# 時間マルチスケール特徴と行動ラベル特徴による オープンボキャブラリ行動区間認識

グエン チュンタイン<sup>1,a)</sup> 川西 康友<sup>2,1,b)</sup> 駒水 孝裕<sup>1,c)</sup> 井手 一郎<sup>1,d)</sup>

# 概要

本研究では,オープンボキャブラリ行動区間認識(Openvocab Temporal Action Detection; Open-vocab TAD)と いうタスクに取り組む.このタスクは,与えられた行動 ラベルに対応する映像中の区間を検出するタスクである. 提案手法は,1段階の行動検出手法を採用し,動画像から 得た時間マルチスケール特徴と行動ラベルから抽出した ラベル特徴を統合することで,多様な時間幅の行動区間 認識を実現する.TAD タスクで一般的に利用されている Thumos14 データセットを用いた実験結果から,提案手法 は Open-vocab の設定だけでなく,学習データに含まれる 語彙を推定するクローズドボキャブラリの設定においても 既存のモデルを一貫して上回り,様々な実験設定で最先端 の結果を達成した.これは動画像から任意の行動区間を切 り出すための有望なツールとしての可能性を示している.

# 1. はじめに

人間が,わずかな例や簡単な説明だけで未知の物体を認 識できるのは,これまでの経験によって蓄積した関連す る知識を応用する能力によるものである.近年,事前学 習済みの大規模な画像・言語(ViL)モデル(CLIP[12]や Align[3]など)の進歩により,この能力をコンピュータ上 でも実現できるようになってきた.これらのモデルは,イ ンターネットから収集された様々な画像とテキストの組 を用いて対照学習することで,それらの共有表現を学習 する.その結果,マルチモーダル理解,クロスモーダル転 移,ゼロショット学習,意味理解などの様々なタスクにお いてテキストの説明から有益な情報を効果的に抽出し,優 れた性能を発揮できる.この進歩により,研究者たちは Open-vocab物体検出,行動認識,行動区間認識(Temporal Action Detection; TAD)などの様々な下流タスクで ViL

# <sup>1</sup> 名古屋大学

- 2 理化学研究所
- <sup>a)</sup> nguyent@cs.is.i.nagoya-u.ac.jp
- $^{\rm b)}~$ yasutomo.kawanishi@riken.jp
- <sup>c)</sup> taka-coma@acm.org
- $^{\rm d)} \quad ide@i.nagoya-u.ac.jp$

を活用する方法を模索している.

本研究では Open-vocab TAD というタスクに焦点を当 てる.TAD タスクは動画像から行動の区間を検出し,その 行動のクラスを識別するタスクである. 従来の TAD は, 訓 練データとテストデータにおいて、テストデータに出現す る行動ラベルは訓練データに必ず出現することを想定して いる.これに対して, Open-vocab TAD では, テストデー タ内に未知の行動が存在し,それを検出することが求めら れる.そのため, Open-vocab TAD は未知の行動の正確な 識別と時系列的な検出が求められる.既存研究 [4], [13] で は, Open-vocab TAD に取り組むために2段階で認識す る方法を採用している.まず,行動区間を検出し,その結 果を用いて行動を識別する.従って,1段階目の誤りが, 2段階目の処理に悪影響を及ぼす可能性がある.また,固 定的な時間スケールで処理をするため,様々な長さの行動 に対処することが難しい.そこで,本研究では2段階のア プローチで生じる問題を避けるために,行動区間の検出の 発見と識別の両方を1段階で実現するアプローチを採用す る.また,様々な長さの行動を検出するために,時間方向 にマルチスケールな特徴を導入する.

本研究の貢献は以下の通りである.

- シンプルな1段階のTAD手法を提案し, Open-vocab TAD における有効性を示す.
- 動画像に対する時間マルチスケール特徴とテキスト特徴の統合法を提案し、多様な長さの行動検出に有効であることを示す。
- 一般的な TAD データセットで評価を行ない, ViL の 特徴を活用して,最先端の性能で未知の行動を検出で きることを示す.

## 2. 関連研究

近年, ViL モデル [3], [11], [12] への関心は急速に高まっている.これらのモデルは画像とテキストの両方の共有表現を学習することで,以前は難しかった多様なタスクを実現可能にしている.インターネットから収集した大規模な



図 1 提案手法の概要.動画像の各フレームは事前学習済みの Video Encoder を通し, マルチ スケール動画像解析とデコーダを介して区間検出をする.テキストは Text Encoder に よって埋め込まれ,その後,動画像特徴と照合することで関係を理解し,行動のラベル を正確に決定する.

データセットを用いて学習した ViL モデルは,特別なファ インチューニングや大量のアノテーションを用意する必要 がなく,高精度なOpen-vocab分類タスクを実現している. 動画像に対する研究では,これらのモデルは行動の分類タ スク[4], [15] やテキストによる動画検索タスク[9] に利用 されている.これらのモデルは, ViL の能力を組み合わせ ることにより,未知の行動や新たな動画像内のシーンでも 認識が可能になるため,実環境での行動認識の可能性を広 げている.特に,最近は行動区間が切り出されていない動 画像から未知の行動区間を検出する Open-vocab TAD タ スク [4], [13] が取り組まれ始めている.これらのモデルは ViL から得られる特徴と動画像から得られる特徴の比較に, 教師付き分類器を用いる2段階モデルを使用している.本 研究はこれらの研究 [4], [13] とはいくつかの側面で異なる. まず,1段目の誤りが2段目へと影響する問題を避けるた めに,行動区間の検出と行動の識別に1段階モデルを採用 する.さらに,様々な長さの行動に対応するために,時間 軸方向のマルチスケール特徴を導入し, Open-vocab の行 動検出の性能を向上させる.

