CH-001

時系列行動セグメンテーションにおける汎用的性能向上のための行動境界識別用チャネル拡張手法の提案

A Channel Extension Method for Action Boundary Identification to Improve General Performance in Temporal Action Segmentation

光岡 日菜子† 堀田 一弘†
Hinako Mitsuoka Kazuhiro Hotta

1. はじめに

近年，動画像中の動作をフレーム単位で識別する，時系列行動セグメンテーションの技術が注目されている．この技術は産業分野，医療分野など様々な応用分野において重要な役割を果たすものである．時系列行動セグメンテーションにおける主要な課題の一つに行動の持続時間を正確に予測する難しさがある．特に各行動の持続時間を過剰に短く分割してしまう“オーバーセグメンテーション”が本タスクの潜在的な問題として存在する．これに対処するため，従来法では行動境界を予測する専用のブランチやモジュールを追加する手法が提案されてきた．しかし，これらの手法はモデル構造の複雑化を招き，汎用性や計算コストの面で課題が残る．

本稿では，上述した問題に対し，出力時のチャネルを 1 つだけ追加するというシンプルな構造変更でありながら，強力な行動境界識別性能を付加できる拡張方法を提案する．非常に簡易的な機構にも関わらず，本タスクではこれまで用いられてこなかった新規性のある手法である．既存の時系列行動セグメンテーション用のモデルに対し，構造変更を最小限にとどめつつ，精度向上を実現できる点が特徴である．

評価実験では，2 つの異なる構造のモデルに本手法を適用し， 複数の公開データセットにおいてベースラインを全ての評価指標において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とは，任意の作業などを撮影した動画像に対して，行動クラスを各フレームに対して割り当てるタスクである．本タスクは時系列情報を扱う行動認識の一種であり，動画像内の行動の種類，持続時間やその順序を機械学習により明示的に分類できる．そのため，作業手順の把握，作業時間の計測の自動化などに応用可能であり，産業・医療の分野で重要度が高まっている．

一方で認識精度を左右する潜在的な課題も存在する．各行動の持続時間や行動境界の時系列的位置が一定でないことや，類似行動が連続することなどにより行動を過剰に細かく分割してしまう“オーバーセグメンテーション”がその代表例である．上述した課題をいかに効果的に解消するかが本タスクの研究の主な目的となっている．

本タスクにおいては動画像からの特徴抽出はほぼ共通して Two-Stream Inflated 3D ConvNet (I3D)[1]と呼ばれる大規模動画像データで事前学習された特徴抽出器を用いて行われるため，I3D の出力を入力とし，時系列情報を扱うモデルの研究がかねてより行われてきた．

これまで様々なモデルが提案されてきたが，代表的なものとして MS-TCN[2]が挙げられる． MS-TCN は時系列データに対し，複数層かつ上位層ほど受容野の広い Temporal Convolutional Network(TCN)を適用し，粗い予測を徐々に洗練させ，より正確な識別を実現する．

また，近年では FACT[3]と呼ばれる，フレーム毎の識別に加え，動画像内のまとまった動作単位での識別も同時に行う非常に高性能なモデルが提案されている．2 種類の識別の学習には構造の異なる 2 つのブランチを使用する．途中でブランチ間の情報交換を行いつつ学習を行うことで出力を洗練させる．フレーム単位のみの識別ではオーバーセグメンテーションが生じやすいため，より大きなまとまりを単位とする行動表現を導入し，長距離依存関係のモデリング能力を向上させている．

2.2 行動境界予測を利用した手法

時系列行動セグメンテーションにおいて，動画像の行動境界を予測する情報を活用し，精度向上を図る手法が提案されてきた．

ASRF[4]では，フレーム毎の予測に加え，行動境界の予測を行うための専用のブランチをモデル内に設けている．これにより行動の切り替わりが強調され，オーバーセグメンテーションの緩和に成功している．

