特別研究報告

題目

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

指導教員
村田 正幸 教授

報告者
安藤 覇人

令和 5 年 2 月 7 日

大阪大学 基礎工学部 情報科学科

令和 4 年度 特別研究報告

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

安藤 覇人

内容梗概

近年のデジタル化技術の一層の発展により，各種機器やセンサから得られた情報を共有しあうことが可能となり，設計や製造などのシステムがデジタル上に統合されつつある．現実世界より得た情報からデジタル化されたシステムを仮想空間上に構築すること，つまりデジタルツイン (Digital Twin, DT) を構築することにより，システムの安全性や保守性を向上させる取り組みが期待されている．DT 実現にあたり，リアルタイムでのフィードバック制御の実現や観測情報の不確実性への対処などの課題が挙げられる．我々の研究グループでは，現実世界のセンサ機器では 100%正確な物体認識を実現することが困難であることから，不確実性を確率的な情報として DT に表現する，確率的 DT の実現を目指す研究を行っている．

複数の不確実な情報をもとに巧みに判断を行うシステムとして，脳の情報処理機構が挙げられる．人間は，目や耳といった感覚器官からノイズを含んだ情報を受け取り，自分の持っている知識や経験と照らし合わせて，意思決定を行う．この仕組みに着想を得た技術として，ゆらぎ学習という技術がある．ゆらぎ学習では，ベイジアンアトラクタモデル（Bayesian attractor model, BAM）と呼ばれる手法を用いて，ノイズを含む情報から物体認識を行う．BAM では，このようなノイズを含む情報をもとに行う意思決定を再現するために，ベイズ推定を利用して脳の情報処理機構をモデル化している．我々の研究グループでは，ゆらぎ学習を利用したマルチモーダル物体認識手法を利用した DT の構築に取り組んでいる．本報告では，モダリティの一つとして，点群データによって表される位置モダリティを用いる．与えられた点群データから特徴量を抽出し，BAM に基づく物体認識を行う手法を提案する．点群データから PointNet を用いてセグメンテーションを行い，物体の位置を抽出し，それを BAM に与えることで物体認識を行う手法を実装し，評価を行った．シミュレーションの結果，点群ベースの位置モダリティにおける物体認識を実現することができ，また，我々がこれまでに用いていた映像モダリティにおける物体認識手法と統合したマルチモーダルな物体認識も実現できることを示した．

主な用語

オブジェクト認識
デジタルツイン
マルチモーダル処理
ゆらぎ学習
ベイズ因果推論
点群データ

目 次

1 はじめに **6**

2 関連研究 **8**

2.1 脳の情報処理モデル 8

2.1.1 ベイジアンアトラクタモデル (Bayesian Attractor Model, BAM) . . . 8

2.1.2 ベイズ因果推論 (Bayesian Causal Inference, BCI) 10

2.2 PointNet 11

2.2.1 位置モダリティの特徴 11

2.2.2 PointNet アーキテクチャ 12

2.3 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 13

3 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 **16**

3.1 全体のアーキテクチャ 16

3.2 物体認識手法の実装 17

3.3 特徴量抽出 18

3.3.1 位置モダリティ 18

3.3.2 映像モダリティ 18

4 提案手法の評価 **19**

4.1 評価環境 19

4.2 利用するデータ 19

4.3 評価結果 20

4.3.1 単一モダリティを利用した場合の物体認識 20

4.3.2 複数のモダリティを利用した場合の物体認識 21

4.4 考察 22

5 おわりに **24**

謝辞 **25**

図 目 次

1 点群の性質 12
2 PointNet の構造，文献 [1] より引用，p3 図 2 13
3 脳のマルチモーダルな情報処理モデルに基づくオブジェクト推定手法，文献 [2] より引用，p4 図 1 14
4 物体認識を行うアーキテクチャの概要 16
5 時系列データ (0 フレーム目) 20
6 時系列データ (10 フレーム目) 20
7 椅子の認識結果（映像モダリティ） 21
8 机の認識結果（映像モダリティ） 21
9 椅子の認識結果（位置モダリティ） 21
10 机の認識結果（位置モダリティ） 21
11 椅子のマルチモーダル認識結果 22
12 机のマルチモーダル認識結果 22

表 目 次

1　点群データ成分 11
2　BAM のパラメータ 17
3　実験環境 19
4　認識精度による物体認識手法の評価 22

1　はじめに

産業のデジタル化が進み，IoT や CPS などの IT 技術を活用するエコシステムの構築が主流となっている [3]．技術の進歩により各種機器がセンサーや他の機器とリアルタイムで情報を相互に送りあうことが可能となり，設計や製造，現在のシステムの状態をデジタル上に統合することができる．また，これらの情報は現実世界の物理システムを仮想空間上で再現することにも役立つ．仮想空間において現実世界のシステムを再現することにより，リアルタイムに故障などの危険予測やシステムの最適化を行うことができ，システムの安全性や保守性を向上させることができる．

製造業では，人間と一緒に作業できる産業用ロボットの数も増えてきている．当初は一つのアームのみ実装され単調な作業しか実行できないものであったが，倉庫内を動きまわり人間と一緒に複雑な作業を行えるようにまで発展してきた．このようなロボットは協調ロボットと呼ばれる．協調ロボットは従来のような人間と区分けされた場所で別々に働くロボットと異なり，同じ環境でともに作業を行う．これを実現するための技術の一つがデジタルツイン (Digital Twin, DT) である．

DT とは，ひとつに定まった定義はないが，広く知られている定義として次のようなものがある．“ DT とは，複雑な製品を作成するために複数の物理法則やマルチスケールな処理，確率的シミュレーションを統合したものであり，また，利用可能な物理モデルやセンサーなどを適切に利用して現実世界と対応する双子を作成するものである” [4]．DT を用いることで製品製造の安全性や品質の向上などを達成することが見込まれる．先の協調ロボットの例では，ロボットにセンサーを内蔵することで，周囲の状況をリアルタイムに収集，解析し意思決定を行うことができ，また，ロボットの関節の状態やロボット内の各種素子の状態に容易にアクセスすることができるため，ロボットの状態についても監視を行うことができる．DT は様々な業界において複数のコンセプトがあるものの，共通認識として，物理世界の物体と対応するものを仮想世界に構築するものであり，それが含んでいる機械の寿命予測や管理のリエンジニアリングであると捉えることができる [5]．

