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

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　〒565–0879 大阪府吹田市山田丘 1–5
†† NEC システムプラットフォーム研究所　〒211-8666 神奈川県川崎市中原区下沼部 1753
E-mail: †{r-seki,d-kominami,h-shimonishi,murata}@ist.osaka-u.ac.jp, ††{fujiwaka,nogami}@nec.com

あらまし　デジタルツインの実現には、実世界の様々なオブジェクトをカメラなどのセンサーを通して瞬時に特定し、その位置を把握し、コンピュータ上に表現することが望まれている。しかし、得られる情報は少なからずセンサー機器のノイズや精度の影響を受けるため、従来検討されてきた正確な情報を前提としたオブジェクト推定手法には限界がある。従って、解像度の低い映像情報などの不確実な観測情報を基に、そのオブジェクトが何であるのかという意思決定を高速かつ高精度に行うことが求められている。本稿では、複数種類の不確実な観測情報を元に意思決定を行っている脳のマルチモーダル情報処理機構に注目し、そのメカニズムを取り入れることで、ノイズを含んだ観測情報からオブジェクト推定を行う手法を提案する。計算機シミュレーションにより、提案手法が不確定な観測情報からも高精度かつ高速にオブジェクト推定を行えることを示した。
キーワード　デジタルツイン, ベイジアンアトラクターモデル, ベイズ因果推論

Proposal and Evaluation of an Object Estimation Method Inspired by Multimodal Information Processing in the Brain

Ryoga SEKI†, Daichi KOMINAMI†, Hideyuki SHIMONISHI†,††, Masayuki MURATA†, Masaya FUJIWAKA††, and Kosuke NOGAMI††

† Graduate School of Information Science and Technology, Osaka University
†† System Platform Research Labs, NEC Corporation
E-mail: †{r-seki,d-kominami,h-shimonishi,murata}@ist.osaka-u.ac.jp, ††{fujiwaka,nogami}@nec.com

Abstract　In order to realize the digital twin, it is desired to instantly identify various objects in the real world through sensor devices, determine their locations, and represent them on the computer. However, the obtained information is affected by noise and traffic of sensor devices to some extent, and object estimation methods based on accurate information that have been considered in the past have their limitations. Therefore, there is a need to make a fast and accurate decision on what the object is based on uncertain observation information such as low-resolution video information. In this paper, we focus on the information processing mechanism of the brain, which makes decisions based on multiple types of uncertain observed information, and propose a method for estimating objects from noisy observed information by incorporating this mechanism. Through a computer simulation, we show that our proposal identifies an object accurately and quickly from uncertain observed information.
Key words　Digital twin, Bayesian Attractor Model, Bayesian Causal Inference

1. はじめに

デジタルツインによる実世界の把握と制御のためには、実世界に存在する様々なオブジェクトをセンサー機器を通して瞬時に理解する必要がある。すなわち、目の前にどのようなオブジェクトが存在するのかを一意に同定し、その位置を把握し、そしてデジタルツイン上に表現することである。近年では CNN などの技術の発達が著しいが、センサ機器のノイズや精度によって認識率には限界があり、また、エッジコンピューティングや端末といった計算リソースの乏しい環境においてリアルタ

イムに画像識別を行うことは依然として困難である。そこで、非常に軽量であるが視覚性の低い画像分析情報、不確実な観測情報を基に、そのオブジェクトが何であるのかという意思決定を高速かつ高精度に行うことが求められている。

このような不確実な観測情報から意思決定を行うシステムの身近な例として、脳の情報処理機構が挙げられる。脳では、目や耳、肌や三半規管などから得られた不確実な観測情報を用いて周囲の状態を推測、最終的な意思決定までをすべて行っている。近年、脳の情報処理機構の数理的なモデル化が進められており、そういったモデルとして Bayesian Attractor Model (BAM) [1] や Bayesian Causal Inference (BCI) [2] がある。

