マルチモーダルセンサ情報の相補性を活用可能なデータドリブン特徴抽出法の開発

Development of data-driven feature extraction methods capable to exploit the complementarity of multimodal sensor information

研究代表者　新居浜工業高等専門学校機械工学科　准教授　田中大介※

Daisuke Tanaka

Autonomous robots that operate in an environment coexisting with humans are expected to recognize the surrounding environment accurately. To achieve object recognition, it is necessary to find the unique features of the objects to be recognized. Prior research results have shown that high accuracy can be achieved when multiple modality information is used. These results suggest that methods that use multiple modalities information are effective to find the unique features. In this research, to discuss about integrating each modality and extracting features, the overview of the system that can extract the features of the objects to be recognized by integrating visual, tactile, and auditory information as multimodal sensor information with the results of verification experiments is shown.

要旨

人間との共存環境下で活動する自律型ロボットは，正確に環境認識することが期待される．ロボットが物体認識を達成するためには，認識対象物体の固有の特徴を見つける必要がある．先行研究では，複数のモダリティに関する情報を利用することにより，高い認識精度を獲得することが示されている．これらの結果は，複数のモダリティ情報を利用する手法が，物体固有の特徴を見つけるために有効であることが分かる．従って，各モダリティの統合及び特徴の抽出方法を議論することが要求される．本稿では，マルチモーダルセンサ情報として視覚，触覚，聴覚に関する情報を統合し，対象物の特徴を抽出するシステムの概要を，検証実験の結果と併せて示す．

1．　まえがき

現在の日本では，Society 5.0 の実現に向け，IoT などにより収集・蓄積されたビッグデータを人工知能が解析し，ロボットを通じて現実世界にフィードバックすることで，様々な社会課題の解決を図っている．そのため，我々人間と共存する環境下で活動する様々な自律型ロボットが開発され続けている．これらのロボットは，周囲の環境を正確に認識することが期待されており．タスクを達成するための適切

＊:新居浜工業高等専門学校電子制御工学科 助教

搭載することが要求されている．

この要求に対し，H. Liu ら(1)は，物体を握っているときの画像及び圧力情報を利用する認識法を提案しており，J. Sinapov ら(2)や K. Noda ら(3)は，ロボットを動かして得た画像及び音響情報とロボットの関節角度情報を時系列データとして利用する認識法を提案している．これらの先行研究から，認識に用いるモダリティ数を増やすことで，対象物の固有の特徴を発見し，高い認識精度を保証していることが分かる．つまり，認識精度を向上させるためには，相補性を活用するために，認識に用いるモダリティ数を増やし，物体固有の特徴を抽出することが必要であると考えられる．

準備実験として，モダリティのうち，視覚及び触覚に関する情報を使用し，特徴抽出するシステムを開発し，その有効性を確認した[1]．本研究では，視覚，触覚，聴覚に関する情報をマルチモーダルセンサ情報として統合し，認識対象物の特徴抽出を可能とするシステムを開発する．視覚に関する情報として画像情報，触覚に関する情報として圧力情報，聴覚に関する情報として音響情報を使用し，これらの情報を，VRAE(Variational Recurrent Auto-Encoder)を利用したモデルによって統合及び低次元特徴抽出を行う．さらに，抽出された特徴量を入力とした MLP(Multi-Layer Perceptron)により物体認識を行うことにより，提案システムの有効性を確認する．

2．問題設定

本研究では，認識対象物の特徴抽出の方法を検討

※ 新居浜工業高等専門学校 電子制御工学科　助教

するため，図1に示すような問題を設定した．

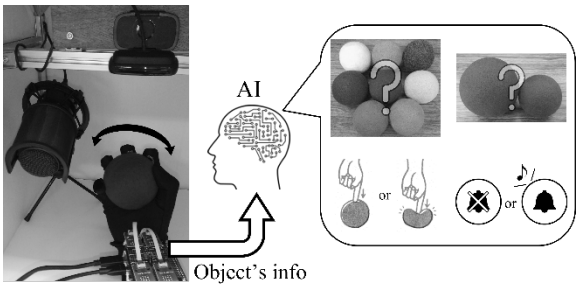
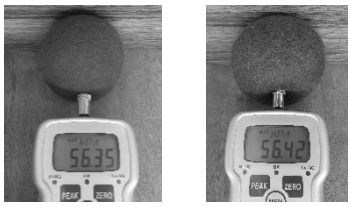