DT の実現にはいくつか課題があり，そのうちの一つとして複数の不確実性を考慮する必要がある点が挙げられる．これには，センサによる測量の不確実性や物体の位置推定における不確実性などが該当する．センサによる物体検出の例を挙げると，データ取得時のノイズなどにより，すべての物体を完璧に検知することは不可能である [6]．したがって，システムにはこのような不確実性を考慮した振る舞いが必要となる．

我々の研究グループでは，この不確実性について焦点を当て，ノイズを含む不確実な入力から対象となる物体を認識する手法を研究している．この手法では，推定結果を確率的に表現することが可能であり，これを用いた確率的 DT の実現を目指している．

複数の不確実な情報をもとに処理を行うシステムとして，脳の情報処理機構が挙げられる [7–9]．人間は，目や耳といった感覚器官からノイズを含んだ情報を受け取り，意思決定を行っている．例えば車を運転する際，目が受け取った光から進行方向の信号が青であるという情報を抜き出し，信号は青から黄色に変わることや止まらなければならないことを考え，通行時に信号が黄色になると推測し，最終的にスピードを落とすという意思決定を行う．このような脳の情報処理モデルを利用したシステムとして，ゆらぎ学習 [7] が挙げられる．

ゆらぎ学習とは人の脳が行っている制御に倣った，ノイズを含む情報から分類，意思決定を行う制御システムのことを指す．ゆらぎ学習における認識は，ベイジアンアトラクタモデル（Bayesian attractor model, BAM）[8] に基づくものであり，我々の研究グループでは，BAM をベイズ因果推論 (Bayesian Causal Inference, BCI) [9] によってマルチモーダル化し，物体認識に応用した研究成果をあげている [2,10]．

BAM では，人の行う確率的な意思決定を再現するため，ベイズ推定を利用して脳の処理機構をモデル化している．また，BCI では，人間が複数のモダリティから取得した情報を結びつける過程を再現するため，因果関係を推論することで脳の処理機構をモデル化している．

先行研究である文献 [2] では，脳の情報処理を模したマルチモーダルな物体認識手法を提案している．当該文献では，物体を映した動画情報（映像モダリティ）と撮影したカメラの向きと深度画像から得た深度情報を統合して求めた位置情報（位置モダリティ）という二つのモダリティを組み合わせた物体認識手法を提案している．提案されている手法では，単一の機器から映像モダリティおよび深度モダリティを取得しているため，各モダリティにおける認識結果に相関が強く現れ，認識精度の低下が両モダリティにおいて同時に発生する場合がある．

そこで本報告では，先行研究 [2] の物体認識手法とは異なる機器から得られる観測情報をもとにした物体認識手法を新たに提案する．また，その認識手法を先行研究 [2] で行うマルチモーダル処理に適用し，精度の向上を図る．まず，異なるセンサ機器から取得するモダリティとして，RGB カメラを用いて取得する動画情報 (映像モダリティ) と LiDAR を用いて取得する点群データ (位置モダリティ) を利用する．映像モダリティを用いた物体認識および複数モダリティの統合は先行研究 [2,10] の手法を利用し，点群データの “位置” を利用する物体認識については，新たに手法を提案する．

まず，取得した点群を既知の物体のクラスに分類する．各物体に属する点の平均座標を各物体の位置とみなし，各フレームにおける物体の位置を BAM に観測値として与える．このように物体の位置を与えた時，その位置にある物体が何であるのかを，BAM は確率的な情報として出力する．このようにして得られた位置モダリティにおける認識結果と先行研究の手法 [10] を利用して得られた映像モダリティにおける認識結果を先行研究 [2] の手法にて統合し，物体認識を行う．

2　関連研究

本報告では，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を提案する．まず，利用する脳の情報処理モデルについて第 2.1 節にて述べる．続いて，新たに設計した位置モダリティを利用した物体認識において，点群から物体の認識を行うために利用する PointNet について第 2.2 節にて述べる．最後に，本報告で提案する物体認識手法の基礎となる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について，第 2.3 節にて述べる．

2.1　脳の情報処理モデル

2.1.1　ベイジアンアトラクタモデル (Bayesian Attractor Model, BAM)

既存研究 [8] では，人が行う確率的な意思決定を行う仕組みをモデル化したベイジアンアトラクタモデル (Bayesian attractor model, BAM) を提案している．BAM では，意思決定をベイズ推定としてモデル化，確率的に表現し，アトラクタモデルと組み合わせることで意思決定行動を説明している．内部に状態変数 $\mathbf{z}$ を持ち，n 個の選択肢に対し，それを選択するに値する証拠を十分に蓄積している状態を表すアトラクタ ϕ_i を同じ状態空間上に設定する．変数 $\mathbf{z}$ を与えられる観測値により更新していき，あるアトラクタ ϕ_i に十分に近づいたとき，BAM は出力として選択肢 i を出力する．

BAM は大きく分けて，(i) 感覚器から得た刺激を特徴ベクトルに変換する部分，(ii) 観測した値から未観測の値を推測する部分，(iii) ベイズ推論を行う部分，(iv) 意思決定を行う部分，の 4 つの構成要素からなる．

まず，刺激から特徴ベクトルへの変換部分 (i) について述べる．時刻 t に観測した刺激の性質 A_t を特徴ベクトル $\mathbf{x}_t$ にマッピングすることで受けとった刺激から特徴ベクトルへの変換を行う．同じ刺激から得られる観測値は同じ確率分布から生成されているものとする．生成される特徴ベクトルは正規分布に従うものとし，$\mathbf{x}_t \sim \mathcal{N}\left(\boldsymbol{\mu}_i, s^2\mathbf{I}\right)$ と表される．ここで s はノイズレベルと呼ばれる入力へのノイズの量 (標準偏差) を表し，$\boldsymbol{\mu}_i$ は，ノイズがない状態で選択肢 i を選択する場合の特徴量を表す．

続いて，刺激の生成モデル (ii) について述べる．BAM では，生成モデルはホップフィールドのダイナミクスに従うと仮定し，時間変化 (時刻 $t - \Delta t$ から t) による意思決定の状態の変化を以下の式 (1) のように定式化する．

$$\mathbf{z}_t - \mathbf{z}_{t-\Delta t} = \Delta t f\left(\mathbf{z}_{t-\Delta t}\right) + \sqrt{\Delta t w_t} \tag{1}$$

