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.

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

久保　快斗† 　関　　良我† 　小南　大智† 　下西　英之† 　村田　正幸†

藤若　雅也††

† 大阪大学 基礎工学部情報科学科　〒 560–8531 大阪府豊中市待兼山町 1–3
† 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: †{k-kubo,r-seki,d-kominami,h-shimonishi,murata}@ist.osaka-u.ac.jp, ††fujiwaka@nec.com

あらまし　実世界上の人や物体などのオブジェクトをセンシングしてリアルタイムに 3 次元のデジタルデータとして仮想世界上に表現するデジタルツインの構築が望まれている。デジタルツインの実現における課題としてリアルタイム、高精度なオブジェクト認識があげられる。しかし現実的には 100% の精度でオブジェクトを認識することは困難である。そこで我々の研究グループでは、誤差を含んだ形で仮想世界上の存在を表現する、確率的デジタルツインの実現を目指している。人の脳が行う情報処理に倣った確率的なデジタルツインの構築のために、これまでに、脳の情報処理モデルの一つである Bayesian attractor model (BAM) に基づくオブジェクト認識技術を開発してきた。本報告では、BAM によるオブジェクト認識を行うために与える特徴量を、動画像から BAM の処理に適した形式で抽出する手法を実装および評価する。結果として、バウンディングボックス位置の推定精度指標である Intersection over Union (IoU) が平均 0.444 であった際、BAM によるオブジェクト認識精度が 54.36% となることを示した。
キーワード　デジタルツイン、ゆらぎ学習、ベイズ推定、機械学習、Siamese Network、Siamese RPN

Implementation and evaluation of an object recognition method for Digital Twin using cognitive mechanism of the Brain

Kaito KUBO†, Ryoga SEKI†, Daichi KOMINAMI†, Hideyuki SHIMONISHI†, Masayuki MURATA†, and Masaya FUJIWAKA††

† Department of Information and Computer Sciences, School of Engineering Science, Osaka University
† Graduate School of Information Science and Technology, Osaka University
†† System Platform Research Labs, NEC Corporation

Abstract　It is desired to construct a digital twin that can sense objects such as people and objects in the real world and represent them in the virtual world in real time. In our previous work, we developed an object recognition method based on the Bayesian attractor model (BAM), which mimics the information processing of the human brain. In this paper, we propose a method for extracting features for the BAM-based object recognition from video images in a format suitable for BAM processing. By using the proposed method, when the intersection over union (IoU), which is the accuracy index of bounding-box estimation, is 0.444, the object recognition accuracy of the BAM is 54.36%.
Key words　Digital twin, Yuragi learning, Bayesian inference, machine learning, Siamese network、Siamese RPN

1. はじめに

近年ロボットの自動化や車の自動運転技術が望まれ、研究が進められている [1]。現在のロボットは pepper や aibo といったコミュニケーションを重視した家庭用製品が主流であり、自動運転技術も高速道路での車線維持、渋滞時の前車の追従と

いったものにとどまっている。将来のロボットは建設現場での自動敷設に使われるといった需要が現れることが考えられ、自動運転も最終的には乗った場所から目的地まで、交差点の右左折など含めてすべて自動で行う技術が現れると考えられる。このとき、一つの行動選択の誤りが重大な事故を引き起こすことが想像できる。自動制御の安全性のためには、正確かつ素早い情報取得、処理、行動判断が不可欠である。このような自動制御の実現方法として、実世界を仮想世界上にリアルタイムに表現し、この仮想世界の情報に基づいて実世界上の制御を行うデジタルツインの構築が検討されている [2]。

デジタルツイン構築の課題として、リアルタイムかつ正確なオブジェクト認識があげられる。カメラによって実世界をセンシングする際、カメラのピントボケ、本体のブレなどによって誤差が含まれることとなる。またセンシングデータの処理は比較的計算処理能力が乏しい機器で行われることが想定され、計算処理にはある程度の時間を要し、精度も低いと考えられる。そもそもが、センシングしたデータからオブジェクトを確実に抽出し、誤ることなくリアルタイムで認識することは非常に困難な課題である。そのため我々の研究グループでは、ノイズが含まれる不確実な入力から、認識誤差を許容し、誤差を含んだ形でオブジェクトを認識し、認識結果を仮想世界上に確率を含んだ形式で表現する確率的デジタルツインの実現を目指している [3]。

