EgoCraft：ネックマウント型カメラを用いたショーケース組み立て作業の一人称映像データセット

三戸　智浩†　　竹内　太法†　　佐藤　可直†　　関　　喜史†

† フェアリーデバイセズ株式会社　〒 113–0034 東京都文京区湯島 2–31–22 湯島アーバンビル
E-mail: †{t.mito,t.takeuchi,sato,y.seki}@fairydevices.jp

あらまし　近年の画像処理技術の進展を背景として，一人称映像理解技術の産業現場への導入が期待されている．本研究では，産業現場向けの一人称動画理解を目的として，ポップアップショーケースの組立作業をネックマウント型カメラを用いて撮影した一人称視点映像データセット EgoCraft を構築した．ネックマウント型カメラは，長時間装着に適し，作業の邪魔になりにくいという利点がある．本データセットは，以下の特徴を持つように設計されている：(1) 作業者の姿勢変化・移動を伴う，(2) 作業手順が定められている，(3) 作業ミスを含む，(4) 同一作業者の習熟度の向上を分析できる．30 人の作業者による合計 60 回の作業が収録されており，合計作業時間は 23 時間である．工程，および手順ミスのアノテーションが付与されており，行動認識・工程認識・誤作業検知モデルの学習・評価に使用可能である．これらの性質により，EgoCraft は産業現場向けの一人称動画理解の技術開発において有用であると期待される．

キーワード　一人称視点映像，データセット，行動認識

1 はじめに

近年，人手不足や働き方改革などの課題を背景として，デジタル技術の導入による産業分野における変革が加速している．特に，製造・建築・農林水産・物流・インフラ保守を始めとする現場作業を伴う産業分野においては，安全管理のための危険予兆・事故予知，生産性向上や技能継承のための作業のデータ化と解析，誤作業・作業漏れ防止や作業効率化のための作業支援，産業機器の動作不良や故障の検知など課題の解決のために，映像理解技術の活用が期待されている．産業分野における映像理解においては，文脈や状況の理解が重要であり，基礎的なタスクとして作業工程の認識が挙げられる．実際の現場作業においては，作業員がしばしば体勢を変えたり動き回りながら作業を行うため，ウェアラブルカメラによる一人称視点映像を理解する技術が重要となる．

一方で，産業現場における作業の映像データは機密性が高いため，一般には公開されていない場合が多い．また分野ごとに作業内容・環境・機器など要素の特殊性が高い傾向にあり，産業分野全体を網羅するデータセットを構築するには難しい．そのため，産業現場における作業の典型的な特徴を兼ね備えた一人称作業映像データセットが求められている．

本研究では，産業現場向けの一人称動画理解を目的として，新しい一人称映像データセット EgoCraft を構築した．EgoCraft はネックマウント型カメラでポップアップショーケースの組み立て作業を録画した映像データセットであり，産業現場における一人称視点映像データを想定した以下の特徴を持つ：(1) 作業者の姿勢変化・移動を伴う，(2) 作業手順が定められている，(3) 作業ミスを含む，(4) 同一作業者の習熟度の向上を分析できる．

アノテーションとして，組み立て作業の行動，および手順ミスの時系列ラベルが付与されている．したがって EgoCraft は時系列行動セグメンテーション，工程認識，手順ミス検知モデルの開発に使用可能である．本研究では，時系列行動セグメンテーションモデルの学習・評価を行った．さらに同一作業者による作業回数と手順ミスの回数および作業時間の変化について分析を行った．

本論文の構成は以下の通りである．2 節では関連研究を紹介する．3 節では EgoCraft の作業内容，動画収録方法，データ構成について述べる．4 節では時系列行動セグメンテーションの実験結果を説明する．5 節ではまとめと今後の展望を論じる．

2 関連研究

ウェアラブルカメラの普及に伴って，さまざまな一人称視点映像データセットが公開されている．調理作業を収録したデータセットとして GTEA [1]，EPIC-KITCHENS [2]，EgoPER [3]，CaptainCook4D [4] が挙げられる．また，Charades-Ego [5]，Ego4D [6] は，日常生活における多様なアクティビティを撮影したデータセットである．これらのデータセットは日常的な動作を対象としているため，産業分野における作業とは性質が異なっている．

産業分野における作業，あるいは産業分野と類似した状況における作業の一人称映像データセットとして IKEA Assembly [7]，Assembly101 [8]，IndustReal [9]，MECCANO [10]，ENIGMA-51 [11]，FineBio [12]，Egocentric-10K [13]，EgoOops [14]，IndEgo [15] などが挙げられる．表 1 に EgoCraft との比較を示す．

表 1: 産業分野向け一人称視点映像データセット．HOI: Hand-object interaction，BB: Bounding box.

