社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

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

安藤　覇人† 　小南　大智† 　関　　良我† 　村田　正幸† 　下西　英之††

† 大阪大学 大学院情報科学研究科　〒565–0879 大阪府吹田市山田丘 1–5
†† 大阪大学 サイバーメディアセンター　〒560–0043 大阪府豊中市待兼山町 1–32
E-mail: †{h-andou,d-kominami,r-seki,murata}@ist.osaka-u.ac.jp, ††shimonishi.cmc@osaka-u.ac.jp

あらまし　Beyond 5G/6G 技術の発展により，デジタルツインの研究が進んでいる．Beyond 5G の特徴である mMTC によって多様なセンサから得られる情報を活用することが可能になり，デジタルツイン構築に向けた現実世界の物体の認識技術が重要となる．しかしながら，センシングには様々な不確実性が伴い，例えば RGB カメラであればその照明条件やカメラの性能に影響を受けるため，ユニモーダルな物体認識には限界がある．本論文では，先行研究であるマルチモーダルな物体認識手法を複数センサに対して扱えるように拡張した．点群を解析するモデル PointNet と脳の認知モデル BAM を組み合わせた位置モダリティの認識手法を新たに開発し，既存モデルに加えた．また，評価のために倉庫内で人とロボットが協調して働く場面を想定したデータセットを新たに構築した．我々のデータセットを用いた検証では，位置情報を活用したマルチモーダル統合を行うことでオブジェクト認識の適合率の向上がみられた．
キーワード　デジタルツイン，物体認識，マルチモダリティ，ベイジアンアトラクターモデル，ベイズ因果推論

Multimodal Object Recognition Method Using Bayesian Attractor Model For 3D Point Clouds and RGB Images.

Haruhito ANDO†, Daichi KOMINAMI†, Ryoga SEKI†, Masayuki MURATA†, and Hideyuki SHIMONISHI††

† Graduate School of Information Science and Technology, Osaka University
1–5 Yamadaoka, Suita, Osaka, 565–0879 Japan
†† Cybermedia center, Osaka University　1–32 Machikaneyama, Toyonaka, Osaka, 560–0043 Japan
E-mail: †{h-andou,d-kominami,r-seki,murata}@ist.osaka-u.ac.jp, ††shimonishi.cmc@osaka-u.ac.jp

Abstract　Beyond 5G/6G, technology is driving the development of digital twins. In recent years, the amount of information that can be used to perceive the environment has increased due to the development of sensors and learning technologies. However, unimodal object recognition is limited because sensing is subject to various uncertainties, for example, RGB cameras are affected by the lighting conditions and camera performance. In this paper, we propose the multimodal object recognition method by expanding our previous work to handle multiple sensors. We developed a new location modality recognition method that combines PointNet and BAM for object recognition using point clouds. For Evaluation, we also developed a new dataset for a situation where humans and robots work together in a warehouse. Validation on our dataset shows that multimodal integration with location information improves the precision of object recognition.
Key words　Digital twin, Object recognition, Multimodality, Bayesian attractor model, Bayesian causal inference

1. はじめに

次世代の情報通信インフラ技術（Beyond 5G/6G）を用いた通信システムでは，現実世界の様々な物体を多様なセンサを通して認識し，仮想空間上で再現する技術が重要な役割を果たすと期待されている [1]．カメラや触覚センサ，LiDAR (Light Detection And Ranging) センサに代表されるセンサ技術の発展と機械学習手法の進歩，そして通信インフラの発達によっ

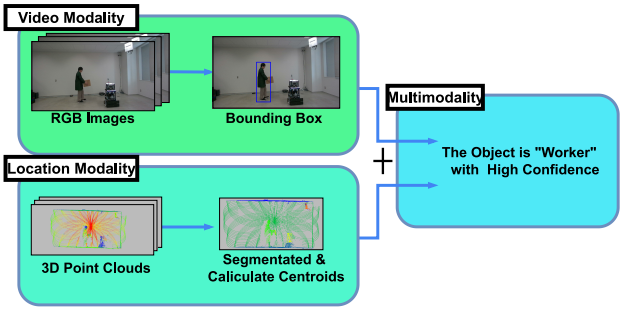


図 1　マルチモーダルな物体認識の例

て，周囲の環境を多角的に認識することが可能になる [2], [3]. 6G における環境の認識技術を用いたアプリケーションの例として，人間の作業員とロボットとの協調作業がある．近年では協調作業中のロボットの安全性と効率性を考慮した確率的制御の研究が盛んであり [4]，そこでは環境を認識した結果を確率的に取り扱うことが必要となる．

