大規模言語モデルによる時系列行動セグメンテーションの精度向上

捧蓮　曲佳　三輪祥太郎
三菱電機株式会社
sasage.ren@dr.mitsubishielectric.co.jp　kyoku.ka@dc.mitsubishielectric.co.jp
miwa.shotaro@bc.mitsubishielectric.co.jp

概要

時系列行動セグメンテーションは、映像内の一連の行動を時間軸に沿って認識し、一連の行動を構成する個別行動区間を検出するタスクであり、行動理解・評価や技術習得支援等への応用が期待されている。既存手法では、全体的な時系列の流れを捉えておらず、ラベルの予測精度が課題となっている。本手法では、多種多様なドメインに対する膨大な一般的知識を持つ大規模言語モデルを汎用的な状態遷移モデルとして活用することで、常識的・論理的な行動手順を考慮し、行動認識の精度を向上させる。提案手法の性能をいくつかのビデオデータセット（50Salads, Breakfast）で評価したところ、既存手法を上回る性能を達成した。

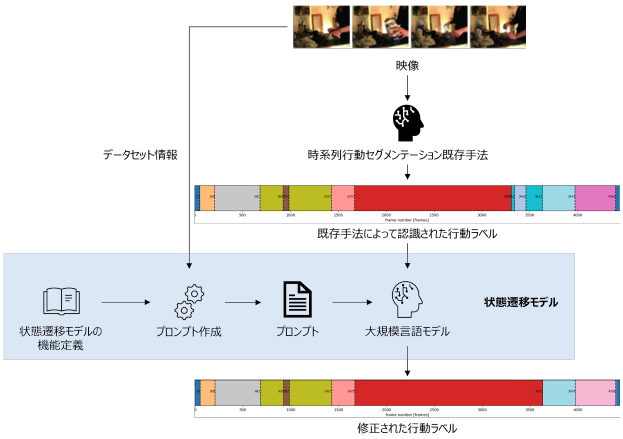

図 1　概要

1　はじめに

時系列行動セグメンテーションは、映像を時間軸に沿って各行動区間に分割することを目的とした映像認識タスクである。映像の自動セグメント化は、人間の前後の行動における相互作用や関係を理解する上で重要な役割を果たす。そのため、人・ロボットの行動理解や周囲環境理解、監視システムなどの映像におけるセキュリティ向上や、製造業における技術的なスキルの習得支援システムの開発などへの応用が期待されている。

時系列行動セグメンテーションには、行動区間とその行動ラベルの認識が必要である。局所的な前後フレームの特徴を捉えることで行動認識を行う手法 [1, 2, 3] が考案されてきたが、全体的な行動手順を捉えることができておらず、依然として行動ラベルの予測精度が課題となっている。一方、全体的な行動手順を考慮した状態遷移モデルとして、マルコフモデルや Recurrent Neural Networks (RNNs)[4] を用いた手法 [5, 6, 7] が提案されてきたが、各ドメインに対応する全体的な行動手順、すなわち状態遷移モデルを設計した上で学習を行う必要がある。それに対して、GPT-4o[8] のような大規模言語モデルは、単なるパターン認識能力を超えて、インターネットスケールの多様な言語・画像データから得たあらゆるドメインに対する一般的知識を持つため、行動手順の正しさを理解しており、汎用的な状態遷移モデルとして機能する。

本論文では、この大規模言語モデルが持つ常識的知識による常識的・論理的な行動手順を考慮した行動認識手法を提案する。具体的には、初めに、既存手法によって時系列行動セグメンテーションを行う。次に、既存手法によって認識された各行動セグメントの行動ラベルを、大規模言語モデルを用いて常識的・論理的な行動手順に修正する。既存手法 [1, 2] と組み合わせた提案手法の性能をいくつかのビデオデータセット（50Salads[9], Breakfast[10]）で評価したところ、既存手法を上回る性能を達成した。