ここで以下のような変数を利用する．状態変数 $\mathbf{z}_t$ は時刻 t における内部状態を表す．各意思決定の選択肢 i に対応する成分 z_i が大きな値をとるとき，それは選択肢 i を採用するに

至るための証拠が十分に集まっていると考えることができる．関数 f は状態変数 $\mathbf{z}$ が表すベクトルに対するアトラクタのダイナミクスを定義する関数である．また w_t は時刻 t におけるノイズであり，正規分布に従うものとし，$w_t \sim \mathcal{N}(0, \mathrm{Q})$ と表される．ここで $\mathbf{Q}$ とは，$\mathbf{Q} = \left(q^2/\Delta t\right)\mathbf{I}$ で表される等方的な共分散であり，この q はダイナミクスの不確実性を表す．ダイナミクスの不確実性とは，予想されるノイズの量を推定する値であり，これが大きいほど，意思決定状態の切り替えが容易となる．特徴ベクトルへ変換後，観測された変数から観測されていない変数の状態を推論する．$\boldsymbol{\mu}_1 \ldots \boldsymbol{\mu}_N$ を行列にまとめたものを $\mathbf{M} = [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N]$ とする．状態 $\mathbf{z}$ が与えられたとき，BAM では，異なる選択肢の状態を補完するために以下の式 (2) に従い特徴量 $\mathbf{x}$ を予測する．

$$\mathbf{x} = \mathbf{M}\boldsymbol{\sigma}(\mathbf{z}) + \mathbf{v} \tag{2}$$

ここで $\mathbf{v}$ は時刻 t におけるノイズであり，正規分布に従うものとし，$\mathbf{v}_t \sim \mathcal{N}(0, \mathrm{R})$ と表される．ここで R は $\mathrm{R} = r^2\mathbf{I}$ で表される予想されるノイズの等方的な分散であり，この r は感覚的な不確実性を表す．感覚的な不確実性とは，意思決定を行う人が予測するノイズの量であり，データにどれほどノイズが含まれているのかを予測して設定される．関数 σ は多次元シグモイド関数を表し，引数の要素を $[0, 1]$ にマッピングする．

次に，ベイズ推論部 (ⅲ) について述べる．観測された刺激や推測された刺激を用いてベイズ推論を行う．ベイズ推論を用いて生成モデルを反転させることで推論の様子をモデル化する．その時点までの観測値 $\mathbf{X}_{\Delta t:t} = \{\mathbf{x}_{\Delta t}, \ldots, \mathbf{x}_t\}$ と生成モデル (式 (1)，(2)) から，時刻 t における決定状態 $\mathbf{z_t}$ を推論する．ここで，生成モデルは観測値がホップフィールドのダイナミクスに従うことを仮定していることから，この推論は時系列データが互いに依存するという仮定を考慮する必要がある．この決定状態 $\mathbf{z}_t$ を求める推論はフィルタリング問題という推定問題に分類され，提案手法では，モデルが非線形であることを考慮し UKF(uncented kalman filter) というカルマンフィルタを採用した．これを用いることで，近似された確率密度関数 $p\left(\mathbf{z}_t \mid \mathbf{X}_{\Delta t:t}\right)$ を得ることができる．

最後に，意思決定を行う部分 (ⅳ) について述べる．意思決定の判断基準は以下の式 (3) で表される．

$$p\left(\mathbf{z}_t = \phi_i \mid \mathbf{X}_{\Delta t:t}\right) \geq \lambda \tag{3}$$

左辺の $p\left(\mathbf{z}_t = \phi_i \mid \mathbf{X}_{\Delta t:t}\right)$ は選択肢 i に対応するアトラクタ ϕ_i で評価される決定状態の事後確率密度であり，つまり選択肢 i が選択肢であるという決定を行った後の確信度である．閾値 λ は信頼度の指標として解釈することができる．

2.1.2 ベイズ因果推論 (Bayesian Causal Inference, BCI)

既存研究 [9] では，人が異なるモダリティから取得した情報を組み合わせて認知を行う仕組みを，ベイズ推論 (Bayesian) と因果推論 (Causal Inference) を組み合わせてモデル化したベイズ因果推論 (Bayesian Causal Inference, BCI) を提案している．例えば，視覚からの情報として救急車が視界の左側に見え，聴覚からの情報としてサイレンの音が左側から聞こえたとき，人はこれらが同じ救急車という物体から得た情報であると認知することができる．このようなマルチモーダルな情報処理をモデル化したものが BCI である (以降二つのモダリティを扱うものとする)．BCI は 2 つのモダリティから情報を得た際，それらの入力が共通の原因から発せられたとするモデル，forced-fusion model と，それらの入力が独立した原因から発せられたとするモデル segregation model，という，2 つのモデルを考慮する．これらのモデルに対し，モデル平均化法 (model averaging) と呼ばれる手法を用いて統合し，物体の認識を行う．それぞれの因果構造の事後確率により重みづけして平均化することで，脳の不確実性を再現している．

BCI は大きく分けて (i) それぞれのモダリティが独立して刺激を取得したとするモデル と，(ii) それぞれのモダリティを統合して活用するモデル ，そしてこれらに対して (iii) モデルの平均化を行い，最終的な推論を行う部分の 3 つの構成要素からなる．以下では文献に倣い，扱うモダリティを視覚と聴覚とする．

まず，2 つの生成モデル (i)，(ii) について解説する．モデルの事前分布を p_{common}，事前分布の平均を μ_P，標準偏差を σ_P，視覚，聴覚のそれぞれの標準偏差を σ_A, σ_V とする．観測した事象が共通の原因 (cause, c) から発生したとするときを変数 c を用いて $c = 1$，独立の原因であるとするときを $c = 2$ とする．これらは二項分布からのサンプリングによって決定されるものと仮定する ($p(C = 1) = p_{common}$)．共通の原因の場合，視聴覚刺激の真の数 N_{AV} は事前分布 $\mathcal{N}(\mu_P, \sigma_P)$ から求まる．独立した原因の場合，刺激の真の聴覚の数 N_A，視覚の数 N_V はこれらの事前分布から独立に求まる．また，パラメータ σ_A, σ_V を持つ真の刺激数を中心とした正規分布からサンプリングした x_A, x_V を用いることで感覚的なノイズを再現する．ベイズ則に基づきこれらを組み合わせることにより，原因となっている因果構造の事後確率を以下の式 (4) のように推論することができる．

$$p(C = 1 \mid x_A, x_V) = \frac{p(x_A, x_V \mid C = 1) p_{common}}{p_(x_A, x_V)} \tag{4}$$