データセット	作業内容	カメラ形状	時間 (h)	シーケンス数	人数	姿勢変化・移動	作業ミス	既定手順	アノテーション	その他の特徴
IKEA Assembly	家具組立	ヘッドマウント	35	371	48	✓		✓	行動，物体のセグメンテーション，姿勢	
Assembly101	玩具組立	ヘッドマウント	513	362	53		✓		行動，作業ミス，ハンドポーズ	
IndustReal	玩具組立	ヘッドマウント	6	84	27				行動，組立状態，工程，ハンドポーズ，視線	
MECCANO	玩具組立	ヘッドマウント	7	20	20				行動，物体・手の BB，HOI，視線	
ENIGMA-51	電気基板修理	ヘッドマウント	22	51	19				行動，物体・手のセグメンテーション，ハンドポーズ，HOI	音声・拡張現実による指示
FineBio	生物実験	ヘッドマウント	14	226	32			✓	プロトコル・工程・動作，物体・手の BB，HOI	階層的アノテーション
Egocentric-10K	工場内作業	ヘッドマウント	10,000	192,900	2,138	✓			なし	
EgoOops	電気回路，化学実験，玩具組立，工作	ヘッドマウント	6	50	4		✓	✓	工程，作業ミス	テキストによる指示
IndEgo	組立，修理，運搬，木材加工	アイウェア	197	3,460	20	✓	✓		行動，工程，作業ミス，ハンドポーズ，視線，ナレーション，要約テキスト	作業者間の協働作業
EgoCraft	ショーケース組立	ネックマウント	23	60	30	✓	✓	✓	行動，作業ミス	

カメラ形状：多くの一人称視点映像データセットではヘッドセット型のカメラが使用されている．また IndEgo ではアイウェア型のカメラが用いられている．これに対して EgoCraft ではネックマウント型カメラを採用した．ネックマウント型カメラは産業現場において作業の邪魔になりにくく，長時間の使用でもユーザーへの負担が少ないという利点がある．また安全管理の観点からも多様な産業現場において受け入れやすい．

姿勢変化・移動：Assembly101，IndustReal，MECCANO は玩具の車・バイクの組み立て， ENIGMA-51 は電気基板修理，FineBio は生物実験を撮影したデータセットである．これらの作業は多数の部品や道具を含む複数の工程により構成されるものであり，産業分野における作業と共通点がある．一方で，作業者の姿勢変化や移動は発生しない．これは EgoCraft が作業者の姿勢変化や移動による画角の大きな変化を含んでいるのと対照的である．

手順の自由度：既存の一人称視点映像データセットは，日常生活における多様なアクティビティ，あるいは手順が作業者に委ねられているような特定の作業を対象としているものが多い．一方で産業現場における作業では，工程の順序はマニュアルなどで既定されている上で，各工程における細かい操作の順序は作業者に委ねられていることが多い．EgoCraft は IKEA Assembly や FineBio と同様に，このような産業現場における作業の典型的な特徴を反映したデータセットになっている．

作業ミス：特定の作業を対象とした映像理解技術の活用において重要な目的の 1 つとして作業ミスの検知が挙げられる．このような技術開発のためには作業ミスのアノテーションが付与されたデータセットが有用である．既存の産業分野向けの一人称視点映像データセットは作業ミスを含まない，あるいはアノテーションが付与されていないものが多い．EgoCraft は Assembly101，EgoOops，IndEgo と同様に，作業ミスに関するアノテーションを含んでおり，作業ミス検知技術の開発に利用可能である．

作業者の習熟度：Assembly101， IndEgo などの既存のデータセットでは，習熟度のばらつきが発生するように作業者が選ばれている．EgoCraft においてはショーケースの組み立て作業が未経験であることを作業者の条件としたにも関わらず，所要時間・手順ミス回数・マニュアル確認回数にばらつきが見られた．本研究が対象とした作業は特別な技能や経験を必要とせず，数回程度の実施により作業を円滑に実施できるようになるように設計されている．したがって，作業の繰り返しに伴う所要時間・手順ミス回数・マニュアル確認回数の軽減を容易に観測可能である．

3 データセット

3.1 作業内容

本研究では，ポップアップショーケースの組み立て作業を撮影した．このポップアップショーケースは本体フレーム，チャネルバー，スクリーン，展示ボックス，天板で構成されていて，

