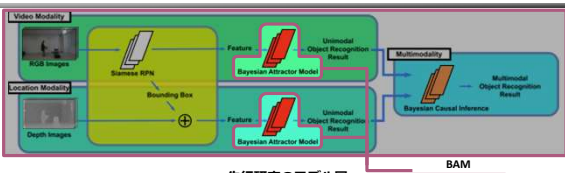
- 追跡対象が含まれる領域を映像から検出するOne-shotの物体検出手法[6]
 - 映像特徴量：物体を検出した領域に対して CNN ベースの特徴量抽出を実施
 - 位置特徴量：深度画像と映像にて抜き出した領域を組み合わせてオブジェクトの特徴量を検出
- 先行研究における問題点
 - 後段の処理で統合される前にSiamese RPN の特徴量抽出の過程で依存関係が発生

[6] S. Wang, S. Liu, C. Fan, and F. Tang, "Survey of the Applications of Intelligent Sensors," *IEEE Access*, vol. 10, pp. 37 988-10 000, 2022.

4

5

Bayesian Attractor Model (BAM) を用いた物体認識



先行研究のモデル図

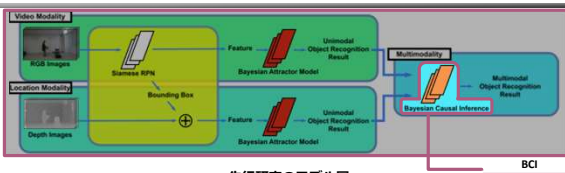
- 不確実性のある情報からでも確かな推論を行える脳に倣ったモデル[7]
 - 状態空間にアトラクターという名オブジェクトの特徴を記憶する定点を設置
 - 観測した特徴量に応じて内部状態を変化させアトラクターとの距離から確信度付きの認識結果を算出
 - 物体の特徴がアトラクター間の境界領域に位置する場合に認識に確信が持てないという状態を表現可能
 - 確信度を伴う認識によりリスクを考慮したシステムを構築可能

[7] S. Wang, S. Liu, C. Fan, and F. Tang, "Survey of the Applications of Intelligent Sensors," *IEEE Access*, vol. 10, pp. 37 988-10 000, 2022.

5

6

Bayesian Causal Inference (BCI) を用いた物体認識



先行研究のモデル図

- 異なるモダリティから得られるデータを統合し因果推論を行うモデル[8]
 - 異なるモダリティを活用することによる認識力の強化
 - 同じ物体を見ているかどうかを判定し、重みづけを決定
 - 各モダリティから得られた確率分布を統合して確信度を算出
 - 不確実性を考慮した上での意思決定をサポート

[8] S. Wang, S. Liu, C. Fan, and F. Tang, "Survey of the Applications of Intelligent Sensors," *IEEE Access*, vol. 10, pp. 37 988-10 000, 2022.

6

先行研究における課題と本研究の目的

7

- **先行研究における課題：マルチモーダル統合を行う前の依存関係**
 - BCIでマルチモーダル統合する前に Siamese RPNにて深度画像を扱うときにRGB画像の情報を利用
 - 深度画像単体では認識するのが難しいためRGB画像から切り出された領域を使用
 - 映像モダリティでの認識精度の低下による位置モダリティでも認識の低下
 - 信頼性の低下
- **本研究の目的：特徴量抽出部分の依存関係を解決**
 - 映像とは異なり単独で解析に用いることができるモダリティとして点群を利用
 - PointNetをセマンティックセグメンテーションのために採用
 - 映像と位置の情報が独立に認識された後に統合されるように設計
 - 信頼性の向上

7

先行研究と本研究の違い

8

- **先行研究[5]**
 - 映像モダリティ
 - RGB 画像
 - Siamese RPN(s)
 - 位置モダリティ
 - 深度画像
 - Bounding Box と統合
- **本研究**
 - 映像モダリティ
 - RGB 画像
 - Siamese RPN(s)
 - 位置モダリティ
 - 3次元点群
 - PointNet(s)

先行研究との比較

8

点群を利用した物体認識手法

9

PointNet with BAM

- **PointNet[9]とBAM[7]を組み合わせた物体認識手法**
 - 点群を用いた認識にPointNet[9]を使用
 - ラベルなし点群に対してセマンティックセグメンテーションを実行
 - 得られたラベル付き点群を集約し各オブジェクトの重心を計算
 - 各オブジェクトの重心をBAM[7]を通して認識
 - 最初のオブジェクトの位置をアトラクターとして記憶し物体認識を行い確信度を計算