確率的ロボット制御において，周囲環境を認識することは重要なタスクであり，ユニモーダルセンサでの認識は限界がある．ユニモーダルなシステムでは，様々な不確実性に影響され，例えば RGB カメラでは，照明条件が撮影データに影響を与える．また，情報量の欠落もユニモーダルな物体認識の問題であり，例えば RGB カメラからの情報は物体との距離やテクスチャといった情報が欠落し，LiDAR センサからの情報は色彩に関する情報が欠落する．これらの理由から異なる種類のセンサを用いたマルチモーダルな物体認識に関する研究が多く進められている．[3].

我々の先行研究では，脳の認知モデルに基づいたマルチモーダルな物体認識手法を提案した [5]．人の脳はベイズ推論に基づき外界の認識を行っている説が近年唱えられており，この手法では，ベイズ推論に基づく脳の情報処理モデルである Bayesian Attractor Model (BAM) [6] と，Bayesian Causal Inference (BCI) [7] を組み合わせ，確率的な物体認識を行う．BAM はノイズを含む観測情報から人の行う意思決定をモデル化したものであり，ユニモーダルな物体認識に用いる．BCI は人の視覚や聴覚の情報を組み合わせた推論をモデル化したものであり，異なるモダリティから得られた認識結果の統合に用いる．移動するカメラが静止物体を撮影した公開データセットを用いた評価を通して，先行研究では高精度な物体認識を実現した．

脳の情報処理システムが物体認識に活用できることを示した一方で，先行研究 [5] にはいくつかの課題が残っている．一つは 2 つのモダリティでの認識結果を統合する際に認識結果間に依存関係があることである．先行研究での特徴量抽出においては深度画像の解析に RGB 画像の解析結果を用いるため，認識結果の統合前の段階で依存関係が発生し，RGB 画像での認識誤りが深度画像の解析に影響を与える．また，深度画像の測定可能範囲が比較的短いことが認識の制限となっていた．

本研究の目的は，脳の情報処理に基づいたマルチモーダルな物体認識手法 [5] を拡張し，独立した二つの特徴量抽出ネットワークを採用することで，複数の情報源を活用し確率的な物体認識手法を提案することである (図 1)．本研究では，先行研究におけるモダリティ間の依存関係を解消するために，新たに LiDAR を用いたモダリティを採用し検証する．レーザー光を直接物体に当てて距離を測定する LiDAR センサから得られる 3 次元の点群は，広範囲に高精度な計測が可能である．点群に対して PointNet [8] を用いたセマンティックセグメンテーションを行い，識別結果毎のオブジェクトの中心座標を算出し，認識に用いる．また，本手法を評価するために，LiDAR センサと RGB カメラを用いて 3 次元点群と RGB 画像からなるデータセットを作成した．

本稿の研究の成果は次の通りである．

- 3 次元点群をもとに得た位置情報を活用した物体認識手法を提案した．
- 実環境を想定したデータセットの開発を行った．
- 従来手法で生じていたモダリティ間の依存関係による認識精度の低下を解消した．
- RGB 映像と 3 次元点群を用いたマルチモーダルな物体認識手法がユニモーダルな物体認識手法よりも高い適合率を達成できることを示した．

2. では我々の先行研究および本研究の元となる研究について述べる．提案手法は 3. にて説明し，その有効性を新たに撮影したデータセットを用いて 4. にて検証する．最後に 5. にて本研究の結論をまとめる．

2. 関 連 研 究

ここでは，本研究の基礎となる先行研究について述べる．まず，Bayesian attractor model (BAM) [6] を用いたユニモーダルな物体認識について説明する．続いて，Bayesian Causal Inference [7] を用いた 2 つのモダリティから得られる認識結果を統合する処理を述べ，最後に BAM と BCI を組み合わせたマルチモーダルな物体認識手法 [5] を説明する．