観測された刺激が (i) 共通の原因である場合その推定値 $\hat{N}_{AV,C=1}$ は以下の式 (5) で求められる．

$$\hat{N}_{AV,C=1} = \frac{\frac{x_A}{\sigma_A^2} + \frac{x_V}{\sigma_V^2} + \frac{\mu_P}{\sigma_P^2}}{\frac{1}{\sigma_A^2} + \frac{1}{\sigma_V^2} + \frac{1}{\sigma_P^2}} \tag{5}$$

観測された刺激が (ii) 独立の原因である場合それぞれの推定値 $\hat{N}_{A,C=2}, \hat{N}_{V,C=2}$ は以下の式 (6)，(7) で求められる．

$$\hat{N}_{A,C=2} = \frac{\frac{x_A}{\sigma_A^2} + \frac{\mu_P}{\sigma_P^2}}{\frac{1}{\sigma_A^2} + \frac{1}{\sigma_P^2}} \tag{6}$$

$$\hat{N}_{V,C=2} = \frac{\frac{x_V}{\sigma_V^2} + \frac{\mu_P}{\sigma_P^2}}{\frac{1}{\sigma_V^2} + \frac{1}{\sigma_P^2}} \tag{7}$$

続いて，統合処理を行う部分 (iii) について述べる．モデル平均化は以下の式 (8)，(9) で表される．

$$\hat{N}_A = p(C=1 \mid x_A, x_V)\hat{N}_{AV,C=1} + (1 - p(C=1 \mid x_A, x_V))\hat{N}_{A,C=2} \tag{8}$$

$$\hat{N}_V = p(C=1 \mid x_A, x_V)\hat{N}_{AV,C=1} + (1 - p(C=1 \mid x_A, x_V))\hat{N}_{V,C=2} \tag{9}$$

BCI では，3 つの推定値 ($\hat{N}_{AV,C=1}$,$\hat{N}_{A,C=1}$, $\hat{N}_{V,C=2}$) を組み合わせることで，最終的な推定値 $(\hat{N}_A, \hat{N}_V)$ を求める．

2.2 PointNet

2.2.1 位置モダリティの特徴

LiDAR にて測定された，今回位置モダリティとして扱う点群データ (point cloud) は次のような成分を持つ．今回はこれらのうち点群の座標を表す x，y，z 成分を利用する．

表 1: 点群データ成分

変数名	内容
x,y,z	取得した点の座標
intensity	強度

点群データは順序不変性，剛体運動に対する不変性，そして局所性という性質を持つ [1]. これについて整理したものを図 1 に載せる．図中の丸は点群を表し，A–C は各性質を図式化したものである．各性質について紹介する．

順不変性 (Unorderd, A) とは，点群を入力として与えたとき，その順序がいかなる場合でも出力が不変であるという性質である．点群は 2 次元の画像データのように規則ある並びをしておらず，不規則に各点の情報が格納されている．そのため，与えられる順序は一定でないため，与えられる順序に関係なく一定の出力を得られるようなモデルを構築する必要がある．

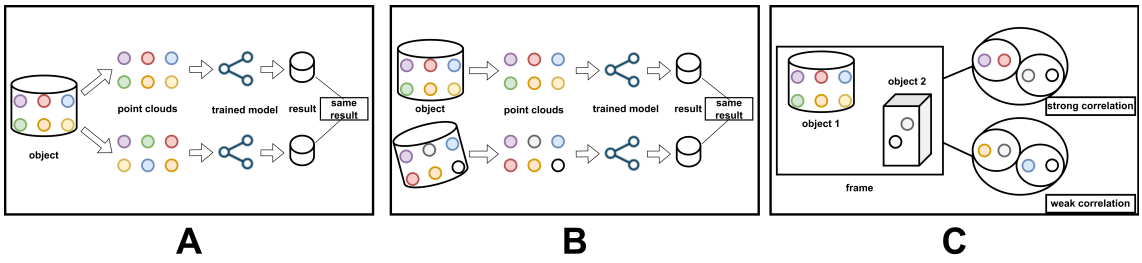

図 1: 点群の性質

剛体運動に対する不変性 (Invariance under transformations, B) とは，与えられた点群に対して，平行移動や回転などの処理を施した場合でも出力は不変であるという性質である．点群はある物体の一種の表現であるため，回転や移動などの操作に対して，その点群が表していた構造は不変であるべきである．例えば，点を回転させたときにその点が属するカテゴリ，ラベルが変更されるべきではない．

局所性 (Interaction among points, C) とは，座標が近い点群間には密接な関係があり，遠い点群間には関係が薄いという性質である．各点は孤立しているものではなく，近傍の点の集合はなんらかの意味ある部分集合を形成していると捉えることができる．

点群データを扱う際にはこれらの条件を満たしたシステムを採用する必要がある．本報告では，測定した点群データを処理するために，PointNet を採用した．

2.2.2 PointNet アーキテクチャ

既存研究 [1] では，点群データを直接処理するニューラルネットワークとして，PointNet を提案している．このネットワークでは，物体の分類やシーンの意味的な解釈などを行うことができる．PointNet は順不変である点群を入力として与えられ，前述した点群に関する性質を満たすようなネットワークである．

PointNet のアーキテクチャは以下の図 2 のように表される．なお図中の mlp とは多層パーセプトロン (Multi-Layer Perceptron, MLP) の略である．PointNet は Classification Network と Segmentation Network からなる．Classification Network では n 個の点を入力として受け取り，k 個のクラスに対するスコアを出力する．Segmentation Network はこれを拡張したものであり，点ごとのスコアを出力する．まず input transform の部分にて入力点群に対して，アフィン変換を行う．次にその点群に対して mlp にて特徴量の変換処理を行い，再び feature transform にてアフィン変換を行う．最後に mlp で処理をして max pooling を行い出力を得る．

