テキスト情報を活用した組立作業のTemporal Action Segmentationにおけるアノテーションコスト削減検討

Investigation of Annotation Cost Reduction in Temporal Action Segmentation for Assembly Tasks Using Text Information

岸 祐輔　森田 亮　香西 諒也
Yusuke Kishi　Ryo Morita　Ryoya Kozai

コニカミノルタ株式会社
KONICA MINOLTA, INC.

In assembly manufacturing lines, work analysis is required to understand and improve productivity. This involves measuring work time and checking the accuracy of work procedures. Traditionally, these measurements have been done manually through visual observation, which requires significant effort and makes automation of measurement and analysis challenging. Recently, research on Temporal Action Segmentation, which segments work videos into basic action units, has been actively conducted, and many deep learning models have been proposed. However, these models require high-cost annotations with action labels for each frame, posing a challenge for practical implementation. In this study, we examined the performance of action segmentation in assembly processes using a CLIP-based model called Caseg, which can input action labels as linguistic information. By providing action labels extracted from assembly standards as prompts, we achieved a certain level of accuracy in inference. This suggests the potential to contribute to the practical implementation of action segmentation while reducing annotation costs, and we report these findings.

1. はじめに

製造業の組立作業現場では，生産性向上，不良品削減，不安全行動の把握を目的として，作業者の動作解析の需要がある．多くの組立作業現場には，作業順序や標準時間，注意事項等が記載された組立基準書と呼ばれるマニュアルが存在し，これに準拠した作業が実施されることが望ましい．特にライン生産方式を採用する組立作業現場では，複数の工程がライン上に配置されており，各工程において正しい順序と一定の作業時間を守りながら，円滑に作業が進行されることが理想的である．このような背景から，各工程で作業者の行動を管理・解析する必要性が高まっている．しかし，現状として多くの組立作業現場ではこのような動作解析が未だ目視によって行われており，コストや効率の面で大きな課題となっている．

近年，Temporal Action Segmentation(時系列行動セグメンテーション，以下 TAS) と呼ばれる，動画中の一連の行動を動作内容に応じたセグメントに分割し，動作順序や各動作の継続時間を推定する技術について深層学習手法を中心に多くの研究が行われている [1, 2, 3, 4]．TAS 研究の多くのモデルは教師あり学習によって学習されるが，このような学習を実行するためには動画の各フレームに対して動作ラベルを付与するという非常に高コストなアノテーションが必要となる．特に製造業の組立作業のような動画に対するアノテーションは専門知識が求められるため，多くのアノテーションデータが必要となる手法は産業応用に対する難易度が高い．

本研究ではそのような課題に対して，CLIP[5] をベースとした Caseg[3] という従来の TAS 手法とは異なり動作ラベルをテキストプロンプトとして与えることが可能な構造を持っているモデルを用いることで，組立作業動画の TAS タスクにおけるゼロショット推論が可能か，あるいは少量の教師ありデータのみで学習した際の推論精度が既存手法と比較して優位になるかどうかを検証した．

連絡先: 岸祐輔，コニカミノルタ株式会社，〒 100-7015 東京都千代田区丸の内 2-7-2 JP タワー，03-6250-2111，yusuke.kishi1@konicaminolta.com

2. 関連研究

教師あり TAS 手法には多くの研究が存在する [1, 2, 3, 4]．Temporal Convolutional Network(TCN) を用いた手法として代表的な MS-TCN[1] は時間畳み込み層を複数ステージ積み上げ，各ステージを経るごとに予測を洗練化していく構造を取っている．さらに，MS-TCN の複数ステージ構成を発展させ，前段を予測生成ステージ，後段を予測洗練化ステージに分けてネットワーク構成を改良した MS-TCN++[4] という手法も提案されている．このような従来の研究手法は，動作ラベル集合を単なるカテゴリとして与えるためラベルの高次の意味理解は低く，新しい動作に対するゼロショット推論やクロスデータセットでの推論能力が制限されることが多い．そのため新たな動作ドメインに対しては都度モデルを学習する必要性が生じる．

