

複合ウェアラブルセンサによる物体操作認識の検討

藤重 凱人^{1,a)} 北森 迪耶^{1,b)} 林 瀚^{1,c)} 下坂 正倫^{1,d)}

概要：日常のライフログの獲得は高齢者の体調評価や行動最適化の観点で重要であり，中でも物体操作中の手動作認識は詳細なライフログの獲得のため注目されている．従来の研究では，単一種類のセンサを用いた物体操作中の手動作認識が多く行われてきたが，獲得可能な情報の制限により認識性能に課題が存在する．本研究では，操作対象を認識するイメージセンサ，手指の形状，手の動きを認識する超音波センサ，腕の軌道，手の動きを認識する IMU センサの 3 種のセンサを統合し，物体操作に関する主要な情報を満たすことで，多様な物体操作の認識を可能とするシステムの提案と，本システムにおけるセンサ配置の影響の調査を行い，その研究の現況を共有する．提案システムのセンサ構成と，そのセンサ構成の組み合わせからなる他センサ構成で認識精度の比較を行い，物体操作中の手動作認識における提案するセンサ構成の優位性を示す．加えて，提案システムにおいて，設置箇所の依存度が特に大きいと考えられるイメージセンサに着目し，複数のイメージセンサ配置での提案システムの認識精度を比較することで，提案システムにおけるイメージセンサ配置と認識性能の関係を示す．

1. 序論

1.1 背景

ウェアラブルセンサを用いた物体操作認識は，従来，盛んに行われてきた全身動作の認識と比較して，詳細なライフログを獲得可能であるため，高齢者の体調評価やリハビリテーション，行動最適化への応用の観点から，近年，注目されている [9][5]．これは，全身を認識対象にするよりも細かな動作に着目できることや，手が日常において多くの物体を扱い，多様な役割を果たしていることに起因する．例えば，手で薬品を掴んで飲むことや，蓋を開けるなどの動作を認識し，正常に動作が行えているか評価することで，体調の客観的評価やリハビリテーションの効果測定に活用できる．

中でも手首に装着するウェアラブルセンサシステムが，ユーザの日常行動の阻害にならないことや，動作認識手に近く情報を獲得しやすいなどの利点から多く提案されている [8][15][11][10][1][13][6][16]．例えば，手首と頭部のイメージセンサ配置での物体操作の認識性能を比較する研究 [13] では，手首装着型の方が対象物を中心で大きく捉えられることで高精度での認識を達成することを示した．超

音波センサを手首に装着する研究 [8] では，手からの直接的な反射信号を獲得することができることで，スマートグラスへの装着する研究 [12] と比べて，より詳細な手動作に認識ができる．このように，手首装着型のセンサシステムは物体操作認識に適した構成である．

物体操作認識に取り組む研究では，単一種類のセンサを用いたアプローチが多く行われるが，獲得可能な情報の限界により，認識可能な動作の種類や，その認識性能に課題が存在する [8][15][7][13]．例えば，先行研究 [8] では，超音波センサを用いて物体操作の認識を行ったが，手形状を基に物体操作を認識するため，手の握り方が類似する物体の識別が困難であるという課題が存在した．先行研究 [15] では，IMU センサを用いた物体操作の認識に取り組んだが，手の動きと腕の軌道の限られた情報に基づく予測であり，類似する物体や手指で行う動作の認識に課題を抱える．この課題の解決のため，センサの統合により幅広い情報を獲得することで，認識できるタスクを拡大することが求められる．

従来の複合センサを用いたアプローチも存在するが，この多くは全身動作のような動きの大きな動作を対象とした ADL (Activities of Daily Living) の認識が中心であり，手形状のような細かな部位の認識には着目していない [11][10][1]．先行研究 [11] では，イメージセンサ，IMU センサ，マイクログフォンを複合した手首装着型センサシステムにより，歯を磨く，皿を洗うなどの 15 種類の ADL の認識に取り組んだが，魚に餌をやる，サプリメントを取る

¹ 東京科学大学 情報理工学院 情報工学系
School of Computing, Department of Computer Science, Institute of Science Tokyo

a) fujishige@miubiq.cs.titech.ac.jp

b) kitamori@miubiq.cs.titech.ac.jp

c) lin@miubiq.cs.titech.ac.jp

d) simosaka@miubiq.cs.titech.ac.jp

といった短時間の手の動きのみの動作の認識が低精度であった。このことから、先行研究で用いられた複合センサシステムは幅広い物体操作認識には最適なシステム構成であるとは言い難い。

ウェアブルセンサシステムにおけるセンサ配置は獲得情報に影響を及ぼすため、多くの研究において認識性能を高めるセンサ配置の探索が行われている [8][7]。[8] では超音波センサ、[7] はイメージセンサを手首のどの配置に装着すべきかの議論がなされた。このようにセンサは配置はウェアラブルセンサシステムにおいて重要な要素である。複合センサを用いたアプローチでは、各センサ間の情報の関係性も考慮することが求められ、この要素がより複雑化する。

以上のことから、本研究では、操作対象の認識に優れた RGB センサ、手指の形状、手の動きの認識に優れた超音波センサ、腕の軌道、手の動きの認識に優れた IMU センサの 3 種のセンサを統合し、物体操作に関する主要な情報を満たすことで、多様な物体操作の認識を可能とする手首装着型のウェアラブルセンサシステムを提案し、提案システムの RGB センサの配置と認識性能の関係を調査する。RGB センサの配置に着目する理由は、配置による隠蔽の変化が大きいことから、提案システムの認識性能への影響が明確であることが想定されるためである。これらを通じて、提案する複合ウェアラブルセンサシステムの物体操作認識性能と RGB センサの配置の影響を明らかにする。