BAM は、人が観測した情報を基に意思決定を行うまでの振る舞いを、ベイズ推定を用いてモデル化したものである。BAM は確率的な意思決定の状態を表す内部変数（状態変数）を持ち、外部からの観測情報を用いて状態変数の推定を繰り返し行う。特定の選択について閾値を超える確率密度となった時点で、その選択を取るという意思決定を行う。このような BAM を映像分析に活用することで、時系列的に揺らぎが大きい不確実な観測情報から、オブジェクトの識別を高精度に行うことが期待される。

BCI は、人間が複数のモダリティを用いて（例えば視覚と聴覚）知覚対象を認知する過程を数理モデルとして表したものである。この認知モデルでは、2 つの入力刺激が同じ刺激源から出たものであるかどうかを確率的に推論し、その確率をもとにそれぞれの入力刺激を統合して最終的な認知判断を行う。

本稿では、映像解析のオブジェクト推定システムに BAM の意思決定モデルを用いて識別を行う。複数のモダリティごとに、BAM による識別を行い、それぞれから得られた出力を BCI で統合することでオブジェクト認識に脳の情報処理機構を利用する方式を提案する。提案方式では、1 枚の画像を 128 次元に圧縮した情報を画像特徴量として入力データとする。ここでは、あらかじめ用意しておいた参照用画像データの画像特徴量をアトラクターとし、フレームごとに入力される入力データの系列を BAM に入力して、脳の内部状態がどのアトラクターに落ちるかということで、入力データがどの参照データに相当するのかを判定する。一般的な CNN は、識別精度が高い代わりに、数百枚以上の参照用データを用いて事前学習させる必要がある。そのため事前学習のコストが非常に高く、今回ターゲットしているデジタルツインのように順次新しいオブジェクトを識別するためには使えない。提案方式では、あるオブジェクトを最初に見たその 1 フレーム目のみを参照データとして用いてアトラクターを生成するため、事前学習のコストが不要になる。一方、当然ながら識別精度が低下することが予想される。そのため、映像モダリティだけでなく、カメラの位置から推測されるオブジェクトの位置モダリティからもオブジェクト推定を行い、各モダリティでの認知の結果を BCI で統合して最終的な意思決定をすることで、精度の向上を図る。

本稿の以降の構成は次の通りである。2 章では、BAM と BCI の数理モデルについて述べ、3 章では、映像から各モダリティの特徴量を抽出して BAM に入力した結果を BCI に応用する手法について述べる。4 章では、提案手法の評価を行い、5 章では、本稿のまとめと今後の課題について述べる。

2. 関 連 研 究

2.1 Bayesian Attractor Model (BAM)

BAM では、観測した情報を用いたベイズ推定により、観測対象が事前に記憶した選択肢のどれに該当するのかを推定する。BAM は意思決定の状態 $\mathbf{z}_t$ を内部状態として持っており、外部からの観測値 $\mathbf{x}_t$ を受けることで状態 $\mathbf{z}_t$ を更新する。ベイズ推定に基づく状態更新により、$\mathbf{z}_t$ は一点として扱われるのではなく、観測や脳の状態の不確実さを反映した確率分布 $P(\mathbf{z}_t)$ として表現される。

選択肢の個数（n）に応じた数だけ、$\mathbf{z}$ の存在する状態空間の中に安定点（アトラクター）$\phi_1, \ldots, \phi_n$ を準備しておき、$\mathbf{z}_t$ が ϕ_i に十分近づいたとき、i 番目の選択肢を取るという意思決定を行う。$\mathbf{z}_t$ は確率的に表現されているため、$\mathbf{z}_t = \phi_i$ である確率密度（確信度）を導出し、確信度を用いた意思決定が用いられる。以降では、状態の更新と意思決定の詳細について述べる。

2.1.1 状 態 更 新

状態の更新は、観測値 $\mathbf{x}_t$ を得たときに、意思決定状態 $\mathbf{z}_t$ の事後分布 $P(\mathbf{z}_t|\mathbf{x}_t)$ をベイズ推定により求めることで行われる。$\mathbf{x}_t$ と $\mathbf{z}_t$ には、次の生成モデルが仮定されている。

$$\mathbf{z}_t - \mathbf{z}_{t-\Delta t} = \Delta t f(\mathbf{z}_{t-\Delta t}) + \sqrt{\Delta t}\mathbf{w}_t \quad (1)$$