不確実な入力から意思決定を行うよく知られたものとして人間の脳がある。人の脳の行う情報処理では、聴覚や視覚といった感覚器から情報を得たのち、その情報と自身の知識や経験を照合し、観測したものが何であるのかを判定する。BAM (Bayesian Attractor Model) [4] は、このような意思決定をベイズ推定に基づいて数理モデル化したものである。我々の先行研究では、この BAM を用いることで、人の脳が行うユニモーダル情報処理を模したオブジェクト認識手法を提案し、有効性を示している [5]。文献 [5] は、動画像におけるオブジェクト認識を BAM によって行う際の、BAM 自体の認識精度、計算速度を評価したものであり、BAM に与える入力には理想的な状況を想定して取得した連続データを用いている。すなわち、動画像中の認識したいオブジェクトを囲むバウンディングボックスが正確に推定できていることを仮定し、その上でバウンディングボックスに囲まれた画像に対する特徴量を抽出している。また計算時間削減のために、特徴量抽出には 4 層の CNN (Convolutional Neural Network) を用いている。

本稿では、BAM に与える入力特徴量を取り出すために、上記の特徴量抽出方法を拡張する。オブジェクトの存在する領域推定に Siamese RPN [6] を採用し、同時に、Siamese RPN において用いる Siamese Network によって特徴量を抽出する。以上のオブジェクト検出、特徴量抽出を含めた形で、BAM によるオブジェクト認識手法の実装を行い認識精度を評価する。文献 [5] では 4 層の CNN を用いて比較的小さな次元の特徴量を抽出していたが、これは前段のオブジェクトの検出が確実にできていることを前提としていたものであり、オブジェクトの検出および特徴量抽出、BAM によるオブジェクト認識までの一連の処理の実行に要する計算時間についても評価を行う。

2. 関 連 研 究

文献 [5] におけるオブジェクト認識手法に関して、映像モーダル処理の一連の流れを図 1 に示す。

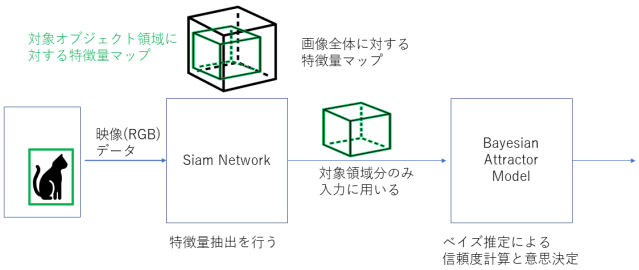


図 1　文献 [5] における映像モーダル処理

文献 [5] ではバウンディングボックスが正確に推定された上で特徴量抽出を行っている。さらに後述の BAM では認識を行うオブジェクトに関する特徴量が状態空間上のアトラクターに紐づけて記憶されるが、文献 [5] ではこの記憶する特徴量を動画の最初の 1 フレーム目に写ったオブジェクトについて取り出したものとしている。認識の手順として、動画像の連続フレームについて、探索対象であるオブジェクトを正確に囲むバウンディングボックス領域内の画像に関する特徴量が取り出され、BAM に入力される。入力された特徴量が記憶したどのオブジェクトの特徴量と近いのかを判断する。

2.1 Bayesian Attractor Model

文献 [4] では、人が知覚情報に対してどのように意思決定を行っているのかを、ベイズ理論を用いてモデル化した Bayesian Attractor Model (BAM) を提案している。BAM について、以下に説明を行う。