2 関連研究

2.1 時系列行動セグメンテーション

時系列行動セグメンテーションにおいて、複雑な時間的依存関係を捉えた高精度行動認識を行うため Transformer[1] や Temporal Convolutional Network (TCN)[2, 3] などのアーキテクチャが多用されてきた。Fangqiu ら [1] が提案した ASFormer は、Transformer[11] を時系列行動セグメンテーションタスクに応用させる際の課題に取り組むため、局所的な接続性の誘導バイアスや事前定義された階層的な表現パターン、エンコーダーの初期予測を段階的に再調整する複数のデコーダが導入されている。Shijie ら [2] が提案した MS-TCN++は、長いビデオの時系列行動セグメンテーションにおける過剰セグメンテーションエラーを解決するため，生成された予測を段階的に再変換する多段階時系列畳み込みネットワークが用いられている。しかし、これらの手法は、全体的な時系列の流れを捉えていないために論理的な行動順序を考慮することができていない。

一方、全体的な時系列の流れを考慮したマルコフモデルや Recurrent Neural Networks (RNNs)[4] を用いた手法 [5, 6, 7] が提案されてきた。Hilde ら [5] は、縮小されたフィッシャーベクトルに基づく視覚表現と隠れマルコフモデルを組み合わせて行動認識を行った。Kevin ら [6] は、最大マージンフレームワークで学習させた可変時間隠れマルコフモデルを利用することで、映像の識別的で興味深いセグメントを自動的に発見した。Hilde ら [7] は、本質的に階層的である人間の活動を学習するために、複数の動画特徴記述子と組み合わせた隠れマルコフモデルを用いた。しかし、これらの手法は、各ドメインに対応する全体的な行動手順、すなわち状態遷移モデルを設計した上で学習を行う必要がある。

2.2 大規模言語モデルの常識推論

常識的知識とは、人間の認知の基礎であり、世界に対する生得的な理解とその中で推論する能力を包含する。近年、大規模言語モデルが、推論や文脈理解、思考の連鎖を含む広範な自然言語処理タスクにおいて大きな進歩を遂げており、これらの成果は、大規模言語モデルがある程度の常識的知識を持つことを示唆している [12]。Jain ら [13] は、イベントの所要時間や一般的な順序などの時間に関する常識的な理解を必要とする MC-TACO データセット [14] などを用いて大規模言語モデルの時間推論能力の分析を行い、ゼロ・少数ショット学習における GPT-3.5 の優れた性能を示した。

3 手法

3.1 大規模言語モデルの常識推論による時系列行動セグメンテーション

本手法の概要を図 1 に示す。本手法では、初めに既存手法によって時系列行動セグメンテーションを行う。次に、既存手法によって認識された各行動セグメントの行動ラベルを、大規模言語モデルを用いて常識的・論理的な行動手順に修正する。3.2 節及び 3.3 節では、大規模言語モデルを汎用的な状態遷移モデルとして機能させるために行った 2 つの工程についてそれぞれ説明する。

3.2 状態遷移モデルの機能設計

大規模言語モデルを各データセットに対応した状態遷移モデルとして機能させるために必要なデータ分析の観点は次の通りである。

- **タスク予測とルールの考慮**: 人間が行う行動は階層的であり、目的となるタスクによって必要となる行動の種類は異なる。そのため、まず初めにスーパークラスとして各映像のタスクを認識することが重要である。
- **論理的な行動手順の考慮**: 各タスクを行う一連の行動群には、組み合わせに特徴がある。そのため、特定の行動ラベルの組み合わせがどのタスクに結びつくかを考慮することで、行動ラベルの予測精度を高めることが可能となる。
- **常識的な所要時間の考慮**: 各行動に要する時間も異なる。そのため、各行動の持続時間がどの程度かを考慮することで、より正確な行動ラベルの予測が可能となる。また、短いセグメントは誤認識されやすいため、それらのラベルを隣接するセグメントと一致させることで一貫性が保たれる。

