

動作特徴の表現学習を活用した スマートフォンによる倉庫内作業認識

渡邊 企章¹ 加納 一馬¹ Tahera Hossain¹ 片山 晋¹ 浦野 健太¹ 米澤 拓郎¹ 河口 信夫^{1,2}

概要：近年、労働力不足が深刻化する物流業務の効率化に向けて、現場を仮想空間に再現するデジタルツインの活用が注目されている。特に荷物や環境のデジタル化が進む一方で、作業者の行動のデジタル化は、現場への負担や環境変化への対応の難しさから依然として困難である。これに対し本研究では、広く普及していて導入も容易なスマートフォンの加速度・角速度センサを用いて、深層学習モデルにより倉庫内作業を自動認識する枠組みを提案する。また、センサベースの行動認識における個人差やセンサ装着位置のばらつきによって認識性能が低下する課題に対して、ラベルなしデータを活用して汎化性能を向上させる手法を検討する。提案する作業認識モデルは、数秒ごとに動作を抽出する CNN と、その系列から作業のコンテキストを認識する Transformer Encoder から構成される深層学習モデルである。訓練は二段階で実施され、まず CNN にラベルなしデータを用いた SimCLR による表現学習を適用して個人差やセンサ装着位置の違いに頑健な特徴表現を獲得する。その後、ラベル付き正解データを用いて後続の層のみを訓練することで、特徴抽出の汎化性能を保ったまま動作と作業の関係性を学習する。実際の物流倉庫において 29 人・計 196 時間分のセンサデータを収集し、10 人・計 6 時間分をラベル付して交差検証による評価を行った結果、提案モデルは物流倉庫における 3 つの作業（検品・仕分け・搬送）を F1 スコア 0.89 で分類した。特に CNN を表現学習しない場合と比べてスコアが約 7 ポイント改善し、提案手法の有効性が示された。

Smartphone-Based Warehouse Task Recognition Using Representation Learning of Motion Features

KISHO WATANABE¹ KAZUMA KANO¹ TAHERA HOSSAIN¹ SHIN KATAYAMA¹
KENTA URANO¹ TAKURO YONEZAWA¹ NOBUO KAWAGUCHI^{1,2}

1. はじめに

電子商取引（EC）の普及に伴い、物流市場の拡大と作業量の増加が続いており、物流倉庫における労働力不足が深刻な課題となっている [1]。こうした背景から、物流倉庫では作業の効率化が求められており、近年では倉庫内をデジタル化し、可視化やシミュレーションに基づいて業務効率の改善に利用する「デジタルツイン」の技術が注目されている [2], [3]。実際に、カメラやセンサを駆使して倉庫内の荷物や人の位置をデータ化する研究が進んでいる [4], [5]。

一方、作業者の行動データの自動取得には依然として課題がある。特に実用化のためには、低コストかつ作業対象や環境の変化に対応しやすく、既存設備への影響が小さい手法が必要となる [6]。このような要求に対し、加速度や角速度のセンサデータを用いて作業者の行動を自動で認識するセンサベースの手法が近年注目されている。なかでもスマートフォンは、加速度センサをはじめとする多様なセンサを備え、特別な装着や設置を要しないため、コスト面・柔軟性の両面で優れた選択肢と言える [7]。

そこで本研究では、スマートフォンの加速度・角速度センサを用いて倉庫内作業を自動認識する手法を提案する。スマートフォンは個人が常時携帯するセンサデバイスとしても有用であり、各作業者のセンシングに活用できる。センサベースの行動認識においては、加速度・角速度セ

¹ 名古屋大学大学院 工学研究科
Graduate School of Engineering, Nagoya University

² 名古屋大学 未来社会創造機構
Institutes of Innovation for Future Society, Nagoya University



図 1: 物流倉庫における荷物入荷から格納までの流れ

ンサなどのセンサデータを数秒単位でウィンドウ処理し、その数秒内の動作を分類する枠組みが広く用いられている [7], [8]. しかし、実際の作業は単一動作での定義が困難で、短時間のセンサデータでは作業を十分に特定できない。そのため、本研究では作業認識に適した長期的な行動の意味合いを考慮できる深層学習モデルを提案する。

また、各現場に適した教師データを大量に収集していくことは現実的ではなく、少ない教師データでも高い汎化性能が得られる手法が求められる。認識モデルの汎化性能を高める手段としては、近年では自己教師あり学習やドメイン適応などの特徴表現に着目する手法が成果を上げている [9], [10]. そこで本研究では、提案する作業認識モデルの訓練時にラベルなしデータを用いた表現学習を取り入れて汎化性能を向上させる手法について検討する。

本手法の認識対象は、物流倉庫における作業である。物流倉庫に荷物が届いてから格納されるまでの流れは図 1 に示す通りで、人が行う作業は主に以下の 3 つに分類される。

- (1) **検品**: 入荷商品の品目や数量を注文書と照合する作業
- (2) **仕分け**: 商品の箱を適切な場所に振り分ける作業
- (3) **搬送**: 荷物を乗せた台車を運ぶ作業

また、含まれる主な動作の例について表 1 に示す。これらの作業は一定の役割分担が存在するものの、荷物の滞留状況や当日の進捗に応じて柔軟に交替される。さらに、作業が数十分間継続する場合もあれば、数分単位で切り替わる場合もある。このような条件下では、作業内容の手動データ化は現実的でなく、センサデータに基づく自動認識手法が求められる。本研究の主な貢献は以下のとおりである。

