1

脳の情報処理モデルに基づく LiDAR と RGB カメラ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

大阪大学 基礎工学部
情報科学科計算機科学コース
村田研究室
安藤 覇人

特別研究報告会 2023/02/14

1

研究背景

2

- **デジタルツイン (*Digital Twin, DT*)[1] への注目**
  - 物理モデルやセンサを利用し，**現実世界のシステムと対応する双子**を仮想空間上に作成
    - 現実世界のシステムを反映させた仮想空間で試作できることによる設計コストの削減
    - 現実世界のシステムを仮想空間で監視できることによる安全性の向上
  - 設備保全や自動運転，災害予測など**様々な分野での応用**
- **デジタルツイン実現への課題**
  - センサやシステムの**不確実性への対応**
    - センサによる測量の不確実性やシステムによるオブジェクト推定の不確実性
    - 誤った測量結果を利用することに由来する誤ったシステム制御の危険
  - センサの測量結果を100%にすることは困難
  - **不確実性にロバストなシステム**が必要

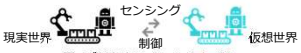


図 : デジタルツインのイメージ

[1] E. Glaessgen and D. Stargel: "The digital twin paradigm for future nasa and us air force vehicles", 53rd AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics and materials conference 20th AIAA/ASME/AHS adaptive structures conference 14th AIAA, p. 1818 (2012).

2

研究目的および研究手段

3

- **研究目的**
  - **入力が備えるノイズによる不確実性を容認した物体認識の実現**
- **研究手段**
  - *ノイズにロバストな処理の実現に***ゆらぎ学習**[2] を活用
    - **脳の処理**に倣った学習を行い，入力のノイズや変化に強いモデルを生成
    - 不確実な観測情報から意思決定を行う処理に倣った ***Bayesian Attractor Model (BAM)*** [3] を応用
  - 認識精度の向上に ***Bayesian Causal Inference (BCI)*** [4] を応用
    - 脳が複数モダリティからの情報を統合する処理に倣ったモデル
    - あるモダリティでの認識結果を他のモダリティでの認識結果を用いて補強
    - **映像モダリティ** (RGB画像)と**位置モダリティ**(点群データ) を統合

[2] M. Murata and K. Leibnitz: "Fluctuation-induced network control and learning: Applying the yuragi principle of brain and biological systems", Springer Singapore (2021)
[3] S. Bitzer, J. Bruineberg and S. J. Kiebel: "A bayesian attractor model for perceptualdecision making", PLOS Computational Biology, 11, 8, pp. 1–35 (2015).
[4] T. Rohe, A.-C. Ehlis and U. Noppeney: "The neural dynamics of hierarchical Bayesian causal inference in multisensory perception", Nature Communications, 10, 1, p. 1907 (2019)

3

先行研究と本研究の位置付け

4

- **先行研究**
  - 脳のマルチモーダルな情報処理に着想を得た物体推定手法 [5]
    - 映像・位置モダリティによる推定結果を ***BAM*** で識別，***BCI*** を用いて統合する物体認識手法

図 : 先行研究による提案手法の概要図

- **本研究の位置付け**
  - 先行研究の課題
    - 先行研究では**RGBDカメラ**から映像/位置モダリティを取得し統合
    - **同一センサからの取得**に由来する**モダリティ毎の認識結果の強い相関**
  - 新規性
    - **異なるセンサ**を用いて取得したモダリティを組み合わせることで解決

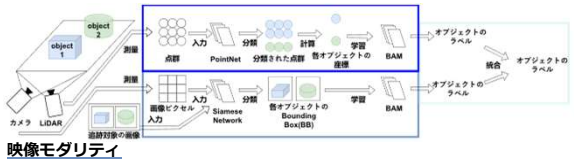
[5] 関 良我, 小南 大智, 下西 英之, 村田 正幸, 藤若 雅也, 野上 耕介 : "脳のマルチモーダルな情報処理に着想を得た物体推定手法の提案と評価", 電子情報通信学会 技術研究報告(CQ2021-14), 121, pp. 59–64 (2021).