9

評価手法概要

10

- **RGB-DカメラとLiDARセンサを用いてデータセットを作成**
 - 屋内での人とロボットが協調作業を行うシーンを想定して撮影
 - RGB-D画像と3次元点群からなる時系列データで表現
- **マルチモーダルな認識を評価するためのタスク**
 - 対象 : Worker / Robot
 - タスク : 各オブジェクトの特徴が与えられたときにどちらのオブジェクトであるか判定

評価シーン

10

評価方法

11

- **入力と出力**
 - 入力 (映像) : 1フレームごとの各オブジェクトを表す画像特徴量
 - 入力 (位置) : 1フレームごとの各オブジェクトを表す位置特徴量
 - 出力 : 各オブジェクトに対する確信度
 - Workerに関する特徴量が与えられたときの出力例
- **定性的な評価および定量的な評価**
 - 定性的な評価：確信度の推移を観測
 - 各モダリティおよびマルチモーダル統合での確信度の推移を比較
 - 定量的な評価：適合率
 - あるオブジェクトを認識したときに真に正解であるものを認識できている割合
 - 予測の確実性を示す指標

Workerを与えた場合の出力例

11

評価結果：確信度の推移

12

- **統合した結果正しい選択肢の確信度を高く保持**
 - 映像モダリティで確信度が高く位置モダリティで確信度が高いとき認識対象であるロボットの確信度を高く認識
 - 映像モダリティで確信度が低く位置モダリティで確信度が高いとき認識対象ではない人の確信度が高くてロボットの確信度も高く認識

12

評価結果：適合率

13

適合率 (Precision)		
Modality	Worker	Robot
Single modality (video)	0.875	<u>0.979</u>
Single modality (depth)	0.804	1.000
Single modality (location)	1.000	1.000
Multi-modalities (video based with depth)	0.814	<u>0.987</u>
Multi-modalities (video based with location)	0.838	<u>0.996</u>

- マルチモーダル統合することによるロボットに対する適合率が向上
 - ロボットに対しては映像モダリティのみを用いた場合より高い適合率を検出
 - 人に対しては映像モダリティで大きな確信度をもって間違えてしまっていたため適合率が減少

13

今後の課題

14

- 認識対象の拡張
 - 1クラスにつき1オブジェクトしか同一フレームに存在しないと仮定
 - 実用には多クラス多オブジェクトの認識を行えるようにする必要
- 同じオブジェクトに対する認識結果の組み合わせ
 - 現状は両モダリティでの認識結果が同じ対象に関する入力であるとして仮定
 - 実用には各モダリティでの認識結果を組み合わせる仕組みが必要

14

まとめ

15

- 提案手法
 - 点群を入力とする確信度を用いた位置モダリティの認識手法の提案
 - 脳の仕組みを活用したマルチモーダルな物体認識手法の検証
- 実験結果
 - 実環境を想定したデータセットを用いた検証
 - 映像モダリティでの認識結果を位置モダリティでの認識結果で補強することで適合率が向上
- 今後の課題
 - 認識対象の拡大
 - 各モダリティでの認識結果の組み合わせ

15

付録

16

- LiDAR (Light Detection And Ranging) の動作原理
- Bayesian Attractor Model[7]
- Bayesian Causal Inference[8]
- PointNet[9]
- ユニモーダルでの認識結果
- マルチモーダルでの認識結果

16

LiDAR (Light Detection And Ranging) の動作原理

17

● 三次元点群を計測するセンサ(Time of Flight 方式)

- レーザー光パルスの放出
 - レーザー光を物体に当て反射光を測定
- 距離の計算
 - 放出から測定までに要した時間を用いて距離を計算
- 点群マップ生成
 - パルスの放出と距離計算を繰り返してマップを生成



図10：使用したLiDAR[10]

● LiDARの代表的メーカー

- Velodyne
 - 様々な3Dデータセットの測定で利用
- Robosense
 - 車載LiDARとして著名
- Livox
 - 安価で入手しやすいLiDARを提供

17

Bayesian Attractor Model (BAM)[7]