また，BMUTCN[5]では，行動境界近傍の領域に注目した Boundary-Match モジュールを導入し，行動境界に特化した処理を行うことによりセグメンテーション精度を向上させている．これらの手法はいずれも行動境界予測を補助的なタスクとして活用し，時系列行動セグメンテーションを高精度化することを目的としている．しかし，専用のブランチやモジュールの追加を必要とするため，モデル構造が大幅に拡張され，設計・実装のコストが増大する課題が存在する．さらに，その追加構造分の計算コスト増大が懸念

†名城大学 Meijo University

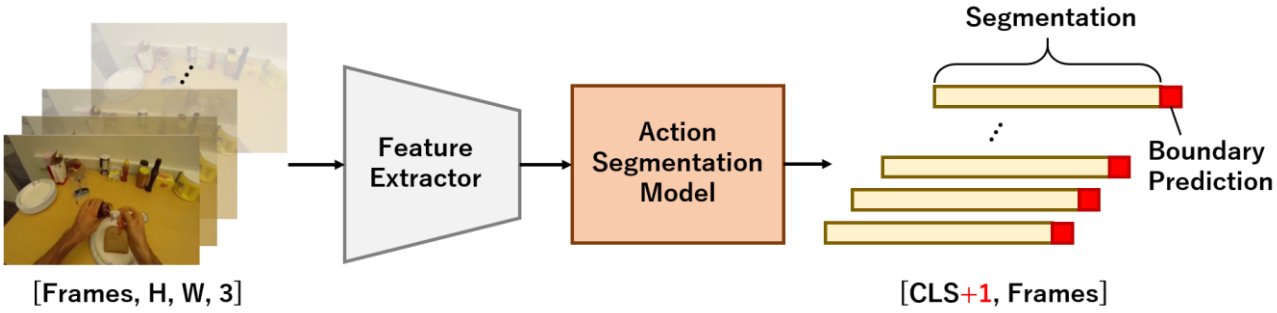


図 1　提案手法の概略図

されるほか，既存の高性能なモデルに容易に導入できる機構ではなく汎用性が低いという問題もある．

本稿ではこの問題に対処するため，出力チャネルを 1 つ追加するのみで行動境界の予測を可能とするシンプルかつ汎用的な拡張手法を提案する．

3. 提案手法

提案手法の概要を図 1 に示す．本稿では，出力チャネルを 1 つ追加することにより行動境界を予測する，シンプルな拡張手法を提案する．時系列行動セグメンテーションモデルの出力に境界識別チャネルを付加することで，行動の切り替わりに対する感度を向上させ，オーバーセグメンテーションの緩和を図る．また，本手法は実装が非常に簡易であり，既存の高性能なモデルにも容易に適用可能であり，汎用性に優れる．

3.1 境界チャネルの設計

本稿では，時系列行動セグメンテーションにおける行動境界情報を明示的に学習させるため，各フレームにおける出力チャネルに行動境界識別用のチャネルを 1 つ追加する拡張を行う．具体的には，通常の行動クラス予測に用いるクラス数分の出力チャネル(CLS)に加え，境界識別用のチャネルを付加し，最終的な出力チャネルを CLS+1 とする．追加されたチャネルは各フレームが行動境界であるか否かを 2 値的に予測するために使用する．

使用する公開データセットには境界識別用のラベルは含まれないため，[4]と同様の方法で行動境界にあたるフレームを 1，それ以外のフレームを 0 とするバイナリラベルを生成し使用する．

3.2 損失関数設計

提案手法では行動セグメンテーションと行動境界の識別の 2 つのタスクを同時に学習させるため損失関数の追加が必要になる．行動セグメンテーションの学習には，基本的に Cross Entropy Loss およびオーバーセグメンテーション緩和のための追加損失が用いられる傾向があるが，モデル構造と共に独自の損失関数が提案され，同時に用いられることが多い．