4

マルチモーダルな物体認識手法

5

- **位置モダリティ**
  - ***PointNet*** [6] を点群データのクラス分類に用いた物体認識
  - 点群情報から重心を特徴量として抽出，*BAM*を用いた識別
- **モダリティの統合**
  - ***BCI***[5] による統合



- **映像モダリティ**
  - *Siamese-RPN* を特徴量の抽出に用いたユニモーダルな物体推定 **[7]**

[6] C. R. Qi, H. Su, K. Mo and L. J. Guibas: "Pointnet: Deep learning on point sets for 3d classification and segmentation", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 652–660 (2017)
[7] 久保 快斗, 関 良我, 小南 大智, 下西 英之, 村田 正幸, 藤若 雅也 : "デジタルツイン構 築のための脳の認知機構を用いたオブジェクト認識手法の実装及び評価", 電子情報通 信学会 技術研究報告 (CQ2021-125), 121, 421, pp. 6–11 (2022).

5

位置モダリティで用いる特徴量

6

- **セマンティックセグメンテーション (*Semantic Segmentation*)**
  - シーンを表す点群に対して，**各点に**用意されたクラスを関連付けるタスク
  - ***PointNet*** [7] を利用
    - 3D 点群データを扱うネットワークの元祖

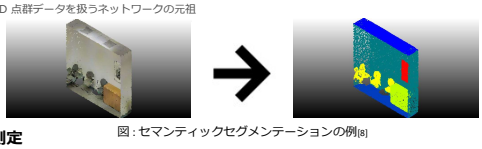


図 : セマンティックセグメンテーションの例[8]

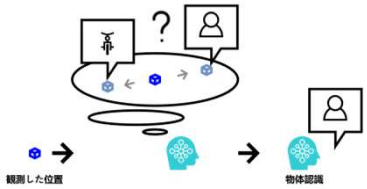
- **位置測定**
  - 位置モダリティの特徴量として物体の**座標**を利用
    - *BAM* に与える情報量の削減，計算量の削減などの利点
    - 分類された点群を用いて，**各クラスについて**重心を測定

[8] I. Armeni, O. Sener, A. R. Zamir, H. Jiang, I. Brilakis, M. Fischer and S. Savarese: "3d semantic parsing of large-scale indoor spaces", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016).

6

位置モダリティの特徴量を用いた物体認識 7

- **ゆらぎ学習[2]** を利用
  - 脳が**連続的で変化する入力**から**記憶している様々な知識や経験**に照らし合わせて**意思決定**を行う処理
    - 連続的で変化する入力　：観測された物体の**位置**
    - 様々な知識や経験　：それまでに観測された追跡対象の**位置**
    - 意思決定　：入力された位置にあるものが**どの追跡対象であるか**の認識



図：ゆらぎ学習を用いた物体認識

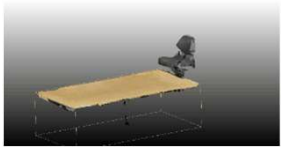
7

評価手法 8

- **評価方法**
  - 公開データセット[8]を利用したシミュレーション
    - 点群で表現：各点には**座標**と**色**，**属する物体のラベル**の情報
    - ノイズを含んだ情報 / マルチモダリティ / 時系列データ が必要
  - データセットの拡張
    - 対象となる物体：椅子，机
    - ノイズの付与
    - 時系列データに拡張
    - 用意された色付き点群を撮影し映像モダリティを取得
- **評価指標**
  - 各フレームでの物体の特徴量を与えた場合の確信度
    - 物体のラベルを認識しているかの正答率

表：モダリティと点群の要素の対応

モダリティ	点群の要素
位置モダリティ	座標(x y z)
映像モダリティ	色(r g b)