- 倉庫内作業のような複数の動作を含む行動を認識するための CNN, Transformer Encoder, MLP からなる作業認識モデルを提案した。
- ラベルなしデータを用いた表現学習を活用し、作業認識モデルの汎化性能を向上させる手法を提案した。
- 実環境データを用いた検証において提案手法はベースラインを上回るスコアを達成し、その有効性を示した。

表 1: 各作業に含まれる代表的な動作の例

作業	含まれる主な動作
検品	段ボールを開封する, 商品を取り出す, スキャンする
仕分け	荷物を持ち上げる, 運ぶ, 置く, 移動する
搬送	台車を押す, 歩く, 方向転換する, 立ち止まる

2. 関連研究

2.1 センサベースの行動認識

スマートフォンに内蔵された加速度センサや角速度センサを用いたセンサベースの行動認識 (Human Activity Recognition, HAR) は、歩行や座位、階段昇降といった日常的な動作の認識に広く利用されてきた [11]. 初期の研究では、平均や分散などの統計的特徴量を手設計し、SVM や GBDT といった機械学習手法を用いるアプローチが主流であった [11]. 近年では、CNN, LSTM, Transformer などの深層学習モデルにより、センサデータから直接特徴を抽出して分類する手法の研究が進んでいる [12], [13]. また、行動認識の応用範囲は日常生活にとどまらず、物流や製造業などの産業現場にも広がっており、作業記録や業務分析、安全管理を目的とした活用が進んでいる [6].

センサベースの行動認識モデルは導入が容易であるが、個人差やセンサの装着位置の違いによる影響を受けやすいという課題がある。実際に、異なるユーザやデバイス、装着位置といったドメインの違いがモデルの性能に大きく影響することが示されており、特に深層学習モデルは訓練データと異なるドメインに対して性能が大きく低下すると報告されている [14]. このような個人差や環境変化への対応は、行動認識の実用化における重要な課題である。

2.2 自己教師あり学習 (表現学習) の活用

ラベル付きセンサデータの収集には多大なコストがかかるため、自己教師あり学習 (Self-Supervised Learning, SSL) を用いてラベルなしデータから有用な特徴表現を学習する手法が注目されている。この手法では、まずラベル

なしデータを用いて動作の特徴を抽出する特徴抽出器（エンコーダ）を事前に学習し、得られた特徴表現から行動を分類する単純な線形層を訓練することで、少量のラベル付きデータでも高精度な認識を可能とする [15].

センサベースの行動認識における自己教師あり学習手法は、主に以下の4つのパラダイムに分類される [15]. 1つ目は Augmentation Recognition 型 であり、適用されたデータ拡張の種類を識別して特徴抽出器を学習する手法である. 代表例としては, MultiTaskSSL [16] が挙げられる. 2つ目は Reconstruction 型 であり、マスクされたセンサ系列を再構成して波形の時系列的特徴を学習する. 代表的な手法には, LIMU-BERT [17] がある. 3つ目は Joint-embedding 型 であり、この手法はある同一サンプルの異なるビュー (例: 異なる時間領域の切り出しや異なるデータ拡張) から得られた埋め込みベクトルが一致するように学習を行う. SimCLR [18], [19] や CPC [20], [21] などがこれに該当し、特に広く利用されている. 4つ目は Hybrid 型 であり、複数のパラダイムを組み合わせた手法であり, ARSSL [22] がその代表例である.

これらの手法を用いると、ラベルなしのセンサデータから有力な特徴抽出器を獲得できるが、多くは数秒程度の単一動作の認識を主な対象としている. そのため、複数の動作を含む物流倉庫における作業を認識するためには、より長期的な動作の文脈を捉えられる手法が求められる.

3. 提案手法

3.1 モデルの概要

提案手法では、各作業を複数の基礎動作の組み合わせとして解釈し、それに基づいて認識モデルを設計する. 例えば、仕分け作業は「屈む」「持ち上げる」「向きを変える」「歩く」といった基礎動作によって構成されている. このような動作の系列を捉えるために、数秒ごとに動作特徴を抽出する CNN、および動作特徴の系列から作業を認識する Transformer Encoder と MLP からなる深層学習モデルを提案する. 図2に作業認識モデルのネットワークを示す. 提案する作業認識モデルは、数秒程度を対象とする一般的な行動認識モデルと比べ、長期的な時系列情報を考慮できる. 一方で、このような比較的大きいネットワークは訓練データが少ないと過学習を引き起こしやすく、十分な汎化性能の確保が難しいという課題がある. そこで本研究では、CNN に対してラベルなしデータを用いた表現学習を適用し、モデルの汎化性能を向上させる手法を検討する.