(a) スクリーンなし　　(b) スクリーンあり

図 1: ポップアップショーケースの外観

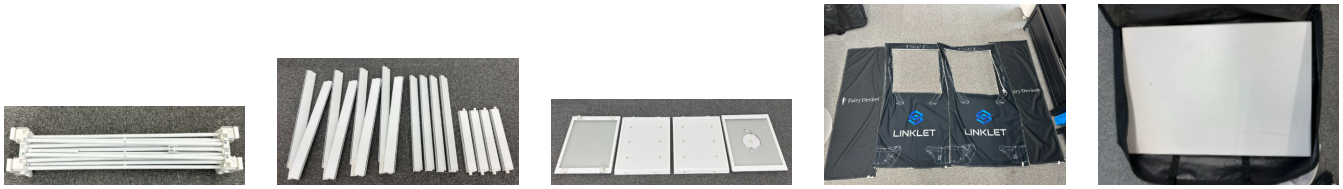
工具を使わずに組み立て・解体が可能である．組み立て作業は特定の技術や経験が要求されるものではない．EgoCraft では，マニュアルに従って作業者が 1 人でポップアップショーケースの部品を取り出し，組み立てを行い，所定の場所に移動させ，展示用デバイスとクリップライトを設置する一連の作業をウェアラブルカメラで撮影した．図 1 は EgoCraft で用いたポップアップショーケース，図 2 はポップアップショーケースの部品，図 3 は展示用の部品の外観写真である．展示品には動画撮影に用いたものと同型のデバイスを使用した．

作業手順は表 2 の通りである．組み立て作業の順番には多少の自由度があるが，EgoCraft では作業者が準拠すべき手順を定め，作業マニュアルに明記した．例えば，必要になった部品をその都度取り出すことも可能であるが，作業マニュアルではすべての部品を最初に取り出しておくように定めている．展示ボックスの組み立ては本体フレームへの取り付け前であればどのタイミングでも実施可能であるが，本体フレームへの取り付け直前に行うこととした．なお，展示ボックスを本体フレームに取り付ける際に上面のチャネルバーが邪魔になるため，チャネルバーは上面と上面以外に分けて取り付けを行う必要がある．工程の遷移時に必ず印刷された作業マニュアルを読み，次工程の作業内容を確認することとした．さらに，作業者の必要に応じて任意の時点で作業マニュアルを確認してよいこととした．

産業分野における一人称映像理解技術の開発を目的として，EgoCraft の作業は以下の特徴を持つように設計されている．

姿勢変化・移動　扱う部品のサイズが比較的大きいため，固定された机上における手元の作業ではなく，さまざまな姿勢変化や移動が発生する．例えば，作業者は屈む・しゃがむ・座り込むなど姿勢を変化させたり，ショーケースを上下左右さまざまな方向から覗き込んだり，ものを持ったまま会議室内を移動したりする．そのため一人称視点映像において大きな画角の変化が発生する．このような画角の変化に伴ってショーケース・展示用の部品・マニュアルなど物体がフレームアウトする状況が頻繁に発生する．

既定手順　作業マニュアルにおいて工程の手順が定められている．一方で，個々の部品の具体的な取り付け・設定方法の詳細は指示されていない．したがって各工程における操作の順番には自由度があり，作業者に委ねられている．



(a) 本体フレーム (b) チャネルバー (c) 展示ボックス (d) スクリーン (e) 天板

図 2: ポップアップショーケースの部品

(a) 展示用デバイス (b) クリップライト

図 3: 展示用の部品

表 2: 作業手順

工程	作業内容
部品取り出し	袋や箱から部品を取り出す
フレーム組み立て	本体フレームを広げ，ロックを固定する
チャネルバー取り付け (上面以外)	上面以外のチャネルバーを取り付ける
展示ボックス取り付け	展示ボックスを組み立て，本体フレームに取り付ける
チャネルバー取り付け (上面のみ)	上面のチャネルバーを取り付ける
スクリーン取り付け	スクリーンを取り付ける
運搬	設置場所へショーケースと展示用部品を移動させる
天板設置	天板を本体フレームに載せる
展示物設置	展示用ウェアラブルカメラを設置する
照明設置	クリップライトを設置し，点灯させる

作業ミス　作業ミスがあると作業完了が困難になり，一部の作業のやり直しが必要となる場合がある．例えば展示ボックスより前に上面チャネルバーやスクリーンを取り付けてしまうと，展示ボックスが設置できず，チャネルバーやスクリーンの取り外しが必要となる．またチャネルバーが適切に取り付けられていないと，後続の工程の最中にチャネルバーが外れてしまい，スクリーンの取り付けができなくなる可能性がある．本作業に不慣れ，あるいは不注意な作業者はこのような作業ミスを起こしやすい．

習熟度の向上　本作業は特別な技能を必要とせず，数回程度繰り返すことで，作業を円滑に実施できるようになる．そのためポップアップショーケース組み立て作業を行ったことがない作業者が複数回同じ作業を行うことで，習熟度の向上を観測可能である．

