

## [研究助成 (A)]

## AI センサ融合基盤が拓く次世代の日常行動認識

2251002

研究代表者  
(助成金受領者)

大阪公立大学 研究推進機構

教授 石丸 翔也

## 1. 研究の目的

本研究の目的は、安価で日常的に装着可能なセンサを用いて認識できる人間の日常行動の種類を大幅に拡大し、機械が一人一人の生活全体を継続的に把握できる基盤を構築することにある。これは、機械の支援によって人間が無意識のうちに合理的な選択を行える環境を実現するうえで、その前提となる「個人の日常生活の全体像の把握」を可能にするものである。

人間の日常行動は、歩行や睡眠といった身体的行動と、読書や会話といった認知的行動に大別できる。このうち身体的行動については、身体に装着するモーションセンサによる認識手法が広く確立されている。一方、認知的行動は顕著な身体運動を伴わないため、モーションセンサのみでの認識が困難であり、従来はカメラ・マイクによる計測や、眼球運動・脳波の解析といった高機能なセンシング手法が用いられてきた。しかし、これらの手法は装置上の制約やプライバシーへの配慮から実験室の外での利用が難しく、結果として、深層学習に堪えうる大規模データの収集や、実利用を想定したアプリケーション開発が進みにくいという課題が存在する。

そこで本研究では、AI技術による性能強化を通じて、安価で日常使いが可能なセンサがどの程度多様な行動を認識できるようになるかを明らかにすることを目指す。具体的には、物体認識や自然言語処理の分野で機械学習の性能を

飛躍させた基盤モデル (foundation model) の考え方を人間の行動認識タスクに適用する。対照学習 (contrastive learning) や自己教師あり学習 (self-supervised learning) に基づく事前学習によって、加速度センサ等から得られる情報により豊かな文脈を持たせ、これまで捉えることが難しかった微細かつ文脈依存的な人間の動きの認識を可能にする。

センサの性能と利便性は一般にトレードオフの関係にあるが、利便性の高いセンサには、多様なデータを長時間にわたり多人数から収集できるという利点がある。本研究は、この利点が大規模な事前学習を前提とする基盤モデルの構築と高い親和性を持つという着想に基づいており、利便性の高いセンサの弱点をAI技術によって補うアプローチの有効性を検証する。これにより、高機能センサに依存せずとも認知的行動を含む幅広い日常行動を認識できる道筋を示し、日常生活に溶け込む形で人間の意思決定を支援する技術の実現に貢献することを目的とする。

## 2. 研究の内容、成果

研究目的の達成に向けて、本研究では自己教師あり学習・対照学習・大規模言語モデルといったAI技術を人間の行動認識に適用し、安価で日常的に装着可能なセンサが認識できる行動の種類を拡大するための一連の検討を行った。得られた成果は、(1) モダリティ欠損に頑健なマ

ルチモーダル自己教師あり学習の基盤構築、(2) ラベルを必要としない言語モデル駆動のゼロショット行動認識、(3) 装着型デバイスによる認知的状態の推定、(4) 安価な音響センサによる摂取行動のセンシング、(5) 一般的な機器への視線推定モデルの展開、の5つに大別される。以下、それぞれについて詳細を述べる。

## 2.1 マルチモーダル自己教師あり学習による 認知状態推定基盤の構築

認知的行動の認識には、加速度センサに加えて視線や音声といった複数のモダリティを活用することが有効である。しかし実世界では、センサの装着不良やバッテリー制約、遮蔽などによって一部のモダリティが欠損することが頻繁に生じる。我々の研究グループでは、まずこの課題に取り組むための概念的枠組み CognifySSL を提案した [Cisse 2025]。この枠組みは、視線追跡、環境音、慣性運動 (IMU) の3つのモダリティを入力とし、それぞれを専用のエンコーダ (視線位置・瞬目・瞳孔径を扱う CNN ベースの GazeNet, 生音声・スペクトログラムを扱う 1 次元 CNN の WaveEncoder, 逐次的な運動データを扱う Transformer ベースの IMU エンコーダ) で符号化したうえで、対照学習、マスク信号モデリング、クロスモーダル予測を組み合わせて共有表現空間に整合させる。予測ヘッドはモジュール化されており、人間行動認識 (HAR)、精神的疲労の検出、エンゲージメント推定といった多様な下流タスクに対応する。また、学習時に一部のモダリティをランダムにマスクするドロップアウト型の訓練により、モダリティ欠損への頑健性を確保する設計とした。