行動境界の識別は二値分類タスクとなるため，Binary Cross Entropy Loss を適用する．これにより，追加した境界識別チャネルの出力と，生成した境界ラベルとの間で損失を最小化するよう学習を行う．行動セグメンテーションのための損失を$\mathcal{L}_{seg}$，境界識別のための損失を$\mathcal{L}_{boundary}$としたとき，最終的な損失関数$\mathcal{L}_{total}$は，以下のように表される．

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \lambda \mathcal{L}_{boundary} \quad (1)$$

ここで，λは両損失間のバランスを調整するためのハイパーパラメタである．本稿では予備実験に基づき設定した．

3.3 既存モデルへの適用方法

提案手法は，出力層に境界識別用チャネルを 1 つ追加するのみで適用でき，基本的にどんな既存の時系列行動セグメンテーションモデルにも容易に導入可能である．

MS-TCN のような比較的単純なモデル構造に対しては，最終分類層の出力チャネル数を従来の行動クラス数 CLS から CLS+1 に拡張し，追加したチャネルを境界識別用の出力として用いればよい．また，FACT のように既に複数のブランチが含まれるモデル構造に対しては，主となるタスクを学習しているブランチの最終分類層にのみ拡張を行う．上述した以外のモデル構造には一切変更を加えず，境界識別用の損失についても元々の損失に加算するのみである．この簡易な拡張により，複雑な設計変更を伴うことなく，従来の時系列行動セグメンテーションに加え，行動境界の識別を行うことが可能になる．

4. 評価実験

本節では，提案手法を MS-TCN および FACT に適用した場合の性能について評価する．比較対象には，各ベースラインモデル(MS-TCN，FACT)および行動境界予測専用のブランチを使用する手法である従来法である ASRF を用いる．

4.1 実験条件

本稿では，提案手法の有効性の検証のため，50Salads[6]，GTEA[7]の異なる 2 つデータセットを使用する．

50Salads はサラダ調理の動作を撮影した 50 本の動画像で構成され，17 種類の行動クラスが含まれる．各動画は平均 6.4 分の長さであり，1 本あたり平均 20 個の行動のまとまりが存在する．25 名の被験者が各々2 種類のサラダを調理しており，5 分割交差検証により評価を行う．

表 1　50Salads データセットにおける定量的結果

Method	F1 @ {10, 25, 50}			Edit	Acc
For Simple Architecture					
MS-TCN	73.96	71.42	62.12	66.29	79.11
MS-TCN + **Ours(λ =1.0)**	**76.30**	**73.59**	**64.39**	**69.18**	**79.40**
For Complex Architecture					
ASRF	82.58	81.22	74.97	75.61	81.39
FACT	82.89	80.94	75.52	76.70	85.47
FACT + **Ours(λ =1.0)**	**83.91**	**81.50**	**76.54**	**78.00**	**87.51**

表 2　GTEA データセットにおける定量的結果

Method	F1 @ {10, 25, 50}			Edit	Acc
For Simple Architecture					
MS-TCN	86.47	83.73	71.20	79.92	76.53
MS-TCN + **Ours(λ =0.5)**	**86.89**	**85.24**	**71.29**	**81.17**	**77.91**
For Complex Architecture					
ASRF	86.52	85.27	73.68	78.59	76.01
FACT	92.49	90.36	79.83	89.89	79.82
FACT + **Ours(λ =0.5)**	**93.05**	**90.72**	**81.51**	**89.95**	**81.31**

GTEA は頭部装着型カメラにより撮影された 28 本の動画から構成され，コーヒーやチーズ，サンドイッチの調理などの活動が 4 名の被験者により行われている．各動画には背景クラスを含む 11 種類の行動クラスが付与されており，1 動画あたり平均 20 個の行動のまとまりが存在する．4 分割交差検証により評価を行う．

いずれのデータセットを用いる場合においても各フレームに対して I3D により抽出した特徴量を使用し，入力としてモデルに与える．また，フレームレートは 50Salads では元の 30fps を 15fps にダウンサンプリングし，GTEA では元々の 15fps を維持している．学習時の設定やハイパーパラメタについては，エポック数を 50 に統一するほかは各ベースラインモデルの設定に準拠する．