本研究の貢献は以下の 3 点にまとめられる。

- 操作対象の認識に優れた RGB センサ、手指の形状の認識に優れた超音波センサ、腕の軌道の認識に優れた IMU センサを複合し、物体操作に関する主要な情報を満たすことで、多様な物体操作の認識を可能とするシステムを提案する。
- 提案システムのセンサ構成が、提案手法のセンサの組み合わせから成る、各センサ構成と認識精度の比較を行うことで、センサ複合により認識性能が向上することを示す。
- 提案システムにおいて、複数の RGB センサ配置での認識精度を比較することで、提案システムにおける RGB センサの配置と認識性能の関係を示す。

2. 関連研究

2.1 単一種類のウェアラブルセンサを利用した物体操作認識

近年、高齢者支援やリハビリテーションなどの領域での応用を目的として物体操作認識に関する研究が進められており、特に単一種類のセンサを用いるアプローチが多数検討されてきた。これは、単一種類のセンサでのアプローチが、ウェアラブルシステムの重要な要素である消費電力の

軽減や、システム小型化に優れる傾向にあるためである。例えば、超音波センサを用いた [8] では一般的なスマートウォッチの 300mAh バッテリーで 19 時間の連続使用が可能で、手首から厚さ 5mm の小型なリストバンド型システムを実現している。

しかし、単一種類のセンサを用いるアプローチでは、認識可能な動作の種類やその認識性能に課題が存在する [7][6][8][2][3][14][15]。例えば、イメージセンサを用いる [7][6] では、視野の制限や、手と物体が互いに障害物になることで認識精度の低下を引き起こす場合がある。また、超音波センサを用いた [8] では、握り方の似た物体や手形状動作の似た動作の区別が難しい。このように、各センサごとに認識を苦手とする動作が存在する。

さらに、これまでの研究においては「物体を持つ動作」や「握った手形状で腕で扱う動作」が操作が主な対象であり、手指による細かなインタラクションの評価は十分に行われていない。例えば、イメージセンサを用いる [6] では、スマートフォン、本、リング、マーカーの 4 種類の物体から、どの物体をどう持つかに着目しており、その物体に対する動作に着目していない。高齢者支援やリハビリテーションへ応用において、より正確なライフログの獲得が求められる。そのため、上記のようなセンサを組み合わせることで互いに情報を補完し、より幅広くかつ細かな物体操作を認識可能とすることを旨とする。

2.2 複合ウェアラブルセンサを用いた ADL 認識

複合ウェアラブルセンサを使用した動作認識のアプローチは、ADL の認識に関する研究で多く行われてきた [11][10][1]。先行研究 [11] では、イメージセンサ、IMU センサ、マイクロフォンを複合した手首装着型センサシステムが歯を磨くや皿を洗うなどの 15 種類の ADL の認識に取り組んだ。先行研究 [1] では、汎用スマートウォッチに搭載された IMU センサーとマイクロフォンを利用して、ADL の認識を試みており、IMU センサーが動作のリズムや方向を捉え、音響センサーが周辺環境音を捉えることで、23 種類の動作分類を行った。

しかし、これらのシステムでは、魚に餌をやる、サプリメントの摂取や、スマートフォンを用いたタイピングとブラウジングのような動作の誤認識が多く存在しており、細かな手指の形状の違いが認識ができていない。このように、先行研究における ADL 認識の複合センサの構成はあくまで全身動作のような動きの大きな動作を得意としており、細かな違いの識別を求められる手の動作認識には最適な構成ではない。このことから、手の動作認識に適した複合センサ構成を提案することが求められる。

3. 物体操作認識の問題設定とその課題

本章では、物体操作認識の一般的な問題設定を数式で定

義し、その課題を議論する。

3.1 物体操作認識の問題設定

物体操作認識の問題は、複数のセンサデータを入力とし、動作クラスを分類する問題として定式化できる。ここで、入力データは時系列データとして与えられ、特徴量抽出およびモデルの処理を通じて最終的に動作クラスに分類される。 $n = 1, \dots, N$ を試行のインデックス、 $t = 1, \dots, T_n$ を時間空間のインデックス、 $s = 1, \dots, S$ をセンサのインデックスとする。時間空間のインデックス t におけるセンサ s の観測データを

$$\mathbf{x}_{n,t}^{(s)} \in \mathbb{R}^{d_s}$$

とし、センサ s における試行 n の時系列データを

$$\mathbf{X}_n^{(s)} = [\mathbf{x}_{n,1}^{(s)}, \mathbf{x}_{n,2}^{(s)}, \dots, \mathbf{x}_{n,T_n}^{(s)}] \in \mathbb{R}^{T_n \times d_s}$$

と定義する。ここで、 d_s はセンサ s の観測データが持つ次元数を表す。さらに、 S 個のセンサデータを統合した入力データを

$$\mathbf{X}_n = \{\mathbf{X}_n^{(1)}, \mathbf{X}_n^{(2)}, \dots, \mathbf{X}_n^{(S)}\}$$

とする。この入力データに対して、各センサ固有の特徴量抽出関数 $\phi^{(s)}$ を適用すると、センサ s ごとの特徴量ベクトル

$$\mathbf{z}_n^{(s)} = \phi^{(s)}(\mathbf{X}_n^{(s)}) \in \mathbb{R}^{p_s}$$

を得られる。ここで、 p_s はセンサ s に対応する特徴量次元数を表す。全センサの特徴量を連結して統合したベクトルを

$$\mathbf{z}_n = [\mathbf{z}_n^{(1)}, \mathbf{z}_n^{(2)}, \dots, \mathbf{z}_n^{(S)}] \in \mathbb{R}^P$$