BAM では、時刻 t における観測値を特徴量空間での変数 x_t として表現する。人の意思決定は状態変数 z_t で表され、観測値 x_t を受けて z_t を更新するモデルとなっている。ある時刻 $t-\Delta$ から次の時刻 t となった際、意思決定状態 z_t は次式で更新される。

$$z_t - z_{t-\Delta} = \Delta f(z_{t-\Delta}) + \sqrt{\Delta} w_t \tag{1}$$

ここで $f(z)$ はホップフィールドダイナミクスである。このダイナミクスは複数のアトラクター $\phi_1 \dots \phi_n$ を持つように設計されており、アトラクターの数は意思決定の選択肢の数に応じた数だけ準備する。w_t はノイズを表し、正規分布 $N(0, \frac{q^2}{\Delta} I)$ に従う。重要なパラメータとして、q はダイナミクスの不確実性を表し、これが大きいほど z_t がアトラクターの間で切り替わり、意思決定状態が変更される可能性が高くなる。

また、式 (2) により観測値 x の確率分布を予測する。

$$x_t = M\sigma(z) + v_t \tag{2}$$

ここで $M = [\mu_1, \dots \mu_n]$ は各選択肢（$i = 1 \dots n$）がノイズなしで提示されたときの特徴量ベクトルを並べた行列であり、$\sigma(z)$ は z の各要素を 0 ～ 1 に変換するシグモイド関数である。

v はノイズを表し、正規分布 $N(0, r^2 I)$ に従う。r は観測の不確実性を表し、r が大きいほどノイズが大きくなる。

式 (1) と式 (2) で表される生成モデルを逆方向に推定することで時刻 t における観測値 x_t から意思決定状態 z_t の事後確率分布 $P(z_t|x_t)$ が得られる。z_t が非線形なホップフィールドダイナミクスに従うことを仮定しているため、UKF (Unscented Kalman Filter) が用いられている。得られた事後確率分布 $P(z_t|x_t)$ に対して、以下の条件を満たすときに意思決定状態が i であると決定する。

$$p(z_t = \phi_i | X_{\Delta:t}) \geq \lambda \tag{3}$$

ただし $X_{\Delta:t} = x_\Delta \dots x_t$ はその時点までに行われたすべての観測値である。$p(z_t = \phi_i|X_{\Delta:t})$ は i に対応するアトラクター ϕ_i で評価された事後確率密度であり、λ は信頼度の基準となるしきい値である。

3. 提 案 手 法

3.1 全体アーキテクチャ

文献 [5] では、バウンディングボックスが正しく推定できているという理想的な状況下で特徴量を取り出した。その結果、BAM による認識精度は比較的高い値が得られていた。しかしながら、バウンディングボックスの推定結果には、多くの場合誤差が含まれる。仮に推定が誤った場合には、そのバウンディングボックスに対応して取り出される特徴量は、BAM によって認識したいオブジェクトの持つ特徴量と異なるものとなってしまう。すなわち特徴量抽出の結果はバウンディングボックスの推定精度に大きく依存する。本稿では、バウンディングボックスの推定とバウンディングボックスに対応する領域の特徴量抽出を同時に行うために、Siamese RPN [6], [7] を採用する。実装するシステム全体アーキテクチャの構成を図 2 に示す。

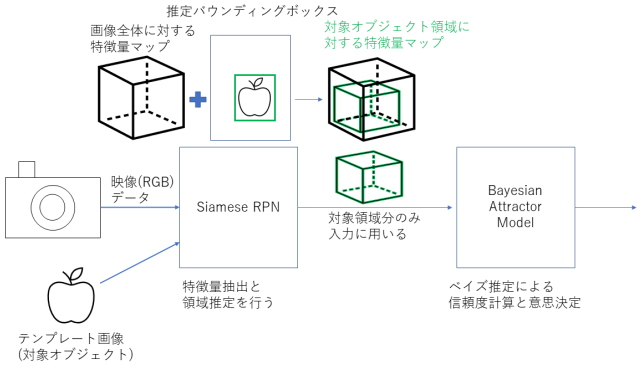


図 2　実装するシステム全体アーキテクチャ

3.2 Siamese RPN

Siamese RPN のネットワーク構造を図 3 に示す。

Siamese RPN は二つのネットワークアーキテクチャにより構成される。一つ目は Siamese Network（3.2.1 節）である。Siamese Network では探索対象のオブジェクトの画像と、そのオブジェクトが写っているのかを探索する全体画像から、特徴量抽出器を使って特徴量マップを取り出す。前者をテンプレート画像、後者をオリジナル画像と呼ぶ。二つ目は Region Proposal Network である。ここでは Siamese Network の出力であるテンプレート画像に対する特徴量マップ $\phi(Z)$ と、オリジナル画像に対する特徴量マップ $\phi(X)$ を受け取り、それぞれに対して畳み込み層による処理を行ったのち、それぞれの結果を合わせて畳み込み演算を行う。結果として検出対象オブジェクトが画像のどの位置に写っているのかを推定し、出力する。