図1 本研究で取り組む課題

まず，認識対象物として，33種類のフェルトで作られた直径8cmのボールを用意した．この33種類には，8種類の色（赤・黄・橙・緑・水・青・桃・茶）や，図2に示すような2種類の柔らかさ，物体内に鈴の有無の違いを持つボール32種類と，直径6cmのボール1種（青，硬い，鈴無し）が含まれている．これらの物体に関して，カメラやマイク，触覚センサを取り付けた測定環境により，物体を握っている間の情報を取得し，人工知能を内包する提案システムに送る．そして，提案システムが，外観や柔らかさ，物体内部の構造に関する特徴が発見することが可能か否かについて検討する．



(a) Hard　　(b) Soft

図2 圧力をかけた際の凹み方の違い

また，本研究で提案するシステムは特徴抽出することを目的としている．本研究で行う特徴抽出とは，認識対象物に関する情報を，提案システムに入力した時，特徴を表現するような低次元の数値，これを特徴量として出力することを指す．さらに，特徴量をプロットした時の位置関係から，物体に関する様々な特徴の傾向が読み取れることを想定している．

3．提案手法

前章で設定した問題に取り組むために，本研究では図3に示すようなモデルを提案する．

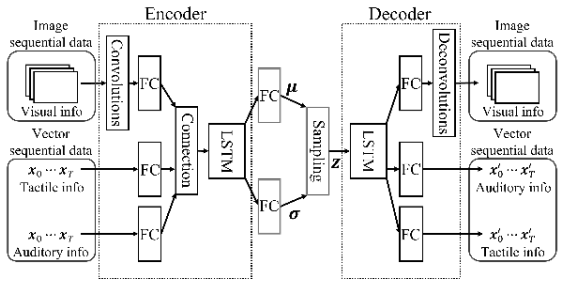


図3 提案するVRAEモデル

また，G. E. Hintonら[4]は，Deep Auto-EncoderがPCA(Principal Component Analysis)よりも優れた特徴圧縮法であると示していることを踏まえ，システム内では，VRAE[5, 6]を採用した．VRAEは，図4に示すように，時系列情報に含まれる特徴を抽出するように学習するVAE(Variational Auto-Encoder)である．

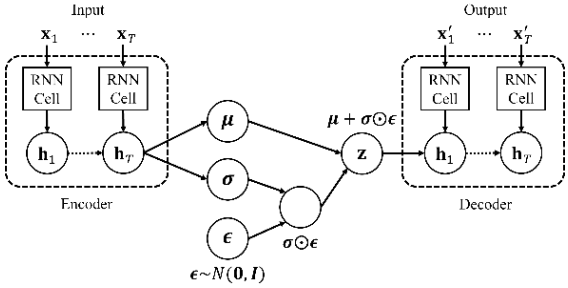


図4 VRAEの概要

本研究で使用する圧力・音響情報は過渡的な情報であり，時間依存性を考慮する必要がある．また，物体を振ることで中の鈴を鳴らすため，画像情報の中には振っているという情報を含める必要があると考えたため，時系列データを入力とした特徴抽出を行うシステムを提案する．

しかし，入力する各モダリティの情報を統合し，VRAEに入力するためにはベクトル化する必要がある．そこで，カメラによって連続的に撮影された画像の時系列データは，CNN(Convolutional Neural Network)を介してベクトル化を行い，触覚センサ及びマイクから得られたベクトルの時系列データは，全結合層を介してベクトル化する．そして，出力された各ベクトルを行方向に連結し，LSTM(Long-Short