と定義する。ここで P は全センサの特徴量次元数の和であり、 $P = \sum_{s=1}^S p_s$ で表される。最終的に、分類モデル f を用いて、特徴量ベクトル $\mathbf{z}_n$ を動作クラス $y_n \in \{1, \dots, C\}$ に分類する。この過程を次式で表す。

$$\hat{y}_n = f(\mathbf{z}_n) = f([\mathbf{z}_n^{(1)}, \mathbf{z}_n^{(2)}, \dots, \mathbf{z}_n^{(S)}]) \quad (1)$$

3.2 単一種類のセンサによる物体操作認識の課題

従来研究ではセンサ $M = 1$ の場合が一般的であり、単一センサデータに基づく特徴量抽出関数 $\phi^{(m)}$ と分類モデル f の工夫によって、物体操作認識の精度向上を行ってきたが、単一種類のセンサでは獲得可能な情報が限定的で、認識できないタスクが各センサに存在するという課題がある。手形状認識を実現するセンサシステムは多く存在するが、物体とのインタラクションを含んだ場合の認識は単一センサから得られる情報のみからは認識が難しいことが想定される [8][18][4]。例えば、超音波センサでは、手形状の認識に優れており、高精度の認識を実現するが、握り方の似た物体の識別は困難である。 [8][18]。

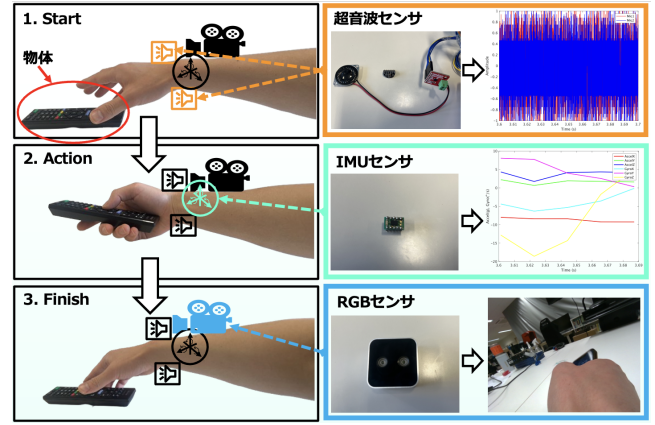


図 1: 提案システム構成図

3.3 複合センサによる物体操作認識の課題

本研究では、センサ数 $M = 3$ に設定し、イメージセンサ、超音波センサ、IMU センサを組み合わせることで、認識可能な動作の幅や精度向上を目指す。しかし、各センサが取得する情報が相互に影響を及ぼすため、それぞれのセンサの配置による認識への影響を理解することが課題として挙げられる。本研究では、この第一段階として、イメージセンサの配置による物体操作認識への影響を調査に取り組む。イメージセンサに着目する動機は、配置による隠蔽の変化が大きいため、提案システムの認識性能への影響が明確であることが想定されるためである。

4. 提案手法: 複合センサシステムによる物体操作認識

4.1 複合センサシステムにおけるセンサの選定指針

複合センサシステムに用いる各センサの選定指針として、物体操作認識における主要な情報と各センサの関係を示す。

4.1.1 物体認識における主要な情報

物体操作における主要な情報である、操作対象物体、手指の形状、手の動き、腕の軌道の 4 種類の情報を獲得することが、物体操作の認識において重要である。この 4 種類の情報の獲得により、スマートフォンへのタップとリモコンのボタン操作といった、操作対象物体だけが異なる動作や、スマートフォンでのタップとスクロールのような、手指の形状や手の動きだけが異なる動作、スマートフォンを持つと、通話するといった、同じ操作対象物体、類似する手指の形状で、腕の軌道だけが大きく異なる動作の識別が可能になるためである。以上のことから、操作対象物体、手指の形状、手の動き、腕の軌道の 4 種類の情報を獲得することで、より幅広く細かな物体操作の認識が実現できると考えられる。

4.1.2 物体認識における主要な情報と各センサの関係

本研究では、イメージセンサ、超音波センサ、IMU センサを提案する複合センサデバイスに採用する。これは、先

行研究で用いられたセンサを比較した結果、以下の特性からイメージセンサ、超音波センサ、IMU センサを組み合わせることで物体認識における主要な情報を網羅できると考えたためである。具体的には、イメージセンサは操作対象物体の検出に優れており、超音波センサは指の形状や手の動きを高精度で認識できる [8][18]。IMU センサは着用位置の動き（本研究では腕）を詳細に計測可能で、また手の動きの認識もできる [17][15]。このことから、イメージセンサ、超音波センサ、IMU センサの複合センサシステムは、高精度な物体操作の認識を実現できると考えられる。

4.2 複合センサシステムの設計

本研究では、図 1 の構成図に示される複合センサシステムを用いて実験を行う。各センサは Raspberry Pi4 を用いて制御を行なった。超音波センサ（スピーカー：CMS-28528N-L152、マイクロフォン：MSM261S4030H0、アンプ：MAX98357A）は先行研究 [8] で高い精度での手形状と物体操作の認識を実現した配置を利用し、手首の両面に設置した。IMU センサ（BNO055）は高精度で腕の軌道の認識を実現した先行研究 [17] と同様の手首背面に設置を行う。イメージセンサは最も一般的な RGB センサ（RealSense D405）は利用し、これを手首の背面の中心に設置する。この配置は、一般に普及するスマートウォッチと同様に手首背面が大きくなる設計であり、日常利用時の装着感や快適性を考慮している。