3.2 作業認識モデルのアーキテクチャの詳細

作業認識モデルのアーキテクチャの詳細を図3に示す. 入力は加速度（重力成分を除く）と角速度からなる6チャネルの時系列データであり、本研究では 100 [Hz] にリサンプリングしたものを用いる. この信号を 2.56 [s] (256 サン

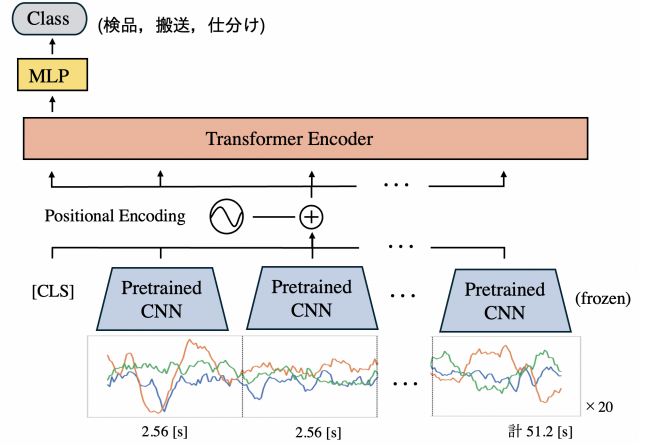


図 2: 作業認識モデルのネットワーク

プル) ごとにセグメント化する. 各セグメントは、入力信号 $\mathbf{x} \in \mathbb{R}^{6 \times 256}$ として表され、CNN によって 128 次元の特徴ベクトルに埋め込まれる.

畳み込み層は3層構成であり、それぞれ Conv1d(6→64, kernel=10, stride=4, padding=3), Conv1d(64→64, kernel=6, stride=2, padding=2), Conv1d(64→128, kernel=4, stride=2, padding=1) から構成される. 各層にはバッチ正規化と ReLU 活性化関数を用いるとともに、最後に Dropout($p = 0.3$) と Global Average Pooling を適用し、特徴ベクトル $\mathbf{h} \in \mathbb{R}^{128}$ を得る.

この処理を 20 セグメント、計 51.2 [s] にわたって行い、得られた 20 個の特徴ベクトル $[\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_{20}]$ を時系列順に並べる. その先頭に分類用の CLS トークンを追加し、全てのベクトルに Positional Encoding を加えて、Transformer Encoder に入力する. Transformer Encoder の各層は、4 ヘッドの Multi-head Attention と、Linear(128→256→128) の構成を持つ Feed Forward Network (FFN) からなる.

最後に、Transformer Encoder によって時系列情報が集約されたコンテキストベクトル $\mathbf{C}_0 \in \mathbb{R}^{128}$ を 2 層の線形層からなる MLP に入力し、検品・仕分け・搬送の 3 クラスに分類する. ここで MLP は Linear(128→64) → ReLU → Dropout($p = 0.3$) → Linear(64→3) で構成される.

3.3 CNN の表現学習

3.3.1 SimCLR を用いた特徴表現の獲得

提案ネットワークにおいて、初めの基礎動作を抽出する CNN には、被験者間の個人差や同一被験者内での動作の強度・速度のばらつき、さらにはセンサ装着位置の違いといった変動要因に対して頑健に、センサデータから動作の意味合いを抽出することが求められる. 限られた教師データでこのような特徴表現を獲得するのは困難であるため、本研究では表現学習手法の一つである SimCLR を用いて、ラベルなしデータから予め動作の特徴表現を獲得しておく手法を提案する.

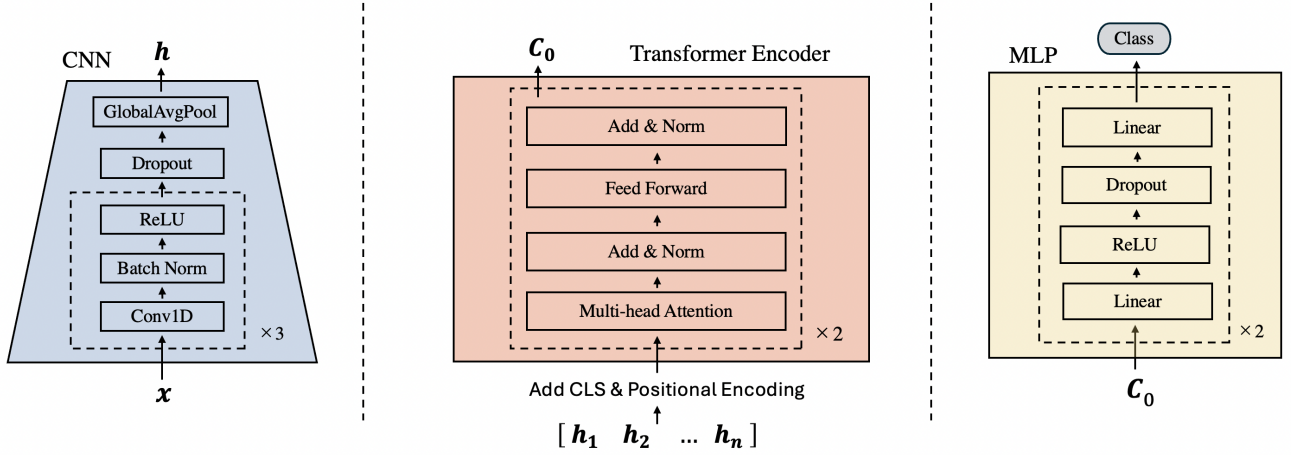


図 3: アーキテクチャの詳細

SimCLR は、元々画像データに対する自己教師あり表現学習の手法として提案されたものであり [18], ラベルなしのデータから意味的に有用な特徴表現を獲得することを目的とする。同一の画像に異なる変換（例：切り抜きや色調変化）を施して 2 つのビューを生成し、それぞれをエンコーダに通して特徴ベクトルを抽出し、同一画像に由来するペアは近づけ、異なる画像のペアは遠ざけるように学習する。この学習により、意味的に類似するデータが特徴空間上で近接し、ラベルなしのデータからも下流タスクに有効な特徴表現が得られる。この枠組みはセンサデータにも応用されており、図 4 のように、加速度や角速度データに対して、動作の意味合いに大きく影響しない変換（ノイズ付加、スケーリング、時間シフトなど）を適用して類似データを生成し、それらを同じ動作とみなして特徴ベクトルが近づくように学習する手法が提案されている [19]。本研究ではこの手法を用いて CNN をラベルなしデータで訓練し、センサ値の揺らぎや装着条件の違いに対して頑健な特徴表現を獲得する。