2.1 Bayesian attractor model

Bayesian attractor model [6] は，脳が意思決定を行う様子を，アトラクターモデルとベイズ推定を組み合わせることでモデル化したものである．このモデルでは，時刻 t における意思決定状態を表現する変数 $\mathbf{z_t}$ が状態空間上で定義され，人の脳はこの変数に応じて観測値 $\mathbf{x_t}$ を予測するという生成ダイナミクスを前提とする．人が観測を得たときの意思決定変数の変化を，ベイズフィルタによる変数の更新で表現しており，$\mathbf{z_t}$ の事後分布 $P(\mathbf{z_t})$ を用いた意思決定過程が定義される．

内部状態 $\mathbf{z_t}$ のダイナミクスは式 (1) で表される．f はホップフィールドダイナミクスであり，アトラクターと呼ばれる状態空間上の定点を N 個内部に持つ ($\{\phi_1 \dots \phi_N\}$)．

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

ここで Δt は時刻変化を表し，w_t は正規分布 $\mathcal{N}(0, (q^2/\Delta t)\mathbf{I})$ に従うシステム雑音を表す．q はダイナミクスの不確実性を表し，この値が大きいほど内部状態がアトラクター間で変動しや

すくなり意思決定状態の切り替えが起こりやすくなる．

観測方程式は式 (2) で与えられる．

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

ここで $\mathbf{v}_t$ は $\mathcal{N}(0, r^2\mathbf{I})$ に従う時刻 t での観測雑音を表し，r は観測の不確実性を表す．$\mathbf{M} = [\mu_1 \dots \mu_N]$ は特徴量行列と呼ばれる．また，σ はシグモイド関数であり，$\mathbf{z}_t$ の各要素を $[0, 1]$ へ変換する．文献 [6] では $\boldsymbol{\sigma}(\phi_i)$ を i 次元で 1, それ以外の次元で 0 の N 次元ベクトルに近くなるように定義しているため，$\mathbf{z}_t = \phi_i$ の時に $\mathbf{x}_t \simeq \mu_\mathbf{i} + \mathbf{v}_t$ となる．

式 (1) と (2)，ベイズの定理と実測値 $\mathbf{x}_t$ から $P(\mathbf{z}_t|\mathbf{x}_t)$ を推定できる．BAM が μ_i を観測するとき，$\mathbf{z}_t$ は ϕ_i に近づき $P(\mathbf{z}_t = \phi_i|\mathbf{x}_t)$ は大きな値を取る．このとき，i 番目の特徴量に関する証拠が十分に収集されたと判断できるため，この $P(\mathbf{z}_t = \phi_i|\mathbf{x}_t)$ を確信度と呼ぶ．

BAM は観測におけるノイズを考慮し，その内部状態と予測を用いてモデルを更新し意思決定を行う．BAM を活用することで，不確実性を考慮した認識モデルを実現できる．

2.2 Bayesian causal inference

Bayesian causal inference (BCI) は，ベイズ理論を用いた因果関係を推定する統計モデルであり，異なるモダリティからの入力が同じ発生源から得られたものかどうかを推論する脳の情報処理のモデル化に用いられている．

以降ではモデルの説明のために，モダリティ A ($audio$) と V ($visual$)，そして情報源の関係性を示す C を用いる．BCI では各モダリティでの観測値 (x_A, x_V) が同じ情報源から発生した $(C = 1)$ のか異なる情報源から発生した $(C = 2)$ のかを推論する．文献 [7] では，$C = 1$ となる場合の事前確率が p_{common} で定義され，通常 0.5 に設定される．文献 [7] のモデルでは，$C = 1$ の場合，共通の音源の位置に対して各モダリティの観測値分布が，$C = 2$ の場合，2 つの音源の位置に対する各モダリティの観測分布が定義される．

これらを事前分布とし，ベイズ理論を用いてその因果構造，つまり $C = 1$ または $C = 2$ どちらかを表す事後確率を推測することができる（式 (3)）．

$$p(C|x_A, x_V) = \frac{p(x_A, x_V|C)p_{common}}{p(x_A, x_V)} \tag{3}$$

二つのモダリティからの信号を観測したときに，BCI では式 (4) に基づき観測値の統合を行う．

$$\hat{x}_A = P(C = 1)\hat{x}_{A,C=1} + P(C = 2)\hat{x}_{A,C=2} \tag{4}$$

ここで，$C = 1$ のときのモダリティ A の推定値を $\hat{x}_{A,C=1}$，$C = 2$ のときのモダリティ A の推定値を $\hat{x}_{A,C=2}$ と表している．これらは前述の観測値分布より求められる．$\hat{x}_V$ も同様に推論することができる．