Term Memory)層へ入力する．本研究では抽出した特徴を可視化するため，LSTM 層から平均 $\boldsymbol{\mu}$ と標準偏差 $\boldsymbol{\sigma}$ に対応する 2 つの全結合層へ入力する次元を，3 次元に低次元化し，これらの値を使用した正規分布から潜在変数 $\mathbf{z}$ をサンプリングする．また，サンプリングされた潜在変数 $\mathbf{z}$ を使用して，Decoder で再構成し，元のデータへ復元するように学習を行う．

また，この VRAE モデルの学習後，抽出した 3 次元特徴量を，分類モデルとして作成した MLP に入力し，33 種に多クラス分類することを目的に学習させ，その精度を検証した．

４．検証実験

４．１　実験準備

本研究では，各モダリティに関する情報を取得するため，ウェブカメラ，触覚センサ，コンデンサマイクを用意した．そして，物体の特徴を正確に記録するため，各センサを自作した測定環境に取り付けて使用し，これらのセンサを，Raspberry Pi 4 で制御することで各モダリティの情報を取得した．

４．１．１　画像及び音響情報の取得環境

画像情報を取得するために使用したウェブカメラは，画像サイズが 640×480 ピクセルの画像を，キャプチャ速度 30 fps で撮影可能である．また，音響情報を取得するコンデンサマイクは，サンプリング周波数が 48 kHz のものを使用した．これらの機器を使用して，図 5 に示すような測定環境を作成した．

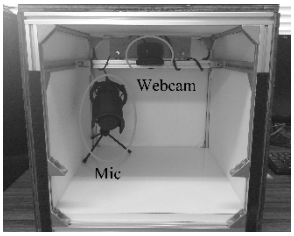


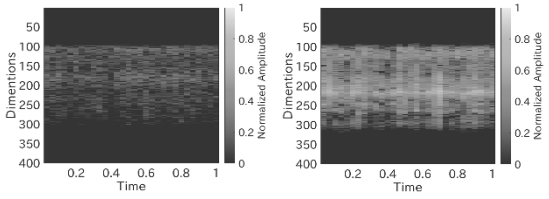
図 5 ウェブカメラとコンデンサマイクを設置した測定環境

この測定環境内で，物体を握っている間の物体の外観に関する情報及び把持する様子をウェブカメラで記録した．実際に，大きさの異なる 2 種類について取得した画像を図 6 に示す．また，VRAE モデルの計算の高速化のため，画像サイズを 10 分の 1 に縮小し，視覚に関する情報として VRAE モデルに入力した．

(a) 直径 8cm　　(b) 直径 6cm

図 6 実際に取得した画像情報

コンデンサマイクは，測定環境内で物体を一定の方向に小さく振った際の音を記録した．また，録音された音声は BPF(Band Pass Filter)により，ノイズ除去を行った．ここで使用した BPF は，通過域端周波数を 1500～4500Hz，阻止域端周波数を 500～6000Hz に設定している．そして，ノイズ除去された音声を，オーバーラップ率を 50%，窓関数をハニング窓に設定した短時間フーリエ変換を行った．短時間フーリエ変換を行う際，ウェブカメラのキャプチャ速度に合わせてスペクトログラムが出力されるように，1 フレーム 3200 サンプルに設定して変換を行った．さらに，用意した物体の中にある鈴の音の周波数が約 3000Hz であることは分かっていたので，出力スペクトログラムのうち 400 次元を抽出し，聴覚に関する情報として VRAE モデルに入力した．実際に，内部構造の異なる 2 種類について取得したスペクトログラムを図 7 に示す．



(a) 鈴無し　　(b) 鈴有り

図 7 実際に取得した音響情報

４．１．２　圧力情報の取得環境

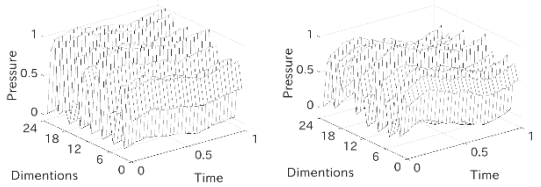
触覚センサは，タッチエンス社から提供されているショッカクチップを使用した．このセンサは，3 軸の圧力と各軸の力のモーメント (計 6 次元データ) を I^2C 通信により取得することが可能である．

この触覚センサを用いて，物体を握ることで圧力情報を取得する手袋型デバイス (図 8) を作成した．ここで，取り付けられている 4 つの触覚センサは，S.