表 3: 作業者ごとの作業回数の分布

作業回数	1	2	3	4
人数	9	13	7	1

3.2　データ収集

30 人の作業者による合計 60 回の作業を収録した．作業者ごとの作業実施回数（シーケンス数）は 1〜4 回とした．作業回数の分布を表 3 に示す．作業者の募集において，ポップアップショーケースの組み立て，および類似の作業の経験を持たないことを条件とした．作業実施前に印刷した作業マニュアルを渡し，その内容を口頭で説明する時間を設けた．口頭説明は監督者（本論文の著者）が行った．組み立て済みショーケース，ショーケースの部品，展示用の部品の模式図・写真を作業マニュアルに掲載したが，実物を用いた説明は実施しなかった．各作業者は，監督者を含む他人の助けを借りずに，全工程を作業者が 1 人で行うこととした．

3 つの会議室を作業場所としてランダムに割り当てた．同じ会議室内で組み立て作業用の場所とショーケースの設置場所を別々に定めた．

動画撮影には THINKLET[1] を用いた．このデバイスは 1 個の RGB カメラと 5 チャネルマイクを内蔵したネックマウント型のウェアラブルデバイスである．図 4 に外観図を示す．動画の解像度は 1080 × 1920（幅 × 高さ），フレームレートは 30 fps である．

ネックマウント型カメラは，ユーザーへの負担が少なく長時間の利用に適しているという利点がある．また，産業現場において作業の邪魔になりにくく，安全管理の観点からも受け入れやすい．ヘッドマウント型と比較すると，視線データの収集には対応しておらず，画角が必ずしもユーザーの視野と一致しない．一方で，ネックマウント型のカメラは頭部の動きによらず作業対象物体をフレーム内に捉えていることが多く，画角のぶれが少ない傾向にある．

3.3　アノテーション

撮影した動画に（1）時系列行動クラス，(2）手順ミス，の 2

1：https://mimi.fairydevices.jp/technology/device/thinklet/

図 4: ウェアラブルカメラデバイス THINKLET

種類のアノテーションを付与した．すべてのアノテーション作業は本論文の著者が実施した．

3.3.1 行　　動

本研究において採用した行動クラスの定義を表 4 に示す．作業マニュアル中の工程（表 2）をベースとして，以下の変更を加えている．工程においては上面と上面以外のチャネルバーの取り付けは区別されていたが，行動においては単一のクラス「チャネルバー取り付け」に統合されている．さらに「マニュアル確認」「その他」の 2 つの行動クラスが追加されている．工程の遷移時にマニュアルを確認するように指示されている上に，工程中でも必要に応じてマニュアル確認が許容されているため，マニュアル確認は任意の時点で発生し得る．どの行動にも該当しない時間フレームは「その他」の行動クラスに割り振られる．本研究で採用した行動クラスは，単位操作レベルではなく工程レベルの分類であると言える．

行動クラスは互いに排他的であり，時系列に対して密であると仮定した．すなわち，すべての時間フレームがいずれかの行動クラスに属するようにアノテーションを行った．ただし，行動クラスの境界を厳密に定める合理的な基準があるとは言いがたい．そこで以下の手順ですべての時間フレームに対して行動クラスを決定した．

- 始めに，映像データから作業者の意図が読み取れるフレームに行動ラベルを付与する．
- 次に，隣接する行動の時間的間隔が 2.0 秒以下の場合は，中間点を境界と定めて，前後の行動クラスを中間点まで拡張する．隣接する行動の時間的間隔が 2.0 秒を超える場合は，ラベルが欠損している区間は「その他」クラスに属するものとする．

3.3.2 手 順 ミ ス

行動クラスに加えて，手順ミスを表すアノテーションを付与した．手順ミスは既定の手順からの逸脱として観測される．ここで，作業クラス「マニュアル確認」「その他」は除外し，上面と上面以外のチャネルバーの取り付けの工程は区別しないものとする．ある工程から想定外の工程への遷移には，既定の工程をスキップしている場合と，実施済みであるべき工程に戻っている場合が考えられる．前者の遷移を「スキップ」，後者の遷移を「手戻り」と呼ぶこととする.「手戻り」が発生した場合に，正常な手順に「復帰」するまでの時間セグメントを「修正作業」と呼ぶこととする．修正作業と同じクラスの工程が過去に実施されていた場合，それらの時間セグメントを「不正作業」と呼ぶこととする．上記の手順にしたがって,「不正作業」「手戻り」「修正作業」「復帰」の 4 種類の作業ミスに関するアノテーションを付与した．特に「手戻り」の発生回数は，手順ミスの度合いを表していると言える．なお，本研究においては「スキップ」は観測されなかった．行動と手順ミスのアノテーションと併用することにより，各時点における組み立て作業の進捗具合を推定することが可能となる．