4.3 物体操作認識におけるモデルの構成

本研究で用いる、分類モデルの入力にあたる特徴量と分類モデルの構成について示す。

4.3.1 特徴量

RGB データの特徴量は、ImageNet データセットで事前学習された、MobileNet モデルを用いて獲得する。MobileNet は、計算効率に優れた軽量な畳み込みニューラルネットワークであり、エッジデバイスやリアルタイム処理に適している。本研究では、MobileNet の出力の特徴量 $\mathbf{z}_n^{(s)} \in \mathbb{R}^{1024}$ を使用して、RGB データから視覚的特徴を抽出する。

超音波データについては、スピーカーから 19kHz~22kHz の C-FMCW チャープ信号を出力し、その反射信号をマイクロフォンで取得する。この信号は、反射対象の位置情報を含む。本研究では、取得した信号に対して STFT（Short-Time Fourier Transform）を適用し、時間周波数領域の特徴量 $\mathbf{z}_n^{(s)} \in \mathbb{R}^{500}$ を抽出する。その後、2つのマイクロフォンの特徴量を結合した $\mathbf{z}_n^{(s)} \in \mathbb{R}^{1000}$ をモデルの入力とする。これにより、対象物や動作による反射特性の変化をモデルに学習させる。

IMU データについては、加速度、および角速度の生データを直接入力特徴量として利用する。そのため、 $\phi^{(s)} = 1$ であり、 $\mathbf{z}_n^{(s)} \in \mathbb{R}^6$ がモデルの入力となる。これらのデー

タにより、動作に伴う動きや姿勢変化をリアルタイムで捉えることが可能である。

4.3.2 分類モデル

本研究では分類モデル f として、マルチブランチ LSTM を採用する。このモデルは、各センサデータ特徴量ベクトル $\mathbf{z}_n^{(s)}$ の時系列情報を効果的に処理できること、異種のセンサデータ $s = 1, \dots, S$ 間の統合を容易に行えること、 $s = 1$ や $s = 2$ のセンサ構成との公平な比較が可能であることから採用した。また、各センサデータを個別に処理するブランチ構造を持ち、それぞれの特徴を学習後に統合することで、複合センサの認識精度を高めることが期待される。

4.4 複合センサシステムにおける懸念点

複合センサシステムにおける動作認識においては、各センサから得られるデータを適切に処理しなければ、期待される精度向上が得られないだけでなく、認識精度が低下する可能性がある。先行研究 [11] では日常での行動認識にイメージセンサ、マイクロフォン、照度計、デジタルコンパスを用いて精度の比較を行ったが、照度計やデジタルコンパスが追加されることで認識精度が低下する様子が確認された。このことから、特定のセンサが他のセンサと干渉し、ノイズとなる場合があることが示される。そのため、各センサデータを適切に処理や統合することが重要である。

5. 実験

本章では以下の 2 つの実験について示す。

- 実験 1: 複合センサシステムにおける認識性能
- 実験 2: 複合センサシステムにおける RGB センサの配置と精度の関係性

5.1 目的

本実験の目的は、提案する複合センサシステムを用いて認識性能が向上することを確認し、提案した複合センサシステムにおいて、RGB センサの配置が物体操作の認識精度に与える影響を明らかにすることである。

5.2 評価対象動作

物体操作の主要な情報である、操作対象物体・手指の形状・手の動き・腕の軌道の 4 種類の情報が獲得可能を評価するために、図 2 に示す評価対象動作を選択した。4 種類の主要な情報の内、操作対象物体のみが異なる動作である、図 2(d) スマートフォンを持つと図 2(g) リモコンを持つにおける互いの誤認識率から、操作対象物体の情報を獲得できているかを確認する。手指形状と手の動きのみが異なる動作である、図 2(a) スマートフォンの電源操作、図 2(b) スクロール、図 2(c) タップの動作における互いの分類結果から、手指形状や手の動きの違いを捉えられるかを確認する。さらに、腕の軌道のみが異なる例として、図 2(d) スマート

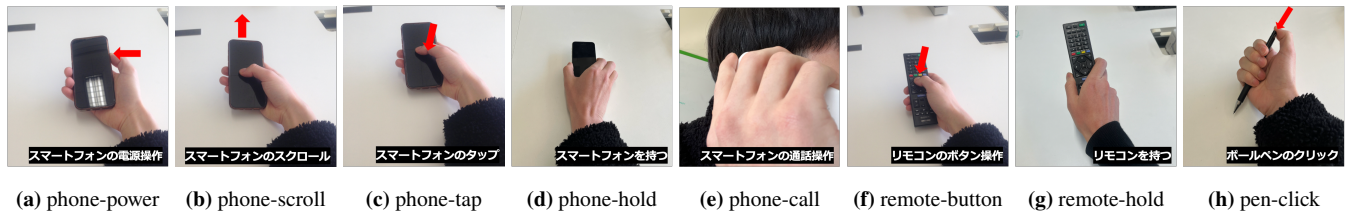


図 2: 評価対象動作

フォンを持つ動作と図 2(e) 通話動作を区別できるか検証することで、腕の軌道の認識性能も評価できる。以上より、これらの動作を総合的に高い精度で分類できれば、物体操作の主要な情報を広く獲得できていると判断でき、より幅広い物体操作の認識が可能であると言える。