これに対して，CLIP を活用した Caseg という手法が提案されている．Caseg では，学習済みの CLIP モデルを利用することで，動作ラベルをプロンプトとして与えることが可能である．具体的には，CLIP を用いることで言語と画像の間でアラインされた特徴を抽出し，それをモデルの入力として活用している．さらに，Caseg ではフレーム単位で抽出された CLIP の画像特徴が動画の時系列情報を捉えるよう，Visual Temporal Prompt Module を導入している．このモジュールにより，フレーム間の時間的関連性を学習することが可能となる．また，CoOp[6] などで提案された prompt learning 手法を取り入れ，プロンプト部分に学習可能なテキストトークンを付与する構造を採用している．テキストトークンを学習することで，画像特徴との整合性を強化し，TAS タスクに特化したプロンプトを生成できるよう設計されている．このような技術によって，動作ラベル集合や動作シーンが異なる場合でもゼロショット推論や少量データでの学習のみで高い精度を実現できる可能性があり，ドメインごとに新たなモデルを一から学習させる必要がなくなることが期待され，組立作業など産業分野での動作解析

におけるアノテーションコストの問題に対して有用であると考えられる．

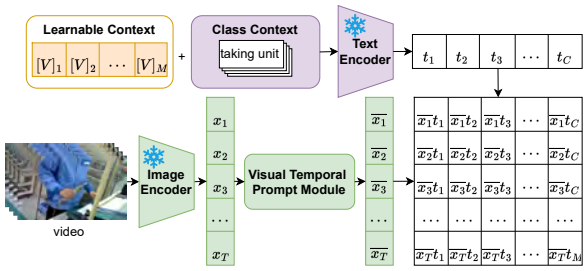

図 1: Caseg のモデル構成．文献 [3] を元に作成．

3. 研究手法

本研究では，製造業における組立作業動画を対象とし，Caseg と従来手法である MS-TCN および MS-TCN++を比較することで，少量データでの学習性能およびゼロショット推論精度を評価する．以下に，各実験の詳細を述べる．

3.1 少量データでの学習

Caseg は，CLIP による大規模事前学習の恩恵を受けており動作ラベルの言語情報と対応する視覚情報に関する表現を事前に獲得していると考えられるため，少量データ環境下においても高い精度が期待される．そこで少量データ環境における学習性能を評価するため，学習に使用するデータ量を複数の水準に分けたデータセットを構築し，各水準において，Caseg と従来手法（MS-TCN，MS-TCN++）の精度を比較する．

3.2 ゼロショット推論性能

Caseg のゼロショット推論性能を検証するため，ある組立作業工程で撮影されたデータセットを用いて Caseg を学習し，学習時に使用していない別工程のデータセットに対してゼロショット推論を行い，精度を検証する．さらに，ゼロショット推論時に使用するプロンプトを複数通り用意し，各プロンプトによる精度を比較する．

4. 実験

4.1 データセット

本研究で使用するデータセットはプリンターのユニット組立ラインで撮影された作業動画を基に独自に構築したものである（図 2）．対象とした組立作業工程は，異なる組立作業を実施している 2 つの工程 (工程 1，工程 2 と呼ぶ) であり，それぞれの工程において組立基準書に準じた動作ラベル定義を行い，動画撮影とアノテーションを実施した．動画データは 1 本あたり 30 フレーム毎秒（fps）で 5000 フレームの長さがあり，各工程ごとに 78 本のアノテーション済み動画から成るデータセットを構築した．動作ラベルは，表 1–3 に示すように定義した．工程 1 と工程 2 は別工程であるため撮影したカメラ画角や作業者などが異なるが，一部共通する作業動作が含まれているものとなっている．また両データセットともに各動作にかかる所要時間の違いなどによりクラス不均衡が存在する．

4.2 損失関数

各モデルの学習に用いる損失関数は，動画の各フレームにおける分類ロス $\mathcal{L}_{cls}$ と過分割抑制のための正則化ロス $\mathcal{L}_{T-MSE}$

図 2: 作業動画例

の重み付き和で表される．Caseg の学習の場合は $\mathcal{L}_{cls}$ には NCE loss

$$\mathcal{L}_{cls} = -\frac{1}{T}\sum_{i}^{T}\left(log\frac{\exp(\overline{x}_i \cdot t_{gt}/\tau)}{\sum_j^C \exp(\overline{x}_i \cdot t_j/\tau)}\right) \tag{1}$$