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

ここで、$f(\mathbf{z})$ はアトラクターモデルの一つである Hopfield Network のダイナミクスを表し、このダイナミクスは複数のアトラクターを持つ。n をベイジアンアトラクターモデルに記憶する選択肢の数とするとき、f が n 個のアトラクター $\phi_1, \ldots, \phi_n$ を持つように設計する。M は各選択肢に対応した観測値を並べた行列であり、$M = [\mu_1, \ldots, \mu_n]$ である。σ は値域が 0 から 1 である多次元シグモイド関数である。$\mathbf{w}_t, \mathbf{v}_t$ はノイズ項であり、それぞれ $\mathbf{w}_t \sim \mathcal{N}(0, \frac{q^2}{\Delta t}I), \mathbf{v}_t \sim \mathcal{N}(0, r^2 I)$ である。$\mathcal{N}$ は正規分布であり、Δt を 1 とするとき、それぞれの標準偏差は q, r である。これらによって生成モデルにおけるダイナミクスと観測のノイズが定まるので、q はダイナミクスの不確かさ、r は観測の不確かさと呼ばれる。また、I は単位行列である。

2.1.2 意 思 決 定

前述の生成モデルをベイズの定理により逆方向に推定することで、意思決定のモデルが得られる。文献 [1] では、生成モデルの非線形性を考慮して Unscented Kalman Filter (UKF) を用いて近似計算を行う。この状態推定によって得られるのは $\mathbf{z}_t$ の事後確率分布 $P(\mathbf{z}_t|\mathbf{x}_t)$ である。そのため脳の内部状態がどのアトラクターに近いのかは確率密度の大きさに基づき判定される。

2.2 Bayesian Causal Inference (BCI)

BCI は、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として表現したものである。例えば、視界の左の方に何かが見え、同じ方向から何かの音が聞こえたとき、これらが同じ刺激源から発せられたものであると判断して視覚と聴覚

を統合して到来方向を判断したり、あるいはそれらが別々の方向から来たものとして到来方向を別々に判断したりすることを表現することができる。文献 [2] の BCI では、オブジェクトが提示されると、両モダリティ（視覚と聴覚）のそれぞれでオブジェクトの位置を知覚し、2 つのモダリティで同じオブジェクトを観測しているか否かを確率的に推論し（Causal Inference）、この推論結果をもとにそれぞれのモダリティの観測値を統合してオブジェクトの位置を認知する（Model Average）。このとき、両モダリティで同じオブジェクトを観測しているとした場合の統合した認識結果と、それぞれ別のオブジェクトを観測しているとした場合の個別の認識結果を、Causal Inference の結果をもとにした重みづけ和が結果として出力される。

3. BAM の BCI 拡張によるオブジェクト推定手法

センサー機器から得られた映像モダリティと位置モダリティそれぞれの特徴量から個別に BAM でオブジェクト推定を行い、それらを BCI で統合して意思決定することで、不確定な観測情報を元にロバストなオブジェクト認識を実現する。

以降では、BAM の意思決定モデルをオブジェクト推定に応用する手法、BAM の確信度を BCI によって統合する手法について述べる。

3.1 BAM のオブジェクト推定への応用

BAM をオブジェクト推定へ応用するにあたり、(1) 各モダリティで BAM が観測する特徴量を決定する、(2) アトラクターに記憶する参照データを決定する、(3) 不確かさのパラメータ q、r の値を決定する必要がある。(3) については、特徴量の分散を一定の値にするように特徴量を変換することで、観測にかかるノイズを決定することができる。以下では、特徴量、変換、アトラクターのそれぞれについて述べる。

