計測自動制御学会東北支部 第 S60 回研究集会 (2024.12.17)
資料番号 S60-21

Perceiverを用いた視覚・聴覚情報に基づく物体認識法の検討

Exploration of Object Recognition Method Based on Visual and Auditory Information Using Perceiver

○西原優空 *，田中大介 *

○ Yura Nishihara*, Daisuke Tanaka*

*新居浜工業高等専門学校

*National Institute of Technology (KOSEN), Niihama College

キーワード： 物体認識 (Object Recognition)，マルチモーダルセンサ (Multimodal Sensor)，Transformer，Perceiver，STFT (Short-Time Fourier Transform)，FFT (Fast Fourier Transform)

連絡先： 〒 792-8580　愛媛県新居浜市八雲町 7-1 新居浜工業高等専門学校
田中大介，Tel.: (0897)37-7775, E-mail: d.tanaka@niihama-nct.ac.jp

1. はじめに

近年，視覚や聴覚といった異なるセンサ情報を組み合わせる物体認識が，ロボット技術をはじめ多くの分野で注目されている．特に，視覚情報は物体の形状や色を，聴覚情報は材質や内部構造などを示す特徴を提供し，これらを統合することで認識精度の向上が期待されている．また，これらの情報を活用することで，従来の単一センサによる認識では困難だった多様な特徴の補完が可能となり，より高精度な物体認識を実現できる．

しかし，これらのセンサ情報を統合する際には，それぞれの情報の相補性を最大限に活かす工夫が求められる．複数のセンサ情報を統合するため，Gao らは CNN を用いて視覚情報と聴覚情報から特徴を抽出し，それらを融合して物体の触覚的特徴を言語で表現するクラス分類が実現している [1]．また，著者らは，Vision Transformer (ViT)[2] を基にした識別モデルである Vision-Auditory Transformer (ViAuT)[3, 4] を用いて，視覚情報と聴覚情報の両方を活用して，ボールの認識タスクを実現している．

これらの例のように，物体認識におけるアプローチでは，センサ情報を事前の仮定に依存せずデータドリブンで取得し，それを基に特徴を抽出して識別を行うアルゴリズムが発展している．

本研究では，Transformer ベースの識別モデルである Perceiver を用いた認識アルゴリズムの可能性を検討する．本稿では先行研究 [4] と同様の分類タスクを実験し，実験的に検討した結果を報告する．

2. 準備：Perceiver

本章では，Perceiver モデル [5] について説明する．Perceiver は，DeepMind によって開発された Transformer ベースの分類モデルであり，点群，音声，画像などの異なるタイプの入力デー

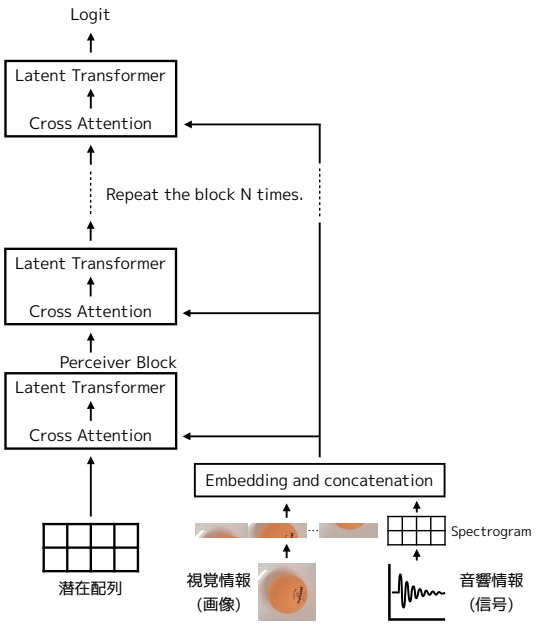


Fig. 1: Perceiver の構造．Perceiver architecture

タを同時に受け取り，処理することができるマルチモーダルモデルである．このモデルは，人間の脳が異なる種類の情報を統合し，柔軟に処理する仕組みに着想を得ており，入力データの種類に依存せず，様々なデータタイプに対応した分類が可能である．

Perceiver の構造は Fig. 1 に示すようになっており，入力データが埋め込まれた入力配列と学習可能なパラメータである潜在配列が，Attention 機構を含む Perceiver Block によって繰り返し処理される．ここでの入力配列は，入力データをトークン化し，その後埋め込み処理を施して構成する．本タスクのように複数種類のデータを扱う場合，それぞれのデータに対して同様の処理を行い，それらを結合することで最終的な入力配列を得る．Perceiver の大きな特徴の一つである潜在配列は，Cross Attention によって入力配列がマッピングされ，入力データを表す特徴を獲得し，クラスの確率を生成する．さらに入力配列長 M に対して潜在配列長 N を $M \gg N$ と設定し，入力配列の代わりに Latent Transformer (Self Attention) で処理することによって，大規