異なる因果構造を持つ 2 つのモデルの予測を平均化することで，より誤差の小さな意思決定を行うことができる．BCI ベースのモデル統合を行うことで，上述のように $\hat{x}_A$ および $\hat{x}_V$ の値が得られ，例えば $\hat{x}_A$ を音声認識に，$\hat{x}_V$ を映像認識に活用する，といった応用が考えられる．

2.3 マルチモーダルな物体認識モデル

BAM と BCI を用いた物体認識手法において，各モダリティにおける物体認識は BAM により行われる．BAM は観測対象の特徴量をアトラクターとして記憶し，観測値と特徴量の類似度合いを確信度として出力する．各モダリティでの認識結果は BCI を用いて統合され，最終的な認識結果を得る．文献 [5] では，認識したい物体の映った RGB 画像および深度画像から特徴量を抽出し，アトラクターに記憶させることで，物体認識を実現している．

本稿では，LiDAR から得られる 3 次元点群データを用いて物体位置を測定するモデルを開発し先行研究のモデルに組み込むことで，前述した，認識結果の統合前に依存関係が発生する課題を解決した．手法の説明は次節で行う．

3. 提 案 手 法

本稿では，これまでの手法で採用していたステレオカメラから得た深度情報ではなく，LiDAR センサから得られる位置情報を用いた物体認識手法を採用し，先行研究のモデルに取り入れた (図 2)．我々の提案手法では，物体認識のために映像モダリティと位置モダリティを利用する．映像モダリティでは RGB カメラをセンサとして使用し，位置モダリティでは LiDAR をセンサとして利用する．LiDAR は高精度の 3D 点群情報を取得することができる機器であり，近年物体認識の分野で広く使用されている [9]．各機器から 2 次元 RGB 画像と 3 次元点群データを収集し，Siamese RPN [10] と PointNet [8] を通してモダリティ固有の特徴量を抽出する．RGB 画像からの特徴量抽出については，文献 [5] と同様の取得手順とした．すなわち，Siamese RPN では物体の参照画像を与えることで，RGB 画像中から物体があると推定される領域を囲むバウンディングボックスを算出しており，この探索に用いるベクトルを映像特徴量として用いた．

3.1 位置モダリティを用いた物体認識

位置モダリティを用いた物体認識では，BAM の特徴量行列に，認識対象の物体の座標を記憶させることで事前学習が行われる．その後，観測した物体の座標値に対して，BAM はどの物体を観測しているのかを確信度とともに出力する．ここでは特徴量行列の更新を伴わないことから，物体が移動するにつれ認識精度が低下する可能性がある．特徴量行列の更新については今後の課題であり，本稿では扱っていない．

位置モダリティの特徴量を 3 次元点群から得るために，PointNet [8] を採用した．PointNet は，点群を入力として，点同士の位置関係を元に一つ一つの点にラベル付けを行うセマンティックセグメンテーションと呼ばれる手法の一つである．本稿では github に公開されている実装を用いた [11]．

まず，1 フレームで計測された 3 次元点群データ（$N \times 3$, N は取得した点の数）を PointNet に与え，セマンティックセグメンテーションを行い，ラベル付き点群（$N \times 4$）を得る．その後検出された各ラベルについて座標の平均をとることで物体の中心座標を求める（$M \times 4$, M は対象とするオブジェクトのラベル数）．BAM は高次元の入力に対して不安定な挙動を示す