a) 特徴量

特徴量は Siamese-RPN [3] を利用して抽出する。Siamese-RPN はテンプレート画像と検出画像の 2 つの画像を入力とし、テンプレート画像と類似する箇所を検出画像から検出し、その位置をバウンディングボックスとして出力する。Siamese-RPN は、SiameseNetwork と Region Proposal Network から構成され、前者は画像の特徴量を抽出し、後者は抽出された特徴量を用いてテンプレート画像と類似する箇所を検出画像から算出する。文献 [3] では Siamese Network として AlexNet [4] などの既存の CNN を利用しているが、提案方式では軽量化を目的として 4 層からなるシンプルな CNN を利用した。BAM の入力となる映像モダリティの特徴量は、Region Proposal Network が出力するバウンディングボックスに対応する SiameseNetwork の出力（画像特徴量）のうち 128 次元のデータとした。位置モダリティの特徴量は、カメラの方向ベクトルと複数フレームを統合した深度情報から算出した 3 次元ワールド座標系データとした。

b) 特徴量の変換

センサー機器から得られた観測値は値域が定まっておらず、ノイズの大きさも未知である。さらに多次元変数である特徴量の各次元ごとに数値のスケールが異なる場合を考慮すると、特徴量の次元ごとに r を調整しなくてはならない。そこで、特徴量のデータをそのまま使用するのではなく、BAM で扱いやすいように変換処理を行う。時刻 t に観測する特徴量ベクトル $\mathbf{x}(t) = (x_1(t), x_2(t), \cdots, x_n(t))$ の各要素を変換する関数 S を以下のように定義する。

$$S(x_i(t)) = \frac{x_i(t) - \mu_i(\mathbf{X})}{s_i(\mathbf{X})} \tag{3}$$

ここで、$\mathbf{X}$ は $\mathbf{x}(t)$ と同サイズのベクトルの集合であり、$\mu_i(\mathbf{X}), s_i(\mathbf{X})$ は、$\mathbf{X}$ に含まれる各ベクトルの i 番目の要素の平均および分散を返す。$\mathbf{X}$ は各識別対象について事前に取得した特徴量ベクトルの集合となる。$\mathbf{X}$ をうまく選ぶことで、式 (3) を計算すると $S(x_i(t))$ の平均は 0、分散は 1 となる。平均を 0 にすることで、値域は正負の両方の符号を含む。このような処理を行う理由は 2 点ある。一つは、観測値の値域に正負の両方の値を含む方が、BAM の推定精度が高まる点である。このことを、符号が正のみである場合に 1 次元の観測値を識別する例を用いて説明する。意思決定状態があるアトラクター付近にあるときに観測値が増加すると、BAM では式 (2) を逆計算して $\mathbf{z}_t$ を求めるが、$\mathbf{x}_t$ が増加しているため $\sigma(\mathbf{z}_t)$ も増加するものだと UKF により推定される。このとき、シグモイド関数 σ が用いられているために、$\mathbf{z}_t$ の各要素の推定結果はシグモイド関数の変曲点から離れる。これはすべてのアトラクターから遠ざかることになるため、確信度が下がるだけで、$\mathbf{z}_t$ は別のアトラクターに近づくことはない。逆に観測値が減少するときは別のアトラクターに近づく。以上より、観測値の符号が正（あるいは負）のみである場合は、$\mathbf{z}_t$ があるアトラクターの付近にあるとき、観測値の増加（あるいは減少）を $\mathbf{z}$ の空間上での別のアトラクターへの移動として捉えられないため、アトラクターに記憶した 2 値の識別を正確に行うことができないケースが生じる。観測値の値域に正負の両方の値を含む場合、アトラクターに記憶した 2 値がいずれも同符号であれば同じことが起きるが、アトラクターに記憶した値が別符号であれば、上記の問題が生じない。二つ目は、適切な r の値を設定する目安が得られる点である。BAM が観測する可能性のある対象オブジェクト全ての情報を $\mathbf{X}$ が含んでいれば、観測値の分散は 1 に近い値となる。そのため、ある特定のオブジェクトを続けて観測している際の分散は 1 よりも小さい値となり、$0 < r < 1$ の範囲で r を選ぶこととなる。

c) アトラクター

アトラクターには、推定の対象となるオブジェクトの特徴量を記憶させる。これは、特徴量行列 M の要素である各ベクトル $\mu_1, \ldots, \mu_n$ に、それぞれのオブジェクトの特徴量を代入することに相当する。本稿では、事前学習のコストを削減するために、特徴量の計算には各オブジェクトを最初に見た 1 フレーム目の映像のみを用いることとする。まず、式 (3) の $\mathbf{X}$ として、n 個のオブジェクトについて、それぞれを最初に見た 1 フレーム目の映像から計算した特徴量である $\mu_1^*, \ldots, \mu_n^*$ を与える。また、M の要素である特徴量ベクトルには、$\mu_1^*, \ldots, \mu_n^*$ を関数 S に