この枠組みの設計にあたっては、IEEE Xplore, ACM Digital Library, arXiv, PubMed を対象に 2020 年から 2025 年までの関連研究を体系的にレビューした。自己教師あり学習の技術を含むこと、2つ以上のモダリティを融合すること、認知的・感情的状態の認識を

扱うことという3条件のうち少なくとも2つを満たす研究を対象とし、使用モダリティ・自己教師ありパラダイム・対象とする認知状態・評価プロトコルといった観点から各研究を分類した。その結果、既存研究の多くが身体的な行動認識に集中しており、視線や音声といったモダリティが十分に活用されていないこと、自己教師あり学習が理想的な完全モダリティの条件下で評価される傾向にあり、欠損や破損したデータ流に頑健な融合手法がほとんど存在しないことを明らかにした。この体系的なギャップ分析が、後述するモジュール型アーキテクチャの設計指針となった。本構想を CognifySSL として公開し、将来的には脳活動や心拍変動といった認知的な基準信号との照合による検証を計画するものとして位置づけた。

次に、上記枠組みを実装・評価・改良し、CognifySSL v2.0 として発表した [Cisse 2026]。CognifySSL v2.0 は、欠損したモダリティをリアルタイムに検出し修復することを目的とし、対照学習とマスクモデリングの目的関数を組み合わせる。本研究では、訓練時のランダムなモダリティドロップアウトが実世界のセンサ故障をどの程度模擬できるか、対照学習による整合とマスク再構成がそれぞれ潜在空間の形成にどう寄与するか、未知のモダリティの組み合わせや異なるセンサ利用可能性に対してどこまで汎化できるか、という3つの問いを設定した。アーキテクチャは、モダリティごとのエンコーダ、ドロップアウトシミュレーションを伴う注意機構ベースの融合層、モダリティの存在を予測する分類器、および欠損チャンネルを再構成するデコーダから構成される。各エンコーダはカーネルサイズ (5, 3, 3)・チャンネル数 (64, 128, 128) の3層からなる1次元畳み込みネットワークとし、全体の損失は重み付けしたマスク再構成損失 (平均二乗誤差) と対照整合損失の和とした (重みは  $\lambda_{\text{rec}} = 1.0$ ,  $\lambda_{\text{con}} = 0.5$ )。学習は Adam (学習率  $1 \times 10^{-3}$ , バッチサイズ 32)・対照学習の温度パラメータ 0.5・固定乱数シードのもと

で行い、(1) MobiAct での IMU エンコーダの単一モダリティ事前学習、(2) WESAD でのモダリティドロップアウトを伴うマルチモーダルマスク事前学習、(3) 両ヘッドを有効化した端から端までの結合最適化、という 3 段階の訓練手順を採用した。