3.3.2 データ拡張の設計

SimCLR におけるデータ拡張では、動作に関する頑健な特徴表現の獲得が目的の場合、動作の意味を保ちつつ、動作のばらつきや環境の違いを模倣できる変換が求められる。本研究では動作の強弱やタイミングの違い、センサ装着方向のずれなど実環境で生じるばらつきを再現するため、以下の 6 種類の変換を用いる。括弧内は適用確率を表し、一部の変換は信号の標準偏差 σ_s を用いる。

- **ノイズ追加 (1.0)**：平均が 0、標準偏差が σ_s の 5% の正規分布に従うノイズを加える。
- **スケール変換 (0.5)**：平均が 1.0、標準偏差が σ_s の 70% の正規分布に従う係数を乗じて振幅を変化させる。
- **回転 (0.5)**：ランダムな回転軸に沿って、ランダムな角度 (-30 度から 30 度) 信号を回転させる。

- **時間シフト (0.5)**：時間軸方向に 1 秒以内のランダムな時間だけ信号をずらす。
- **時間歪み (0.2)**：3 次スプライン補間により、信号の時間軸をランダムに非線形に変形する。制御点は 4 点、歪みの強さは平均 0、標準偏差 1.0 の正規分布に従う。
- **セグメント並べ替え (0.2)**：信号を 4 つのセグメントに分割し、順序をランダムに入れ替える。

3.3.3 NT-Xent 損失の計算

SimCLR では、ランダムな変換を適用して生成した 2 つの系列を共通のネットワークで特徴ベクトルに変換し、それらの特徴表現が近くなるように学習を進める。

具体的には、ミニバッチ内の各サンプル x に対して、異なる 2 種類のランダムな変換を適用し、それぞれから得られた 2 つの系列 \tilde{x}_i, \tilde{x}_j を正例ペアとする。この処理によって、バッチサイズが N の場合、 $2N$ 個の変換済みサンプルが得られる。それぞれのサンプルは共通のエンコーダで特徴ベクトル h_i, h_j に変換された後、さらに小さな空間へ写像するプロジェクション層を通して、最終的な表現 z_i, z_j を得る。各正例ペア (i, j) に対して、他の $2N - 2$ 個の埋め込みを負例として扱い、以下の NT-Xent (Normalized Temperature-scaled Cross Entropy) 損失 [18] を最小化し、正例間の距離を縮め、負例とは遠ざけるように学習を行う。

$$\ell(i, j) = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2N} I_{[k \neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)} \quad (1)$$

ここで、 $\text{sim}(\mathbf{z}_i, \mathbf{z}_j)$ は通常、コサイン類似度によって定義される。 τ はスケーリング係数として働き、値が小さいほど類似度の違いを強調する。また、 $I_{[k \neq i]}$ はインデックス $k \neq i$ の時に 1、 $k = i$ の時に 0 をとる指示関数 (Indicator Function) である。この損失により、動作の識別性を保ちつつ、センサ値の変動に強い表現を学習できる。

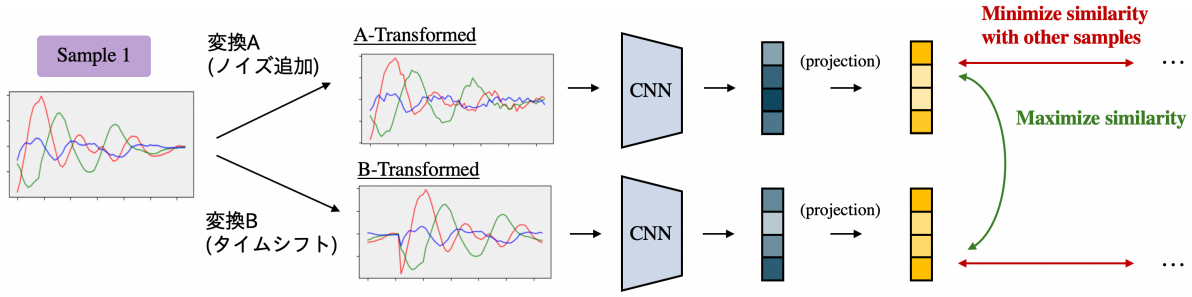


図 4: SimCLR による CNN の事前訓練

3.3.4 SimCLR アーキテクチャの詳細

SimCLR では、CNN と MLP からなるネットワークを設計する。CNN は 3.2 節に示した 3 層の 1 次元畳み込みネットワークを用いる。MLP は、Linear(128→128) → ReLU → Linear(128→64) の構成とし、特徴空間の再マッピングを行う。各入力 $x_i \in \mathbb{R}^{6 \times 256}$ に対して 2 種類のランダムな変換を施し、得られたペアを共通の CNN で、それぞれ 128 次元の特徴ベクトル $h_i, h_j \in \mathbb{R}^{128}$ に埋め込む。これらを MLP により 64 次元の特徴ベクトル $z_i, z_j \in \mathbb{R}^{64}$ に写像した後、NT-Xent 損失に基づきネットワークを学習する。