Sundarametal ら[7]の検証結果を参考に，親指，人差し指，中指，掌の 4 箇所に配置することに決定した．

図 8 手袋型デバイス

この手袋型デバイスにより取得した 24 次元のデータを，触覚に関する情報として VRAE モデルに入力した．但し，ウェブカメラのキャプチャ速度に合わせるため，センサのサンプリング時間が 30sample/s となるように時系列データを取得した．また実際に，柔らかさの異なる 2 種類について取得した時系列データを図 9 に示す．



(a) 硬い　　(b) 柔らかい

図 9 実際に取得した圧力情報

4．2　使用データとネットワーク設定

前節で作成した測定環境を用いて，物体ごとに 1000 個の学習データを用意し，約 1 秒間の時系列データを VRAE モデルに入力した．学習後，VRAE モデルから出力された特徴量を入力とする分類用のモデルを新たに作成し，その精度を確認した．本研究では，　入出力層を含め，3 層で構成される MLP を分類モデルとして使用し，33 種類の多クラス分類を行った．ここで，隠れ層の次元は，1000 次元に設定した．但し，分類モデルに用いる学習データは，VRAE モデルに使用したものとは異なる 1000 個のデータを使用した．

4．3　実験：特徴量を用いた分類精度の検証

最適な VRAE モデルから得た特徴量を使用した際の分類モデルの学習曲線を図 10 に示す．ここで，使用するデータを学習用 70%，検証用 15%，テスト用 15%の割合で分割し，学習を行った．

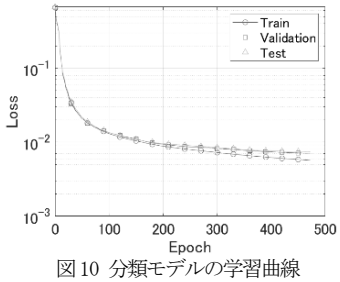
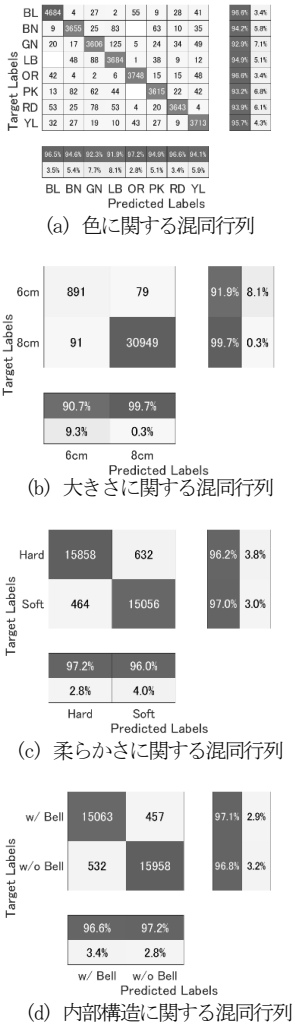


図 10 分類モデルの学習曲線

図 10 より，検証用データに対する誤差が大きくなる前に学習を終えたため，過学習することなく高い精度を獲得することができたと考えられる．

次に，各物体が持つ特徴ごとの正答回数を混同行列のヒートマップ図として図 11 に示す．



(a) 色に関する混同行列

(b) 大きさに関する混同行列

(c) 柔らかさに関する混同行列

(d) 内部構造に関する混同行列

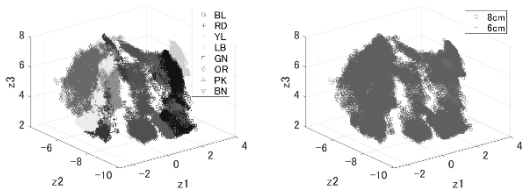
図 11 各物体が持つ特徴ごとの混同行列

図 11 の結果より，どの特徴の混同行列に関しても正答率は90%以上を示しているため，分類モデルは，

特徴量を基に分類ができるよう学習できていることが確認できた．しかし，図 11(b)の結果より，直径 6cm のボールに関する誤答が多いことが分かる．この結果の原因は，直径 6cm のボールのデータが 8cm のものに比べて，非常に少ないことが考えられる．

