1

脳のマルチモーダル情報処理に着想を得た物体推定手法の提案と評価

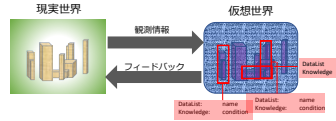
関 良我† 小南 大智† 下西 英之††† 村田 正幸†
藤若 雅也†† 野上 耕介††
†大阪大学 大学院情報科学研究科
††NEC システムプラットフォーム研究所

1

2

研究背景

- デジタルツインへの期待
  - 現実世界のあらゆる物体をリアルタイムに仮想空間上にマッピングすることで環境を再現
  - 仮想空間でのシミュレートによって得られる様々な情報を現実世界へフィードバック
- 想定されるデジタルツインの活用例
  - 自動運転：人や自動車、障害物の動きを収集、自動車の運転制御
  - 公共空間：空港など公共空間内の人の動きを把握、不審者や不審人物を検出
  - 荷物管理：倉庫などの荷物の動きを把握、ARゴーグルによる情報提示や自動搬送



2

3

デジタルツインに必要な物体推定手法

- 高速なセンシングデータ処理
  - 画像分析などによる物体やその状態の識別
  - センサー機器自体の自己位置推定と物体位置推定
  - リアルタイムに複数の物体を個別に追跡して識別／推定
- 従来の物体推定手法
  - CNN などの物体推定手法では高精度な識別が可能であるが データをクラウド上に集約して処理
  - 実世界のすべてのセンサー情報をクラウドに送信して処理することは困難

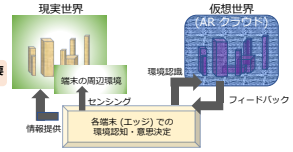
→逐一クラウドに送信せず各端末で物体推定処理を行う

3

4

AR クラウドを利用したエッジ – クラウド型デジタルツイン構想

- それぞれの端末で集めた現実世界の情報を統合して仮想世界を構築
  - 各端末で環境認識を行い、局所的な仮想世界を構築
  - AR クラウド上で各端末の情報を統合し、巨大な仮想世界を構築
  - AR クラウドの情報をもとに各端末は行動を決定
- より小さな時間粒度で処理するため 各端末で物体推定
- 刻一刻と変化する現実世界の環境に 対応するため柔軟性のある設計
  - **軽量・ロバストな端末での処理が必要**



4

5

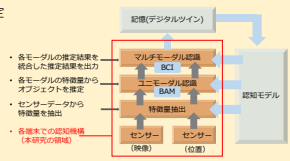
研究の目的・アプローチ

目的

端末単位で処理可能な軽量性と現実世界の環境変化に適応可能な ロバスト性を持った物体推定手法の提案 / 評価

アプローチ

- 脳の優れた認知機構に着想を得た物体推定手法の提案
  - 目や耳、三半規管などから得られた不確実な情報のマルチモーダル統合処理
  - 階層構造の軽量かつロバストな意思決定
- 脳の認知機能の数理モデルを利用
  - 脳や生体の「ゆらぎ学習」による ユニモーダル処理 **B**ayesian **A**ttractor **M**odel (BAM)
  - 脳の知覚による因果推論に基づいた マルチモーダル処理 **B**ayesian **C**ausal **I**nference (BCI)



5

6

BAM の物体推定手法への利用

- 映像モダリティと位置モダリティをもとに意思決定
  - 特徴量抽出
    - SiameseRPN [1] によりテンプレート画像をもとに バウンディングボックスとして位置を出力
    - 抽出された特徴量を 128 次元に圧縮 **→映像モーダル入力**
    - カメラ方向ベクトルとデータセットの深度情報から 算出した 3 次元ワールド座標系 **→位置モーダル入力**

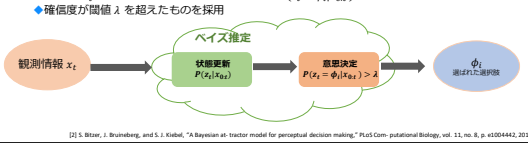
  - 特徴量変換
    - 平均・分散を一定にする標準化処理
    - 映像特徴量の値域が定まっておらず、BAM で処理しやすい形に整形

[1] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8971–8980, 2018.

