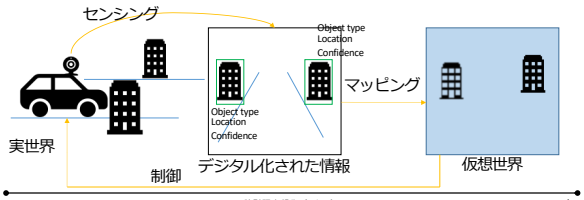
デジタルツイン構築のための脳の認知機構を用いたオブジェクト認識手法の実装および評価

大阪大学基礎工学部情報科学科　村田研究室
久保 快斗

0

研究背景

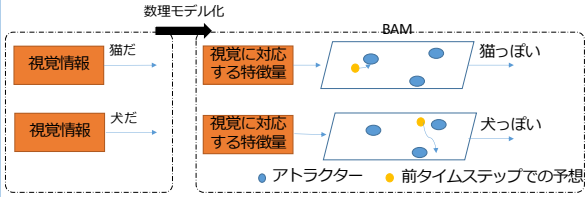
- **高まるデジタルツインへの期待**
  - 実世界のセンシング結果を仮想世界にマッピングして表現
  - 仮想世界上のシミュレーション結果を実世界にフィードバック
  - 物流管理や自動運転技術などへの応用が期待されている
- **リアルタイム性と仮想世界上の情報の精度が重要**
  - 自動制御などでは移り変わる現実情報に即座に対処することが必須
  - センシングしたデータには一般に誤差が含まれており、そこから可能な限り正確な情報を取り出すことが必須



1

先行研究

- **不確実な観測結果を用いて意思決定を行う人の脳に着想を得た物体認識手法[4]**
  - Bayesian Attractor Model (BAM) を用いたユニモーダル認識
    - 観測したものが記憶にある選択肢のいずれであるのかを脳が判断する認識過程をモデル化
    - 特徴量の取得に関しては理想的な状況を想定していた
  - ベイズ因果推論を用いたマルチモーダル意思決定

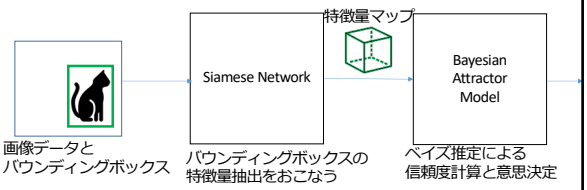


4. 関良我, 小南大智, 下西英之, 村田正幸, 藤若雅也, 野上耕介, "脳のマルチモーダルな情報処理に着想を得た物体推定手法の提案と評価," 電子情報通信学会技術研究報告（CQ2021-14）2021.

2

手法[4]での特徴量の取り出し手法

- **オブジェクト認識を行うための観測特徴量を取得する状況として、理想的な状況を想定していた**
  - オブジェクトを予測したバウンディングボックスから特徴量を抽出
  - 動画中の認識対象物体の検出が**成功したとき**のバウンディングボックス内の画像から、Siamese Network における軽量CNN を用いて特徴量を取得し BAM へ入力
  - 実際にはバウンディングボックスは**誤差を含んで推定**される