4. 評価

提案手法の有効性を検証するために実環境のデータを用いた評価実験を行った。4.1 節でデータの収集方法、4.2 節でデータの前処理について説明する。また 4.3 節では評価方法を説明し、4.4 節で実験の詳細を示す。4.5 節で評価結果として予測スコアとその分布、および混同行列を示す。結果の考察は 4.6 節で行う。

4.1 データ収集

実際の物流倉庫の作業員に、腰にスマートフォンを装着した状態で業務を行ってもらい、その間の加速度・角速度センサの値を計測した。1 日の計測で 29 人、計 196 時間の加速度・角速度データが収集された。また業務中の様子を録画し、一部の被験者について各時刻に対応する実施作業をラベル付けして正解データを作成した。正解データは 10 人、計 6 時分を収集した。収集した正解データにおける作業ラベルの内訳は図 5 に示す通りである。

4.2 データの前処理

収集したデータに施す前処理を図 6 に示す。ラベルなしセンサデータに対しては、全データに対して 2.56 [s] ごとのセグメント化を適用し、SimCLR による CNN の事前訓練に用いる。ラベル付きセンサデータに対しては、1 ウィンドウの長さを 20 セグメントとして、一定のストライドでスライドさせながら切り出し、各ウィンドウの中央時刻に実施されていた作業ラベルを紐付けることで、作業認識の

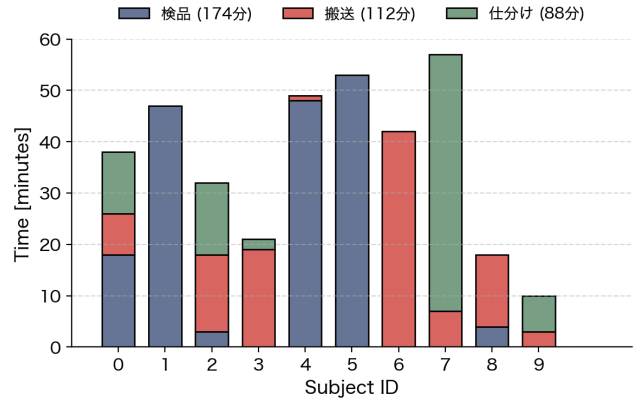


図 5: 正解データの作業ラベル内訳

ためのデータセットを構築する。推論時には、ストライドを 1 セグメントとし、ウィンドウを 1 セグメントずつ動かしながらそれぞれの中央時刻に対応する作業を逐次予測する。一方、訓練時にはデータの冗長性を抑えるため、ストライドを 4 セグメントに設定し、ウィンドウを 80% のオーバーラップでスライドさせながら訓練データを収集する。

4.3 交差検証による評価

提案した作業予測モデルは事前訓練済み CNN と Transformer, MLP で構成され、一定区間の加速度・角速度データを入力としてその間実施していた作業を出力とする。評価実験では、事前訓練として 29 人、計 196 時間の全センサデータのうち 190 時間分のラベルなしデータを用いて SimCLR による CNN の表現学習を行う。その後、10 人、計 6 時間分のラベル付き正解データを用いて予測モデルの交差検証を行う。モデルの性能は被験者単位での一つ抜き交差検証 (Leave One Out Cross Validation, LOOCV) を行なって評価する。10 人分の正解データのうち、9 人を train data, 1 人を test data として訓練、性能評価を行い、これを全被験者に対して実行する。

モデルが予測する 3 分類問題の性能指標として、Precision, Recall, F1 を用いる。それぞれ真陽性を TP 、偽陽性を FP 、偽陰性を FN として、以下の式で定義される。

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

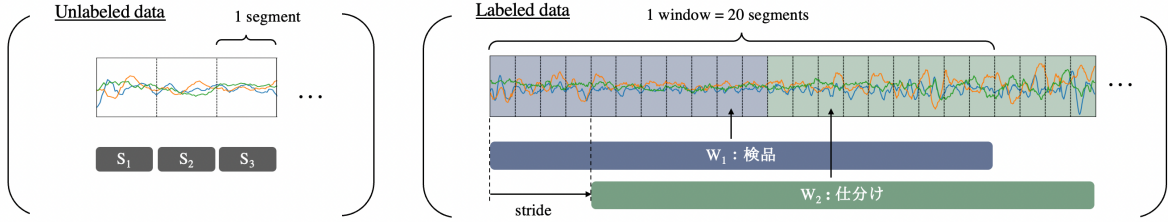


図 6: データ前処理の流れ

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

また、ラベルの不均衡を考慮し、ラベル数の割合に基づく重み付き平均をモデルの性能指標として採用する。

4.4 実験の詳細

4.4.1 ベースラインの設計

評価実験では2つのベースラインを用意し、以下の3つのモデルの認識性能を比較する。

- **GBDT+smoothing**: 統計的特徴抽出と決定木+出力ラベルの平滑化
- **CNN-Transformer**: 提案ネットワークにおいてCNNの事前学習をしない場合
- **PreCNN-Transformer**: 提案手法