6

Bayesian Attractor Model[2]

❖人が観測した情報をもとに意思決定する脳の振る舞いを数理モデル化
- 情報を観測し続け、事前に学習した選択肢のどれに近いかを判断
- **状態更新**：観測情報 x_t に基づいて内部状態 z_t をベイズ推定によって更新
  →ベイズ推定であるため z_t は点ではなく脳の不確実性を反映した確率分布 $P(z_t)$
- **意思決定**：事前に学習した選択肢（アトラクター）のいずれかを採用
  - ◆状態空間上にアトラクター $\phi_1 \cdots \phi_n$ を事前に用意
  - ◆状態値 z_t が選択肢 i である事後確率（確信度 $P(z_t = \phi_i | x_{0:t})$ を算出
  - ◆確信度が閾値 λ を超えたものを採用



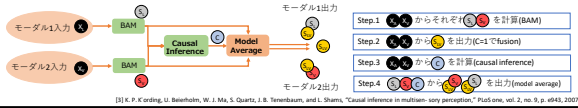
[2] S. Bitzer, J. Bruineberg, and S. J. Kiebel, "A Bayesian at- tractor model for perceptual decision making," PLoS Com- putational Biology, vol. 11, no. 8, p. e1004442, 2015.

7

Bayesian Causal Inference[3]

❖人がマルチモーダルな知覚刺激をもとに行う認知を数理モデル化
- 2つの入力刺激が同じ刺激源から出たものであるかを確率的に識別（**Causal Inference**）
- その確率をもとにそれぞれの入力刺激を統合する（**Model Average**）

1. 2つの入力刺激 x_L, x_V でそれぞれ個別に認知結果 S_L, S_V を計算
   - 本研究では BAM による認知するよう設計
2. 両モーダルの入力刺激を統合した認知結果 S_{LV} を計算
3. 両モーダルで同じものを認知しているか（C=1）
   別のものを認知しているか（C=0）を確率的に計算する（Causal inference）
4. C=1であれば S_{LV} を出力、0<C<1 であればその割合に応じた認知結果を出力 (Model Average)



[3] K. P. K ording, U. Beierholm, W. J. Ma, S. Quartz, J. B. Tenenbaum, and L. Shams, "Causal inference in multisen- sory perception," PLoS one, vol. 2, no. 9, p. e943, 2007

8

Causal Inference

❖観測対象が同じであるか計算
- 従来[3]は刺激源の位置（連続値）を対象としたが
  今回はオブジェクトの識別（離散値）が対象

$$p(C|X,Y) = \frac{p(X,Y|C)p(C)}{p(X,Y)} = \frac{p(X,Y|C)p(C)}{p(X,Y|C=1)p(C=1) + p(X,Y|C=0)p(C=0)}$$

❖離散値への拡張

	❖従来の Causal Inference（連続値）	❖拡張 Causal Inference（離散値）
$p(X,Y\|C=1)$ 同じオブジェクトを観測	$\int p(X,Y\|s)p(s)ds$	$\sum_{k=1}^{K} p(X,Y\|O_k)p(O_k)$
$p(X,Y\|C=0)$ 別のオブジェクトを観測	$\int p(X\|s)p(s)ds \int p(Y\|s)p(s)ds$	$\sum_{k=1}^{K} p(X\|O_k)p(O_k) \sum_{k=1}^{K} p(Y\|O_k)p(O_k)$
s：オブジェクトの位置		O_k：オブジェクト k （識別ラベル）
$p(s)$：位置の確率分布		$p(O_k)$：オブジェクト O_k が観測される確率
$p(X\|s)$：オブジェクトが s にあるときに X として観測される確率		$p(X\|O_k)$：オブジェクト O_k を観測しているときに X として観測される確率

[3] K. P. K ording, U. Beierholm, W. J. Ma, S. Quartz, J. B. Tenenbaum, and L. Shams, "Causal inference in multisen- sory perception," PLoS one, vol. 2, no. 9, p. e943, 2007