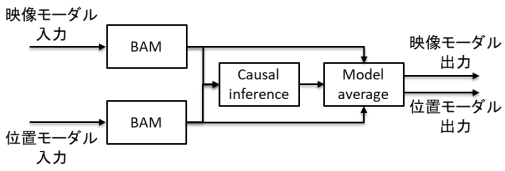


図 1　BCI による BAM の拡張

よって変換した値を与える。

3.2　BCI による BAM の拡張

BAM にて各モダリティそれぞれでオブジェクト推定を行った後、図 1 の通り、BCI による因果推論を行う。ここでは、BAM の確信度を観測値として BCI に入力し、Causal Inference を行って映像モダリティと位置モダリティで同じオブジェクトを観測しているかどうかを推論し、その結果をもって Model Average でマルチモーダル統合して最終的なオブジェクト推定結果を出力する。ただし、従来の BCI は連続値（位置）を対象としているのに対し、ここでは離散値（オブジェクトの識別）を対象としているため、若干の拡張が必要である。以下にその概要を説明する。

3.2.1　Causal Inference

文献 [2] における BCI ではベイズの定理に従って以下の式のように Causal Inference を行う。

$$p(C|X,Y) = \frac{p(X,Y|C)p(C)}{p(X,Y)} = \frac{p(X,Y|C)p(C)}{p(X,Y|C=1)p(C=1)+p(X,Y|C=0)p(C=0)} \quad (4)$$

ここで $p(C)$ は両モダリティで同じオブジェクトを観測しているか否かの確率であり、C は 0（別々のオブジェクトを観測している）と 1（同じオブジェクトを観測している）の 2 値をとる。X と Y はそれぞれ各モダリティの観測値、すなわちここではそれぞれの BAM の確信度の値である。確信度は非常に小さい値を取る場合があるため、閾値以下の値は全て閾値と同値として Causal Inference への入力としている（今回の閾値は 10^{-50}）。また、文献 [2] より $p(X,Y|C)$ は以下の通りである。

$$p(X,Y|C=1) = \int p(X,Y|s)p(s)ds = \sum_{k=1}^{K} p(X,Y|O_k)p(O_k) \quad (5)$$

$$p(X,Y|C=0) = \int p(X|s)p(s)ds \int p(Y|s)p(s)ds = \sum_{k=1}^{K} p(X|O_k)p(O_k) \sum_{k=1}^{K} p(Y|O_k)p(O_k) \quad (6)$$

ここで s はオブジェクトの位置であり $p(s)$ はその分布を表す。$p(X|s)$ はオブジェクトが s の位置にある時に、あるモダリティでは位置が X として観測される確率である。これをオブジェクト推定に用いるため、上式ではオブジェクト O_k が観測される確率を $p(O_k)$ とし、オブジェクト O_k を観測しているときに BAM の確信度の値が X となる確率を $p(X|O_k)$ とする。

3.2.2　Model Average

次に、Causal Inference の結果に基づいて、マルチモーダル統合して最終的なオブジェクト推定結果を出力する。ここでは、以下の式のように、因果推論の結果で重み付けしたコスト関数を計算し、これを最小化するオブジェクト O'_X と O'_Y をそれぞれのモダリティの最終的なオブジェクト推定結果とする。ここで $C=1$ であれば O'_X と O'_Y は同じオブジェクトを出力し、$C=0$ であればそれぞれのモダリティの推定結果がそのまま出力される。

$$Cost_X(O_X) = p(C=1)\sum_{k=1}^{K}|O_X - O_k|^2 p(O_k|X,Y) + p(C=0)\sum_{k=1}^{K}|O_X - O_k|^2 p(O_k|X) \quad (7)$$

$$Cost_Y(O_Y) = p(C=1)\sum_{k=1}^{K}|O_Y - O_k|^2 p(O_k|X,Y) + p(C=0)\sum_{k=1}^{K}|O_Y - O_k|^2 p(O_k|Y) \quad (8)$$