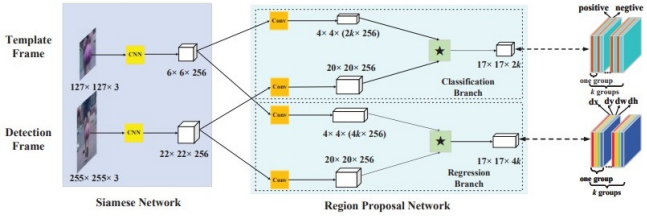


図 3　Siamese RPN のメインフレームワーク (文献 [6] より引用)

3.2.1 Siamese Network

Siamese Network [8] はオブジェクトトラッキングの問題を、認識したいテンプレート画像から抽出される特徴表現と、オリジナル画像から抽出される特徴表現間の、相互相関により得られる汎用的な類似性マップが正解となるように学習することで解決を図る。入力 A をテンプレート画像、入力 B をオリジナル画像としたときに A と B を同一の CNN で畳み込み、特徴量マップを取り出す。この特徴量マップの特徴量空間状の位置が、A と B が同一クラスのものであれば近くなるように、違うクラスのものであるならば遠くなるように学習を進める。このときの距離はユークリッド距離で定義される。

3.2.2 Siamese RPN におけるモデル学習

図 3 の Siamese Network において、CNN の構成、用意する学習用データセットによって特徴量抽出の学習結果は異なる。今回は学習用データセットに Youtube-BB (YTBB) データセット [9] を用いた。このデータセットはアノテーションされ、正解バウンディングボックスを保持している。Siamese RPN が、取り出した特徴量に準じて予測するバウンディングボックスに対応する場所の特徴量マップを取り出す。よって、本手法の CNN の学習は、特徴量の取り出し方のみではなく、間接的に次の段階である RPN に影響する。CNN について、二通りの構造（三層、四層）を用意し、層の次元数も複数種類のものを用意した。

表 1　CNN 構成

モデル	1 層目次元数	2 層目次元数	3 層目次元数	4 層目次元数
A	24	32	16	
B	24	24	8	
C	32	16	16	4
D	16	16	16	16
E	24	32	32	16

三層及び四層の CNN における特徴量抽出部分は、オートエンコーダを構成し、入力画像と出力画像が同じになるような学習を行う。この特徴量抽出器において、classification ブランチと regression ブランチをチューニングする。

また、文献 [6] の元来の手法である、AlexNet [10] [11] を特徴量抽出として用いる学習（モデル F）も試みた。AlexNet は学習済みモデルであるので、特徴量抽出部分は変更せず、classification ブランチと regression ブランチをチューニングする。なお AlexNet によって抽出された特徴量マップの構成は、図 3 の Siamese Network におけるテンソルである。

3.3 BAM における学習

BAM の学習は、アトラクターに対応付けて記憶させる特徴量の選定である。今回は、映像フレームの中で、バウンディングボックスの推定に成功したフレーム番号を手動で選定し、その特徴量を記憶させた。また、記憶させる際に標準化処理が必要となるが、これは文献 [5] と同様の手順で行った。

4. 評価結果

特徴量抽出のための CNN の構成によるバウンディングボックスの推定精度を評価する。また、取り出す特徴量が BAM における認識精度に与える影響を評価する。

Siamese Network の CNN には、表 1 で定義したモデル A から E を用いた。各層数で良好であった結果を以降に示す（モデル A と D）。また、CNN に AlexNet を用いた結果（モデル F と呼ぶ）についても示す。認識精度の評価には、形状、大きさ、質感、重さ、剛性などが異なる日常的なオブジェクトの写ったデータセットである YCB データセット [12] を用いた。今回は YCB データセットの中から、固定点にある 4 オブジェクトが、移動するカメラにより撮影される全 1,111 フレームの映像データセットを用いた。それぞれにおいてアノテーションが行われており、各オブジェクトに関する正解バウンディングボックスもあらかじめ得られている。この正解情報はバウンディングボックスの推定精度の評価に用いる。