PointNet がいかにして前述した 3 つの性質 (i) 順序不変性，(ii) 剛体運動に対する不変性，

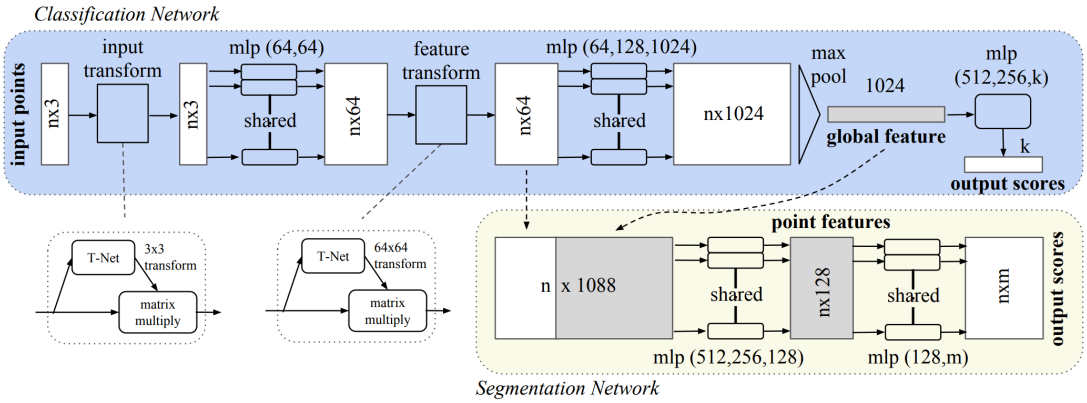

図 2: PointNet の構造，文献 [1] より引用，p3 図 2

そして (iii) 局所性を解決しているのかについて述べる．

まず，順序不変性 (i) について述べる．PointNet では，各点への処理の後に Max pooling を行うことで，前述した順不変性を担保している．これはどのような並びで特徴量が与えられても最大のものを選択するため，与えられた点群の順序によらない結果を得られるからである．

次に，剛体運動に対する不変性 (ii) について述べる．input transform や feature transform は T-Net と呼ばれる，ネットワークを構成要素にもつ．これは，与えられた点群を回転や移動について正規化することを目的としているためである．これに点群を与えることで，そのアフィン行列を得ることができる．これにより，剛体運動に対する不変性を担保している．

最後に，局所性 (iii) について述べる．各点について，その点が持つ特徴量と，全体の特徴量を統合する処理を行う．統合された特徴量をもとに計算した新たな特徴量を利用することで，局所的な特徴量と大域的な特徴量について考慮している．

2.3 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

先行研究 [2] では，不確実な情報をもとに意思決定を行う仕組みを利用したマルチモーダルな物体認識手法を提案している．この手法では，複数のモダリティごとに受け取った入力に対して，BAM による識別を行い，それぞれから得られた出力を BCI を用いて統合し，物体の認識を行う．複数モダリティを用いた手法において，ユニモーダルでは判断できなかった部分，例えば映像モダリティであれば暗い部屋を撮影したことにより視認性が低い映像が入力として与えられた場合などにおいて，互いの認識結果を用いて補いあうことで，より正確な判断が可能となる．提案手法を用いることで，ユニモーダルでは適切に認識できなかった物体に対して，BAM と BCI を用いてマルチモーダル処理を行うことで適切に認識できる

ことが示されている．ここでは，(i) 映像モダリティでの物体認識，(ii) 位置モダリティでの

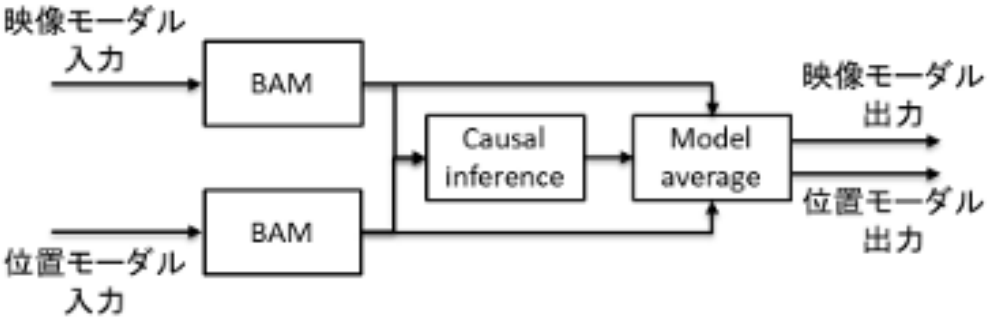

図 3: 脳のマルチモーダルな情報処理モデルに基づくオブジェクト推定手法，文献 [2] より引用，p4 図 1

物体認識，そして (iii)BCI での統合手法の設計について紹介する．

まず映像モダリティ (i) について述べる．各フレームの画像から抽出した特徴量を入力とする．また，あらかじめ用意された追跡対象の画像データから抽出した特徴量を BAM のアトラクタとする．BAM で確信度を求めるにあたり，BAM が観測する特徴量，アトラクタに与えるデータ，そして不確実性のパラメータの設定の 3 つが必要となる．特徴量について，Siamese-RPN を用いて抽出する．これはテンプレート画像と検出画像の 2 つの入力から，類似する箇所を検出し，その位置をバウンディングボックスとして出力するものである．画像の特徴量を抽出する部分とそれを用いて類似する箇所を見つける部分からなり，提案手法では前者に 4 層からなる簡潔な CNN を利用することで軽量化を実現している．BAM の入力として与える映像モダリティの特徴量は，出力される BB に対応する抽出された特徴量のうち 128 次元のデータに設計されている．アトラクタについて，提案手法では最初に見た 1 フレームを参照データとして用いたアトラクタを生成する設計である．不確実性のパラメータについて，生成された特徴量の分散を一定の値にするように正規化を行い，観測にかかるノイズを決定することで，大きさを決定している．

続いて，位置モダリティ (ii) について述べる．センサ機器であるカメラの方向ベクトルと複数のフレームを統合して深度情報から算出される 3 次元のワールド座標系データを BAM に与える特徴量として利用している．アトラクターおよび不確実性のパラメータについて，映像モダリティと同様に特徴量に基づいて設定する．

最後に BCI で確信度を統合する手法 (iii) について述べる．前述したように各モダリティごとに BAM を適用させて物体推定が行われ，それぞれのモダリティにおける入力として BCI に与えらる．これらに対し，Causal Inference を行い同じ物体を観測しているかを推論し，

その結果を基に Model Averaging でマルチモーダル統合することで，最終的な物体推定の結果を出力する．文献 [2] では，BCI で最終的な認識判断を行うために，出力結果はその物体が何であるのか，というラベル情報を出力する．

3　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

本論文では，第 2.3 節で紹介した先行研究における手法 [2] を拡張し，異なる入力から異なるモダリティを取得し，統合処理するマルチモーダルな情報処理モデルを提案する．まず，設計した提案手法の全体像について，第 3.1 節にて述べる．続いて，利用した BAM，BCI の設定について，第 3.2 節にて述べる．最後に，利用する特徴量について．第 3.3 節にて述べる．