ここで、$|O_X - O_Y|$ は、$O_X = O_Y$ であれば 0、さもなければ 1 とする。文献 [2] では、ここでオブジェクトの推定位置の距離誤差を用いているが、オブジェクト推定を行うにあたっては距離は定義できないので一致不一致のみで計算するものとする。

4.　シミュレーション評価

上述の BAM と BCI を応用したオブジェクト推定手法の有効性を確認するために、シミュレーションによる評価を行った。

4.1　シミュレーション環境

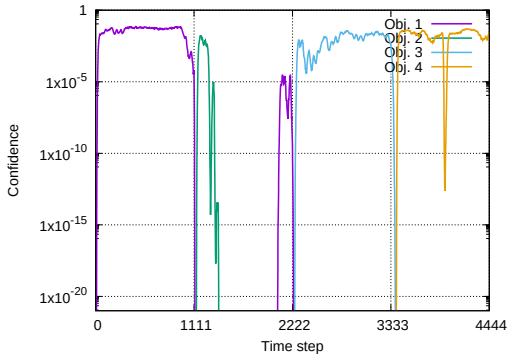
観測機器から得られた映像特徴量と三次元座標である位置特徴量を BAM の観測に与える。BAM は映像と位置それぞれの観測について、時系列順に各フレームで意思決定を行い、それぞれのアトラクターに対する確信度を出力する。確信度を BCI の入力に与え、Causal Inference を行うことで最終的な意思決定を行う。

映像データセットには、様々な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）を用いた [5]。図 2 は入力に使用した映像データである。オブジェクトを移動しながら撮影した映像であり、フレーム数は 1,111 である。フレームごとにこの映像データの 4 つのオブジェクトそれぞれについて、3.1 節 a) の手法で映像モダリティと位置モダリティの特徴量を抽出する。以下の評価では、それぞれのオブジェクトについて、映像モダリティと位置モダリティの特徴量を観測したときに、正しくそのオブジェクトが識別されるかを確認する。そのために、BAM には 4 つのオブジェクトの映像・位置モダリティそれぞれの特徴量を記憶する。また、各オブジェクトについての 1,111 フレーム分の特徴量の時系列データを連結し、計 4,444 フレーム分を提案するオブジェクト認識手法への入力として用いる。