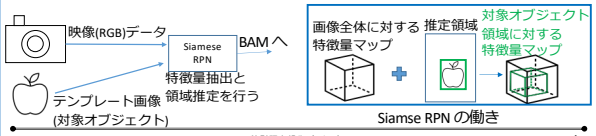


手法[4]における特徴量抽出から意思決定を行うまでのアーキテクチャ

3

研究目的とアプローチ

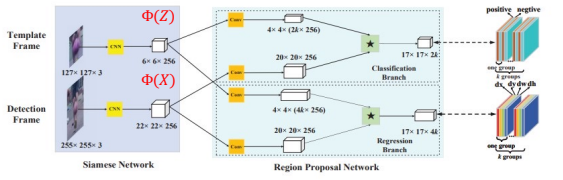
- **研究目的**
  - BAM の処理に適した形式での特徴量抽出アーキテクチャを実装
  - 実際の取得特徴量を BAM に入力したときの精度や計算時間を評価
- **アプローチ**
  - Siamese RPN (Region Proposal Network) を使用した領域推定
    - 「Siamese Network + RPN」アーキテクチャにより認識対象物体の存在する領域位置を推定
  - 推定領域に対応する特徴量を抽出
    - Siamese Network の学習時に作成したモデルを用いて特徴量抽出
  - 取り出した特徴量が有効か BAM に入力し評価



4

SiameseRPNを用いた領域推定・特徴量抽出

- **Siamese Network で取り出したテンプレート画像特徴量マップ $\Phi(Z)$ と 対象画像特徴マップ $\Phi(X)$ を用いて RPN が領域を推定**
- **推定した領域の映像特徴量を $\Phi(X)$ から取り出して物体の特徴量を抽出**



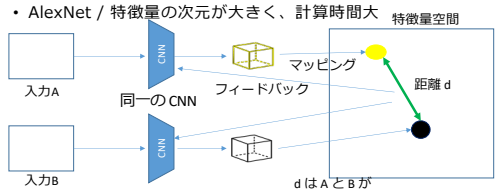
図は [5]より引用

5. L. Bo, Y. Junjie, W. Wei, Z. Zheng, and H. Xiaolin, "High performance visual tracking with siamese region proposal network," in Proc. CVPR 2018.

5

Siamese Network における学習([9]使用)

- **入力 A と 入力 B から同じ CNN で特徴量マップを取得**
- **特徴量空間上にマッピングした時の距離 d が**
  - A と B が同じクラスであれば近くなるように CNN をチューニング
  - A と B が違うクラスであれば遠くなるように CNN をチューニング
- **CNN 構造として３種類の構造を実装した**
  - 3 層、4 層の CNN / 特徴量の次元が小さく、計算時間小
  - AlexNet / 特徴量の次元が大きく、計算時間大



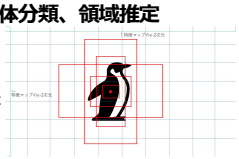
d は A と B が
同じクラスであれば近くなるように、
異なるものであれば遠くなるように学習

[9] "Youtube-boundingboxes dataset," available at https://research.google.com/youtube-bb

6

RPN での学習([9]使用)

- **領域対象選定基準となるアンカー、アンカーボックス設定**
  - $\Phi(X)$ の (x-2) と (y-2) の積の数だけアンカーが画像領域中に生成
  - アンカーを基準とした SCALES*RATIOS サイズのアンカーボックスが ANCHOR_NUM だけ設定
- **$\Phi(X)$ と $\Phi(Z)$ の畳み込みによる物体分類、領域推定**
  - Classification ブランチ
    - 各アンカーボックスが物体か背景を囲っているかの確認
    - IoU 0.7 以上で物体、0.3 以下で背景
      - IoU：$(B_{ans} \cap B_{box})/(B_{ans} \cup B_{box})$
      - 正解(B_{ans})、アンカーボックス(B_{box})
  - Regression ブランチ
    - IoU ≥ 0.7 ∨ IoU ≤ 0.3 のアンカーボックスの正解とのずれを回帰
    - 0.3 < IoU < 0.7 のアンカーボックスは学習に用いない