5.3 データ収集

作成した複合センサシステムを被験者 1 名に装着し、各評価対象動作を行うことで実験用のデータを収集した。各動作は、机上の初期位置に置かれた物体に手を伸ばして掴むことから始め、指定の動作を行い、物体を元の位置に戻し、手をスタート位置に戻すまでを 1 回と定義する。この様子の一部は、図 1 に示される。

この一連の動作を 5 秒以内で完了させ、12 動作を連続で行うことで 1 セットとした。評価対象の各動作について 18 セットを収集し、2:1 の比率で学習データとテストデータに分割して使用した。各センサのサンプリング条件を以下に示す。イメージセンサは 10.5Hz、IMU センサは 46Hz でサンプリングを実施した。超音波センサは、周期 0.012 秒の 19 kHz ~ 22 kHz の C-FMCW チャープ信号を出力信号とし、標準化定理に従い、出力信号の周波数を獲得可能な 48kHz でマイクロフォンによるサンプリングを行った。

5.4 性能の評価基準

本研究の性能評価指標として、Accuracy を用いた。Accuracy は、モデルがテストデータ内のサンプルを正しく分類した割合を示す指標であり、以下の式で定義される。

$$\text{Accuracy} = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}, \quad (2)$$

ここで、 T_p は正例を正しく分類した数、 T_n は負例を正しく分類した数、 F_p は負例を誤って正例として分類した数、 F_n は正例を誤って負例として分類した数をそれぞれ表す。加えて、本研究では詳細な誤認識のパターンを把握するため、混同行列の利用も行う。これにより、各クラスの予測結果と実際のラベルの対応関係を視覚的に示すことで、誤認識の傾向を分析することが可能である。

5.5 実験 1: 複合センサシステムにおける認識性能

5.5.1 比較対象のセンサ構成

比較対象のセンサ構成として、作成した複合センサシ

テムを用いて収集したデータを基に、超音波 + IMU、超音波 + RGB、IMU などの 6 種類のセンサ構成を設定した。これらのセンサ構成に対し、上記の特徴量とモデルを用いて分類を行い、各構成の分類結果を比較に用いた。

5.5.2 実験結果

表 1 に提案手法および比較手法の各センサ構成における 8 クラス分類の分類精度を、図 3 に混同行列を示す。表 1 に示されるように、提案する複合センサシステムは超音波 + IMU の次に高い、分類精度を達成した。これは、センサ複合により、各センサから得られる情報が互いに干渉し、重要な情報にノイズを生んでしまったためと推測される。具体的には、IMU+RGB の精度が IMU 単体の場合と比較して低下していることから、RGB データが、IMU データとの複合で適切に作用してせず、互いの得意とする情報にノイズを生んでいると考えられる。この結果から、RGB を主とする各センサのデータ処理の工夫や、獲得情報の性質変化のために配置の工夫が必要になることが示される。

また、図 3 の混同行列から、提案手法は、比較対象のセンサ構成と比較して、操作対象物体に特徴のある動作の誤分類が抑制されていることが見て取れる。高い分類精度を示した超音波 + IMU、超音波 + RGB のセンサ構成では、手の動作が共通もしくは類似し、操作対象物体の異なる、スマートフォンの電源操作、スクロール、タップの 3 つの動作とリモコンのボタン操作間の誤分類が抑制されている。同様の傾向がスマートフォンの電源操作とリモコンの電源操作にも見られ、2 種類の複合から 3 種類の複合になることで、総合的な認識精度は低下する場合もあるものの、物体操作の主要な情報を広く利用できていることが示された。

表 1: 各センサ構成における認識精度

センサ構成	Accuracy
1. 超音波 + IMU + RGB	0.86542
2. 超音波 + IMU	0.87222
3. 超音波 + RGB	0.85634
4. IMU + RGB	0.50093
5. 超音波	0.68478
6. IMU	0.60370
7. RGB	0.10261

Confusion Matrix								
Estimate label	phone-power	phone-scroll	phone-tap	phone-hold	phone-call	remote-button	remote-hold	pen-click
phone-power	62	2	0	0	1	0	0	0
phone-scroll	0	48	0	0	0	0	0	0
phone-tap	9	16	64	0	0	0	0	0
phone-hold	0	0	0	55	0	0	18	0
phone-call	0	0	0	0	70	0	0	0
remote-button	0	2	1	3	0	66	1	0
remote-hold	0	0	0	14	0	2	41	0
pen-click	0	0	1	0	0	0	0	59
Ground truth	phone-power	phone-scroll	phone-tap	phone-hold	phone-call	remote-button	remote-hold	pen-click

(a) 超音波 + IMU + RGB

Confusion Matrix								
Estimate label	phone-power	phone-scroll	phone-tap	phone-hold	phone-call	remote-button	remote-hold	pen-click
phone-power	66	2	0	0	0	1	0	0
phone-scroll	2	58	1	0	0	0	0	0
phone-tap	0	3	57	0	0	0	0	0
phone-hold	0	0	0	64	0	0	22	0
phone-call	0	0	0	0	71	0	0	3
remote-button	4	7	8	0	0	67	5	0
remote-hold	0	0	1	8	0	0	31	0
pen-click	0	0	0	0	0	0	2	57
Ground truth	phone-power	phone-scroll	phone-tap	phone-hold	phone-call	remote-button	remote-hold	pen-click