を用い，T は各動画の系列長，i は動画の各フレーム位置，C は動作ラベル数，$\overline{x}_i$ は Visual Temporal Prompt Module による i フレーム目の出力値，t_{gt} はフレーム i における正解ラベルのテキスト特徴量，t_j は j 番目の動作ラベルのプロンプトにおけるテキスト特徴量である．

MS-TCN 及び MS-TCN++の学習の場合は $\mathcal{L}_{cls}$ には cross entropy loss

$$\mathcal{L}_{cls} = \frac{1}{T}\sum_{i}^{T} -log(y_{i,c}), \tag{2}$$

を用い，$y_{i,c}$ はフレーム i の正解ラベルに対する予測値である．過分割抑制のための正則化ロス $\mathcal{L}_{T-MSE}$ としては

$$\mathcal{L}_{T-MSE} = \frac{1}{TC}\sum_{i,c}\tilde{\Delta}_{i,c}^2, \tag{3}$$

を用い，

$$\tilde{\Delta}_{i,c} = \begin{cases} \Delta_{i,c} & : \Delta_{i,c} \leq \tau \\ \tau & : \text{otherwise} \end{cases}, \tag{4}$$

$$\Delta_{i,c} = |log\ y_{i,c} - log\ y_{i-1,c}|\,, \tag{5}$$

である．各モデルの予測値に対して重み λ のもとで

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda\mathcal{L}_{T-MSE}, \tag{6}$$

を計算し，これを損失関数とする．また各モデルの中間ステージにおける出力に対しても損失計算を行い，全ステージでの損失関数の総和を最小化するよう学習を行う．

4.3 学習・評価方法・パラメータ設定

学習は動画単位での 5 分割交差検証を用いて実施した．各分割において，検証用に 16 本，学習用に 62 本の動画を使用した．また，評価指標の計算には，検証ロスが最も低かったエポックにおけるモデルを用い，正解率（Accuracy），編集スコア（Edit），F1@{10,25,50} スコアの平均を算出した．さらに，使用する各モデルについては，原著論文に記載されたパラメータ設定を踏襲して実験を行った．

Caseg の Visual Temporal Prompt Module には MS-TCN++を用い，最終的な予測は各フレームの動画特徴量とテキスト特徴量の内積より求めた．Image Encoder には VIT-B-16[7] を使用し，Text Encoder と Image Encoder についてはモデルパラメータを固定し，学習時に重み更新がされないようにした．学習可能なテキストプロンプトについてはコンテキストトークンの数（M）を 4 に設定した．

また，損失関数の設定として，各モデルで λ の値を 0.15 に固定し，学習はすべて 200 エポックにわたって実施した．

表 1: 工程 1 における動作ラベル

作業	プロンプト	動作内容	頻度 [%]
1	taking unit	部品を取り出す	2.5
2	checking unit	欠陥がないか確認する	13.4
3	attaching pins	ピンをとめる	15.5
4	turning unit	組み立てユニットをひっくり返す	2.3
5	attaching tapes	テープを貼る	11.5
6	placing strip	部品を取り付ける	6.6
7	preparing parts	待機・組み立て準備	28.4
8	Inactive assembly line	人がいない状況	1.3
9	screwing ends	ドライバーでねじ止め	14.3
10	passing unit	組み立てユニットを所定位置に置く	2.9
11	adjusting unit	調整作業	1.2

表 2: 工程 2 における動作ラベル - Normal

作業	プロンプト	動作内容	頻度 [%]
1	taking unit	部品を取り出す	2.5
2	placing springs	ばねを設置	4.7
3	placing strip	補助部材を設置	10.1
4	fixing strip	補助部材を固定する	1.7
5	placing rod	ブラシを配置する	3.4
6	placing bearing	ベアリングを設置	7.7
7	screwing ends	ドライバーでねじ止め	10.4
8	attaching E-rings	E 形止め輪を設置	9.8
9	checking unit	欠陥がないか確認する	4.0
10	passing unit	組み立てユニットを所定位置に置く	3.4
11	preparing parts	待機・組み立て準備	40.3
12	Inactive assembly line	人がいない状況	1.9