Siamese Network のテンプレート画像として、612 フレーム目の画像において正解バウンディングボックス情報から切り出した各オブジェクトを設定する（図 4）。612 フレーム目を選定した理由は以下の通りである。切り出したオブジェクト画像から取り出す特徴量は 1,111 フレームのオブジェクト推定の基準特徴量となるため、他のオブジェクトの特徴量が入り込む影響を少なくする必要や、そのオブジェクトから汎用的な特徴量を取り出す必要がある。後者について、仮にオブジェクトを真上からみたものをテンプレートとして与えた場合、そこから取り出す特徴量は真上の情報のみとなってしまい、上と側面で色が全く違う物体などを別のものとして判別してしまう可能性が大きい。こうしたことから、主観による評価で、切り出した時に比較的オブジェクト全体を捉えていた 612 フレーム目から切り出した各オブジェクトをテンプレートとして与えることとした。

バウンディングボックスの検出精度を表す IoU が高いフレームが、実際に見たいオブジェクトを切り出せている箇所である。その領域から得られる特徴量は実際にその物体を表す特徴量に近い可能性が高い。そこで今回は、そのようにして得られた特徴量を BAM のアトラクターに記憶させた。

4.1 三層 CNN からなる特徴量抽出器（モデル A）

Siamese Network の CNN 構成として、三層で層の出力をそれぞれ 24、32、16 とした特徴量抽出器モデル A の各タイムステップにおける IoU を図 5 に示す。

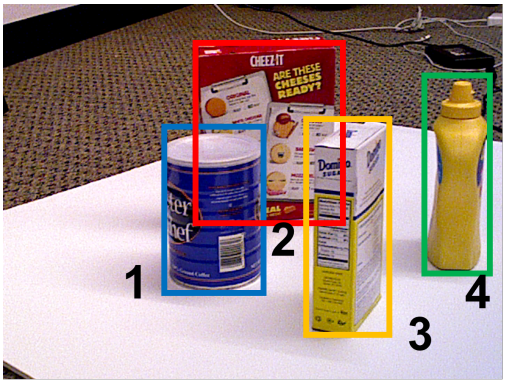
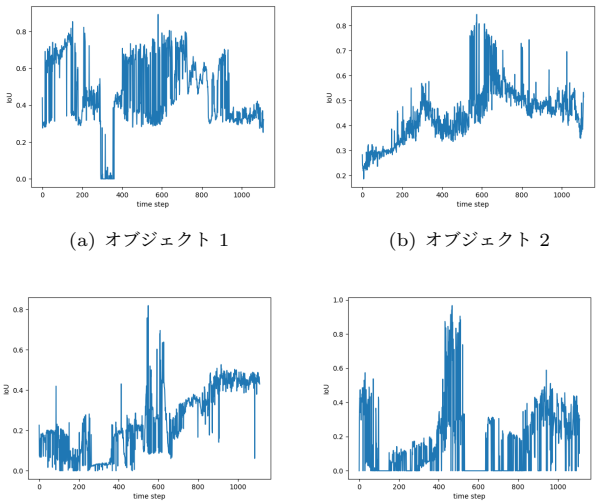


図 4　テンプレート画像（612 フレーム目より選定）

図 5　モデル A における各オブジェクトの IoU



(a) オブジェクト 1　(b) オブジェクト 2

(c) オブジェクト 3　(d) オブジェクト 4

次に Siamese RPN による推定結果と計算時間を表 2 に示す。

表 2　モデル A での Siamese RPN の計算結果

	obj.1	obj.2	obj.3	obj.4	平均
平均 IoU	0.484	0.452	0.241	0.178	0.339
計算時間 [s/f]	0.0293	0.0297	0.0304	0.0294	0.0297