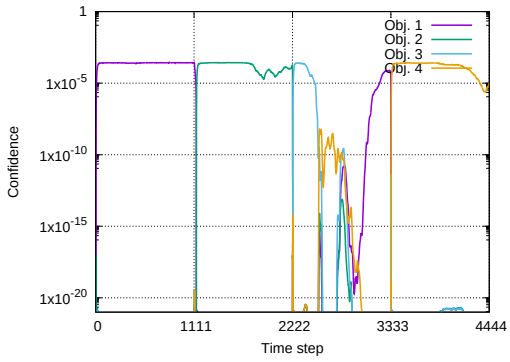
図 2　映像データ

4.2　評価結果

4.2.1　ユニモーダルでのオブジェクト推定



(a) 映像モダリティ



(b) 位置モダリティ

図 3　BAM の確信度

図 3 に BAM の確信度出力をそれぞれ示す。確信度の最も大きいオブジェクトを識別結果とするとき、4,444 個の入力に対して入力ごとに出力される識別結果が、観測したオブジェクトと一致した割合を正答率とするとき、映像モダリティのみでの BAM の正答率は 79.41%、位置モダリティのみでの BAM の正答率は 81.66% であった。図 3 の縦軸は確信度を常用対数で表記してあり、現在どのオブジェクトを観測しているのかを表す判断結果である。横軸はタイムステップであり、1 タイムステップあたり 1 フレーム、1 オブジェクトの特徴量を BAM に入力している。はじめの 1〜1,111 タイムステップでは、1 番目のオブジェクト（Obj. 1）の特徴量を、次の 1,112〜2,222 タイムステップでは 2 番目のオブジェクト（Obj. 2）の特徴量を、と 4,444 タイムステップまでに 4 つのオブジェクトの特徴量が BAM に入力される。図 3(a) は映像モダリティでの結果であり、確信度を見ると、2 つ目のオブジェクトを認識している途中で確信度が下がり、認知できなくなっているが、それ以外のほとんどすべてのフレームで観測したオブジェクトを正しく認識している。図 3(b) は位置モダリティでの結果であり、確信度を見ると、3 つ目のオブジェクトの認識が 2,500 タイムステップ付近からできていないものの、それ以外のオブジェクトは観測したものを正しく認識できていることがわかる。それぞれのモダリティにおけるオブジェクトの認知の失敗は、カメラを移動させながら取得した動画を用いていることで、抽出したオブジェクトの特徴量が時間ともに大きく変化するためである。このような場合、いかなる映像分析手法を用いたとしても、ユニモーダルでは精度よくオブジェクト推定を行うことが困難である。

4.2.2　BCI による因果推論

BCI を用いたマルチモーダル推定の評価を行う前に、まず Causal Inference にて両モダリティで同じオブジェクトを観測しているか否かの判定が正しく行われているかどうかを評価する。Causal Inference への映像モダリティ入力として図 3(a) の確信度を用いる。位置モダリティ入力として図 3(b) の確信度を用いるが、同じオブジェクトを観測する場合と、異なるオブジェクトを観測している場合のそれぞれの結果を比較評価するため、図 3(b) のタイムステップ 1〜2,222 まの確信度を 2 回繰り返して Causal Inference への入力とする。図 4 は Causal Inference の結果である。縦軸の値が 1 に近いほど二つのモダリティで同じオブジェクトを観測していると判断する。前述の入力値を与えた際に、図 4 では、前半のタイムステップ 1〜2,222 で同じオブジェクトを観測し、後半のタイムステップ 2,223〜4,444 では別々のオブジェクトを観測していると判断している。このことから、正しく Causal Inference が行われ、マルチモーダルの観測を処理できることが確認できる。

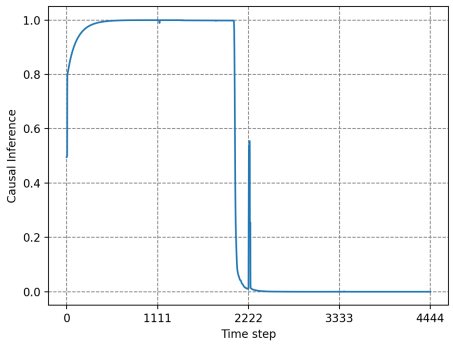


図 4　Causal Inference の結果

4.2.3　マルチモーダルでのオブジェクト推定

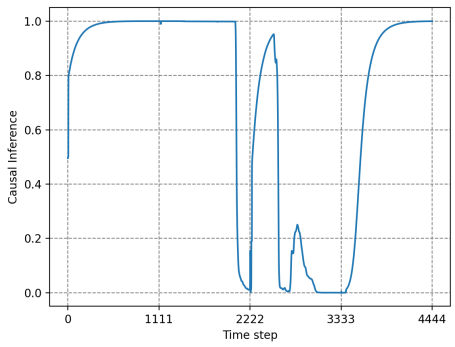
4.2.1 節で得られた、映像と位置のモダリティにおける確信度を BCI に与え、Causal Inference および識別結果を統合した結果を評価する。表 1 に各モダリティの認知の正答率を、図 5

