5Q-01

自動車組立作業映像を対象とした時系列行動セグメンテーションの基礎検討

久保 莞太　玉城 大生 †　久冨 あすか　伊藤 浩隆　東園 雄太 ‡　小野 智司 §
鹿児島大学 †　トヨタ車体研究所 ‡　鹿児島大学 §

概要

近年，工場などにおける作業員の行動解析の要望が高まっており，自動車組立作業においても同様の解析が望まれている．このため本研究では，既存の複数の時系列行動セグメンテーション手法を本問題に適用し，その有効性を検証する．

1　はじめに

近年，製造業における人手不足と作業効率向上の観点から，組立作業の行動解析の要望が高まっている．行動解析を行うことにより，各作業に要する時間の計測の自動化や，作業者がマニュアルと同様の手順で作業を行っているかの確認を行うことが可能となる．そこで，深層学習ベースの時系列行動セグメンテーション，すなわち，映像を構成するフレーム単位で行動クラスの認識を行う技術が広く研究されている [1–5].

本研究では，自動車組立工場における作業員の行動解析に着目する．しかし，一般に時系列行動セグメンテーションタスクとして適用・公開されているデータセットは，行動する人物や行動に伴う物体が映像の画角内における主体となることが多い．これに対して自動車組立における映像は，作業者よりも自動車車体が大きく写りながら画角内を移動しているため，時系列行動セグメンテーション手法が作業者に注目することを阻害するおそれがある．

このため本研究では，複数の時系列行動セグメンテーション手法を自動車組立作業映像に適用することで，その有効性を検証する．4 種類の手法を比較した実験により，本問題においては少量のデータセットから適切に学習を行える ASFormer [4] の性能が最も高いことを確認した．

2　関連研究

時系列行動セグメンテーションに関する研究は広く研究されており，主に深層学習ベースの手法が多く提案されている．これらの時系列行動セグメンテーション手法は「料理」や「おもちゃの組立・分解」,「ネジ締め動作」[1] などのデータセットにおいて検証されている．しかし，これらの一般的なデータセットと比較して，本研究で対象とする自動車組立工場のような環境における適用例はない．

3　方法

3.1　解析対象となる映像の特徴

本研究では，自動車組立作業を行う作業員を撮影し，時系列行動セグメンテーションを試みる．対象とする作業映像は，ライン上を常に移動する車体に対してフロントドアとその付近の部品を取り付ける工程を対象としており，1 台の定点カメラで撮影する．本工程に対して 1 名の作業者が割り当てられており，1 名の作業員が行動する範囲はライン上の前後約 7 メートル程度と幅広いため，画角の広いカメラで撮影を行った．

本研究では，2 名の異なる作業者による作業を撮影し，それぞれ 10 本，5 本の映像を作成した．映像の長さは各々 1 分程度であり，映像の組立作業は「フェンダーを取り出す」,「車体にセットする」,「ボルトを締める」など，35 クラスの行動で構成される．また,「ボルトを作業台に戻す」など一部の映像にのみ含まれる行動が 4 クラスほど含まれる．

3.2　時系列行動セグメンテーションモデル

時系列行動セグメンテーション手法の多くは，図 1 に示すように RGB 映像特徴とオプティカルフロー

A Preliminary Study on Temporal Action Segmentation for Automobile Assembly Work Videos
† Kanta Kubo, Daiki Tamashiro, Kagoshima University
‡ Asuka Hisatomi, Hirotaka Ito, Yuta Higashizono, TOYOTA AUTO BODY Research & Development
§ Satoshi Ono, Kagoshima University