評価には、IMU に基づく行動認識データセットである MobiAct (57 名, 9 種類の活動, 200 Hz) と、生理・運動信号を含む感情状態認識データセットである WESAD (15 名, ECG・EDA・呼吸・皮膚温・加速度を含む, 64 Hz にダウンサンプリングし 10 秒窓・50% 重複で分割) を用い、ユーザ非依存の評価のため WESAD では Leave-One-Participant-Out 交差検証を採用した。30% のモダリティ欠損下において、CognifySSL v2.0 は EDA・皮膚温・ECG・呼吸の再構成誤差 (RMSE) でそれぞれ 0.108, 0.061, 0.096, 0.124 を達成し、下流タスクの分類精度も MobiAct の行動認識で 91.5%, WESAD の感情状態分類で 82.6% と、比較手法である COCOA, MU-MAE, CroSSL を上回った。全モダリティが利用可能な設定での F1 スコアは 92.2% であり (MU-MAE 89.6%, CroSSL 88.7%), IMU のみの場合でも 85.5%, EDA のみの場合でも 84.3% を維持した。アブレーション分析からは、訓練時のモダリティドロップアウトが欠損時の頑健性を高めること (ドロップアウトなしでは 74.1% であった F1 スコアが 85.5% に向上), 対照学習と再構成を組み合わせたハイブリッドな目的関数が再構成のみ・対照のみのいずれをも上回ること (F1 スコアでそれぞれ +3.7%, +4.1%, 再構成誤差では再構成のみに対し平均 RMSE で 0.085 対 0.096), 注意機構ベースの融合が単純な連結を上回ること (部分欠損下の F1 スコアで +2.9%) が確認された。また学習済みエンコーダは MacBook Pro (16 インチ, 2021, Apple M1 Max) の上で 1 入力窓あたり 90 ミリ秒未満の推論遅延で動作し、準リアルタイムでの動作可能性を示した。コードと対話的な可視化ダッ

シュボードも公開している。

## 2.2 大規模言語モデルを活用したゼロショット行動認識

認識できる行動の種類を拡大するうえで大きな障壁となるのが、新しい行動ごとに大量のラベル付きデータを収集・注釈する必要があることである。そこで我々は、この障壁を取り除くため、ラベル付きの訓練例を一切用いずに両手首の加速度信号と自然言語による行動記述とを対応づける、ゼロショットの行動認識枠組みを提案した [Bhatt 2026]。本検討は、ラベルからテキストへの移行を促すコンペティション NHIC-ABC2026 への応答として行ったものである。提案手法は 3 段階のパイプラインから成る。第 1 段階では、大規模言語モデル (Gemini, gemini-2.5-flash) を用いて、各行動の自然言語記述から、エネルギー、規則性、周波数、手の関与、想定継続時間といった属性で構造化された意味的プロファイルを抽出する。第 2 段階では、両手首に装着した加速度センサ (64 Hz) のデータを 5 秒の窓 (50% の重複, ステップ 2.5 秒) に分割し、意味的プロファイルと同型 (isomorphic) になるよう設計された特徴量 (エネルギー、規則性、周波数、手の関与、対称性) を各窓について計算する。データの前処理では、左右の信号を 50 ミリ秒の許容幅で最近傍照合により同期させ、欠損値はリストワイズ削除で処理した。第 3 段階では、これらの特徴量に対し、エネルギー 0.30, 周波数 0.25, 規則性 0.20, 手の関与 0.15, 対称性 0.10 という経験的に定めた重みを付した規則ベースの類似度関数と top-K ランキング ( $K = 3$ , 閾値  $\tau = 0.6$ ) を適用し、各窓に対して複数の有力な候補を保持することで、ゼロショット推論に内在する不確実性を保持する。この枠組みは、自然言語を構造化属性に変換する意味抽象化層、センサ信号を解釈可能な特徴量に変換する特徴変換層、両者を規則ベースで照合する整合層という 3 層構造として実装した。

サラダ調理データセットの8種類の衛生関連動作（3種類の手伸ばし動作、手洗い、すすぎ、乾燥、2種類の着用動作）を対象とした評価では、全長1359.7秒・6チャンネルにわたる87019サンプルから得た542個の解析窓に対し、top-K マッチングが合計1542個のランク付き検出を生成し、窓レベルで94.8%の検出率を達成した（514/542の窓で3つの候補が得られ、閾値を超える候補が得られなかったのは28窓（5.2%）のみであった）。すすぎのような意味的に特徴的な動作は比較的高い確信度（0.740）で照合された一方、運動学的に類似した手伸ばし動作は低い確信度（0.635）にとどまり（両者の差は16.5%）、現在の特徴空間では区別できないことが示された。特に3種類の手伸ばし動作は同一の意味的プロファイル（低エネルギー、散発的、周波数なし、利き手主体、2秒）を共有するため、同一の類似度スコアが付与され、多くの窓で常に同じ順位に並ぶことが観察された。また、スライディング窓の重複に起因する継続時間の系統的な過大評価を補正する手法（検出区間の和集合を求める区間併合）により、過大評価を2.2%低減した。本検討を通じて、ラベルなしで動作する端から端までのゼロショットパイプライン、不確実性を保持するtop-K ランキング、重複を補正する時間解析という方法論的貢献を示すとともに、手首装着型の加速度センサのみでは対象に向かう手伸ばし動作が区別できないという運動学的エイリアシングの問題を実証的に明らかにし、文脈情報を取り入れた将来のマルチモーダルセンシングの必要性を示した。