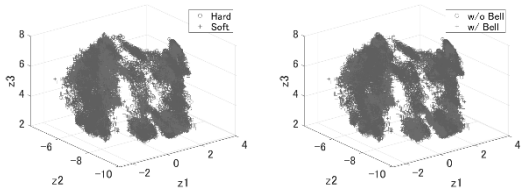
４．４　実験：最適な VRAE モデルを使用した 3 次元特徴量の抽出

最適な VRAE モデルより，図 12 に示すような 3 次元特徴量が出力された．これらの図は，3 次元空間プロットを各特徴に基づいて色分けしている．



(a) 色に関する潜在空間

(b) 大きさに関する潜在空間



(c) 柔らかさに関する潜在空間

(d)内部構造に関する潜在空間

図 12 最適な VRAE モデルから抽出された 3 次元特徴量

図 12(a)の結果より，物体の色が同じ特徴量は，潜在空間における位置関係が近くなっており，逆に色が異なる特徴量は位置が大きく離れていることが分かった．さらに，図中左側には暖色系の特徴量が集まっており，右側には寒色系の特徴量が集まっていることが確認できた．このことから，特徴の傾向まで読み取ることを可能にする特徴抽出ができていることが分かる．また，図 12(b)より，大きさに関する潜在空間においても，同様のことが言える．従って，外観に関する特徴抽出は，所望の結果を得ることができた．

しかし，図 12(c)及び(d)の結果より，柔らかさと内部構造に関する潜在空間では，外観に関するものと比較して，上手く抽出できておらず，位置関係から傾向を読み取ることは困難であることが分かった．

このような結果となった原因の一つとして，VRAE モデルによって抽出する次元を 3 次元に設定したことが考えられる．特徴量を 3 次元に設定することは，特徴を表現するための自由度を低下させていることになる．そのため，VRAE モデルによって抽出する次元を 3 次元よりも高く設定して学習を行い，別の方法で低次元化する必要があると考えた．

また，もう一つの原因として，色に関する特徴の複雑さが挙げられる．本研究では，大きさ，柔らかさ，内部構造に関する特徴は，2 種類に分けて抽出する必要があるが，色に関する特徴は，8 種類に分けて抽出する必要があり，他の特徴と比較して複雑な問題設定になっていることが分かる．

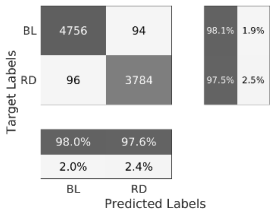
４．５　実験：改良モデルによる抽出及び精度検証

前節での考察を踏まえ，VRAE モデルによって抽出する次元を 80 次元に設定して学習を行い，80 次元の潜在変数を 3 次元に圧縮するため，確率的主成分分析にガウス過程を利用した GPLVM(Gaussian Process Latent Variable Models)を使用することを検討した．また，認識対象物の色を赤と青の 2 種類に限定し，前節までと同様に，分類精度を前節の実験結果と比較して検討した．

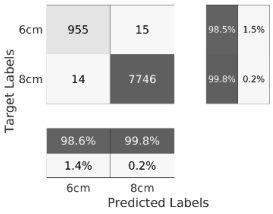
４．５．１　改良モデルを使用した分類精度の検証

本実験設定における最適モデルが獲得した最も高い精度は，94.7%であることを確認した．この結果を，前節の最適な VRAE モデルと比較すると，少し向上していることが分かった．従って，物体の各特徴の種類を揃え，抽出法を改善すると，精度向上が期待できると考えられる．加えて，各物体の特徴ごとの正答回数を混同行列のヒートマップ図として図 13 に示す．