表 3: 工程 2 における動作ラベル - Rich

作業	プロンプト	動作内容	頻度 [%]
1	Taking the unit.	部品を取り出す	2.5
2	Placing springs to both ends of the unit.	ばねを設置	4.7
3	Placing the strip in the unit.	補助部材を設置	10.1
4	Sliding the jigs at both ends of the unit.	補助部材を固定する	1.7
5	Placing the rot in the unit.	ブラシを配置する	3.4
6	Place bearing at both ends of the unit.	ベアリングを設置	7.7
7	Screwing both ends of the unit using a screwdriver.	ドライバーでねじ止め	10.4
8	Attaching the E-rings to both ends of the unit.	E 形止め輪を設置	9.8
9	checking unit	欠陥がないか確認する	4.0
10	Passing the unit.	組み立てユニットを所定位置に置く	3.4
11	waiting or preparing parts	待機・組み立て準備	40.3
12	Inactive assembly line	人がいない状況	1.9

4.4 実験 1：少量データでの学習

実験 1 では，学習に使用するデータ量を複数の水準で設定し，工程 1 における動画データセットを用いてモデルの学習および検証を行った．比較対象としたモデルは，Caseg，MS-TCN，および MS-TCN++の 3 種類である．

学習および検証には表 1 に示した動作ラベルを使用した．学習データ量の水準は 4 水準とし，それぞれ各分割における学習データの 10%（6 サンプル），20%（12 サンプル），70%（43 サンプル），および 100%（62 サンプル）の割合でランダムに動画データをサンプリングして作成した．一方，検証データは各水準において共通である．各水準に対してそれぞれのモデルで学習を行い，精度評価を実施した．その結果を表 4 に示す．

結果から，Caseg は学習サンプル数が少ない水準（6，12 サンプル）で他手法よりも優れたパフォーマンスを示しており，少量データ環境における強みが確認された．一方で，学習サンプル数が増加すると MS-TCN++の精度が Caseg を若干上回ったが，Caseg についても十分高い精度が確認された．

4.5 実験 2：ゼロショット推論性能

実験 2 では，工程 1 の動画データセットを用いて Caseg を学習し，そのモデルを用いて工程 2 の動画データセットに対するゼロショット推論の性能を検証した．

Caseg の学習に使用した工程 1 の動作ラベルは表 1 に示すものである．また，工程 2 のゼロショット推論に使用した動作ラベルは，表 2 と表 3 に示しており，これらはそれぞれ，工程 1 の動作ラベルと同じ形式で設計した Normal プロンプトと，より記述的な文に拡張し精度向上を狙う Rich プロンプトである．

評価対象として，次の 4 つの推論条件を設定した．1 つ目は，Normal プロンプトを用いたゼロショット推論 (Normal)．2 つ目は，Rich プロンプトを用いたゼロショット推論 (Rich)．3 つ目は，予測ラベルを全て発生頻度が最も高いラベルとしたときのベースライン (Naive Baseline)．4 つ目は，直接工程 2 の動画データを用いて Normal プロンプトで学習・推論を行う条件（Full Train）である．

実験 2 における各条件の精度比較を表 5 に示す．結果から，Edit や F1 値といった指標では，Naive Baseline よりもゼロショット推論 (Normal，Rich) のほうが精度が高いことが分かる．しかし，Accuracy に関しては Naive Baseline よりも低い精度となっている．これは，動画内の約 40%が「preparing parts」という動作に該当しており，単一の動作ラベルで全体を埋めるだけでも Accuracy が約 40%に達してしまう，不均衡な動画データセットのためである．

また，Normal プロンプトと Rich プロンプトを入力したときのゼロショット推論結果を比較すると，Rich プロンプトの精度が Normal プロンプトより低い結果となった．これは，Rich プロンプトにおいては，学習時のプロンプトよりも推論時の系列長が長く，動画とテキストのアライメントを上手くとることができなかったためと考えられる．