3.3 プロンプト作成

プロンプトエンジニアリングとは、期待される穴埋め形式の出力を埋め込む入力テキスト・テンプレートの設計を指す。今回のプロンプトエンジニア

表 1 50Salads データセットに対する時系列行動セグメンテーションの結果

50Salads	F1@{10, 25, 50}			Edit
ASFormer (reproduced)	83.4	80.7	74.6	75.7
ASFormer +GPT-4o	**86.1**	**83.6**	**76.4**	**79.5**
MS-TCN++ (reproduced)	76.7	73.6	64.9	76.6
MS-TCN++ +GPT-4o	**78.8**	**75.9**	**66.6**	**79.7**

表 2 Breakfast データセットに対する時系列行動セグメンテーションの結果

Breakfast	F1@{10, 25, 50}			Edit
ASFormer (reproduced)	74.0	68.7	55.0	73.5
ASFormer +GPT-4o	**74.9**	**69.5**	**55.8**	**74.5**
MS-TCN++ (reproduced)	57.8	52.0	39.8	61.6
MS-TCN++ +GPT-4o	**61.2**	**55.2**	**42.5**	**65.3**

リングの目的は、大規模言語モデルを汎用的な状態遷移モデルとして機能させることで、各セグメントの行動ラベルを常識的な行動手順に修正するためのテキスト・プロンプトを作成することである。3.2 節で設計した状態遷移モデルの機能を基に作成したテキストプロンプトの概要は以下の通りである。詳細については付録 A を参照されたい。

1. **タスク予測とルールの考慮**: 以下の 2 つの観点から各映像で行われているタスクを予測する。
   - 映像内の各行動ラベルの出現頻度を数える。最も頻度の高い行動ラベルに基づいてタスクを予測する。
   - 頻繁に一緒に発生する行動ラベルの組み合わせを特定する。これらの行動ラベルの組み合わせに基づいてタスクを予測する。

   また、行動ラベルには活動の開始や終了を示すものが含まれているため、それらのラベルの位置が限定的であることを考慮する。
2. **論理的な行動手順を考慮した行動認識**: 映像内の各行動ラベルがリスト化された潜在的な行動ラベルと一致しない場合は、映像内に既に存在する行動ラベルと関連性の強い行動ラベルに修正する。
3. **常識的な所要時間を考慮した行動認識**: 映像内の各行動ラベルがリスト化された潜在的な行動ラベルと一致しない場合は、映像全体に対するセグメントの長さを考慮して修正する。非常に短いセグメントについては、隣接するセグメントに合わせてラベルを修正し、一貫性を持たせる。

また、本手法では few-shot 推論を用いる。few-shot 推論では、条件付けとして推論時にモデルにタスクのデモをいくつか与えるが、重みの更新は行わない。

4 評価実験

4.1 データセット

本手法をベンチマークとなる 2 つのビデオデータセットによって評価した。

4.1.1 50Salads[9]

25 人の個人がキッチンで 2 種類のサラダを作る 50 本の映像で構成され、合計 19 種類の行動ラベルが含まれる。各映像は約 8000〜18000 フレームで構成される。実験では、5 分割交差検証によって評価を行い、その平均値を算出した。

4.1.2 Breakfast[10]

52 人の個人が 18 種類の異なるキッチンで 10 種類の朝食活動を行う 1,712 本の映像で構成され、合計 48 の行動ラベルが含まれる。各映像は約 800〜4000 フレームで構成される。実験では、4 分割交差検証によって評価を行い、その平均値を算出した。

4.2 評価方法

交差検証による性能評価を行い、F1 スコア（F1@τ）、編集スコア（Edit）の平均を算出した。F1 スコアは、時間軸における予測した各行動セグメントと対応する正解の各行動セグメントとの Intersection over Union（IoU）が閾値 τ/100 を超えた場合に、その予測行動セグメントを True Positive と判定し、他を False Positive と判定して計算した適合度（Precision）