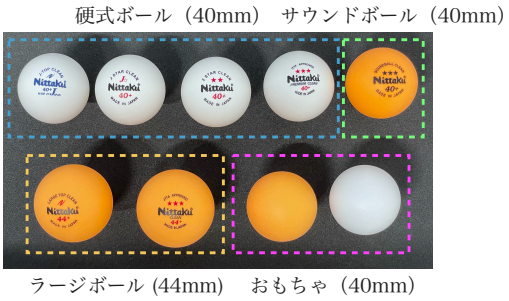


Fig. 2: 実験に用いた卓球ボール．Table tennis balls used in the experiment.

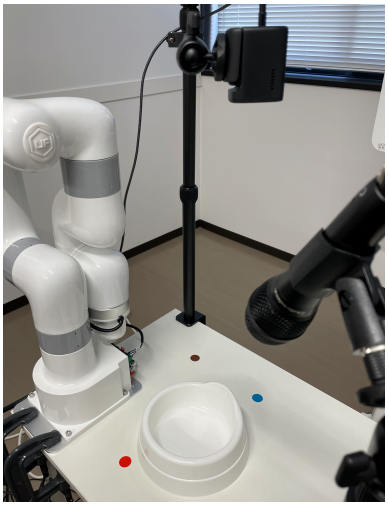


Fig. 3: 実験用データの取得システム．Experimental data acquisition system.

模な入力を扱うことが可能となり計算量の削減につながる．

3. 実験

3.1 実験設定

本研究では，先行研究 [4] と同様に卓球ボールの画像および音声データを用いて認識実験を行う．学習データの詳細は以下の通りである．実験に使用する卓球ボールは，Fig. 2 に示す計 9 種類を用意した．これらのボールは机上に置いた状態のボールの視覚情報（400 × 400 px のカラー画像）を Fig. 3 上部の Web カメラで取得し，ロボットアーム（UFactory Lite6）を使用して同じ高さからボールを落下させた際の聴覚情報（サンプリング周波数 48 kHz，1.6 秒の時

Table 1: 実験に用いた卓球ボールの概要. Overview of the table tennis balls used in the experiment.

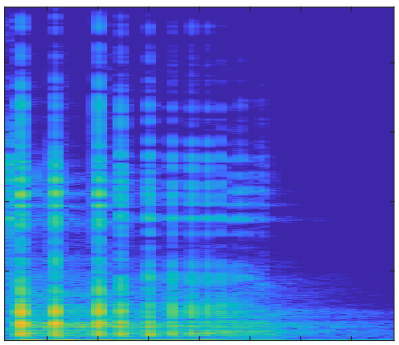
種類		直径	色
硬式ボール	3 スター	40 mm	白
	2 スター	40 mm	白
	J スター	40 mm	白
	練習球	40 mm	白
サウンドボール		40 mm	橙
ラージボール	3 スター	44 mm	橙
	練習球	44 mm	橙
おもちゃ		40 mm	白・橙

(a) 硬式ボール (3 スター). table tennis ball (3 stars).

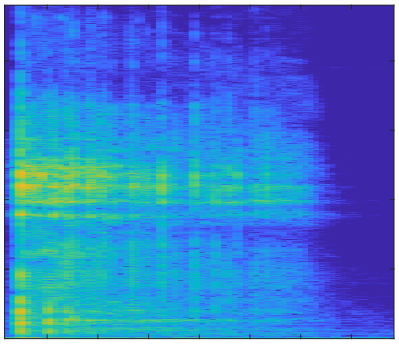
(b) サウンドボール. Sound ball.

(c) おもちゃ. Toy ball.

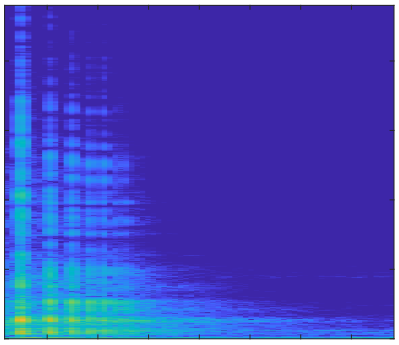
Fig. 4: 画像データの一例. Examples of image data.



(a) 硬式ボール (3 スター). table tennis ball (3 stars).



(b) サウンドボール. Sound ball.