にシミュレーション結果をそれぞれ示す。

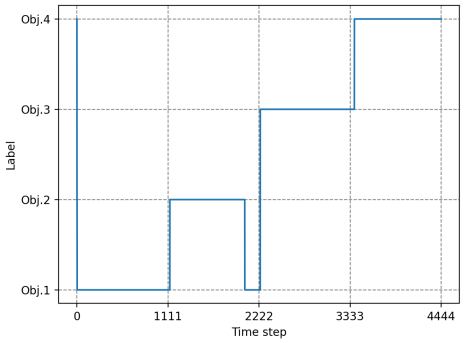
表 1 では、ユニモーダルはそれぞれ BAM の確信度出力のみから意思決定を行った結果の、マルチモーダルではそれぞれのモダリティで Model Average を行った結果の正答率を示している。Model Average では、式 (7) と式 (8) のそれぞれから結果を出力することになり、一方のモダリティを元にしつつ、他方のモダリティが認識結果を補うような出力が得られる。マルチモーダルにおいて、どちらのモダリティを主として扱うかは本稿では議論せず、今後の課題と考えているが、今回の結果ではいずれもユニモーダルの結果と比較して全オブジェクトを推定した場合の正答率が向上している。

表 1　正　答　率

モダリティ	Obj. 1	Obj. 2	Obj. 3	Obj. 4	Total
ユニモーダル 映像	100.00%	25.20%	97.83%	94.59%	79.41%
ユニモーダル 位置	99.63%	97.65%	29.61%	99.72%	81.66%
マルチモーダル 映像	99.63%	82.08%	98.46%	98.46%	93.83%
マルチモーダル 位置	99.63%	97.74%	36.63%	99.72%	83.43%



(a) 図 3 の確信度を入力とした際の Causal Inference



(b) マルチモーダル統合後の出力ラベル

図 5　BAM と BCI の統合結果

図 5(a) は Causal Inference の結果である。タイムステップ 1〜2,222 については、図 4 と同じ入力値であるため、同じ結果となっている。タイムステップ 2,223〜4,444 での位置モダリティでは、はじめは観測したものと同じオブジェクトの確信度（Obj. 3）が高いものの、タイムステップ 2,520 以降は異なるオブジェクトの確信度の方が高くなり、Causal Inference の結果は、二つのモダリティで別のものを観測していると判断される。図 5(b) は 2 モダリティを映像モダリティを元に統合した最終的なオブジェクトの識別結果である。タイムステップ 2,200 付近では、映像モダリティが誤った認識を、位置モダリティが正しい認識を行っているがいずれの確信度の値も高いために、Causal Inference としては別々のものを観測していると判断し、統合の結果誤った判断となっている。それ以外ではオブジェクト推定の結果は正しく、正答率は 93.83% に向上している。以上より複数のモダリティを組み合わせることで、4. 2. 1 節においてユニモーダルでは正しく判断できなかった部分を補いあい、より正確な判断が実現できることを確認できた。提案手法の計算時間については、最も計算量の多いのが、128 次元の映像特徴量を BAM に入力し、$\mathbf{z}_t$ の推定および確信度の出力を行う部分である。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が、1 フレームの入力あたり 1.18 ms であり、今回の評価環境であれば、30fps、60fps の映像に対しても適用可能である。

5. ま と め

本稿では、複数種類の不確実な観測情報を元に、ノイズを含んだ観測情報からオブジェクト推定を行う手法を提案した。提案手法では、脳が観測によって得た情報を処理する数理モデルである BAM と、マルチモーダルな認識を処理する数理モデルである BCI を用いて、各モダリティでは不完全な観測情報を処理して組み合わせることで適切な意思決定を行う仕組みを導入した。計算機シミュレーションにより、提案手法が各モダリティで高い確信度をもって認知している部分を組み合わせ、ユニモーダルよりも高い精度で意思決定を行えることを示した。今後の課題として、事前に学習していないオブジェクトを認識した場合に新たなアトラクターを学習する方法を検討することと、複数のオブジェクトを同時に観測して推定することを検討する。

文　　献

[1] S. Bitzer, J. Bruineberg, and S. J. Kiebel, “A Bayesian attractor model for perceptual decision making,” *PLoS Computational Biology*, vol. 11, no. 8, p. e1004442, 2015.

[2] 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, p. e943, 2007.

[3] 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.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[5] “YCB benchmarks–object and model set.” available at `http://www.ycbbenchmarks.com/`.