GBDT+smoothing は行動認識において広く使われる勾配ブースティング決定木 (GBDT) を用いた手法である。この手法では 2.56 [s] のセグメントごとに加速度・角速度データを区切り、統計的な特徴量を抽出して決定木を訓練する。また、2.56 [s] ごとの予測を時系列順に並べ、直近 20 個の最頻値を採用していく平滑化処理を行う。

CNN-Transformer は提案したネットワークにおいて、CNN の表現学習を行わず、正解データを用いてネットワーク全体をエンドツーエンドで訓練する手法である。一方提案手法の PreCNN-Transformer は CNN の表現学習を事前に行い、正解データで後続の層のみを訓練する。

4.4.2 学習設定

学習設定の詳細を本項に示す。SimCLR を用いた CNN の表現学習では、バッチサイズ 512、学習率 1.0×10^{-3} のもと、NT-Xent 損失を用いて 100 エポック学習した。最適化には Adam[23] を使用した。

作業認識モデルの学習では、全ての手法において、訓練データの 90% を訓練用、残りの 10% を検証用として分割し、検証データを用いてハイパーパラメータを調整した。CNN-Transformer, PreCNN-Transformer の訓練では、バッチサイズ 16、学習率 1.0×10^{-4} 、最大エポック数 20 の設定の

もと、検証損失が連続で改善しない場合に学習を停止する early stopping を適用してモデルを学習した。損失にはラベルの不均衡を考慮して重み付き交差エントロピー損失を用いて、最適化には Adam を使用した。CNN-Transformer は CNN をランダムに初期化してネットワーク全体を訓練し、PreCNN-Transformer では SimCLR によって表現学習された CNN を用いて、CNN 部分は重みを固定して後続の層のみを訓練した。

統計モデルである GBDT では、各セグメントに対して、時間領域の特徴量 (平均、標準偏差、最大値、最小値など) および周波数領域の特徴量 (平均周波数、バンドパワー) を含む、計 128 次元の統計特徴量を抽出して分類に用いた。学習率、木の深さ、サブサンプリング比などの主要なハイパーパラメータは、検証データに基づいて調整した。

4.5 評価結果

10 人分の正解データを用いて交差検証を行った。各モデルの予測スコアは表 2 のとおりである。表中の値は、全被験者に対する 3 分類予測の重み付き平均を示す。また、被験者ごとの F1 スコアの分布を図 7 に示す。提案手法の PreCNN-Transformer は F1 スコア 0.89 を達成し、全ての指標においてベースラインを上回る結果となった。特に CNN の表現学習をしない単純な CNN-Transformer と比較して、F1 スコアが約 7 ポイント向上した。また、ベースラインでは一部被験者で性能が大きく低下するのに対し、提案手法では全被験者で F1 スコア 0.7 以上を維持していた。

続いて、交差検証における全テスト被験者の分類結果を合計した混同行列を図 8 に示す。図中の割合は Recall を表す。まず、3 つのモデルにおいて検品作業の Recall は全て 0.85 以上、搬送作業は全て 0.82 以上と比較的良好な認識結果が得られた。一方、ベースラインによる予測では特に仕分け中の誤認識が目立ち、Recall は GBDT+smoothing, CNN-

表 2: 各モデルの予測スコア

モデル	Precision	Recall	F1
GBDT + Smoothing	0.85	0.83	0.83
CNN-Transformer	0.82	0.82	0.82
PreCNN-Transformer	0.89	0.89	0.89