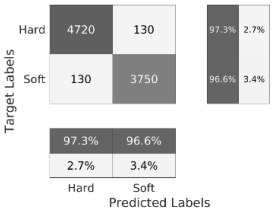
図 13 の結果より，分類モデルは入力された特徴量を基に，特徴ごとに分類するように学習できていることを確認した．また，図 13(b)に示すように，大きさに関する分類において，他の特徴の場合と比較して，誤答が非常に少ないことが分かる．これは，画像情報と圧力情報の両方に，大きさに関する特徴が現れたデータが含まれていることを示唆しており，この 2 つのモダリティを使用することで，大きさに関する特徴はより上手く抽出され，精度も向上することが考えられる．



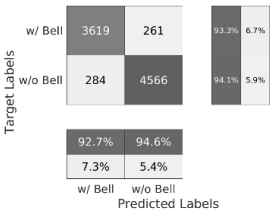
(a) 色に関する混同行列



(b) 大きさに関する混同行列



(c) 柔らかさに関する混同行列

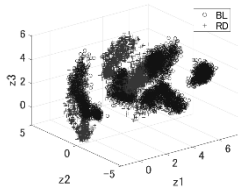


(d) 内部構造に関する混同行列

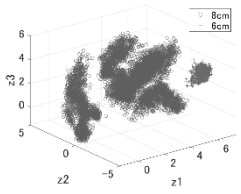
図 13 改良モデルを使用した際の各特徴の混同行列

4．5．2　改良モデルによる 3 次元特徴量の抽出

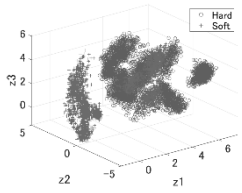
前款で発見したモデルを使用し，図 14 に示すような 3 次元特徴量が出力されたことを確認した．



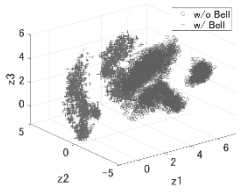
(a) 色に関する潜在空間



(b) 大きさに関する潜在空間



(c) 柔らかさに関する潜在空間



(d) 内部構造に関する潜在空間

図 14 改良モデルから得られた 3 次元特徴量

図 14 の結果より，各モダリティに関する情報に基づいた特徴抽出を行い，その 3 次元特徴量の位置関係から，特徴の傾向がある程度読み取れることが確認できる．これらの結果より，本研究で提案した VRAE モデルでは，使用するモダリティの情報に基づき，入力情報に含まれる特徴を特徴量として抽出することが可能であることが分かった．

4．6　実験：モダリティの組み合わせの比較

特徴抽出に必要なモダリティを検討するため，各モダリティの組み合わせを変更したモデルの精度の比較実験を行う．本実験で使用するモデルや使用データは，4．3及び4．4節で使用したものと同じものを使用する．また，各モデルから抽出された特徴量を使用した精度は，表 1 のようになった．

表 1 使用するモダリティの組み合わせによる精度比較

Input data			Accuracy
Tactile	Auditory	Vision	
○	-	-	32．8%
○	○	-	35．3%
○	-	○	81．9%
-	○	○	84．1%
○	○	○	93．8%

表 1 の結果より，モダリティを増やすことで精度が向上していることを確認し，3 つのモダリティを組み合わせたモデルが，最も高い精度を獲得していることが分かった．この結果より，3 つのモダリティを統合可能な VRAE モデルは，相補性を活用することで特徴抽出可能な手法であり，認識精度を向上させることができることが分かった．また，視覚に関する情報に，補助情報として触覚・聴覚情報を利用することで精度向上が期待できると考えられる．

4．7　実験：計算量を考慮した物体認識法の検討

実用的な物体認識システムとは，高い精度を保証するだけでなく，認識時の計算量も少ない方が望ましい．そのため，別のアプローチとして，計算量の多いニューラルネットを使用しない ViT(Vision Transformer)[8]に発想を得た認識法について検討した．ViT は，自然言語処理モデルである Transformer[9]のエンコーダ部分を利用しており，画像分類タスクに用いることにより，ImageNet 等で SoTA モデルと

同程度またはそれを上回る性能を達成したモデルである.