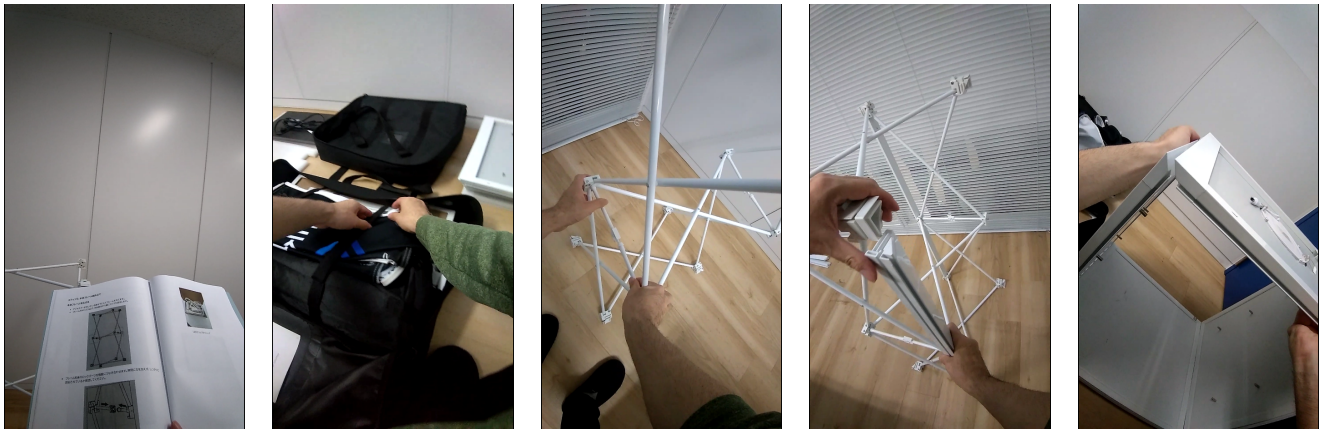
表 4: 行動クラス（11 クラス）

行動	作業内容
マニュアル確認	マニュアルを読む
部品取り出し	袋や箱から部品を取り出す
フレーム組み立て	本体フレームを広げ，ロックを固定する
チャネルバー取り付け	チャネルバーを取り付ける
展示ボックス取り付け	展示ボックスを組み立て，本体フレームに取り付ける
スクリーン取り付け	スクリーンを取り付ける
運搬	設置場所へショーケースと展示用部品を移動させる
天板設置	天板を本体フレームに載せる
展示物設置	展示用ウェアラブルカメラを設置する
照明設置	クリップライトを設置し，点灯させる
その他	その他の作業（休憩，部品の整理，次工程への遷移など）

3.4 分　　析

本研究においては，合計 23 時間の一人称視点作業動画を収録した．1 回あたりの所要時間は 11〜46 分（平均 23 分）であった．行動クラスの構成を表 5 に示す．「チャネルバー取り付け」「スクリーン取り付け」に多くの時間が費やされている一方で,「運搬」「天板設置」「展示物設置」「照明設置」は短時間で終了していることが分かる．また「マニュアル確認」は頻繁に発生している一方で，その他の行動は 1〜2 つのセグメントで終了していた．このように行動クラス間で発生回数・所要時間に不均衡が見られた．産業現場における実際の作業において行動クラスは均一であるとは限らないため，EgoCraft は目的に則した特徴を持っていると言える．

さらに作業の繰り返しに伴う習熟度の変化を分析した．図 6, 7，8 は各作業回数（1〜4 回）ごとのシーケンスにおける作業時間，手戻り発生回数，マニュアル確認回数の分布を図示したものである．これらの値は小さい方が作業を円滑に実施できており，作業者の習熟度が高いと言える．ただし工程遷移時にマニュアル確認をするように指示をしているため，指示通りに作業していれば 10 回以上マニュアル確認が発生する．これらの分析結果において，1 回目の作業では作業者ごとの習熟度のばらつきがあることが確認された．2〜4 回と作業を繰り返すにつれ，作業が効率的に行われていることが観測できる．なお，4 回目を実施したのは 1 人の作業者のみであるため，統計的に信頼できる結果であるとは言えないことに注意が必要である．