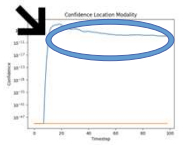


図：拡張したシーンの一部

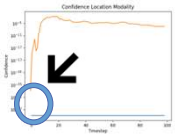
8

評価結果：位置モダリティ 9

- **評価結果**
  - 横軸：時間 [フレーム], 縦軸：入力データに対する確信度
    - オレンジ：椅子の確信度，青：机の確信度
  - 机　：高い精度での認識 [97%]
  - 椅子 ：高い精度での認識 [100%]



図：机の認識結果

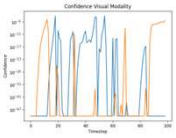


図：椅子の認識結果

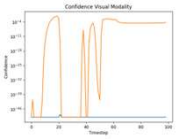
9

評価結果：映像モダリティ 10

- **評価結果**
  - 横軸：時間 [フレーム], 縦軸：入力データに対する確信度
    - オレンジ：椅子の確信度，青：机の確信度
  - 机　：低い精度での認識 [65%]
  - 椅子：高い精度での認識 [79%]



図：机の認識結果

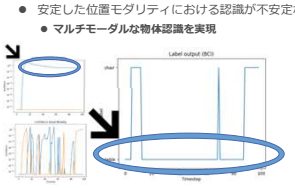


図：椅子の認識結果

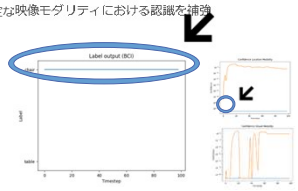
10

評価結果：マルチモダリティ 11

- **評価結果**
  - 横軸：時間 [フレーム], 縦軸：認識しているラベル
  - **位置モダリティ**に関する物体認識手法ではノイズに対応した安定した物体認識
  - 安定した位置モダリティにおける認識が不安定な映像モダリティにおける認識を補強
    - **マルチモーダルな物体認識を実現**



図：机の認識結果



図：椅子の認識結果

11

まとめ 12

- **脳の情報処理モデルに基づくマルチモーダルな物体認識手法の実装**
  - **複数のモダリティ**を組み合わせたノイズにロバストな物体認識の実現
  - 位置モダリティにおける**ゆらぎ学習[2]** と ***PointNet*[7]** を組み合わせたユニモーダルな物体認識の実装
- **今後の課題**
  - 位置モダリティを用いた物体認識手法の改修
    - 実用に向けた計算速度の実現
  - 複数物体への対応
    - より多数の物体が存在する場面での活用

12

Bayesian Attractor Model, BAM[3]

13

- **脳が不確実な情報をもとに意思決定を行う処理に倣ったモデル**
  - **時刻** t に**刺激** A_t を観測し，正規分布に従う**特徴量** x_t にマッピング
  - 時刻 t までにおけるノイズを含む観測情報 $x_{1:t}$ から**内部状態** z_t (状態空間内の点) が変化 (ベイズ推定)
  - ***ノイズを含まない場合の選択肢*** k を表す特徴量 μ_k が状態空間における各**定点** Φ_k に対応
  - 選択肢 iであるという**確信度** $p(z_t = \Phi_k | x_{1:t})$ と意思決定に至るための**閾値** λ を比較
  - 確信度が閾値を超えたとき，**対応する選択肢** k を出力

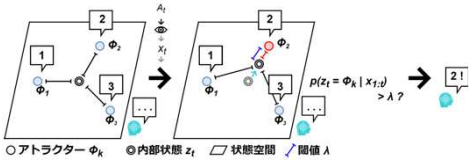


図 : *BAM* の認識モデル

13

Bayesian Causal Inference, BCI [4]