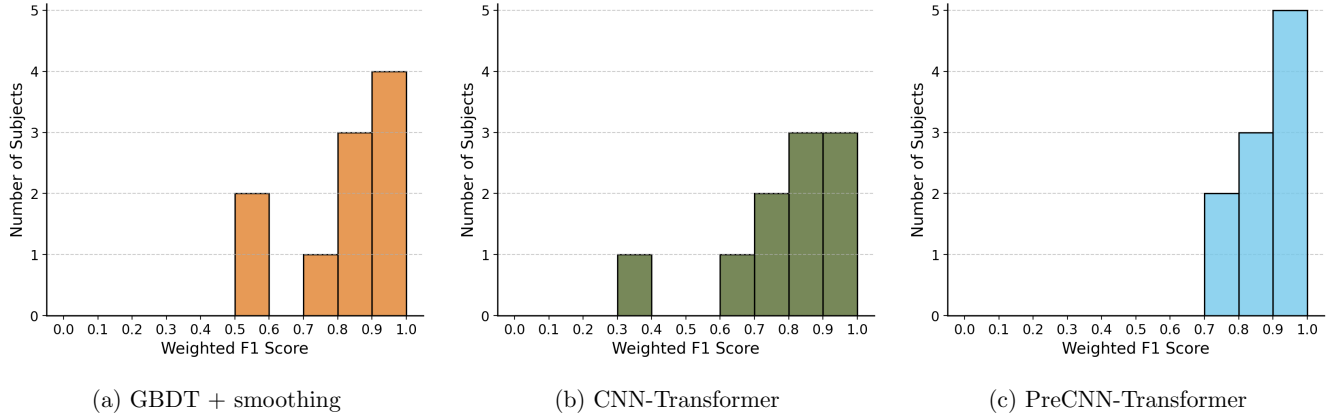


図 7: 各予測モデルの被験者別 F1 スコアの分布

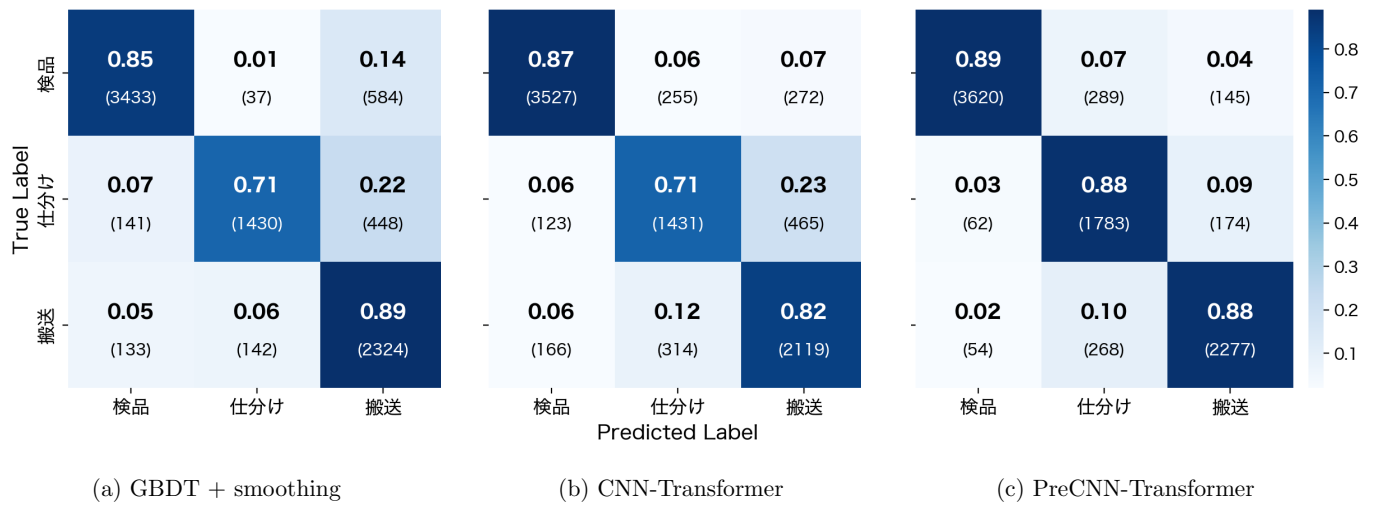


図 8: 各予測モデルの混同行列

Transformer とともに 0.71 となった。提案手法の PreCNN-Transformer ではこれが 0.88 まで向上し、明確な認識性能の向上が見られた。

4.6 結果の考察

ベースラインによる予測では、比較的動作の少ない検品作業や、動作のバリエーションが少ない搬送作業の認識は良好であったが、仕分け作業においては誤認識が多くみられた。一方、表現学習した CNN を用いた提案手法では、仕分け作業の認識性能が大きく改善した。仕分け作業は、「屈む」「持ち上げる」「運ぶ」などの多様な動作を含み、また体を大きく使うことから人による動作の個人差も大きい。このような作業に対しては、訓練データが少ない場合、モデルが限られた教師データに過剰に適合し、汎化性能が低下する傾向がある。一方、SimCLR によって事前に表現学習した CNN を用いると、認識モデルは正解データから汎用的な動作と作業の関係性のみを学習し、特徴抽出の層が訓練データに特化しないため、テストデータでも全体的に安定した性能が得られたと考えられる。

次に、提案手法の課題を考察する。作業認識モデルは以下に示すような場合において誤認識をしていた。

- 検品中に包装ゴミを整理する動作を仕分けと認識
- 仕分け中に移動する動作の一部を搬送と誤認識
- 搬送中に荷物を整理する動作を仕分けと誤認識

誤認識をしている場面では別作業の類似動作を行なっている場合が多く、提案手法ではこれらの区別が課題となった。

また、提案手法では実施中の作業を検品・仕分け・搬送の 3 分類のいずれかとして予測したが、実際の作業場では、いずれの作業にも該当しない期間も存在する。実用に向けては、このような期間を検出する手法の検討も必要となる。

5. まとめ

本研究では、スマートフォンに搭載された加速度・角速度センサを活用し、倉庫内における代表的な作業（検品・仕分け・搬送）を対象とした自動認識手法を提案した。倉庫内の作業のような複数の動作を含む行動を認識するため、動作特徴を抽出する CNN と、その系列からコンテキ

ストを捉えて作業を分類する Transformer Encoder からなる深層学習モデルを提案した。さらに、SimCLR による表現学習を活用し、大規模なラベルなしセンサデータから個人差やセンサ装着条件の違いに対して頑健な特徴表現を獲得してモデル全体の汎化性能を高めた。

実際の物流倉庫で収集したセンサデータに基づく交差検証の結果、提案手法はベースラインよりも高い認識性能 (F1 スコア 0.89) を達成し、特に CNN を表現学習しない場合と比べて約 7 ポイントのスコア向上が確認された。また、全被験者において F1 スコアが 0.7 を上回る安定した予測性能が得られており、提案手法の高い汎化性能が確認された。今後の課題としては、作業間の類似動作の識別や、いずれの作業にも該当しない期間の検出が挙げられる。