(a) マニュアル確認 (b) 部品取り出し (c) フレーム組み立て (d) チャネルバー取り付け (e) 展示ボックス取り付け

(f) スクリーン取り付け (g) 運搬 (h) 天板設置 (i) 展示物設置 (j) 照明設置

図 5: 各行動クラスのサンプル画像

表 5: 行動クラスの構成

行動	時間比率	セグメント数	平均継続長 (s)
マニュアル確認	18.6%	1067	14.7
部品取り出し	9.0%	185	41.2
フレーム組み立て	1.8%	92	16.9
チャネルバー取り付け	24.0%	322	62.9
展示ボックス取り付け	8.3%	158	44.7
スクリーン取り付け	29.2%	118	208.7
運搬	0.9%	64	12.1
天板設置	0.8%	67	10.1
展示物設置	0.6%	66	7.8
照明設置	2.4%	91	22.9
その他	4.0%	371	9.2

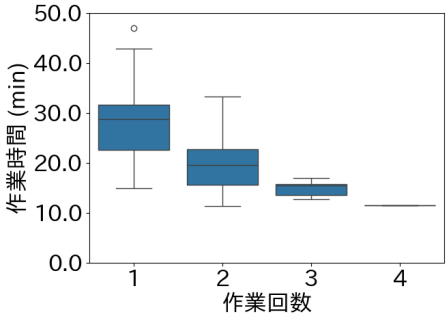


図 6: 作業反復による作業時間の変化

4 実　　験

4.1 実験設定

EgoCraft の行動時系列データ (temporal action segmentation: TAS) の特徴を明らかにするために，時系列行動セグメンテーションモデルの学習・評価実験を行った．

特徴量抽出には Kinetics-400 データセットで事前学習された Inflated 3D (I3D) モデル [16] を用いた．抽出された I3D 特徴量の次元は 2048，フレームレートは 30 fps である．時系列セグメンテーションモデルとして ASFormer [17] を採用した．モデル構造のパラメータ，学習に関するハイパーパラメータの設定値は [17] に準拠した．表 6 に主要なモデル構造のパラメータの設定値をまとめた．モデルの学習・評価時には特徴量と正解ラベルのフレームレートを係数 2 でダウンサンプリングした上で使用した．

損失関数はクラス分類のクロスエントロピーと，過剰セグメンテーションにペナルティを与える truncated mean square error (T-MSE) [18] の和である．クラス不均衡への対策としてクロスエントロピー損失にはクラス頻度の逆数の平方根に基づ

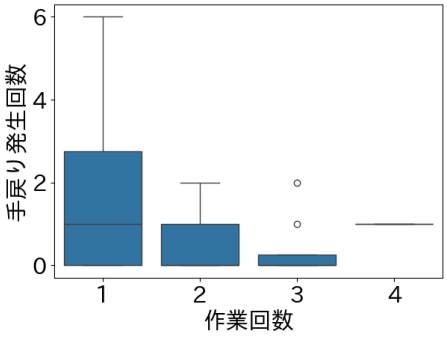


図 7: 作業反復による手戻り回数の変化

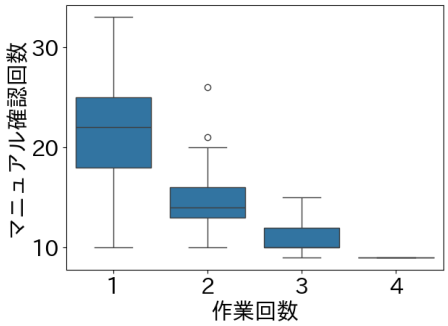


図 8: 作業反復によるマニュアル確認回数の変化

表 6: ASFormer のモデル構造に関するパラメータ設定

入力特徴量の次元	2048
デコーダの段数	4
エンコーダ・デコーダのブロック数	10
エンコーダ・デコーダの特徴量次元	64
自己注意機構のヘッド数	1

く重み付けを導入した．最適化手法には Adam [19] を使用し，学習率は 5×10^{-4}，バッチサイズは 1, T-MSE に対するウェイト λ は 0.25 に設定した．8 分割交差検証によりモデルの学習・評価を行った．データ分割はシーケンス（すなわち一連の組み立て作業の録画）単位で行った．7 セットを学習に，1 セットを評価に用いた．検証用のセットは設けず，200 エポック学習後のモデルを評価に用いた．

4.2 実験結果

時系列行動セグメンテーションの実験結果を表 7 に示す．F1@{10, 25, 50} は intersection over union の閾値が 10%, 25%, 50% の時にセグメント単位の F1 スコアを表す [20]．11 クラス条件では評価用動画クリップ中の全フレームを用いて評価指標の計算を行い，10 クラス条件では正解ラベルが「マニュアル確認」である時間フレームを除外して評価指標の計算を行った．10 クラス条件において，予測クラスが「マニュアル確認」である時間フレームは不正解とみなした．どちらの条件でも評価対象のモデルは同一であり，学習用動画クリップの全フレームを用いて学習されたものである．