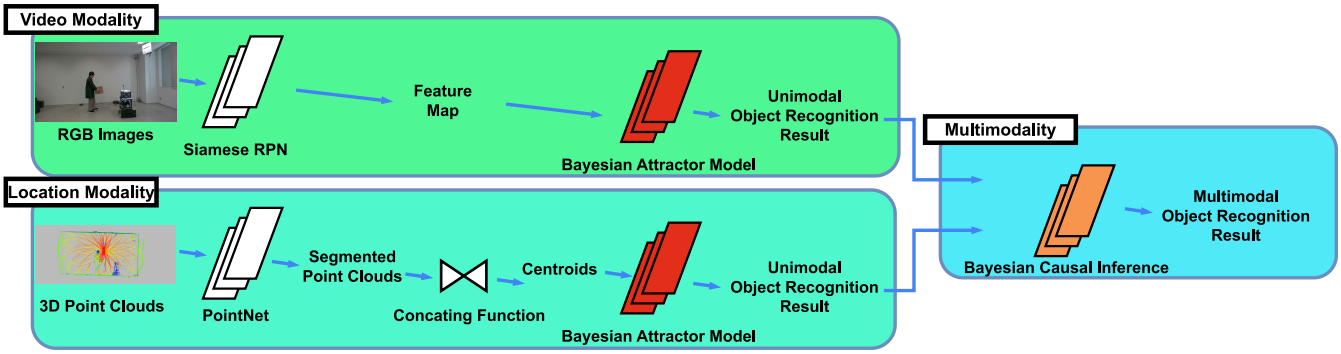


図 2　脳の仕組みに基づいた 2 次元映像データと 3 次元点群データを用いた物体認識手法

ことがわかっているため，BAM に位置モダリティを入力として与える際にはこの集約処理を行っている．

3.2　BCI を用いた統合処理

各モダリティの BAM における物体認識結果（確信度）を観測値として BCI での因果推論を行う．まず，映像モダリティと位置モダリティから得られる観測が同じ物体から得られる観測かどうかを評価し，2.2 にて述べた，異なるモダリティでの観測値が共通の発生源から得られたときの確信度の推定値と，異なる独立した発生源から得られたときの推定値を，式 (4) で表されるモデル平均化により統合し，推論結果を得る．

BCI を実行するためには，確信度の生成モデルが必要である．c_V と c_L を，それぞれ映像モダリティと位置モダリティから得られる確信度とするとき，$P(c_V, c_L|C=1)$ と $P(c_V, c_L|C=2)$ を以下のように定義する．

$$P(c_V, c_L|C=1) = \begin{cases} 0.5 + \sigma_{c=1} & \text{if } L_V = L_L, \\ \sigma_{c=1} & \text{otherwise,} \end{cases} \tag{5}$$

$$P(c_V, c_L|C=2) = \begin{cases} 1 & \text{if } L_V \neq L_L \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

ここで，$\sigma_{c=1} = 0.5\sigma(\max(c_V)) + 0.5\sigma(\max(c_L))$, $\sigma = 1/(1+\exp(-x-\lambda_{bci}))$, $L_V = \text{argmax}(c_V)$, $L_L = \text{argmax}(c_L)$ である．

この生成モデルは，BAM から得られる確信度が同じ物体を観察したときに，少なくとも 1 つのモダリティで増加すること，異なる物体を観測する場合には，2 つのモダリティにおける認識結果が異なることを仮定している．シグモイド関数は、得られた信頼度が閾値 (λ_{bci}) を超えるかどうかを示すために用いられ $\lambda_{bci} = 10^{-3}$ を採用する．

4.　評 価 結 果

4.1　評 価 指 標

はじめに，確信度の時系列変化を示し，入力の変化に対応する認識の変化について，提案手法の定性的な考察を行う．提案手法の定量的な評価には，適合率（Precision）を用いる．誤検出を最小化することを目的とするアプリケーションにおいては，適合率は重要な指標となる．

適合率は式 (7) に基づき算出される．TP は正解オブジェクトの確信度のほうが高いフレームの総数を表し,FP は誤ったオブジェクトの確信度のほうが高いフレームの総数を表す．

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

表 1　設定したパラメータ

Name	Value
Dynamics uncertainty (q)	1.0×10^{0}
Video-modal sensory uncertainty (r_v)	1.0×10^{1}
Depth-modal sensory uncertainty (r_d)	1.0×10^{0}
Location-modal sensory uncertainty (r_l)	3.0×10^{-1}

4.2　データセット

評価を行うにあたり，次の要件を満たす画像と点群からなる時系列データが必要である．

- 倉庫内での作業を想定したシナリオ
- 実機を用いて撮影されたデータ
- 作業員とロボットがオブジェクトとして内在

そこで，Intel RealSense D455 と Livox Avia を用いて画像と点群からなるデータセットを作成した．撮影したデータのうち，オブジェクト同士が近づく部分に焦点をあて，トレーニング用に 60 フレームを，テスト用に 300 フレームを用意した．3 次元点群には “worker,” “robot,” “ceiling,” “floor,” “wall,” “container,” そして “clutter” の 7 種類のオブジェクトについてラベリングを行いデータセットを作成した（図 1）．それぞれ，人，ロボット，天井，床，側壁，ダンボール箱，その他を表す．