本実験では，低計算量でマルチモーダルセンサ情報を活用できる認識法として，図 15 に示すような Transformer モデルをベースにした認識モデルについて検討を行う．また，圧力情報は，時系列データとして利用することで画像情報と同じ次元数となり，ViT をベースにした手法を検討することが可能となる.

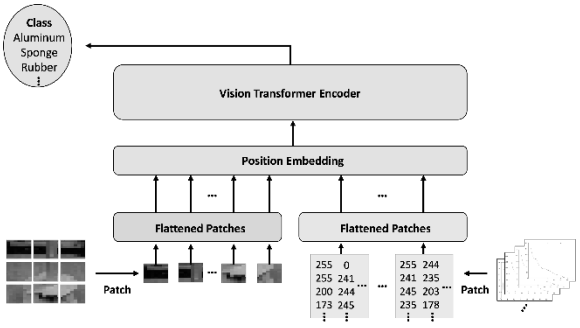


図 15 Transformer モデルに基づく認識モデルの概要

４．７．１　使用データとネットワーク設定

本実験では，53 種類の認識対象物のデータを含む学習データセット PHAC-2(Penn Haptic Adjective Corpus 2)[10]を使用する．PHAC-2 内では図 16 のような画像情報が用意されており，この画像は非常にデータ量が多い．計算の高速化のため，対象物を中心として 32×32 ピクセルに大きさを変更し，モデルに入力した.

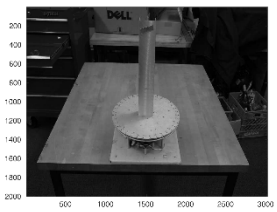


図 16 PHAC-2 内に用意されている画像情報（例）

また，圧力情報は Y.　Gao ら[11]の検証結果を参考に PCA を用いて，32sample/s となるようにデータを変更した．学習に使用する圧力情報のうち 4 種を図 17 に示す.

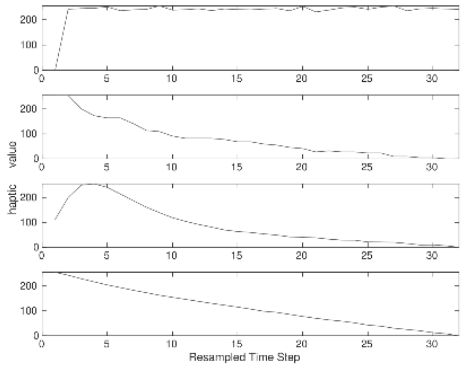


図 17 学習時に使用する圧力情報（例）

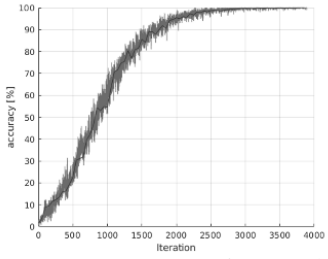
このモデルは表 2 に示すようなハイパーパラメータで作成し，学習を行った.

表 2 Transformer モデルのハイパーパラメータ

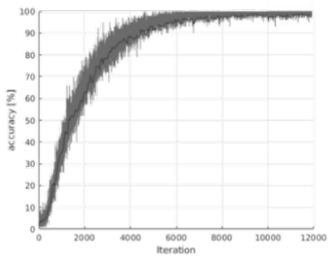
Parameter	Setting
Transformer layers	6
Transformer units	128
Projection dimensions	64
Activation function	GELU

４．７．２　PHAC-2 を使用した分類精度の検証

本実験の分類精度の検証は，パッチサイズを 8×8 及び 16×16 に設定した 2 種類について行った．学習回数と精度の関係を表したグラフを図 18 に示す.