(b) 超音波 + IMU

Confusion Matrix								
Estimate label	phone-power	phone-scroll	phone-tap	phone-hold	phone-call	remote-button	remote-hold	pen-click
phone-power	60	1	1	0	0	0	0	0
phone-scroll	0	55	0	0	0	0	0	0
phone-tap	9	4	62	0	2	1	0	0
phone-hold	0	0	0	49	0	1	22	0
phone-call	1	0	1	0	69	0	0	0
remote-button	1	8	2	5	0	67	0	0
remote-hold	0	0	0	18	0	0	38	0
pen-click	0	0	0	0	0	0	0	59
Ground truth	phone-power	phone-scroll	phone-tap	phone-hold	phone-call	remote-button	remote-hold	pen-click

(c) 超音波 + RGB

Confusion Matrix								
Estimate label	phone-power	phone-scroll	phone-tap	phone-hold	phone-call	remote-button	remote-hold	pen-click
phone-power	9	0	0	0	0	0	0	1
phone-scroll	0	36	17	0	0	2	0	5
phone-tap	12	20	47	1	1	7	0	7
phone-hold	0	0	0	6	0	0	9	0
phone-call	18	4	0	0	67	2	0	2
remote-button	32	1	1	6	2	30	3	15
remote-hold	0	1	1	59	1	18	45	1
pen-click	0	6	0	0	0	9	3	28
Ground truth	phone-power	phone-scroll	phone-tap	phone-hold	phone-call	remote-button	remote-hold	pen-click

(d) IMU + RGB

Confusion Matrix								
Estimate label	phone-power	phone-scroll	phone-tap	phone-hold	phone-call	remote-button	remote-hold	pen-click
phone-power	33	2	1	0	0	3	0	0
phone-scroll	13	37	1	0	0	0	0	0
phone-tap	24	27	63	0	6	2	0	0
phone-hold	0	0	0	23	0	6	17	0
phone-call	1	1	0	0	63	0	0	0
remote-button	1	4	7	8	1	59	0	0
remote-hold	0	1	0	41	1	2	42	2
pen-click	0	0	0	0	1	0	1	58
Ground truth	phone-power	phone-scroll	phone-tap	phone-hold	phone-call	remote-button	remote-hold	pen-click

(e) 超音波

Confusion Matrix								
Estimate label	phone-power	phone-scroll	phone-tap	phone-hold	phone-call	remote-button	remote-hold	pen-click
phone-power	49	12	0	0	3	1	0	5
phone-scroll	8	39	3	0	2	5	0	7
phone-tap	0	7	54	1	0	10	8	5
phone-hold	0	0	0	28	0	4	20	0
phone-call	1	0	0	0	59	0	0	0
remote-button	2	4	8	0	0	27	2	3
remote-hold	0	0	1	43	0	16	30	0
pen-click	12	8	1	0	7	5	0	40
Ground truth	phone-power	phone-scroll	phone-tap	phone-hold	phone-call	remote-button	remote-hold	pen-click

(f) IMU

Confusion Matrix								
Estimate label	phone-power	phone-scroll	phone-tap	phone-hold	phone-call	remote-button	remote-hold	pen-click
phone-power	9	20	23	31	19	8	5	15
phone-scroll	1	6	2	0	0	0	0	3
phone-tap	20	16	17	12	19	5	12	17
phone-hold	0	0	2	1	6	3	0	0
phone-call	3	0	11	0	0	12	0	0
remote-button	21	1	1	0	1	11	11	2
remote-hold	3	15	5	20	22	1	1	12
pen-click	14	10	5	8	4	29	31	10
Ground truth	phone-power	phone-scroll	phone-tap	phone-hold	phone-call	remote-button	remote-hold	pen-click

(g) RGB

図 3: 各センサ構成における分類結果



図 4: 比較する RGB センサの配置

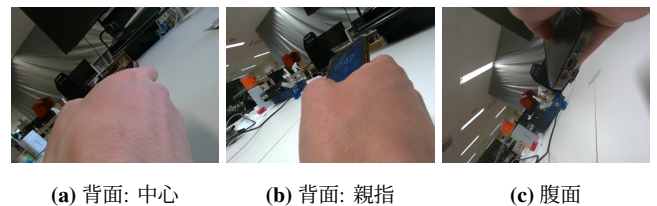


図 5: 各 RGB センサの配置における生データ

5.6 実験2：複合センサシステムにおける RGB センサの配置と精度の関係性

5.6.1 評価対象の RGB センサの配置

比較する RGB センサの設置箇所として、図 4 に示す 3 種類の配置を選択した。図 4(a) の配置は、実験 1 で用いた手首背面の中心に RGB センサを設置する構成であり、スマートウォッチなど、日常生活で使用される装着方式に近いものである。図 4(b) の配置は、手首背面の親指の付け根付近に RGB センサを設置する構成である。この配置は、多くの物体操作で親指が関与する点に着目しており、図 4(a) の配置で想定される、手の甲が視野を遮る問題を軽減できると考えられる。

最後に、図 4(c) の配置は、手首の腹面に RGB センサを設置する構成である。この配置では、物体との間に障害物入りやすく対象物を長期間捉えることができると考えられる。しかし、日常動作においては装着が阻害要因となる可能性がある。これらの配置における物体操作の認識精度を比較することで、提案システムにおける、RGB センサの配置と精度の関係性を評価した。

5.6.2 RGB センサの配置における物体情報取得率の関係

図 5 は、異なる RGB センサの配置における物体写り込みの例を示している。3 つの画像は共通して図 2(c) における、対象物に対して親指でタップ動作を行う際のものである。図 5(a) 背面：中心の場合、手の甲が視界を大きく遮っており、物体の写り込みがほとんどない。