4.3　パラメータ設定

モデル中の主要なパラメータを表 1 に示す．ダイナミクスの不確実性 (q) と観測の不確実性 (r_v,r_d,r_l) は BAM のパラメータである．これらの値は入力されるデータの特性により決定する必要があり，事前実験を行い決定した．また，BAM の特徴量行列には，はじめの 5 フレームの，“worker” と “robot” の特徴量の平均値を各モダリティにおいて設定した．

4.4　実 験 結 果

はじめに，各モダリティにおけるユニモーダル物体認識手法の比較を行った (図 3(a)–3(f))．以降，先行研究の手法での結

表 2　適合率 (Precision)

Modality	Worker	Robot
Single modality (video)	**0.875**	**0.979**
Single modality (depth)	0.804	1.000
Single modality (location)	1.000	1.000
Video-based multi-modalities (depth)	0.814	0.987
Video-based multi-modalities (Location)	0.838	0.996
Depth-baed multi-modalities (video)	0.838	1.000
Location-based Multi-modalities (video)	0.958	1.000

果を深度モダリティとして記述する．グラフの縦軸は対数表記された確信度を，横軸はフレーム番号を表す．また，グラフの青色の線は “worker” に対する確信度を，オレンジ色の線は “robot” に対する確信度の時系列変化を表している．

映像モダリティでの “worker” の特徴量を入力する際は，ほぼ全てのフレームで “worker” の確信度が “robot” の確信度よりも高くなっている．深度モダリティ，位置モダリティでの認識結果でも，映像モダリティでの認識結果と同様の結果が得られている（図 3(a)–3(c)）．

映像モダリティでの “robot” の特徴量を入力する際の認識結果では，後半のフレームにおいて “robot” を認識する確信度の低下がみられた．これは，“worker” と “robot” の Bounding Box が重なることで生じていた．深度モダリティの認識結果でも，映像モダリティでの認識結果と同様に低下している（図 3(d)–3(e)）．位置モダリティでは，ほぼ全てのフレームで “robot” である確信値が高くなっている（図 3(f)）．

次に，先行研究で用いていた，映像モダリティと深度モダリティのマルチモーダル統合による物体認識を行った (図 4)．BCI では，マルチモーダル（映像），マルチモーダル（深度）のように，統合後に 2 つの結果が得られる．前者は映像による認識を深度で補正したもの，後者は深度による認識を映像で補正したものと捉えることができる．先行研究の手法では，マルチモーダル（映像），マルチモーダル（深度）のいずれでも，“robot” を認識するタスクにおいて，両モダリティで同時期に確信度が低下するため，マルチモーダル統合を行った場合でも正しいオブジェクトに対する確信度が向上することはなかった (図 4(b),4(d))．

次に，我々の提案手法に基づき，映像モダリティと位置モダリティで認識を用いた物体認識を行った (図 5)．我々の提案手法では，“robot” に対する物体認識を行った場合における正しいオブジェクトに関する認識精度の低下を補正できることを確認した (図 5(b), 5(d))．また，一方のモダリティで誤ったオブジェクトへの確信度が高く観測されている場合でも，他方のモダリティで正しいオブジェクトへの確信度が高ければ，マルチモーダル統合後に，正解オブジェクトに関する確信度を高く保持することができており，後段のシステムにどちらの確信度も高い，注意する必要がある，という情報を伝えることができる．

最後に，定量的な評価のため，適合率での評価を行った (表 2)．確信度のグラフで示したように，映像モダリティ単体の精度に対して位置モダリティで補正した場合の適合率が改善されていることが確認できた．

4.5　手法の制限

ここでは，我々の手法の制限について議論する．

一点目として，今回我々の提案した PointNet を採用した手法では，クラスごとのセマンティックセグメンテーションを行い，その後インスタンス認識するというモデルであるため，1 クラスについて 1 オブジェクトしか認識することができない．そのため，評価テストでは，各クラスに属するオブジェクトは一つという制限があった．実応用を想定すると，多クラス多オブジェクトのデータセットへの対応を考慮する必要があり，この点に関しては，既存のクラスタリング手法などによって，同一クラスの点群をさらに分類する方法で解決できると考えている．