この CNN はオートエンコーダで学習しているため、テンプレート画像と同等である 612 フレーム付近の IoU は他のタイムステップより比較的高くなる傾向にある。このように取り出した特徴量を標準化したのち BAM に入力として与えた。このときの BAM での認識結果を図 6 に示す。

各ステップにおける確信度が一番高い選択肢を意思決定結果とするとき、認識精度を表 3 に示す。

4.2 四層 CNN からなる特徴量抽出器（モデル D）

Siamese Network の CNN 構成として、四層で層の出力をそれぞれ 16、16、16、16 とした特徴量抽出器モデル D の各タイムステップにおける IoU を図 7 に示す。

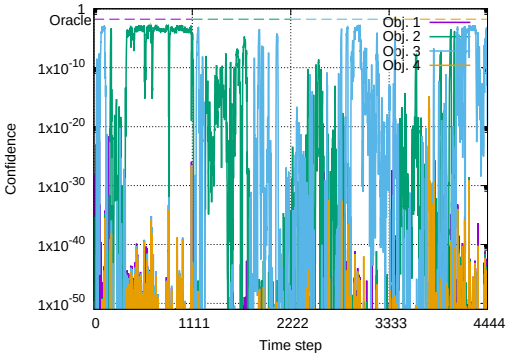
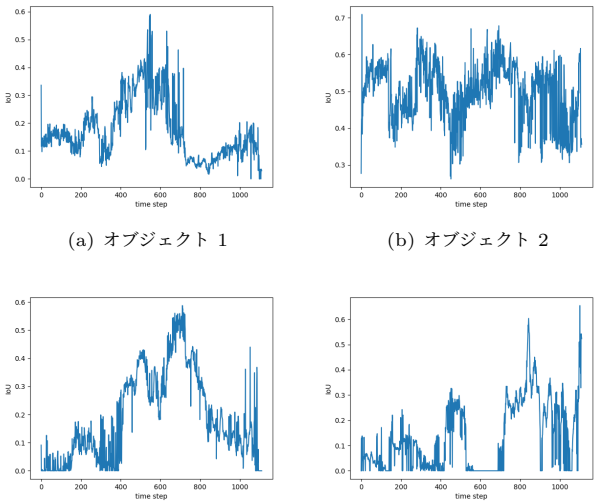


図 6　BAM の確信度（モデル A の特徴量を使用）

表 3　BAM の認識精度（モデル A の特徴量を使用）

	obj.1	obj.2	obj.3	obj.4	平均
認識精度 [%]	5.94	55.4	62.1	6.93	32.6
計算時間 [s/f]	0.0539				



(a) オブジェクト 1　(b) オブジェクト 2

(c) オブジェクト 3　(d) オブジェクト 4

図 7　モデル D における各オブジェクトの IoU

表 4　モデル D での Siamese RPN の計算結果

	Obj.1	obj.2	Obj.3	Obj.4	平均
平均 IoU	0.279	0.502	0.196	0.131	0.277
計算時間 [s/f]	0.0243	0.0238	0.0233	0.0236	0.0238

次に Siamese RPN の推定結果と計算時間を表 4 示す。

三層のときと同様、612 フレーム付近の IoU は他のタイムステップより比較的高くなる傾向にある。このときの BAM での認識結果は図 8 のようになった。また、認識精度は表 5 に示す通りである。