14

- **脳が複数のモダリティを組み合わせて認知を行う処理に倣ったモデル**
  - 2つのモダリティからの入力x_a, x_vを考える (a=audio, v=visual)
  - 与えられた入力について，二つのモデルを利用して推測
    - **共通の原因から発生**$(c=1)$: ***forced-fusion model*** 原因 $N_{av,c=1}$ 確率 $p(c=1) = p_{common}$
    - **独立の原因から発生**$(c=2)$: ***segregation model*** 原因 $N_{a,c=2}$, $N_{v,c=2}$ 確率 $p(c=2) = 1 - p(c=1)$
    - **二項分布**に従いモデルが決定
  - ***Model averaging*** を利用した統合処理
    - 2つのモデルの結果を組み合わせて，その推測される数 $\bar{N}_a$, $\bar{N}_v$を算出
      - $\bar{N}_a = p(c=1|x_a,x_v)\,N_{av,c=1} + (1-p(c=1|x_a,x_v))N_{a,c=2}$
      - $\bar{N}_v = p(c=1|x_a,x_v)\,N_{av,c=1} + (1-p(c=1|x_a,x_v))N_{v,c=2}$

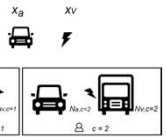


図 : 因果推論のモデル

14

脳のマルチモーダルな情報処理に着想を得た物体推定手法の提案と評価[5]

15

- **複数のモダリティを利用した不確実な情報にロバストな物体認識手法**
  - 各モダリティにて ***BB*** が与えられ，その *BB* のオブジェクトが何であるかを推測
  - 各推定結果を時系列順に ***BAM*** に与え，各フレームで認識を行い，**確信度**を算出
  - 確信度を ***BCI*** **を用いて統合**することで最終的な意思決定を実行

図 : 先行研究による提案手法の概要図

- **評価結果 / 課題**
  - *BB* 内のオブジェクトを正しく認識できるか評価
  - どのオブジェクトの場合に対してもユニモーダルの結果より優れた精度を確認
  - 学習していない物体を認識した場合のアトラクターの更新，複数オブジェクトの推定 といった課題

15

デジタルツイン構築のための脳の認知機構を用いたオブジェクト認識手法の実装及び評価[6]

16

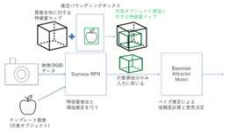


図 : *Siamese RPN* を特徴量抽出に用いたユニモーダルオブジェクト認識アーキテクチャ（[6]より引用）

- ***Siamese RPN* と *BAM* を組み合わせたユニモーダルなオブジェクト認識**
  - *Siamese RPN*
    - 追跡対象の画像(**テンプレート画像**)と探索を行う全体の画像（**オリジナル画像**）から追跡対象がオリジナル画像のどの位置に映っているのかを推定
  - ***BAM*** の利用
    - 1フレーム目の追跡対象についての *BB* をアトラクターに設定
    - 各フレームごとの *BB* から取り出される特徴量を与え，どのオブジェクトに近いのかを判断

16

PointNet [7]

17

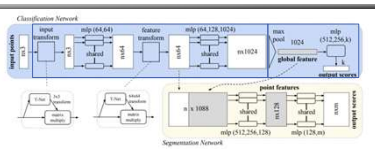


図 : PointNetのアーキテクチャ([2]より引用)

- **点群を直接処理するニューラルネットワーク**
  - 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

17

PointNet[7] 採用理由

18

- **先行研究との位置モダリティにおける相違点**
  - 脳のマルチモーダルな情報処理に着想を得た物体推定手法の提案と評価
    - センサ機器：**RGBカメラ**
      - **同一センサから複数のモダリティを取得，利用していたこと**が課題
    - 位置モダリティ：**深度画像**
      - 深度画像による**画像ベース**の物体認識手法
  - 本研究
    - センサ機器： ***LiDAR***
      - **位置モダリティのみ**を取得，先行研究の課題を解決
    - 位置モダリティ：**点群**
      - 点群をそのまま扱う**点群ベース**の手法に着目
      - 後進の手法のモデルとなった***PointNet***を利用
- **位置モダリティを利用した物体認識に必要とした要件**
  - 点群をクラスに分類可能であり，各物体の位置情報が取得可能なネットワーク
    - **物体の座標（重心）を用いた認識**ができることが主張

18