18

● アトラクターモデルとベイズ推定を組み合わせたモデル

- 感覚器官の情報から脳が意思決定を行うプロセスに倣ったモデル
 - アトラクターモデルの記憶モデルとベイズ推定の状態更新モデルを融合
- 連続的に観測を行い観測しているものは何かという認識結果を確信度とともに出力

● アトラクターモデル

- 脳の記憶と認知モデルに基づいた認識モデル
 - 状態空間上に観測対象に一対一対応するアトラクター (安定点) を設置
 - 内部状態を入力に応じて変化
- 同じ対象に対する入力を受け続けるとその対象を表すアトラクターに収束

● ベイズ推定

- 事前確率と観測値を用いて事後確率を算出
- 新しい入力があるたびに情報を蓄積し意思決定を更新

18

Bayesian Causal Inference (BCI)[8]

19

● **ベイジアン理論と因果推論を組み合わせた統計モデル**

- マルチモーダルな情報から脳が意思決定を行うプロセスに倣ったモデル
 - 二つの異なるモーダルで認識している物体が同じものかどうか推論し統合
- 利点の異なるモダリティを組み合わせてより正確な予測を提供
 - 異なるモダリティからの感覚入力と同じ外部事象によって引き起こされる確率を評価
 - 評価された因果関係に基づき感覚入力を統合するか別のものとして扱うかを決定

● **確信度の生成式**

- c_V, c_L : 映像モダリティおよび位置モダリティにおける確信度

$$P(c_V, c_L | C = 1) = \begin{cases} 0.5 + \sigma_{c=1} & \text{if } L_V = L_L \\ \sigma_{c=1} & \text{otherwise,} \end{cases}$$

$$P(c_V, c_L | C = 2) = \begin{cases} 1 & \text{if } L_V \neq L_L \\ 0 & \text{otherwise} \end{cases}$$

$$\sigma_{c=1} = 0.5\sigma(\max(c_V) + 0.5\sigma(\max(c_L)))$$

$$\sigma = 1 / (1 + \exp(-x - \lambda_{bcI}))$$

$$L_V = \operatorname{argmax}(c_V)$$

$$L_L = \operatorname{argmax}(c_L)$$

19

PointNet[9]

20

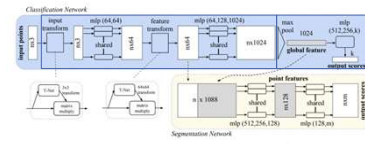


図5: PointNetのアーキテクチャ ([9]より引用)

● **3次元点群を扱う深層学習モデル**

- 各点に対し同じ *Multilayer Perceptron (mlp)* を適応, 後に *Max-Pooling* を適応
 - 入力される点群の順序によらず同じ結果を出力
- T-Net と呼ばれるサブネットワーク
 - 入力に対して, アフィン変換行列を出力, 点群の回転や移動などに対応
- 大局的な特徴量の利用
 - 局所的な特徴量と大域的な特徴量を統合して MLP を適応, セマンティックセグメンテーションの実現

20

実験結果 (ユニモーダル認識)

21

● **ユニモーダル**

- 深度画像を用いた位置モダリティを区別のため深度モダリティと表記
- 先行研究の課題である映像モダリティと深度モダリティでの相関を観測

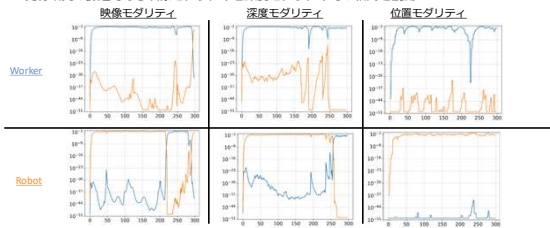


図5: ユニモーダルでの評価結果

21

実験結果 (マルチモーダル認識)

22

● **マルチモーダル**

- 深度画像を用いた場合に相関がでていた部分は修正できていない
- 位置情報を用いることでこの問題を修正

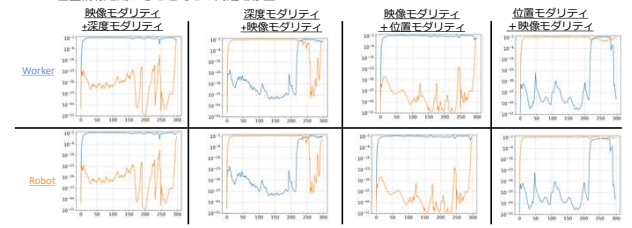


図6: マルチモーダルでの評価結果

22