表 5　BAM の認識精度（モデル D の特徴量を使用）

	obj.1	obj.2	obj.3	obj.4	平均
認識精度 [%]	13.2	75.5	94.5	22.2	51.4
計算時間 [s/f]	0.0393				

4.3　AlexNet からなる特徴量抽出器（モデル F）

Siamese Network の CNN 構成として、AlexNet を用いた特

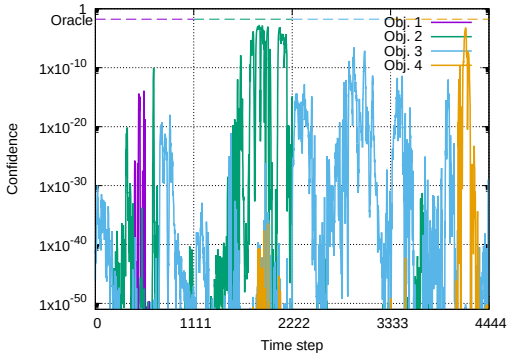
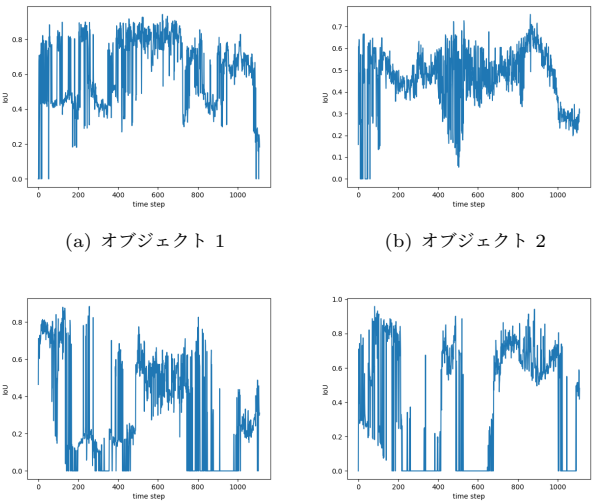


図 8　BAM の確信度（モデル D の特徴量を使用）

徴量抽出器モデル F の各タイムステップにおける IoU を図 9 に示す。



(a) オブジェクト 1　(b) オブジェクト 2

(c) オブジェクト 3　(d) オブジェクト 4

図 9　モデル F における各オブジェクトの IoU

Siamese RPN の推定結果と計算時間は表 6 に示す通りであり、平均的な IoU がモデル A や D と比べて大きいが、計算時間は約 2 倍となっている。このように取り出した特徴量を標準化したのち BAM に入力として与えた。

表 6　モデル F の Siamese RPN の計算結果

	Obj.1	Obj.2	Obj.3	Obj.4	平均
平均 IoU	0.626	0.461	0.306	0.382	0.444
計算時間 [s/f]	0.0509	0.0488	0.0484	0.0490	0.0493

BAM での認識精度は図 10 に示すとおりである。

各ステップにおける信頼度が一番高い選択肢を意思決定結果とするときの認識精度を表 3 に示す。

表 7　BAM の認識精度（モデル F の特徴量を使用）

	obj.1	obj.2	obj.3	obj.4	平均
認識精度 [%]	69.9	82.7	40.1	24.8	54.36
計算時間 [s/f]	0.0657				

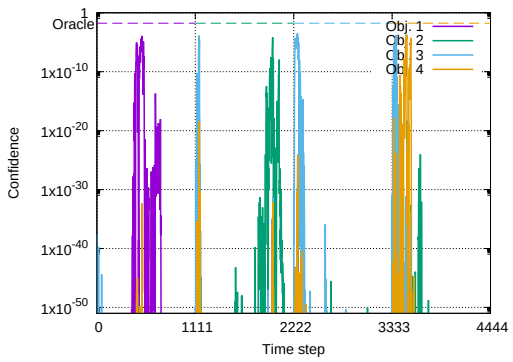


図 10　BAM の確信度（モデル F の特徴量を使用）

4.4 考　　察

総じて IoU が大きいモデルを用いた際に BAM の認識精度が比較的高い結果となっている。しかしながら、モデル D におけるオブジェクト 3 の認識結果や、モデル A におけるオブジェクト 1 の認識結果など、平均 IoU が低い際にも BAM の精度が高い結果も得られた。この理由に関しては現在調査を行っている。

IoU が比較的高い際にも BAM の認識精度が低下する結果も見られた。この原因としては、推定されたバウンディングボックスが正解バウンディングボックスに対して大きいとき、IoU が 0.5 程度あったとしても、隣り合った物体が映り込みその特徴量を取り出してしまうことである。また BAM のアトラクターに記憶させる特徴量として、IoU が高いフレームでの特徴量を設定しているが、図 7(d) などをみても分かる通り、IoU が最大でも 0.6 程度のときには、背景などの別の情報を大きく含んだ特徴量をアトラクターとして与えていることになる。先行研究 [5] における理想的な状況を想定した評価では、BAM を用いた認識精度は 8 割程度であり、IoU の精度が重要であることが分かる。今回の結果では、領域予測および特徴量抽出に関して、Siamese Network に AlexNet を利用した Siamese RPN が最も有効であった。