### 2.3 装着型 AR デバイスの運動データによる認知負荷の推定

認知的な状態の一つである認知負荷は、従来NASA-TLXのような主観評価や、侵襲的になりがちな生理センサによって測定されてきた。本研究グループでは、拡張現実(AR)ヘッドセットに内蔵されたセンサが取得する運動データの

みから認知負荷を推定できるかを検討した[Melo 2026]。産業的な組立作業と認知的・運動的特徴を共有する課題として、難易度を操作したタングラムパズルを用い、高低2水準の認知負荷を誘発する実験を設計した。パズルはTangram Builder を用いて47種類作成し、ピースの輪郭の見えやすさで難易度を定め、低難易度40種・高難易度7種とした。実験はPsychoPyで構築し、HoloLensの生データ取得にはHL2SSプラグインを用いた。参加者はMicrosoft HoloLens 2（頭部・手・視線の運動データを名目30 Hz、実効21.34 Hzで取得）とEmpatica E4（血液量脈波を64 Hz、EDAと皮膚温を4 Hzで取得）を装着し、30秒の窓（10秒の重複）に区切ってデータを解析した。運動データからは瞬目・注視・サッカードに関する特徴量を、生理データからはNeuroKit2を用いて心拍変動やEDAのトニック・フェイジック成分を抽出した。技術的問題による欠損を除いた最終的な分析対象は23名であり、運動データのみのデータセットでは21名分を用いた。

主観評価(NASA-TLX)と課題完了時間は、認知負荷の実験的操作が成功したことを裏付けた。高難易度パズルの平均完了時間は202.53秒(SD 127.79)、低難易度は31.84秒(SD 8.45)であり、Wilcoxon検定によって両者の差は統計的に有意であった( $W = 136.0$ ,  $p < .001$ )。NASA-TLXでは、精神的要求・時間的要求・達成度・努力・フラストレーションのいずれの下位尺度でも高難易度条件で有意に高いスコアが報告された一方、身体的要求には条件間で有意差が見られなかった( $p = .501$ )。これは、パズルの複雑さを増しても解くために必要な身体活動は本質的に変わらないという想定と整合する。分類にはロジスティック回帰・サポートベクタ分類器・ランダムフォレスト・XGBoost・LightGBM・AdaBoostの6種類のモデルを用い、leave-one-participant-out交差検証によって未知の参加者への汎化を評価し

た。その結果、運動データのみで学習したランダムフォレストモデルが最も高い F1 スコア 0.886 を達成し、これは生理データのみ（最大 59%）や両者を統合した場合（F1 スコア 0.831）を上回るものであった。特徴量の重要度を分析した結果、視線運動が全体の 54.1% を占め（最も重要な特徴量は注視距離の平均）、次いで手の運動（23.5%）、頭部の運動（22.4%）の順であり、視線運動が認知負荷分類において最も重要な要素であることが示された。先行研究 [Hou et al.] では Transformer モデルでより高い F1 スコア（96%）が報告されているが、これは計算負荷の大きいアーキテクチャに依存し、データ漏洩の影響を受けていた可能性がある一方、本成果は軽量のランダムフォレストと厳密な交差検証によって資源制約環境に適した性能を示した点に意義がある。構築したデータセットは公開している。