11 クラス条件における重み付き正解率は 0.71 であり，フレーム単位で見ると一定水準の精度で行動が認識できている．

表 7: 時系列行動セグメンテーションの評価結果

	11 クラス	10 クラス
重み付き正解率	0.71	0.81
重みなし正解率	0.57	0.60
マクロ平均 F1	0.51	0.61
編集スコア	0.50	0.52
F1@10	0.44	0.48
F1@25	0.36	0.42
F1@50	0.22	0.26

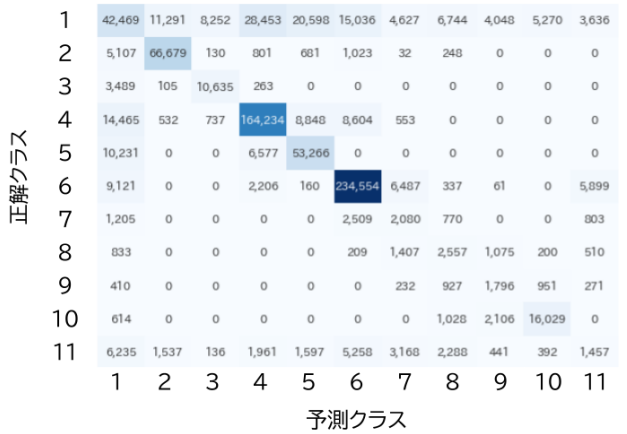


図 9: 時系列行動セグメンテーションの混同行列

これと比較して重みなし正解率は 0.57 に留まっており，クラス不均衡の影響やクラス間で認識の難易度の差があることが示唆される．F1@50 が 0.22 と低い値であることから，セグメント境界，すなわち行動の切り替わりのタイミングの推定が不正確であることが分かる．10 クラス条件，すなわち「マニュアル確認」クラスを除外した評価の方が 11 クラス条件より高いスコアが達成された．この結果は突発的に短時間だけ発生する「マニュアル確認」の認識が難しいことを示唆している．図 9 に示した混同行列から，継続時間が短い「マニュアル確認」「運搬」「天板設置」「照明設置」は検知が難しいことが分かる．特に「マニュアル確認」の誤認識が顕著であるが，この理由として他の行動セグメントの中で散発的に「マニュアル確認」が発生することや，マニュアルを読んでいる間は映像に動きがないため画角によっては状況が分かりにくいなどの要因が考えられる．後者は，ネックマウント型のカメラではユーザーの視野を追えないという特性に起因している．

図 10 はタイムラインの例である．正解ラベル（上段）と比較して，予測ラベル（中段）は工程の遷移は概ね正しく捉えているものの，時間アライメントがずれている箇所が散見される．すなわち ASFormer モデルによる時間アライメントの認識が不正確であり，正しいアライメントに修正する仕組みの必要性が示唆される．

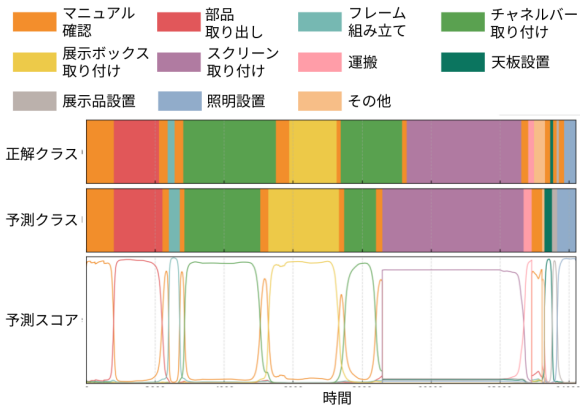


図 10: タイムラインのサンプル

5 まとめと展望

本研究では産業現場における一人称視点映像理解を目的として新しいデータセット EgoCraft を構築した．このデータセットはポップアップショーケースの組み立て作業をネックマウント型のウェアラブルカメラで撮影したものであり，(1) 作業者の姿勢変化・移動を伴う，(2) 作業手順が定められている (3) 作業ミスを含む (4) 同一作業者の習熟度の向上を分析できる，などの特徴を持つ．本研究では 11 クラスの行動の時系列ラベル，および手順ミスのアノテーションを付与した．

作業の繰り返しに伴って，作業者の習熟度が向上していくことを，作業時間・手順ミスの回数・マニュアル確認の回数の観点から検証した．さらに ASFormer による時系列行動セグメンテーションの実験を行った．実験結果から行動の時系列パターンは概ね正しく予測できるものの，継続時間が短い行動クラスの誤認識が多く，時間アライメントが不正確であるという問題が明らかになった．産業現場における一人称視点映像理解技術の開発において，本データセットが有用であると期待される．