5. おわりに

本稿では、人の脳の備える、不確実な観測データから高精度な意思決定を行う特徴を用いたオブジェクト認識手法の実装を行った。我々の先行研究において用いていた、あらかじめ与えられた映像領域から特徴量を取り出すアーキテクチャを拡張し、映像領域から認識を行いたいオブジェクトの存在する領域の推定と、領域からの特徴量抽出とを同時に行うアーキテクチャの実装を行った。取り出した特徴量を BAM の入力としてオブジェクト信頼度を計算したのちオブジェクト推定を行った。公開データセットである YCB データセットを用いた評価の結果，BAM におけるオブジェクト認識精度は 54.36% であった．計算時間については，1 フレームの映像に対して Siamese RPN による領域推定と特徴量取り出しに平均 0.0473 秒を要し，BAM による認識に平均 0.0666 秒を要した．合計では 0.1 秒以上となり，10fps 以下である．一般に動画像に用いられるフレームレートである 30fps や 60fps よりも遅いため，実用に向けて計算時間の削減が必要である．認識精度に関して、機械学習を用いたフレームワークで向上することは、計算時間の観点から現実的ではなく、映像以外の情報を用いた認識結果とのマルチモーダル統合認識が有力であると考えており、今後取り組む予定である。

謝　　辞

本研究の一部は、総務省の委託研究開発「Beyond 5G を活用した安全かつ効率的なクラウドロボティクスの実現」により実施したものである。

文　　献

[1] N. Nikolakis, V. Maratos, and S. Makris, “A cyber physical system (CPS) approach for safe human-robot collaboration in a shared workplace,” Robotics and Computer-Integrated Manufacturing, vol.56, pp.233–243, 2019.
[2] B. He and K.-J. Bai, “Digital twin-based sustainable intelligent manufacturing: A review,” Advances in Manufacturing, vol.9, no.1, pp.1–21, 2021.
[3] 下西英之，大下裕一，小南大智，関良我，村田正幸，吉田裕志，野上耕助，藤若雅也，中野谷学，金友大，“確率的デジタルツイン，” 電子情報通信学会技術研究報告（CQ2021-57），vol.121，no.173，pp.97–101，2021.
[4] 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.
[5] 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），vol.121，no.15，pp.59–64，2021.
[6] L. Bo, Y. Junjie, W. Wei, Z. Zheng, and H. Xiaolin, “High performance visual tracking with siamese region proposal network,” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), vol.1, no.1, pp.8971–8980, 2018.
[7] G. Yao, C. Rui, T. Ying, C. Xuehong, and L. Ruiyu, “Combining siamese network and regression network for visual tracking,” IEICE Transactions on Information and Systems, vol.E103.D, no.8, pp.1924–1927, 2020.
[8] G. Koch, R. Zemel, R. Salakhutdinov, et al., “Siamese neural networks for one-shot image recognition,” ICML deep learning workshop, vol.2, no.1, p.8, 2015.
[9] “Youtube-boundingboxes dataset”. available at https://research.google.com/youtube-bb/,Accessed : 2021-12-20.
[10] K. Alex, S. IIIya, and E.H. Geoffrey, “Imagenet classification with deep convolutional neural networks,” Advances in Neural Information Processing Systems, vol.25, no.1, p.1097–1105, 2012.
[11] R. Olga, D. Jia, J. Krause, S. Sanjeev, M. Sean, H. Zhiheng, K. Andrej, K. Aditya, B. Michael, C. Alexander, and F.-F. Li, “Imagenet large scale visual recognition challenge,” International Journal of Computer Vision (IJCV), vol.115, no.3, pp.211–252, 2015.
[12] “YCB benchmarks–object and model set”. available at http://www.ycbbenchmarks.com/,Accessed : 2021-12-20.