## 2.4 二つのマイクロフォンを用いた飲料摂取の音響センシング

いつ・何を・どれだけ飲むかという飲料摂取は、日常的で生理的に重要な行動であるが、従来の容器装着型センサやカメラベースの分類器は、量と種類のいずれか一方しか扱えなかった。我々の研究グループでは、イヤフォン型マイクロフォンと咽喉マイクロフォンを融合させ、飲料の種類と量を同時に認識する小型の装着型システム SipSense を提案した [Ueshima 2026]。10 名の実験参加者（男性 9 名、女性 1 名、平均年齢 24.3 歳）を対象に、10 種類の飲料（リンゴジュース、コーヒー、コーラ、エナジードリンク、緑茶、オレンジジュース、プロテイン飲料、炭酸水、トマトジュース、水；このうちコーラ・エナジードリンク・炭酸水は炭酸あり、緑茶とコーヒーは温かい飲料）について各 10 回、合計 1000 回の飲水イベントを収集した。収録の際には時刻同期の基準として実験参加者に明瞭な「ah」を発声させ、摂取量は飲用前後のタンブラーの質量を電子はかりで測定して

算出し、各飲水イベントの音響データに対応づけてグラウンドトゥルースとした。摂取量はあらかじめ固定せず、実験参加者が自然な量を飲むものとし、平均摂取量は 11.58 mL であった。

咽喉マイクロフォン（48000 Hz）と AirPods Pro によるイヤフォン型マイクロフォン（16000 Hz）の音声から、メル周波数ケプストラム係数（13 次元、librosa で算出）を特徴量として抽出し、分類にはランダムフォレスト・サポートベクタ分類器（RBF カーネル、標準化前処理）・XGBoost（10 クラスのソフトマックス）を、量の推定にはランダムフォレスト回帰（決定木 100 本）・サポートベクタ回帰を用いた。評価には leave-one-out 交差検証を採用し、量の推定には平均絶対誤差（MAE）・平均誤差率（MPE）・平均絶対誤差率（MAPE）を用いた。その結果、両マイクロフォンを統合したランダムフォレスト回帰が飲料量の推定で最も高い精度を示し、MAE は 3.89 mL を達成した（咽喉のみ 3.95 mL、イヤフォンのみ 4.28 mL）。10 クラスの飲料種類分類では、両マイクロフォンを統合したランダムフォレストが F1 スコア 0.56 を達成し、これは咽喉マイクロフォン単独の性能と同等であった。粘性の高いプロテイン飲料は量の推定で最も小さい誤差（3.09 mL）を示し、種類分類でも統合構成で 68% と最も高い精度を示した一方、オレンジジュースは 44% にとどまり、温度の高いコーヒーと緑茶は相互に誤分類されやすい傾向が見られた。これらの誤分類は、温かい飲料における口に含む量や嚥下速度の類似性に起因すると考えられる。参加者ごとの Macro F1 スコアは 0.24 から 0.74 まで幅があり、飲水行動や咽喉の構造などの個人差が音響特徴に影響することが示唆された。本成果は、装着型の音響センサによって飲料の種類と量を同時に把握できる可能性を示すものである。

## 2.5 軽量のオンデバイス微調整による視線推定の汎化

視線は認知的状態を反映する重要なモダリティであるが、外観ベースの視線推定モデルは、訓練データに含まれない機器上で展開すると精度が低下するという課題を抱える。そこで我々は、スマートフォンとタブレットから収集された GazeCapture データセットのみで訓練された iTracker が、ノートパソコン上では下向きの視線が訓練分布外となるため画面下部で誤差が増大することに着目し、事前学習済みモデルを未知の機器構成へ適応させる軽量の微調整手法を提案した [Kokado 2026]。iTracker は顔画像 (224 × 224 画素)、左右の目の画像 (各 224 × 224 画素)、顔の位置を表すフェイスグリッド (25 × 25 の二値マスク) を入力とするモデルであり、本手法ではこれを Core ML 形式に変換したうえで、最終全結合層のみを微調整する。キャリブレーションには、画面中央・四隅・8つの中点からなる 13点グリッドを用い、各点で 10 フレームずつ、計 130 枚の顔画像を約 30 秒で収集する。微調整は Apple Core ML に搭載の MLUpdateTask API によってデバイス上で完結して実行され、確率的勾配降下法と平均二乗誤差損失を用いた (学習率は FT-Self で 0.005, FT-Others で 0.0001, バッチサイズ 32, エポック数 10)。比較対象として、キャリブレーションの点对応から 3 × 3 行列を最小二乗推定する幾何補正であるホモグラフィ変換も実装した。