3.1　全体のアーキテクチャ

本論文で提案する物体認識手法について述べる．全体の流れは図 4 のようになる．図 4 ではモダリティを取得する機器として，位置モダリティに対しては LiDAR を，映像モダリティとしてはカメラを設定している．

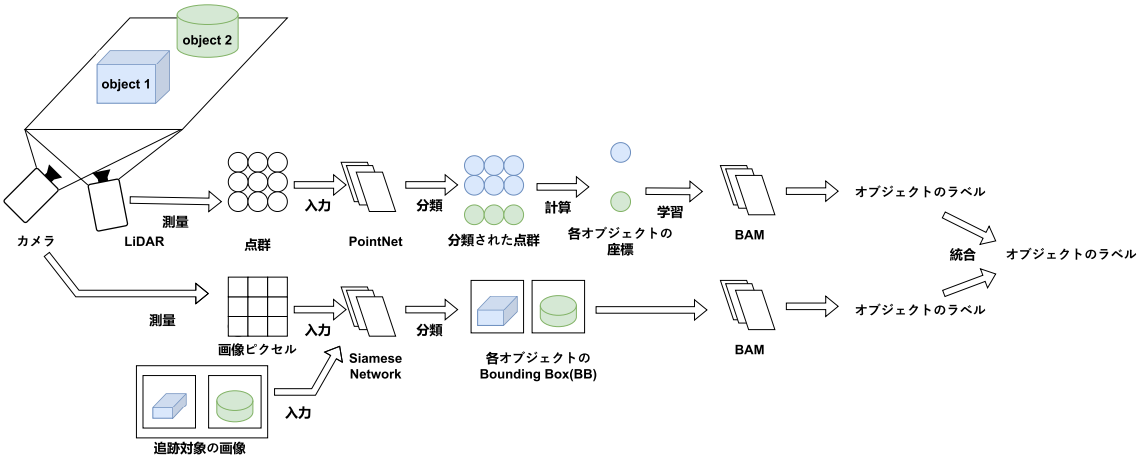

図 4: 物体認識を行うアーキテクチャの概要

まず，位置モダリティにおける物体認識手法について述べる．始めに，物体の位置に関する情報を取得する．この時，取得するデータは点群データ (point cloud) として表現され，物体の位置を表す座標に関する情報を利用する．続いて，点群データを追跡対象の物体にクラス分類する．本報告では，クラス分類を行うために，PointNet [1] を利用する．PointNet を用いたクラス分類により，各点群に対しそれが属する物体のラベルを取得することができる．次に，各物体の位置を求める．今回は特徴量として，物体の重心を物体の座標と定める．各物体に属する点群を取得し，各物体の重心となる座標を算出する．それらから得た情報を用いて BAM の学習を行う．最終的に得られるモデルでは，物体の位置を表す座標を入力とし，それが属する物体を確率的な形で表現したものを出力とする．

続いて，映像モダリティの取得方法について述べる．今回は前述したように先行研究 [10] の手法を利用する．まず，物体の位置に関する情報を取得する．この時，取得したデータは画像ピクセルとして表現され，対象を表す RGB データやその座標を利用する．続いて，取得した画像ピクセルの中から追跡対象を検出する．本報告では，SiameseRPN と呼ばれるネットワークを利用した先行研究 [10] の手法を利用し特徴量抽出を行う．この時，Siamese Network には先ほど取得した画像の他に，事前に撮影した追跡対象となる物体の画像も与える．これにより，各物体が存在する場所を表すバウンディングボックス (Bounding Box, BB) を取得することができる．それらから得た情報を用いて BAM の学習を行う．最終的に得られるモデルでは，物体が捉えられている画像を入力とし，その中にある物体を確率的な形で表したものを出力とする．

最終的にこれらの二つのモダリティから得られた情報を既存手法 [2] を用いて統合し，物体認識を行う．

3.2 物体認識手法の実装

先行研究 [2] と同様に，脳のマルチモーダル情報処理システムとして BAM を採用した．BAM の不確実性を表すパラメータ設定について述べる．BAM の持つパラメータとして，ダイナミクスの不確実性 q，感覚の不確実性 r があげられる．ダイナミクスの不確実性は選択肢間の切り替えの傾向として解釈でき，q の値が高いほど，状態が切り替わる可能性が高くなる．また，感覚の不確実性は選択肢に収束するための情報の精度として解釈でき，r の値が高いほど，入力から内部状態の推定が受ける影響は小さくなる．今回はこれらを以下の表 2 のように設計した．

表 2: BAM のパラメータ

パラメータ名	値
r	1.0
q	0.1

BAM によって物体の認識を行うためには，その物体をセンサで観測したときに得られる代表的な特徴量をあらかじめ，BAM のもつアトラクタに記憶しておく必要がある．本報告では，文献 [2] で行っている方法と同様に，分析を行う映像の 1 フレーム目から抽出した特徴量を BAM に記憶させることとする．

また，映像モダリティより抽出した特徴量と，位置モダリティより抽出した特徴量をそれぞれ BAM に与えることで，映像モダリティにおいて観測した物体が何であるのかという確

信度と，位置モダリティにおいて観測した物体が何であるのかという確信度のそれぞれが得られる．これらの統合についても，第 2.3 節で説明した，文献 [2] の方法を用いることとする．

3.3 特徴量抽出

3.3.1 位置モダリティ

本手法では，位置モダリティとして，物体を表す点群を利用する．これを PointNet に与えることで，各点がラベルつけされたものが得られる．これを利用して各物体ごとの位置を求め，BAM の特徴量として与える．この時，位置は各物体に属する点群の座標の平均値であると定義する．

3.3.2 映像モダリティ

本手法では，映像モダリティとして，物体を撮影した画像を利用する．映像モダリティで用いるネットワークとして Siamese RPN を採用しているため，追跡対象を表す画像を推論の際に与える必要があるが，これには文献 [2] と同様に，1 フレーム目の映像から抽出するものとする．撮影した画像を Siamese RPN に与えることで，入力画像内の追跡対象が位置している BB が得られ，その内部の映像を Siamese RPN が使用している Siamese network によってエンコードしたものを BAM の特徴量として与える．