謝辞

本研究は国立研究開発法人新エネルギー・産業技術総合開発機構 (NEDO) の委託業務 (JPNP23003), 科研費挑戦的研究 (開拓) JP22K18422, トラスコ中山株式会社に支援いただいている。

参考文献

- [1] “経済産業省: 令和 5 年度 電子商取引に関する市場調査, 経済産業省 (オンライン),” <https://www.meti.go.jp/press/2024/09/20240925001/20240925001-1.pdf>.
- [2] Kaiblinger, A. and Woschank, M. : State of the Art and Future Directions of Digital Twins for Production Logistics: A Systematic Literature Review, *Applied Sciences*, Vol.12, No.2, Article 669 (2022).
- [3] Leng, J., Zhang, H., Yan, D. et al. : Digital twin-driven manufacturing cyber-physical system for parallel controlling of smart workshop, *Journal of Ambient Intelligence and Humanized Computing*, Vol.10, No.3, pp.1155–1166 (2019).
- [4] Yokoyama, K., Katayama, S., Urano, K. et al. : Digitization and Analysis Framework for Warehouse Truck Berth, *Proc. Int. Conf. on Mobile Computing and Ubiquitous Network (ICMU)*, pp.1–4 (2023).
- [5] Kano, K., Yoshida, T., Hayashida, N. et al. : Smartphone Localization with Solar-Powered BLE Beacons in Warehouse, *Proc. Int. Conf. on Human-Computer Interaction (HCII), Lecture Notes in Computer Science (LNCS)*, Vol.13303, pp.291–310 (2022).
- [6] Niemann, F., Lüdtke, S., Bartelt, C. and ten Hompel, M. : Context-Aware Human Activity Recognition in Industrial Processes, *Sensors*, Vol.22, No.1, Article 134 (2022).
- [7] Straczekiewicz, M., James, P. and Onnela, J.-P. : A systematic review of smartphone-based human activity recognition methods for health research, *NPJ Digital Medicine*, Vol.4, Article 148 (2021).
- [8] Ferrari, A., Micucci, D., Mobilio, M. et al. : Trends in Human Activity Recognition Using Smartphones, *Journal of Reliable Intelligent Environments*, Vol.7, No.3, pp.189–213 (2021).
- [9] Chakma, A., Faridee, A. Z. M., Ghosh, I. et al. : Domain Adaptation for Inertial Measurement Unit-based Human Activity Recognition: A Survey, *arXiv preprint arXiv:2304.06489* (2023).
- [10] Yuan, H., Chan, S., Creagh, A. P. et al. : Self-supervised learning for human activity recognition using 700,000 person-days of wearable data, *npj Digital Medicine*, Vol.7, Article 91 (2024).
- [11] Lara, O. D. and Labrador, M. A. : A Survey on Human Activity Recognition using Wearable Sensors, *IEEE Communications Surveys & Tutorials*, Vol.15, No.3, pp.1192–1209 (2013).
- [12] Ordóñez, F. and Roggen, D. : Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition, *Sensors*, Vol.16, No.1, Article 115 (2016).
- [13] Dirgová Luptáková, I., Kubovčík, M. and Pospíchal, J. : Wearable Sensor-Based Human Activity Recognition with Transformer Model, *Sensors*, Vol.22, No.5, Article 1911 (2022).
- [14] Bento, N., Rebelo, J., Barandas, M. et al. : Comparing Handcrafted Features and Deep Neural Representations for Domain Generalization in Human Activity Recognition, *Sensors*, Vol.22, No.19, p.7324 (2022).
- [15] Logacjov, A., Alzantot, M. and Stantic, B. : Self-supervised Learning for Accelerometer-based Human Activity Recognition: A Survey, *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, Vol.8, No.4, pp.1–42 (2024).
- [16] Saeed, A., Ozcelebi, T. and Lukkien, J. : Multi-task Self-Supervised Learning for Human Activity Detection, *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, Vol.3, No.2, pp.1–30 (2019).
- [17] Xu, H., Zhang, Y., Srivastava, M. et al. : LIMU-BERT: Unleashing the Potential of Unlabeled Data for IMU Sensing Applications, *Proc. ACM Conf. on Embedded Networked Sensor Systems (SenSys)*, pp.103–116 (2021).
- [18] Chen, T., Kornblith, S., Norouzi, M. and Hinton, G. : A Simple Framework for Contrastive Learning of Visual Representations, *Proc. International Conference on Machine Learning (ICML)*, pp.1597–1607 (2020).
- [19] Tang, C. I., Perez-Pozuelo, I., Spathis, D. and Mascolo, C. : Exploring Contrastive Learning in Human Activity Recognition for Healthcare, *arXiv preprint arXiv:2011.11542* (2021).
- [20] van den Oord, A., Li, Y. and Vinyals, O. : Representation Learning with Contrastive Predictive Coding, *arXiv preprint arXiv:1807.03748* (2018).
- [21] Haresamudram, H., Essa, I. and Plötz, T. : Contrastive Predictive Coding for Human Activity Recognition, *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, Vol.5, No.2, pp.1–26 (2021).
- [22] Xu, C., Zhang, L. and Wang, M. : Augmentation Robust Self-Supervised Learning for Human Activity Recognition, *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp.1–5 (2023).
- [23] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, (2014).