11 名の参加者 (男性 8 名, 女性 3 名, いずれも日本人, 20 ~ 30 歳) を対象に, MacBook Pro (14 インチ, M4 チップ, 画面解像度 1512 × 945, カメラ 1280 × 720・30 fps, 視距離 30 ~ 40 cm) 上で, ベースライン, ホモグラフィ変換, 他ユーザのデータによる微調整 (FT-Others), 対象ユーザ自身のデータによる微調整 (FT-Self) の 4 条件を比較した。ベースラインの平均誤差は 6.17 cm であり, モバイル機器で報告される 1.7 ~ 2.5 cm と比べて大幅に

大きかった。これに対し FT-Self は平均推定誤差を 2.91 cm へと 52.8% 低減し, ホモグラフィ変換 (3.38 cm, 45.2% 低減) を上回った。特に誤差が大きかった画面下部の領域では, FT-Self は誤差を 8.83 cm から 2.68 cm へと 69.6% 低減し, 同領域でのホモグラフィ変換の低減率 (61.8%) を上回った。また, 対象ユーザのデータを用いない FT-Others でも誤差を 4.47 cm まで低減し (27.6% の低減), 必要な適応のかかなりの部分がユーザ固有ではなく機器固有であり, 機器に適応させたモデルをユーザ間で部分的に転用できる可能性を示した。補足的な汎化評価として, 周辺 8 点のみで較正し中央 5 点で評価した場合でも, FT-Self は 3.59 cm (22.3% 低減) を達成した。キャリブレーションデータの収集は約 30 秒, 微調整は約 10 秒で完了し, 全工程が 1 分未満でデバイス上において実行できることから, 特別なハードウェアを必要とせずに実世界で展開できる実用的な手法であることを示した。

### [成果の発表, 論文など]

[Cisse 2025] Aboul Hassane Cisse and Shoya Ishimaru. "From Signals to Cognition: Self-Supervised Multimodal Learning for Cognitive State Inference from Wearable Sensor Streams". In Adjunct Proceedings of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2025 ACM International Symposium on Wearable Computers (UbiComp '25 Adjunct), pp. 105-109, 2025.

[Cisse 2026] Aboul Hassane Cisse and Shoya Ishimaru. "Detect and Repair: Robust Self-Supervised Wearable Sensing Under Missing Modalities". Sensors 26 (8), p. 2419, 2026.

[Bhatt 2026] Ankur Bhatt and Shoya Ishimaru. "From Language to Sensors: Zero-Shot Activity Recognition via LLM-Driven Semantic Alignment and Top-K Ranking". International Journal of Activity and Behavior Computing 2026 (2), p. 1-22, 2026.

[Melo 2026] Javier Melo, Leyla Akinci, Ko Watanabe, Nicolas Großmann, Shoya Ishimaru and Andreas Dengel. "SensCogAR: Cognitive Load Estimation Via Movement Data in Assembly Tasks". International Journal of Activity and Behavior Computing 2026 (1), p. 1-20, 2026.

**[Ueshima 2026]** Ryugo Ueshima, Shunta Sato, Ko Watanabe, Liu Mengxi, Sungho Suh, Paul Lukowicz, Andreas Dengel, Shoya Ishimaru, Shingo Ata and Manato Fujimoto. “SipSense: Acoustic Sensing of Beverage Type and Volume Using Dual-Microphone Wearable System”. *International Journal of Activity*

*and Behavior Computing* 2026 (1), p. 1-18, 2026.

**[Kokado 2026]** Soki Kokado, Ko Watanabe and Shoya Ishimaru. “Expanding the Range of Pre-Trained Gaze Estimation via Lightweight On-Device Fine-Tuning”. In *Proceedings of the 2026 Symposium on Eye Tracking Research and Applications (ETRA '26)*, 2026.