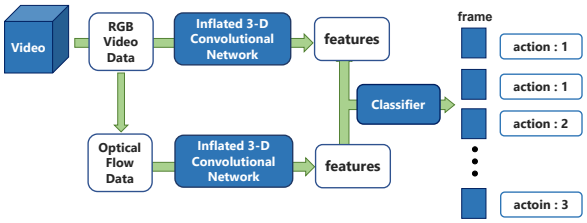


図 1　時系列行動セグメンテーション手法の構成と処理手順

特徴とを入力とする．すなわち，特徴抽出器として Inflated 3D ConvNet（I3D） [6] を用いることを前提としている．本研究では図 1 における行動分類器として，4 手法を用いて比較を行う．まず，Temporal Convolutional Network（TCN）ベースの手法である Multi-Stage TCN（MS-TCN） [2]，MS-TCN を改良した MS-TCN++ [3]，Transformer ベース手法である ASFormer [4] ，ASFormer のデコーダ部分を改良した Unified Video Action Segmentation model via Transformers（UVAST） [5] を使用する．

3.3　実験設定

3.1 節で述べた映像に I3D を適用して特徴抽出を行い，抽出した特徴量を行動分類器に入力してフレーム毎に作業動作の識別を行った．また，汎化性能を評価するために 15 本の映像に対して一つ抜き交差検証を行い，4 手法の正解率（Accuracy），編集スコア（Edit-score），F1 スコアの平均を算出した．

4　結果

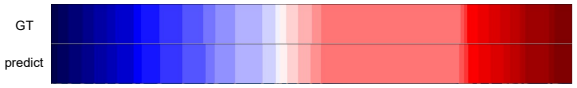
表 1 に各手法の正解率，編集スコア，F1 スコアの値を示す．結果から，ほぼ全ての手法において 9 割程度の正解率を示すことがわかる．これは，本実験において単一の車種かつ単一の作業工程を撮影したデータのみで学習を行っており，15 本の映像の間で各クラスの分布の偏りが少なかったためと考える．

4 手法のうち最も正解率が高かった ASFormer の認識結果の例を図 2（a）に示す．各クラスは正解において青から赤に徐々に変化するように色分けされており，滑らかに変化していない箇所がエラーとなる．図 2（a）よりほぼ全ての行動セグメントが正解に類似しており，詳細な行動を高精度で認識できたことがわかる．

一方，図 2（b）に ASFormer の正解率が最も低かったテスト映像の識別結果を示す．後半の行動において誤認識が生じており，前後の行動の順序が入れ替わっていた．これは，特にサンプル数の少ないマイナークラス周辺の行動で発生しており，クラス間の不均衡性が原因であると考える．

表 1　比較実験結果（テストデータにおける性能）

method	Accuracy	Edit-score	F1@10%	F1@25%	F1@50%
MS-TCN	90.06	95.95	95.97	93.99	88.12
MS-TCN++	89.98	95.95	96.17	94.45	88.11
ASFormer	91.33	96.99	97.16	96.29	90.66
UVAST	90.43	96.76	96.78	95.24	88.86



（a）　最も Accuracy が高かった Fold におけるテスト結果

（b）最も Accuracy が低かった Fold におけるテスト結果

図 2　ASFormer によるテスト結果の分析

5　結論

本研究では，自動車組立作業における時系列行動セグメンテーション手法の比較検討を行った．実験の結果，単一の車種かつ単一の作業工程における行動セグメントを高精度に識別することが可能であり，行動解析の自動化の可能性が示唆された．また，Transformer をベースとした，ASFormer の性能が最も高くなることを確認した．今後は，損失関数の工夫によるクラス不均衡性への対処を行う．

参考文献

[1] T. Kobayashi, et al. Fine-grained action recognition in assembly work scenes by drawing attention to the hands. SITIS, 2019.
[2] Y.A. Farha, et al. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. CVPR, pp. 3575–3584, 2019.
[3] S.J. Li, et al. Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. PAMI, 2020.
[4] F. Yi, et al. Asformer: Transformer for action segmentation. BMVC, p. 236, 2021.
[5] N. Behrmann, et al. Unified fully and timestamp supervised temporal action segmentation via sequence to sequence translation. ECCV, pp. 52–68, 2022.
[6] J. Carreira, et al. Quo vadis, action recognition? a new model and the kinetics dataset. CVPR, pp. 6299–6308, 2017.