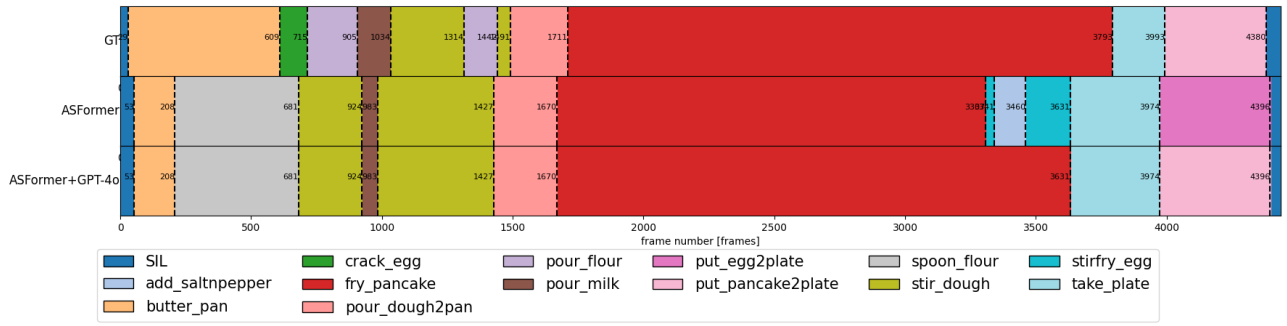

図 2　Breakfast データセットにおける定性評価の結果

と再現度（Recall）の調和平均を表す。編集スコアは、行動セグメントを単位とするレーベンシュタイン編集距離を用いて計算され、行動セグメントの開始終了時間の差異を無視して行動順序のみの正確さを評価する。

4.3　結果

2 つのビデオデータセット（50Salads, Breakfast）に対して提案手法を最先端手法と比較した結果をそれぞれ表 1, 2 に示す。表に示すように、我々の手法はいずれのデータセットにおいても最先端手法を上回る性能を達成した。また、Breakfast データセットにおける定性評価の結果を図 2 に示す。図の上部は正解、中央は既存手法の結果、下部は提案手法の結果を示している。図より、既存手法では'fry pancake' に続く行動ラベルを'stir fry egg' や'add salt and pepper'、'put egg onto plate' と誤って認識しているのに対して、我々の手法では、まず初めに'pancake' を作るタスクであることを認識した上で、そのタスクに関連しない行動ラベルを修正対象とし、'fry pancake' に続けて'take plate'、'put pancake onto plate' と論理的な行動手順に修正できていることが分かる。

5　おわりに

本論文では、大規模言語モデルが持つ常識的知識を活用した常識的・論理的な行動手順を考慮した時系列行動セグメンテーション手法を提案する。我々の手法では、大規模言語モデルを多種多様なドメインに対応した汎用状態遷移モデルとして機能させることで、全体的な行動手順を考慮した行動認識が可能となる。2 つのビデオデータセットにおける優れた評価結果から、我々の手法の有効性が実証された。

参考文献

[1] Fangqiu Yi, Hongyu Wen, and Tingting Jiang. ASFormer: Transformer for Action Segmentation. **CoRR**, abs/2110.08568, 2021.

[2] 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, 2023.

[3] Colin Lea, Michael D. Flynn, René Vidal, Austin Reiter, and Gregory D. Hager. Temporal Convolutional Networks for Action Segmentation and Detection. **CoRR**, abs/1611.05267, 2016.

[4] De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Connectionist Temporal Modeling for Weakly Supervised Action Labeling. **CoRR**, abs/1607.08584, 2016.

[5] Hilde Kuehne, Juergen Gall, and Thomas Serre. An end-to-end generative framework for video segmentation and recognition. In **2016 IEEE Winter Conference on Applications of Computer Vision (WACV)**, 2016.

[6] Tang Kevin, Fei-Fei Li, and Koller Daphne. Learning latent temporal structure for complex event detection. In **2012 IEEE Conference on Computer Vision and Pattern Recognition**, 2012.