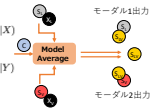
9

Model Average

❖0<C<1 の場合の重み付け処理
- BCI ではどちらのモダリティが優先されるべきかは定義しない
- 映像 / 位置モダリティを優先した場合の２通りの結果を出力

$$Cost_X(O_X) = p(C=1)\sum_{k=1}^{K}|O_X - O_k|^2 p(O_k|X,Y) + \quad p(C=0)\sum_{k=1}^{K}|O_X - O_k|^2 p(O_k|X)$$

$$Cost_Y(O_Y) = p(C=1)\sum_{k=1}^{K}|O_Y - O_k|^2 p(O_k|X,Y) + \quad p(C=0)\sum_{k=1}^{K}|O_Y - O_k|^2 p(O_k|Y)$$

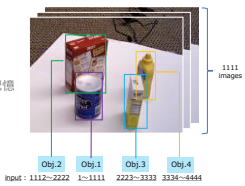


- $|O_X - O_Y|$： $O_X - O_Y = 0$ ならば 0、そうでなければ 1
  従来の *BCI* では推定位置の距離誤差を用いている

10

シミュレーション評価方式

1. 公開データセットの映像データをもとに
   映像モーダル・位置モーダルの特徴量を抽出
   - 1111 枚の映像フレームと 4 つのオブジェクト
   - 4444 枚をオブジェクトごとに直列で入力
   - アトラクターには各オブジェクトのはじめの1枚を記憶
2. ユニモーダルによる物体推定
   - BAMのみの推定精度を評価
3. マルチモーダルによる物体推定
   - BCI によって拡張された BAM の推定精度を評価



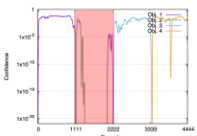
[4]オブジェクトの実計測公開 データセット(Yale-CMU-Berkeley (YCB) Object and Model set)

11

ユニモーダルによる物体推定

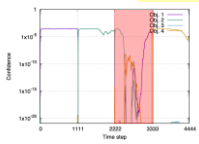
❖1 モダリティのみを BAM に入力したときの確信度出力
- 縦軸：BAM の確信度
- 横軸：映像フレーム

- 映像モダリティ



- 正答率：79.41 %
- オブジェクト 2 を上手く認知できず

- 位置モダリティ

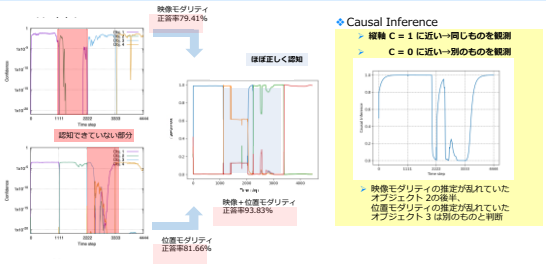


- 正答率：81.66%
- オブジェクト 3 を上手く認知できず

12

13

マルチモーダル物体推定手法の精度評価



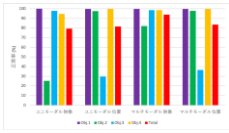
- Causal Inference
  - 縦軸 C = 1 に近い→同じものを観測
  - C = 0 に近い→別のものを観測
  - 映像モダリティの推定が乱れていた オブジェクト 2の後半、 位置モダリティの推定が乱れていた オブジェクト 3 は別のものと判断

13

14

推定手法の性能評価

- ユニモーダルでは判断できなかった部分を補い合うことで 正しい意思決定を行うことが可能
  - ユニモーダル → BAM 単体
  - マルチモーダル → BCI 拡張 BAM
    ※Model Average の結果なので、 映像重視 / 位置重視 の２通りを出力



- デスクトップ型計算機(CPU: Core-i7 8700, RAM: 16.0 GB)における 実計算時間 1 フレームの入力あたり 1.18 ms
  - 30fps、60fps の映像に対しても適用可能

14

15

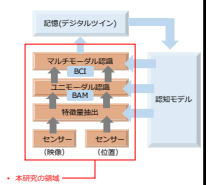
まとめ

本研究の成果

- 複数種類の不確実な観測情報を元に物体推定を行う手法を提案
- 端末レベルでの**軽量**な物体推定が可能
- 複数モダリティを組み合わせることで**ロバスト性**を獲得

今後の課題

- 事前学習のない新たなアトラクターを学習する方法の検討
- 複数のオブジェクトを同時に観測して推定することの検討
  - 今回は単一オブジェクトの直列での入力
- Model Average で重視するモダリティの判断基準
  - 今回は 2 モダリティの平均を使用
- BAM のチューニングの自動化



15