今後の課題として以下のような点が挙げられる．ポップアップショーケースの組み立て作業に加えて解体作業の撮影を実施しており，アノテーションが必要な状態である．組み立て作業と解体作業は短時間の映像からは区別が難しく，長時間の時系列パターンの理解の重要度が増すと考えられる．さらに，類似したショーケースも対象とすることで，より多様な部品・操作・工程をデータセットに含めることも有用である．また，本研究で用いた行動の定義は基本的に工程をベースとしているが，一部のクラスは操作に近いものであった．工程レベルと操作レベルの 2 階層のクラス構造を定義し，アノテーションを付与することも今後の研究の方向性の 1 つである．階層的アノテーションにより，詳細な作業ミスの分析が可能となる．多様な作業シナリオやアノテーションへの拡充や分析・評価を行った上で，本データセットの公開を計画している．

謝　　辞

本研究の一部は東京都による「ゼロエミッション東京の実現等に向けたイノベーション促進事業」の助成を受けて実施されたものである．

文　　献

[1] A. Fathi, X. Ren, and J. M. Rehg, “Learning to Recognize Objects in Egocentric Activities,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3281–3288, 2011.

[2] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, “Scaling Egocentric Vision: The EPIC-KITCHENS Dataset,” in Proc. European Conference on Computer Vision (ECCV), pp. 720–736, 2018.

[3] S. P. Lee, Z. Lu, Z. Zhang, M. Hoai, and E. Elhamifar, “Error Detection in Egocentric Procedural Task Videos,” in Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 18655—18666, 2024.

[4] R. Peddi et al., “CaptainCook4D: A Dataset for Understanding Errors in Procedural Activities.” in Proc. of Neural Information Processing Systems (NeurIPS), 2024.

[5] G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, and K. Alahari, “Actor and Observer: Joint Modeling of First and Third-Person Videos,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7396–7404, 2018.

[6] K. Grauman et al., “Ego4D: Around the World in 3,000 Hours of Egocentric Video,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18995–19012, 2022.

[7] Y. Ben-Shabat, X. Yu, F. Saleh, D. Campbell, C. Rodriguez-Opazo, H. Li, and S. Gould, “The IKEA ASM Dataset: Understanding People Assembling Furniture Through Actions, Objects and Pose,” in Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 846–858, 2021.

[8] F. Sener, D. Chatterjee, D. Shelepov, K. He, D. Singhania, R. Wang, and A. Yao, “Assembly101: A Large-Scale Multi-View Video Dataset for Understanding Procedural Activities,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 21096–21106, 2022.

[9] T. J. Schoonbeek, T. Houben, H. Onvlee, P. H. N. de With, and F. van der Sommen, “IndustReal: A Dataset for Procedure Step Recognition Handling Execution Errors in Egocentric Videos in an Industrial-Like Setting,” in Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 4353–4362, 2024.

[10] F. Ragusa, S. Livatino, A. Furnari, and G. M. Farinella, “The MECCANO Dataset: Understanding Human-Object Interactions from Egocentric Videos in an Industrial-like Domain,” in Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021.

[11] F. Ragusa, R. Leonardi, M. Mazzamuto, C. Bonanno, R. Scavo, A. Furnari, and G. M. Farinella, “ENIGMA-51: Towards a Fine-Grained Understanding of Human Behavior in Industrial Scenarios,” in Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 4549–4559, 2024.

[12] T. Yagi, M. Ohashi, Y. Huang et al., “FineBio: A Fine-Grained Video Dataset of Biological Experiments with Hierarchical Annotation,” International Journal of Computer Vision, Vol. 133, pp. 7352-–7367, 2025.

[13] Build AI, “Egocentric-10K,” Hugging Face Datasets (Dataset Card), 2025. Available: https://huggingface.co/datasets/builddotai/Egocentric-10K.

[14] Y. Haneji et al., “EgoOops: A Dataset for Mistake Action Detection from Egocentric Videos referring to Procedural

Texts," in Proc. IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2025.
[15] V. Chavan, Y. Imgrund, T. Dao, S. Bai, B. Wang, Z. Lu, O. Heimann, and J. Krüger, "IndEgo: A Dataset of Industrial Scenarios and Collaborative Work for Egocentric Assistants," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2025.
[16] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[17] F. Yi, H. Wen, and T. Jiang, "ASFormer: Transformer for Action Segmentation," in Proc. British Machine Vision Conference (BMVC), 2021.
[18] Y. A. Farha and J. Gall, "MS-TCN: Multi-Stage Temporal Convolutional Network for Action Segmentation," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[19] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in Proc. International Conference on Learning Representations (ICLR), 2015.
[20] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal Convolutional Networks for Action Segmentation and Detection," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.