いずれの条件も Full Train と比較すると，全ての精度指標が低くなっており，学習データセットと検証データセットのドメインが異なる場合ではゼロショット推論精度が大きく劣化することが分かる．

図 3 に正解と Normal の予測の系列を示す．予測系列を見ると，「screwing ends」のような，学習と評価データに共通するラベルは，比較的正しく予測を行うことができている．また，予測した動作ラベルそのものが正解と異なる場合でも，正解の動作境界と Normal の動作境界の位置については一致するケースが確認された．このことから学習データセットと検証データセットのドメインが異なる場合でも，Visual Temporal Prompt Module が動作の切り替わりに関する表現については抽出できる可能性がある．

これらのことから，Caseg を用いたゼロショット推論は，一定の予測能力を有しているものの，実際の現場適用という観点に関しては精度に課題がある結果となった．

図 3: 工程 2 の 1 サンプルに対するゼロショット推論（Normal）の予測系列の例

5. 結論

本研究では，製造業の組立作業における動作解析を目的に，プリンターの組立作業工程を対象とした独自データセットを用い，Temporal Action Segmentation（TAS）におけるアノ

表 4: トレーニングサンプル数ごとのモデル精度比較

# Train Samples	Model	Accuracy	Edit	F1@{10,25,50}		
6 (10%)	MS-TCN	55.21	41.20	41.15	37.96	31.29
	MS-TCN++	56.07	45.09	48.79	45.75	38.54
	Caseg	**58.55**	**47.41**	**50.14**	**46.43**	**38.94**
12 (20%)	MS-TCN	80.86	72.39	71.97	69.14	62.64
	MS-TCN++	81.45	**78.41**	74.77	72.58	64.72
	Caseg	**82.66**	77.52	**76.07**	**73.90**	**66.25**
43 (70%)	MS-TCN	83.67	77.58	76.35	74.24	68.17
	MS-TCN++	**84.48**	**84.90**	**80.03**	**77.19**	**71.45**
	Caseg	83.65	82.24	76.95	74.25	65.74
62 (100%)	MS-TCN	85.42	82.38	79.54	78.33	73.20
	MS-TCN++	**85.83**	**84.51**	**82.71**	**81.44**	**75.00**
	Caseg	85.69	84.03	81.75	80.10	74.83

表 5: 実験 2 における各条件の精度比較

Type	Accuracy	Edit	F1@{10,25,50}		
Normal(Zero-shot)	37.27	20.73	12.03	8.75	4.32
Rich(Zero-shot)	29.02	20.25	9.00	6.08	2.79
Naive Baseline	40.04	14.25	5.66	3.37	2.05
Full Train	81.18	84.00	75.05	73.70	70.86

テーションコスト削減に向けて CLIP ベースの Caseg モデルの性能を従来手法（MS-TCN，MS-TCN++）と比較検討した．

Caseg は少量データでの学習において従来手法を上回る性能を示した．これは，CLIP の大規模事前学習の恩恵を受け，限られたアノテーションデータ量でも高い精度を維持できるためであり，アノテーションコスト削減への寄与が期待される．一方，学習データセットと異なる組立動作を含む別工程のデータセットに対するゼロショット推論では，動作境界や学習データと検証データで共通の動作ラベル関して一定の予測能力を示したものの，実現場への適用という観点で精度に課題があることが確認された．今後はゼロショット推論の精度向上に向けた手法の改良が求められる．

参考文献

[1] Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

[2] Fangqiu Yi, Hongyu Wen, and Tingting Jiang. Asformer: Transformer for action segmentation. In The British Machine Vision Conference (BMVC), 2021.

[3] Suyuan Huang, Haoxin Zhang, Yanyu Xu, Yan Gao, Yao Hu, and Zengchang Qin. Caseg: Clip-based action segmentation with learnable text prompt. In 2024 IEEE International Conference on Image Processing (ICIP), pages 2201–2207, 2024.

[4] Shijie Li, Yazan Abu Farha, Yun Liu, Ming-Ming Cheng, and Juergen Gall. Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(6):6647–6658, 2023.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pages 8748–8763. PMLR, 18–24 Jul 2021.

[6] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. International Journal of Computer Vision (IJCV), 2022.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations, 2021.