[7] Hilde Kuehne, Ali Arslan, and Thomas Serre. The Language of Actions: Recovering the Syntax and Semantics of Goal-Directed Human Activities. In **Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition**, 2014.

[8] OpenAI. Hello GPT-4o, 2024. https://openai.com/index/hello-gpt-4o/.

[9] Sebastian Stein and Stephen J. McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In **Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing**, 2013.

[10] Hilde Kuehne, Ali Arslan, and Thomas Serre. The Language of Actions: Recovering the Syntax and Semantics of Goal-Directed Human Activities. In **Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition**, 2014.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. **CoRR**,

abs/1706.03762, 2017.

[12] Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, Ben He, Shanshan Jiang, and Bin Dong. ChatGPT Is a Knowledgeable but Inexperienced Solver: An Investigation of Commonsense Problem in Large Language Models. In **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, 2024.

[13] Raghav Jain, Daivik Sojitra, Arkadeep Acharya, Sriparna Saha, Adam Jatowt, and Sandipan Dandapat. Do Language Models Have a Common Sense regarding Time? Revisiting Temporal Commonsense Reasoning in the Era of Large Language Models. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, 2023.

[14] Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. ''Going on a vacation'' takes longer than ''Going for a walk'': A Study of Temporal Commonsense Understanding. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, 2019.

A 参考情報

A.1 テキスト・プロンプト

A.1.1 50Salads

Objective: Refine the action labels of existing segments in a video portraying a person making one of two specific types of mixed salads.
Video {index}: Video {index} spans from frame 1 to {frame_count}.
Current Segments and Labels of Video {index}: {seg_label_str}
Think and refine the action labels based on the following step-by-step instructions.

1. **Refine 'action_start' and 'action_end' Labels**: Always use 'action_start' and 'action_end' exclusively for the start and end segments of the video, respectively. Correct 'action_start' and 'action_end' segments in between to their appropriate actions. Again, consider the segment length relative to the entire video when determining the appropriate action.
2. **Refine Incorrect Actions**: If a segment is extremely short, adjust its label to match the adjacent segments for consistency.

Output Format: Provide the refined labels in the following example format, with your step-by-step thinking. Do not deviate from this structure under any circumstances. Make sure to conclude with: "Therefore, the answer is..." followed by the refined labels.

A.1.2 Breakfast

Objective: Refine the action labels of existing segments in a video portraying a person engaged in one of ten specific breakfast activities.
Video {index}: Video {index} spans from frame 1 to {frame_count}.
Current Segments and Labels of Video {index}: {seg_label_str}
Think and refine the action labels based on the following step-by-step instructions.

1. **Predict the Breakfast Activity**: Use the following two approaches simultaneously to predict only one breakfast activity:
   - Count which actions correspond to each meal based on the reference. Predict the dish with the highest match count.
   - Identify pairs or combinations of actions that frequently occur together based on "Reference: Action Units for Breakfast Activities." Predict the dish based on the action combinations in the reference.
2. **Verify all Potential Actions**: List all potential actions associated with the predicted breakfast activity based on "Reference: Action Units for Breakfast Activities."
3. **Refine Incorrect Actions**: Check each segment label. Only if a label does not correspond to any potential actions listed, consider the length of the segment in relation to the total video length to determine the most likely action for that duration. If a segment is extremely short, adjust its label to match the adjacent segments for consistency.
4. **Refine 'SIL' Labels**: Always use 'SIL' for the start segments of the video. If the label for the last segments of the video is already 'SIL', do not change it. Correct 'SIL' segments in between to their appropriate actions. Again, consider the segment length relative to the entire video when determining the appropriate action.

Output Format: Provide the refined labels in the following example format, with your step-by-step thinking. Do not deviate from this structure under any circumstances. Make sure to conclude with: "Therefore, the answer is..." followed by the refined labels.