一方で、図 5(b) 背面：親指、および図 5(c) 腹面の場合、隠蔽が軽減され、物体の写り込みが多くなっている。この結果は、センサの配置位置が物体情報の取得性能に直接的な影響を与えることを示している。特に図 5(a) 背面：中心の配置は手の甲による隠蔽の影響が大きいため、実験 1 における低精度やセンサ複合の分類精度の低下の一因になっていると考えられる。

5.6.3 実験結果

表 2 に、図 4 に示す 3 種類の RGB センサ配置における複合センサシステムの 8 クラス分類の分類精度を示し、図 6 にその際の混同行列を示す。この表には、各 RGB センサ配置における、実験 1 で用いた比較センサ構成の場合の分類精度も示す。これは、RGB センサの影響評価を各配置で平等に評価する目的で追加した。表 2 において、RGB センサのデータを用いない、共通の配置のセンサのみの構成での分類精度が異なるが、これはカメラ設置箇所を変更したことによる、動作への影響や、セッション依存性によるものであると考えられる。

実験の結果、各 RGB センサの配置における RGB センサ単体での認識精度は図 4(c)、図 4(b)、図 4(a) の順に高くなった。これは、図 5 に示された、物体の認識に障害物が生じない順であり、獲得情報が影響することが示される。ただし、この認識精度の差はそれぞれ 0.01 程度と非常に小さ

く、この結果だけで配置 4(c) が優れていると断言することは難しい。

そこで、図 6 の混同行列を用いて、配置による誤分類の傾向を分析した。図 4(a) では、手指の形状と腕の軌道が一致する、スマートフォンを持つ動作とリモコンを持つ動作において多くの誤認識が見られた。一方、図 4(b) および図 4(c) では、この誤認識が大きく抑制されており、物体情報の取得においてこれらの配置が優れていることが示された。

また、提案する 3 種類の複合センサ構成にすることで、共通の RGB センサ配置における、他のセンサ構成と比較して最も精度が向上したのは、図 4(c) の配置であった。この RGB センサ配置では、センサの種類を増やすことで認識精度が向上したこのことから、他センサデータへのノイズの影響が最も小さくなる RGB センサデータが獲得できる配置であることが示される。以上の分析から、提案システムにおける RGB センサの配置として、図 4(c) が最適であると結論づけられる。

表 2: 各 RGB センサ配置における認識精度

センサ構成	Accuracy		
	配置 a	配置 b	配置 c
1. 超音波 + IMU + RGB	0.86916	0.93273	0.89821
2. 超音波 + IMU	0.87222	0.94316	0.89007
3. 超音波 + RGB	0.85634	0.95423	0.86964
4. IMU + RGB	0.50093	0.60498	0.53929
5. 超音波	0.68478	0.88889	0.81250
6. IMU	0.60370	0.61101	0.59929
7. RGB	0.10261	0.13181	0.13929

6. 結論

本研究では、日常における様々な物体操作を認識可能な RGB センサ、超音波センサ、IMU センサの複合センサシステムの提案と、提案システムにおける RGB センサの配置と認識精度の関係について調査を行った。実験結果から、図 4(c) の RGB センサの配置の場合において、提案センサシステムは比較対象のセンサ構成と比較して最も高い、0.89821 の認識精度を達成することが示された。この結果から、図 4(c) で示す、RGB センサの配置が、本研究で比較した提案システムの RGB センサの配置の内、物体操作認識に最も優れることが明らかになった。

しかし、その他の RGB センサの図 4(a)、図 4(b) の配置においては、比較対象のセンサ構成の超音波 + IMU や超音波 + RGB より、劣った認識性能を示した。この原因として、RGB センサと IMU センサが互いに獲得に優れる、手の動き、腕の軌道と操作対象物体の情報に対して、RGB センサであれば手の動き、腕の軌道、IMU センサであれば操作対象物体に当たる獲得を苦手とする情報をもたらすこと

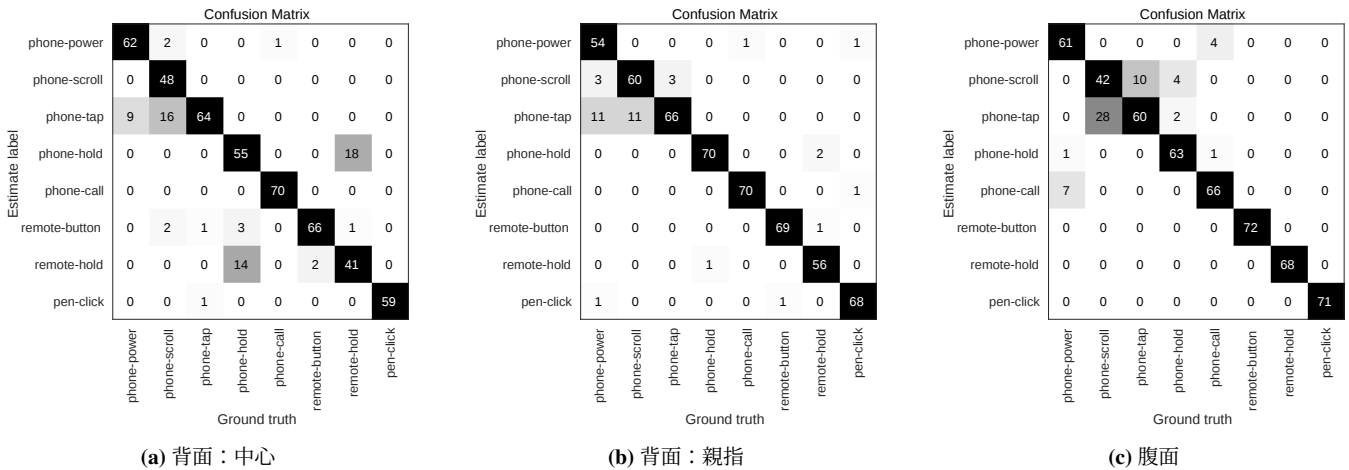


図 6: 各 RGB センサの配置における提案システムでの分類結果

で、逆にノイズを生んでしまっていることが考えられる。そのため、RGB センサに物体認識を追加し、判定した操作対象物体のラベルのみ学習に扱うといった、各センサが役割に特化した情報のみ獲得するようにすることで、他センサに負の影響を与える情報を利用せず、高精度で獲得できる情報のみを利用するアプローチが、提案する複合センサシステムでより適切に情報を活用するために求められる。