設定変数	使用目的
RATIOS	縦横比
SCALES	縦の長さ
ANCHOR_NUM	アンカーボックス数

7

BAM での推定

1. **時刻tにおける意思決定状態z_tを保持し、観測特徴量x_tを受け状態z_tを更新**

   i. $z_t - z_{t-\Delta t} = \Delta t f(z_{t-\Delta t}) + \sqrt{\Delta t}\, w_t$
   - f は勝者総取りのホップフィールドダイナミクス
   - 意思決定状態の選択肢 i に対応するアトラクター φ_i が z 空間に設定
   - w_t はノイズ項

   ii. $x_t = M\sigma(z) + v_t$
   - σは値域を0～1にするシグモイド関数
   - Mは各選択肢のノイズなしの特徴量を並べたもの
   - v_t はノイズ項
2. **i、ii式を逆に推定し事後推定確率$P(z_t|x_t)$を計算**
3. **$P(z_t = \varphi_i | X_{0:t}) \geq \lambda$ を満たすφ_i の対応選択肢を意思決定**
   - これまでの観測値から、意思決定の確信度を計算
   - $X_{0:t}$は時刻0からtまでのすべての観測値集合
   - λは閾値

8

性能評価

- **先行研究で使用していた公開データセット[10]を用いた評価**
  - Siamese RPN を用いた領域推定
    - Siamese Network の CNN 構造に 3、4層の CNN、AlexNet の３種
    - テンプレート画像は 612 フレーム目
  - 推定した領域内の映像から特徴量を抽出し、BAM への入力として用いて認識精度を評価
    - BAMのアトラクターは Siamese RPN で得られた領域推定結果のうち、IoU が高いものから取り出した特徴量を設定

テンプレートとして用いる612フレーム目の画像

[10] "YCB benchmarks–object and model set," available at http://www.ycbbenchmarks.com/

9

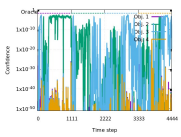
評価結果：軽量 CNN を用いた場合

3 層 CNN [24,32,16]

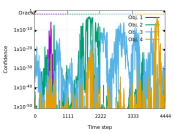
オブジェクト	1	2	3	4
平均 IoU	0.484	0.452	0.241	0.178
RPN の平均計算時間 (s/f)	0.0293	0.0297	0.0304	0.0294

4 層 CNN [16, 16, 16, 16]

オブジェクト	1	2	3	4
平均 IoU	0.279	0.502	0.196	0.131
RPN の平均計算時間 (s/f)	0.0243	0.0238	0.0233	0.0236



3層構造における BAM の信頼度



4層構造における BAM の信頼度

BAM の推定結果

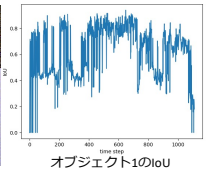
	精度[%]	計算時間[f/s]
3層	32.6	0.0539
4層	51.4	0.0393

10

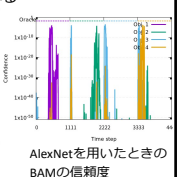
評価結果：AlexNet を用いた場合

オブジェクト	1	2	3	4
平均 IoU	0.626	0.461	0.306	0.382
RPN の平均計算時間 (s/f)	0.0509	0.0488	0.0484	0.0490

- ほとんどの場合において、 軽量 CNN の IoU を上回る
  - 一方で計算時間は 4層 CNN を用いた際の 2倍程度
- どの時間においても 比較的して、高い IoU が得られている



オブジェクト1のIoU



AlexNetを用いたときの BAMの信頼度

- BAM における認識精度 54.36%
- BAM における計算時間 0.0657s/f

11

まとめと今後の課題

- **まとめ**
  - デジタルツインの実現を目指し、脳の情報処理にならう物体認識手法を実装
  - BAM への入力として与えるための特徴量抽出器を実装
    - 領域推定と特徴量抽出を同時に実施
  - Siamse RPN の精度は特徴量マップの取り出し方に依存
    - IoU の精度が高いことが必要で、特徴量次元の大きな AlexNet が有効
      - 領域内に含まれる物体の特徴を抽出する可能性が高くなる
  - BAM のアトラクター設計が認識精度にとって重要

- **今後の課題**
  - マルチモーダル認識への拡張
    - 映像モーダルだけで得られる認識精度には限界がある
    - 映像モーダルのみならず他のモーダルの認識による精度向上を図る