4　提案手法の評価

上述した物体認識手法の性能を評価するため，認識精度の評価を行った．まず，第 4.1 節にて，評価環境について述べる．次に，第 4.2 節にて，撮影したデータについて述べる．続いて，第 4.3 節にて，評価手法および評価結果について述べる．最後に，第 4.4 節にて，結果に関する考察について述べる．

4.1　評価環境

評価を行った計算機環境について述べる．使用した計算機の性能は以下の表 3 の通りである．

表 3: 実験環境

項目	値
CPU	Core i9 10940X
GPU	NVIDIA RTX A5000
メモリ	128G

4.2　利用するデータ

評価を行うにあたり，The Stanford 3D Indoor Scene Dataset (S3DIS) [11] を利用した．提案する物体認識手法は，時系列データに対して，観測の蓄積を行いながら，認識判断を行うものである．そのため，データセットも時系列ごとに準備されたものが必要となる．広く使われている点群のデータセットは，シーンごとに準備されたものであるため，本報告では，用意されたシーンの物体にノイズや移動処理を加え擬似的な時系列データとして使用することとする．

このデータセットでは各シーンごとのオブジェクトが rgb 情報を付与された点群として用意されているため，この点群の xyz 座標を位置モダリティとして利用する．また rgb 情報を用いてこれらを撮影して取得した画像を映像モダリティとして利用することとした．生成した各フレームごとにモダリティを取得し，位置推定および，その精度の評価を行う．撮影したデータのうち 1 フレームを以下の図 5，6 に載せる．このシーンでは，二つの物体（椅子と机）を認識対象とする物体としている．椅子は元の座標から動かない物体として，机は x 軸方向に常に動いてる物体として設定し，どちらもその座標成分にノイズを加えてデータセットを生成した．これにより，ノイズを含んだマルチモーダルな時系列データセットになっている．

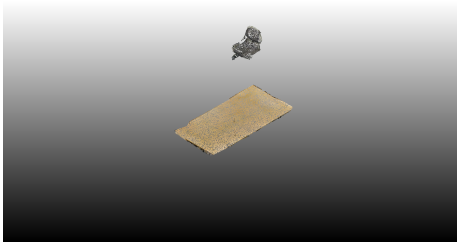

図 5: 時系列データ (0 フレーム目)

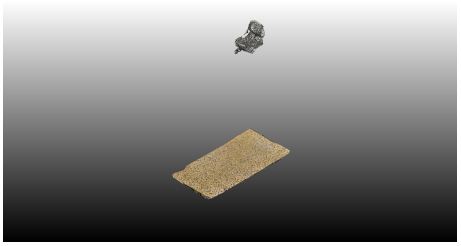

図 6: 時系列データ (10 フレーム目)

以下の評価では，このように用意したデータセットに対して，映像モダリティと位置モダリティを取得した際に，その認識結果と正解ラベルを比較し，正しく物体認識を行えているのかを確認する．

4.3 評価結果

4.3.1 単一モダリティを利用した場合の物体認識

まず，映像モダリティを用いた認識結果について述べる．認識結果を以下の図 7，8 に載せる．縦軸は観測対象に対する確信度であり，横軸はフレーム数である．

BAM には 1 フレーム目にて観測された椅子および机の映像特徴量をアトラクタとして記憶する．図 7 では入力として椅子を観測した場合の映像特徴量を，図 8 では入力として机を観測した場合の映像特徴量を与え，それぞれの確信度の推移を示したものである．椅子である確信度はオレンジの折れ線，机である確信度は青色の折れ線で表される．図 7 では，椅子であると判断する確信度が高い値を示すか，あるいはいずれの確信度も低い値である．図 8 では，椅子あるいは机であると判断する確信度が変動して，いずれかが高い値を示していることがわかる．確信度の高い側を認識結果とすると，椅子の認識精度は 98%であり，机の認識精度は 65%であった．映像については我々の先行研究 [10] において Siamese RPN での認識結果に大きく影響を受けることが分かっており，今回も認識結果が誤る状況は，同様の原因であると考えられる．

続いて，位置モダリティを用いた認識結果について説明する．認識結果を以下の図 9，10 に示す．縦軸は観測対象に対する確信度であり，横軸はフレーム数である．

BAM には 1 フレーム目にて観測された物体の位置を表す特徴量をアトラクタとして記憶し，図 9 では入力として椅子を観測した場合の特徴量を, 図 10 では入力として机を観測した場合の特徴量を与えて，それぞれの確信度を示している．椅子である確信度はオレンジの折れ線，机である確信度は青色の折れ線で表される．確信度の高い側を認識結果とすると，椅子の認識精度は 100%であり，机の認識精度は 97%であった．映像モダリティの際の認識

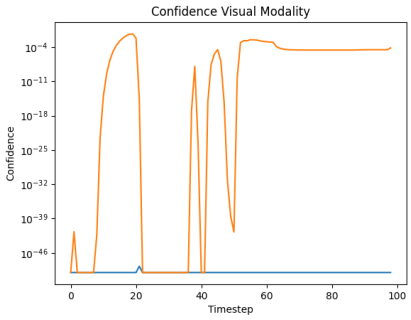

図 7: 椅子の認識結果（映像モダリティ）

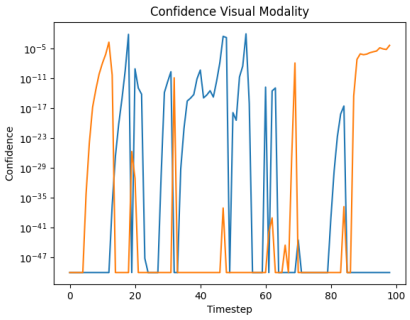

図 8: 机の認識結果（映像モダリティ）

結果と比べて，どちらも安定して，正しい認識結果が得られており，点群データを用いたゆらぎ学習を利用した物体認識が有用であることがわかる．

図 9: 椅子の認識結果（位置モダリティ）

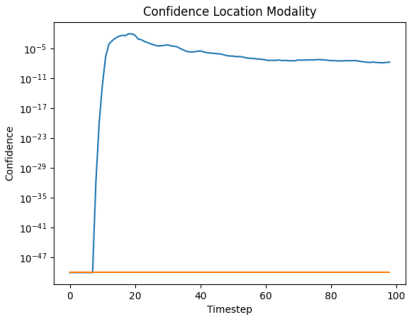

図 10: 机の認識結果（位置モダリティ）

4.3.2 複数のモダリティを利用した場合の物体認識

複数のモダリティを利用した場合の物体推定の結果を以下の図 11,12 に載せる．文献 [2] の統合方式は，認識結果として物体のラベルを出力するものとなっている．そのため図の縦軸は認識している物体のラベルを表し，横軸はフレーム番号を表す．マルチモーダル認識によって，椅子の認識精度は 100%であり，机の認識精度は 79%であった．