次に，独立した認識結果の組み合わせに用いた前提条件について述べる．マルチモーダル統合処理では，映像モダリティと位置モダリティの二つから，それぞれ独立して人とロボットを表す認識結果，すなわち確信度が得られる．それぞれのモダリティから得られた確信度を統合する際には，適切な認識結果の組み合わせを決定し，その後統合が行われる必要がある．今回の実験では，観測対象が人かロボットの識別を行う二値分類タスクとなっており，それぞれのモダリティにおいて特徴量抽出を行う段階で，人やロボットに関する参照情報を使用している．そのため，認識結果の組み合わせ方が一意に定まっていた．この点は特徴量抽出の手法に依存しないように対応する必要がある．

5. おわりに

脳の情報処理システムに倣った統合方法は，異なるセンサから得られるデータを活用した物体認識に応用できる．我々は異なるセンサから得られるマルチモーダルデータへの先行手法の適用可能性に着目した．本稿では先行研究を拡張し，3 次元点群を計測する LiDAR を活用する物体認識手法を提案した．提案手法では PointNet を用いて点群から物体の位置座標を算出し，位置情報を元に認識を行う．実験より，映像モダリティでの解析結果に位置モダリティでの解析結果を統合することで，適合率が向上することを確認した．

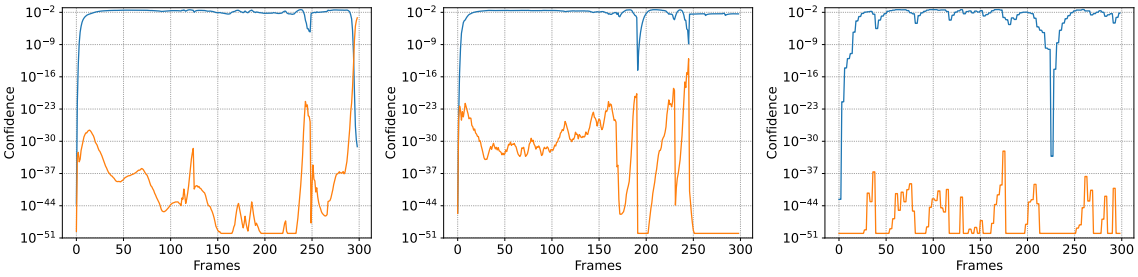
4.5 で述べたように，多クラス多オブジェクトを対象とした認識タスクへの対応，異なるモダリティにおける認識結果の組み合わせ方法の検討が今後の課題である．

謝　辞

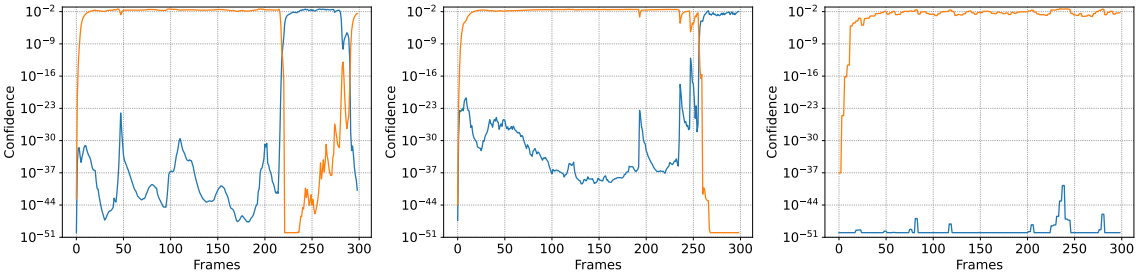
本研究は独立行政法人情報通信研究機構（NICT）の助成金 JPJ012368C00701 の助成を受けて行った．

文　献

[1] H. Viswanathan and P.E. Mogensen, “Communications in the 6g era,” IEEE Access, vol.8, pp.57063–57074, 2020.
[2] X. Ding, J. Guo, Z. Ren, and P. Deng, “State-of-the-art in perception technologies for collaborative robots,” IEEE Sensors Journal, vol.22, no.18, pp.17635–17645, 2022.
[3] Y. Wang, Q. Mao, H. Zhu, J. Deng, Y. Zhang, J. Ji, H. Li, and Y. Zhang, “Multi-modal 3d object detection in autonomous driving a survey,” International Journal of Computer Vision, pp.1–31, 2023.
[4] S. Yasuda, T. Kumagai, and H. Yoshida, “Cooperative transportation robot system using risk-sensitive stochastic