また，提案手法適用時の境界識別用損失の重みλは，予備実験を行い学習時の損失の値を元に決定した．本稿の評価実験においては 50Salads では 1.0，GTEA では 0.5 に設定する．

4.2 評価指標

評価指標には Accuracy，Edit Distance，F1 Score の 3 種を使用する．Accuracy は，全フレームに対する正解率を表すが，長時間の行動クラスが影響を与えやすいこと，およびオーバーセグメンテーションが小さく評価されがちである点に留意する．

Edit Distance は，予測した行動列と正解の行動列の一致度を，挿入・削除・置換といった編集操作数に基づいて評価する指標である．特に，短い区間の過剰分割や行動クラスの順序を厳しく評価でき，Accuracy の補完指標として用いられる．

F1 Score は，予測と正解の重なり率である IoU（Intersection over Union）を 10%，25%，50%の閾値で評価し，閾値以上の重なりがあるものを正しく検出されたと見なすものである．

4.3 定量的結果

表 1 に 50Salads データセットにおける定量的結果を示す．MS-TCN に提案手法を適用した場合，全ての評価指標で性能向上が確認された．特に F1@10 は 2.34%，Edit Distance は 2.89%と大幅に改善した．また，FACT に対し提案手法を適用した場合では，Accuracy が 2.04%向上し，FACT 単体を大きく上回る結果を達成した．MS-TCN ベースにおいては，境界識別用ブランチを設ける ASRF に若干及ばない結果となったが，チャネルの拡張のみで ASRF に近づく性能を達成した．

表 2 に GTEA データセットにおける定量的結果を示す．50Salads の場合と同様に，提案手法を MS-TCN および FACT に適用したとき，全ての評価指標で性能向上が確認された．さらに，GTEA データセットの場合，MS-TCN に提案手法を適用しただけで ASRF を凌駕する性能を達成した．これらの結果から，出力チャネルの拡張のみという軽微な構造変更であっても行動境界予測による性能向上が可能であり，複雑な機構や境界識別用にわざわざ専用のブランチを用意せずとも高精度な時系列行動セグメンテーションが行えることが示された．

4.4 定性的結果

提案手法の有効性を直感的に確認するため，予測結果の可視化による定性的評価を行った．図 2 および図 3 に，それぞれ 50Salads および GTEA データセットにおける定性的結果を示す．

50Salads データセットにおいては，MS-TCN の予測において，短時間の行動区間が過剰に分割されるオーバーセグメンテーションが確認された．これに対し，提案手法適用後は，行動区間のまとまりが改善され，行動境界付近の精度が向上していることが視覚的に確認できた．FACT に対しては，単体ですでに高い精度を有しているため，提案手法による可視的な変化は小さい．しかし，行動境界付近のクラスの切り替わりがわずかに正確になっている点が確認できた．

GTEA データセットにおいても同様に，MS-TCN ではオーバーセグメンテーションが見られたが，提案手法適用後はこれらが大きく改善された．FACT においても目立った大きな違いは見られないものの，短時間の行動クラスがより正確に予測できていることが確認できた．

4.5 考察

本稿では時系列行動セグメンテーションにおいて出力チャネルを 1 つ追加するだけで行動境界を学習させ，予測の正確化を図るシンプルな拡張手法を提案した．MS-TCN および FACT という異なる構造を持つベースラインに適用し，かつ複数のデータセットを用いて評価を行った．定量的評価の結果，提案手法は MS-TCN ベースにおいて，境界識別専用ブランチを持つ従来法 ASRF と比較しても同等以上，FACT ベースにおいては ASRF および FACT を超える性能