図 11 では，高い精度を安定して出している位置モダリティの精度が不安定な映像モダリティでの認識結果を補っている結果となっている．

図 12 では，椅子の場合と同様に，認識結果を補っている箇所もある一方で，位置モダリティのみであれば正しい認識結果が得られていた箇所が，映像モダリティが誤った認識を高く確信しているときに，そちらに強い影響を受けて誤ったラベルに分類されている場合があった．

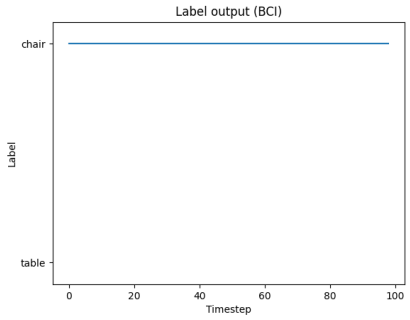

図 11: 椅子のマルチモーダル認識結果

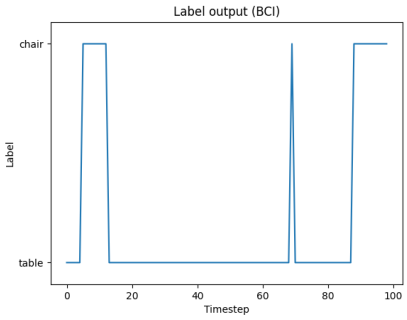

図 12: 机のマルチモーダル認識結果

4.4 考察

それぞれの認識精の評価結果として得られたものを以下の表 4 にまとめる．モダリティの項目は物体認識に用いたモダリティを表し，椅子，机の項目は追跡対象とした物体の特徴量を入力した結果得られた認識精度を表す．位置モダリティにおける物体認識手法は高い認識率を確認できた．マルチモーダルな物体認識手法について，映像モダリティにおける物体認識では低い認識結果が得られていたが，位置モダリティにおける物体認識では高い認識結果が得られており，当初のマルチモダリティを採用した目的である，あるモダリティでの精度が低下した場合に他のモダリティでの認識結果で補う，という処理を実現できていることが示せた．今回のテストでは位置モダリティにおいて高い精度を確認できているため，映像モダリティでの認識結果がその精度を低下させているとも考えられるが，これについては，先行研究では精度を向上させている例も確認できており，モダリティを追加して認識手法を拡張した際など，これらのモダリティが与える影響についての今後検討したい．

表 4: 認識精度による物体認識手法の評価

モダリティ	椅子	机
位置モダリティ	100%	97%
映像モダリティ	98%	65%
マルチモダリティ	100%	79%

先行研究 [2] では同一機器から取得した映像・位置モダリティを使用しており，映像モダリティにおける認識と位置モダリティにおける認識において，一部相関関係がみられていた本報告で提案した位置モダリティは，文献 [2] で用いている機器とは異なる機器から取得するため，映像モダリティを用いた認識と位置モダリティを用いた認識に相関が強く現れる問題が解決できる．

第 4.3.2 節では，あるモダリティにおける認識結果の低下を他のモダリティの認識結果を用いて補うというマルチモダリティの利点を生かすという点で，有用であることが示せた．

また，ノイズを加えたデータセットを評価に用いたことで，位置モダリティに関する提案手法がノイズに対してロバストであることを示せた．今後は，今回新たに設計した位置モダリティにおける物体認識手法について，点群データを用いた物体認識手法をゆらぎ学習と組み合わせるということを目的に設計したため，その計算速度について考慮していないことから，実用に向けた計算速度の向上について検討する．また，実際に測量した時系列データを用いた有効性の評価についても取り組む．

5　おわりに

本報告では，不確実な入力情報から意思決定を行う脳の仕組みに着想を得て，ゆらぎ学習を利用したマルチモーダルな物体認識手法を提案した．また，位置モダリティを利用した物体推定において，PointNet を利用して特徴量の抽出を行い，ゆらぎ学習を利用してノイズへのロバスト性を持たせた物体認識手法を提案した．既存手法の映像モダリティにおけるゆらぎ学習を利用した物体認識手法と組み合わせることで，より正確な物体認識を実現することができた．今回設計した位置モダリティにおける物体認識手法について，物体認識手法にロバスト性を持たせるということを目的として実装したため，その計算速度について考慮していない．今後は実用に向けた計算速度の向上について取り組んでいきたい．

謝辞

本報告を終えるにあたり，日頃より熱心にご教授いただきました大阪大学大学院情報科学研究科の村田 正幸教授に深謝いたします．また，大阪大学大学院情報科学研究科の小南 大智助教授には日頃からご指導やご助言をいただき，研究を進めることができました．心より感謝申し上げます．最後に，大阪大学大学院情報科学研究科村田研究室の皆様には本研究の遂行および本論文作成のために多大なご助言，ご協力いただきました．ここに感謝の意を表します．

参考文献

[1] 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).

[2] 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**, 15, pp. 59–64 (2021).

[3] A. K. Ramasubramanian, R. Mathew, M. Kelly, V. Hargaden and N. Papakostas: “Digital twin for human-robot collaboration in manufacturing: Review and outlook”, Applied Sciences, **12**, 10, p. 4811 (2022).

[4] 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).

[5] E. J. Tuegel, A. R. Ingraffea, T. G. Eason and S. M. Spottswood: “Reengineering aircraft structural life prediction using a digital twin”, International Journal of Aerospace Engineering, **2011**, pp. 1–14 (2011).

[6] G. P. Meyer and N. Thakurdesai: “Learning an uncertainty-aware object detector for autonomous driving”, 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 10521–10527 (2020).

[7] M. Murata and K. Leibnitz: “Fluctuation-induced network control and learning: Applying the yuragi principle of brain and biological systems”, Springer Singapore (2021).

[8] S. Bitzer, J. Bruineberg and S. J. Kiebel: “A bayesian attractor model for perceptual decision making”, PLOS Computational Biology, **11**, 8, pp. 1–35 (2015).

[9] 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).

[10] 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).

[11] 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, pp. 1534–1543 (2016).