(a) パッチサイズ 8×8 に設定した際の精度



(b) パッチサイズ 16×16 に設定した際の精度

図 18 Transformer モデルベースの認識モデルの認識精度

図 18 の結果より，学習するごとに精度は向上しており，最終的にパッチサイズが 8×8 の場合で 89.6%，16×16 の場合で 98.9%の精度を獲得したことを確認した．また，相補性を検討するため，入力情報にノイズを加え，再度同じ実験を行った．画像情報にノイズを加えると，精度は 91.0%，圧力情報にノイズを加えると，精度は 97.3%であることを確認し，所望のシステムを獲得できていることが分かった．

５．まとめ

本稿では，物体認識における対象物の特徴の抽出方法を検討するため，視覚，触覚，聴覚に関する情報を，マルチモーダルセンサ情報として統合し，特徴抽出する手法を，VRAE モデルを用いて提案した．また，MLP による多クラス分類の精度を指標としてモデルを評価し，モデルの有効性を検討した．自作した測定環境によりデータを取得し，提案システムの検証実験を行い，認識対象物の固有の特徴を捉えることが可能なセンサを使用することで，非常に高い分類精度を確認することができた[1,3]．

さらに，実験での考察を踏まえ，GPLVM をシステム内に用いることで，より上手く特徴抽出を行い，精度向上が期待できることが分かった．また，計算量を考慮した別のアプローチでは，計算時間を大幅に短縮し高い分類精度を獲得したことを確認した[4]．

今後の展望として，本研究で得た知見を融合させることでさらに低計算量でマルチモーダルセンサ情報を活用できる認識法について検討するとともに，実際のロボットに適用することで，計算時間及び精度を検証する必要がある．

謝辞

本研究の遂行に当たり、実験等に協力していただいた本校専攻科の林和輝氏、森田俊平氏に感謝申し上げます。また本研究は公益財団法人マツダ財団助成金により行われました。関係各位に心よりお礼を申し上げます。

口頭発表、受賞等

[1] 林和輝, 田中大介, VRAE モデルによる触覚・視覚情報の統合と物体認識への応用, 電気学会全国大会, 2021 年 3 月 11 日.

[2] 林和輝, 電気学会全国大会, 優秀論文発表賞.

[3] K. Hayashi. and D. Tanaka., Integration of Multimodal Sensor Information Using VRAE and Application to Object Recognition, RISP International Workshop on Nonlinear circuits, Communications and Signal Processing, Feb. 28, 2022.

[4] S. Morita. and D. Tanaka., Development of an Object Recognition Method Based on Transformer Architecture, RISP International Workshop on Nonlinear circuits, Communications and Signal Processing, Feb. 28, 2022.

参考文献

（1） H. Liu. et al., Visual-Tactile Fusion for Object Recognition, IEEE Transactions on Automation Science and Engineering, Vol.14, 2016, p.996-1008.

（2） J. Sinapov. et al., Grounding semantic categories in behavioral interactions: Experiments with 100 objects, Robotics and Autonomous System, Vol.62, 2014, p.632-645.

（3） K. Noda. et al., Multimodal integration learning of robot behavior using deep neural networks, Robotics and Autonomous System, Vol.62, 2014, p.721-736.

（4） G. E. Hinton. et al., Reducing the Dimensionality of Data with Neural Networks, Science, Vol.313, 2006, p.504-507.

（5） O. Fabius. et al., Variational Recurrent Auto-Encoders, arXiv:1412.6581, 2014

（6） Oriol Vinyals. et al., Generating Sentences from a Continuous Space, Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, 2016, p10-21.

（7） S. Sundarametal. et al., Learning the signatures of the human grasp using a scalable tactile glove, Nature, Vol.569, 2019, p.698-702.

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

（9） A. Vaswani. et al., Attention Is All You Need, arXiv, 2020.

（１０）V. Chu. et al., Robotic learning of haptic adjectives through physical interaction, Robotics and Autonomous Systems, 2015.

（１１）Y. Gao. et al., Deep learning for tactile understanding from visual and haptic data, Proceeding of 2016 IEEE International Conference on Robotics and Automation, 2016