(c) おもちゃ. Toy ball.

Fig. 5: パワースペクトログラムの一例. Examples of a power spectrogram.

間波形）を Fig. 3 の右部のマイクで録音し，それぞれ 500 個ずつ取得した．これらの取得された各ボールのデータを訓練用，検証用，テスト用にそれぞれ 70%，20%，10%に分割した．このうち訓練用データを用いて，Table 1 に示す 8 クラス分類を行うモデルを学習した．なお,「おもちゃのボール」については，白と橙の 2 色が存在するが，先行研究 [4] と同様，これらは同一のクラスとして扱うこととした．

ここで，これらのデータの前処理について補足する．まず，視覚情報としての画像の前処理は，224×224 px へのリサイズのみとし，必要最小限としている．それに対して聴覚情報としての音声データについては，周波数成分の時間変化を考えるための STFT と，信号全体から詳細な周波数成分を得る FFT の 2 通りの変換を検討した．STFT ではウィンドウ幅 4096 (85.3 ms)，オーバーラップ 75%で得られたパワースペクトログラムを使用し，FFT ではウィンドウ幅 76800 (=信号長）を用いて解析を行った．先行研究 [4] では，得られたパワースペクトログラムにデシ

ベル変換を施すことで，変換を行わない場合に比べて精度が向上することが確認されている．これを踏まえ，本研究でも同様にデシベル変換を適用した．

一例として，Figs. 4, 5 に硬式ボール (3 スター)，サウンドボール，およびおもちゃのボールのそれぞれの視覚情報，聴覚情報の例を示す．ボールのロゴが写る向きにある時には視覚情報をもとに認識することも可能であるが，そうでない場合には聴覚情報も加えることが有効であると考えられる．

本実験では，聴覚情報に対して STFT と FFT を行う条件 2 つに加えて，視覚情報のみを用いた場合の計 3 条件のモデルを学習した．最大エポック数は 1000 とし，検証用データに対する損失関数が最も低いときのモデルを得た．

3.2 実験結果

テストデータに対する識別精度を Table 2 に示す．なお，乱数を用いてパラメータの初期値を与えてモデルを学習したため，初期値依存性を軽減するために 3 回学習した平均値を用いている．Table 2 に示している ViAuT は先行研究 [3, 4] で用いられたモデルであり，実験で使用する Perceiver モデルと ViAuT モデルとが同程度のパラメータ数になるようにハイパーパラメータを調整している．Table 2 より，この 2 つは同程度の性能を発揮していることが確認できる．一方，NVIDIA Quadro RTX 8000 を用いた時の推論時間を Table 3 に示す．推論においては，Perceiver のほうが 10 倍程度の時間が必要となることが確認された．そのため，入力する情報やタスクによっては Perceiver の計算量が問題になることも考えられ，引き続き検証を行いたい．

Table 2: 正解率 [%]．Accuracy scores [%].

Input type	Perceiver	ViAuT
STFT-dB	99.99	100.00
FFT-dB	99.66	99.93
image only	60.59	65.93

Table 3: 推論時間 [s]．Inference time [s].

Input type	Perceiver	ViAuT
STFT-dB	2.8005	0.3024
FFT-dB	2.7096	0.2538
image only	2.6819	0.2344

4. おわりに

本稿では，マルチモーダルセンサ情報に基づく物体認識において，事前の仮定に依存せずセンサ情報から特徴抽出を行えるアルゴリズムとして Perceiver を検討し，実験的にその特性を確認した．今後更にデータ量を増加させたり，ハイパーパラメータの設定によるモデルの複雑性の変更により，物体認識タスクにおける他手法との差を更に検証したい．

謝辞

本研究は JSPS 科研費 JP22K17918 の助成を受けたものです．

参考文献

1) Y. Gao *et al.*, Deep learning for tactile understanding from visual and haptic data, 2016 IEEE International Conference on Robotics and Automation (ICRA), pp. 536–543, 2016.

2) A. Dosovitskiy *et al.*, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, *arXiv:2010.11929*, 2021.

3) 田中ら, Transformer アーキテクチャに基づく視覚・聴覚情報の統合システムの開発, 第 37 回信号処理シンポジウム, 新潟, pp. 368–371, 2022.

4) 田中, Transformer アーキテクチャに基づく視覚・聴覚情報からの特徴抽出の検討, 第 43 回計測自動制御学会九州支部学術講演会, 熊本, pp. 204–205, 2024.

5) A. Jaegle *et al.*, Perceiver: General Perception with Iterative Attention, Proceedings of the 38th International Conference on Machine Learning, pp. 139:4651–4664, 2021.