また、本研究では比較対象が構成要素のより少ない種類のセンサ構成だけであり、提案システムの価値を示すために十分でない。そのため、多くのセンサを利用する構成や複雑なモデルを活用する構成との比較が必要である。加えて、ベンチマークについても類似する動作の数が2、3個と少なく、より正確な性能評価のためにより多くの動作に対する分類が求められる。さらに、被験者も1名であり、指や手の形状の違いから個人差の影響を受けることが想定されるため、被験者を跨いだ認識精度の検証も求められる。

参考文献

- [1] S. Bhattacharya, R. Adaimi, and E. Thomaz. Leveraging sound and wrist motion to detect activities of daily living with commodity smartwatches. *Proc. of ACM on IMWUT*, 6(2):1–28, 2022.
- [2] K. Choi, H. Ryu, and J. Kim. Deep residual networks for user authentication via hand-object manipulations. *Sensors*, 21(9):2981, 2021.
- [3] K. P. Fishkin, M. Philipose, and A. Rea. Hands-on rfid: Wireless wearables for detecting use of objects. In *Proc. of ISWC*, pages 38–43, 2005.
- [4] R. Fukui, M. Watanabe, T. Gyota, M. Shimosaka, and T. Sato. Hand shape classification with a wrist contour sensor: Development of a prototype device. In *Proc. of UbiComp*, pages 311–314, 2011.
- [5] G. Garcia-Hernando, S. Yuan, S. Baek, and T. K. Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proc. of CVPR*, pages 409–419, 2018.
- [6] F. Hu, P. He, S. Xu, Y. Li, and C. Zhang. Fingertrak: Continuous 3D hand pose tracking by deep learning hand silhouettes captured by miniature thermal cameras on wrist. *Proc. of ACM on IMWUT*, 4(2):1–24, 2020.
- [7] V. Kakaraparthi, M. Goldberg, and T. McDaniel. Wrist view: Understanding human activity through the hand. In *Proc. of UAHCI*, 2023.
- [8] C. J. Lee, R. Zhang, D. Agarwal, T. C. Yu, V. Gunda, O. Lopez, J. Kim, S. Yin, B. Dong, and K. Li. Echowrist: Continuous hand pose tracking and hand-object interaction recognition using low-power active acoustic sensing on a wristband. In *Proc. of CHI*, pages 1–21, 2024.
- [9] J. Ma and D. Damen. Hand-object interaction reasoning. In *Proc. of AVSS*, pages 1–8, 2022.
- [10] T. Maekawa, Y. Kishino, Y. Yanagisawa, and Y. Sakurai. Wristsense: Wrist-worn sensor device with camera for daily activity recognition. In *Proc. PerCom*, pages 510–512, 2012.
- [11] T. Maekawa, Y. Yanagisawa, Y. Kishino, K. Ishiguro, K. Kamei, Y. Sakurai, and T. Okadome. Object-based activity recognition with heterogeneous sensors on wrist. In *Proc. of Pervasive Computing*, pages 246–264, 2010.
- [12] S. Mahmud, V. Parikh, Q. Liang, K. Li, R. Zhang, A. Ajit, V. Gunda, D. Agarwal, F. Guimbretiere, and C. Zhang. Act-sonic: Recognizing everyday activities from inaudible acoustic wave around the body. *Proc. of ACM on IMWUT*, 8(4):1–32, 2024.
- [13] K. Ohnishi, A. Kanehira, A. Kanezaki, and T. Harada. Recognizing activities of daily living with a wrist-mounted camera. In *Proc. of CVPR*, pages 3103–3111, 2016.
- [14] M. Philipose, K. P. Fishkin, M. Perkowitz, D. J. Patterson, D. Fox, and H. Kautz. Inferring activities from interactions with objects. 3(4):50–57, 2004.
- [15] J. Ranjan and K. Whitehouse. Towards recognizing person-object interactions using a single wrist wearable device. In *Proc. of UbiComp*, pages 722–731, 2016.
- [16] J. C. R. Rudolph, D. Holman, B. De Araujo, R. Jota, D. Wigdor, and V. Savage. Sensing hand interactions with everyday objects by profiling wrist topography. In *Proc. of TEI*, pages 1–14, 2022.
- [17] S. Wang, T. Zhang, Y. Li, P. Li, H. Wu, and K. Li. Continuous hand gestures detection and recognition in emergency human-robot interaction based on the inertial measurement unit. *Transactions on Instrumentation and Measurement*, 73:1–15, 2024.
- [18] T. C. Yu, G. Hu, R. Zhang, H. Lim, S. Mahmud, C. J. Lee, K. Li, D. Agarwal, S. Nie, J. Oh, F. Guimbretière, and C. Zhang. Ring-a-pose: A ring for continuous hand pose tracking. *Proc. of ACM on IMWUT*, 8(4):1–30, 2024.