図 2　50Salads データセットにおける定性的結果

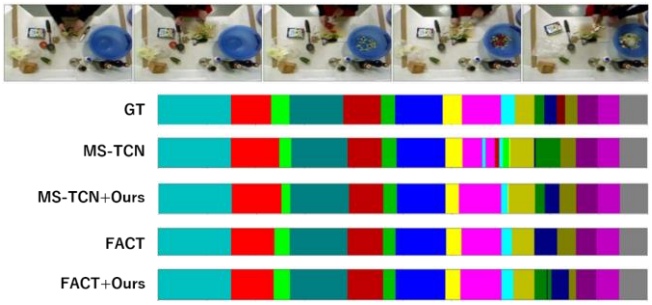
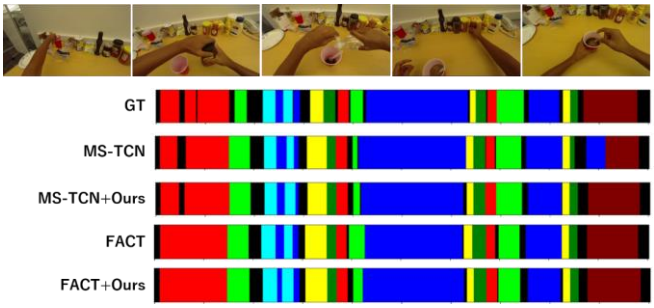


図 3　GTEA データセットにおける定性的結果



を達成した．特に Edit Distance や F1 Score の向上が顕著であることから，オーバーセグメンテーションの緩和に寄与していることが示された．定性的評価においても視覚的にオーバーセグメンテーションの緩和を確認できた．さらに，2.2 節で示した従来法では行動境界の情報を活用するためにブランチを増やしたり複雑なモジュールを設計するコストが課題となっていたが，提案手法はその課題を構造的に回避できる点でも有用である．

これらの結果から，提案手法は時系列行動セグメンテーションにおいて，軽微な構造変更のみで汎用的に高精度化を図ることができると考えられる．また，行動境界を学習するため，オーバーセグメンテーションの緩和に特に寄与すると考えられる．したがって，今後本タスクにおいて性能面に課題を感じた際の第 1 手として，提案手法の導入が推奨できるであろう．

5. おわりに

本稿では，時系列行動セグメンテーションにおける認識精度向上を目的として，出力チャネルを 1 つ追加するだけで行動境界を予測可能にするシンプルな拡張手法を提案した．提案手法は既存の時系列行動セグメンテーションモデルに対して，モデル構造の大幅な変更を必要とせず，容易に導入可能であり，境界識別専用のブランチなどの複雑な改変を必要としないという利点がある．複数ベースライン，複数データセットを用いた評価実験においてその有効性を示した．

今後の課題としては，損失部分に用いている手動で設定しているハイパーパラメタの最適化や，より難易度の高いデータセットにおける汎用性の検証が挙げられる．さらに，提案手法と他手法の組み合わせにより，学習データ量が限られる状況でも高精度な行動理解を実現する可能性についても検討していく．

参考文献

[1] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299-6308, 2017.

[2] Yazan Abu Farha and Jurgen Gall, "MS-TCN: Multi-stage temporal convolutional network for action segmentation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp.3575-3584, 2019.

[3] Zijia Lu, and Ehsan Elhamifar, "FACT: Frame-action cross-attention temporal modeling for efficient action segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.18175-18185, 2024.

[4] Yuchi Ishikawa, et al. "Alleviating over-segmentation errors by detecting action boundaries." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp.2322-2331, 2021.

[5] Zhengwei Shen, et al. "Boundary-Match U-Shaped Temporal Convolutional Network for Vulgar Action Segmentation." Mathematics, Vol.12, No.6, pp.899, 2024.

[6] Sebastian Stein, and Stephen J. McKenna, "Combining embedded accelerometers with computer vision for recognizing food preparation activities." Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing, pp.729-738, 2013.

[7] Alireza Fathi, Xiaofeng Ren, and James M. Rehg, "Learning to recognize objects in egocentric activities." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.3281-3288, 2011.