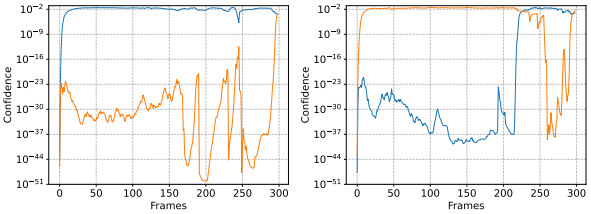


(a) 映像モダリティ：Worker　(b) 深度モダリティ：Worker　(c) 位置モダリティ：Worker

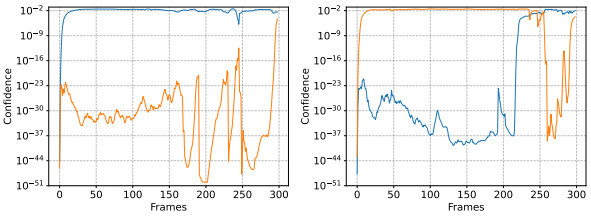


(d) 映像モダリティ：Robot　(e) 深度モダリティ：Robot　(f) 位置モダリティ：Robot

図 3　各モダリティにおけるユニモーダルな物体認識結果

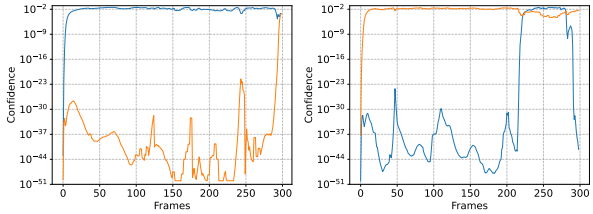


(a) マルチモーダル（映像）：Worker　(b) マルチモーダル（映像）：Robot

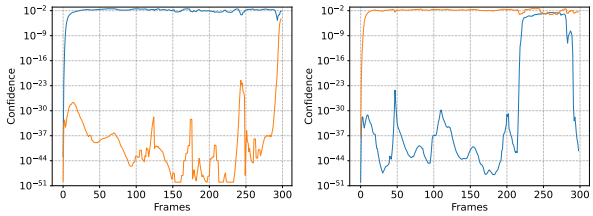


(c) マルチモーダル（深度）：Worker　(d) マルチモーダル（深度）：Robot

図 4　マルチモーダルな物体認識結果（映像と深度モダリティ）



(a) マルチモーダル（映像）：Worker　(b) マルチモーダル（映像）：Robot



(c) マルチモーダル（位置）：Worker　(d) マルチモーダル（位置）：Robot

図 5　マルチモーダルな物体認識結果（映像と位置モダリティ）

control," 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp.5981–5988, 2021.

[5] R. Seki, D. Kominami, H. Shimonishi, M. Murata, and M. Fujiwaka, "Multi-object recognition method inspired by multimodal information processing in the human brain," 2022 IEEE Globecom Workshops (GC Wkshps)IEEE, pp.569–574 2022.

[6] S. Bitzer, J. Bruineberg, and S.J. Kiebel, "A Bayesian Attractor Model for Perceptual Decision Making," PLOS Computational Biology, vol.11, no.8, pp.1–35, august 2015. https://doi.org/10.1371/journal.pcbi.1004442

[7] K.P. Körding, U. Beierholm, W.J. Ma, S. Quartz, J.B. Tenenbaum, and L. Shams, "Causal Inference in Multisensory Perception," PLOS ONE, vol.2, no.9, pp.1–10, 09 2007. https://doi.org/10.1371/journal.pone.0000943

[8] 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, pp.652–660, 2017.

[9] J. Janai, F. Güney, A. Behl, and A. Geiger, "Computer vision for autonomous vehicles: Problems, datasets and state of the art," Foundations and Trends® in Computer Graphics and Vision, vol.12, no.1–3, pp.1–308, 2020. http://dx.doi.org/10.1561/0600000079

[10] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High Performance Visual Tracking With Siamese Region Proposal Network," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.8971–8980, June 2018.

[11] C.R. Qi, H. Su, K. Mo, and L.J. Guibas, "pointnet," `https://github.com/charlesq34/pointnet`, 2016.