## 3. 提案手法

## 3.1 概要

Open-vocab TAD の目標は入力動画フレーム $I = (I_1, I_2, ..., I_T)$ に対して,一連の行動区間及びそのクラ ス $Y = \{y_1, y_2, ..., y_N\}$ を推定することである(NはX内の行動区間の数である).まず,入力動画フレームIから,特徴ベクトル $X = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T)$ を抽出する.ここ で,離散時刻  $t = \{1, 2, ..., T\}$ の最大値 T は動画像の長さによって異なる.出力される各  $\mathbf{y}_i = (s_i, e_i, a_i)$ は,開始時刻  $s_i$ ,終了時刻  $e_i$ ,および行動ラベル $a_i$ である.ここで, $a_i \in D_{\text{train}} \cup D_{\text{test}}$ は学習時またはテスト時に用いる行動のラベルの1つである.なお,本研究で主に対象とするOpen-vocabの設定では $D_{\text{train}} \cap D_{\text{test}} = \emptyset$ である.一方,従来の Closed-set の設定では $D_{\text{test}} \subset D_{\text{train}}$ である.

図1に提案手法の全体像を示す.本研究は Closed-set で の先行研究 [5], [16] に倣い, Open-vocab TAD で行動の ローカリゼーションにおける1段階の検出をする.重要な 点として,提案手法は(1)マルチスケール動画像解析モ ジュールと(2)動画像-テキスト統合モジュールの2つか ら構成される.マルチスケール動画解析モジュールは行動 の開始時刻と終了時刻を決定し,それが行動区間であるか 否かを評価する.動画像-テキスト統合モジュールは動画 像特徴とテキスト特徴の関係を学習し,行動のラベルを決 定する.

## 3.2 マルチスケール動画像解析モジュール

入力動画フレーム *I* から抽出した特徴量列 *X* =  $(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T)$ をエンコードしてマルチスケール特徴表 現 *Z* =  $\{Z^1, Z^2, ..., Z^L\}$ を出力する.ここで,*L* はスケー ルの数であり,ネットワークの階層数に相当する.最初 に,浅いニューラルネットワークを用いて入力の各特 徴  $\mathbf{x}_t$ を *D* 次元の空間に埋め込む投影関数 *E* を適用し,  $Z^0 = (E(\mathbf{x}_1), E(\mathbf{x}_2), ..., E(\mathbf{x}_T))$ を得る.ここで, $E(\mathbf{x}_i)$ は  $\mathbf{x}_i$ を *D* 次元空間へ埋め込む.次に, Transformer( $f_{l-1}^l$ )を 用いて埋め込まれた特徴 *Z*<sup>*l*-1</sup> をひとつ上のスケールの特 徴表現 *Z*<sup>*l*</sup> に変換する.

$$Z^{l} = f_{l-1}^{l}(Z^{l-1}).$$

Transformer のセルフアテンション(自己注意機構)によ り,動画像内のフレーム間の類似度を計算し,重み行列を 算出することで,重要なフレームを自動的に抽出できる. これにより,このモデルは行動を含む区間に注目し,関係 ないノイズを除去することができる.その結果,動画像内 の行動の正確な区間検出を実現できる.本研究では,6層 のTransformerを使用し,階層間の時間方向ダウンサンプ リングには効率的な処理のためにストライド付きの深さ方 向の1D畳み込みを使用する.これにより半分のダウンサ ンプリングをする.

次に,マルチスケール特徴  $Z^* = \{Z^1, Z^2, ..., Z^L\}$ を 抽出した後,デコーダを用いて行動クラス及びその区間  $\hat{Y} = \{\hat{y_1}, \hat{y_2}, ..., \hat{y_N}\}$ を推定する.このデコーダには,回帰 ヘッドと背景分類ヘッドを備えた軽量な畳み込みネット ワークを用いる.回帰ヘッドは動画像内の各行動の開始お よび終了時刻を推定する.同時に,背景分類ヘッドは各フ レームが行動を含むか否かを判断する.

#### 3.3 動画像-テキスト統合モジュール

行動のラベルはテキストで表して,テキストの埋め込み には事前学習済みの ViL モデルを使用する.テキストエン コーダへの入力は  $A = \{a_1, a_2, ..., a_N\}$ であり,ここで Nは訓練行動の数である.その後,テキスト特徴と動画像か ら得た特徴  $Z^0 \ge Z^* \ge k$ 統合する.ここで,動画像の特徴 は行動が発生する位置だけを考慮する.検出した各行動の 区間には,その区間から抽出した特徴量に対して平均プー リングを適用する.その結果得られた特徴は,テキスト特 徴と組み合わせて行動とテキストの対応付けに利用する.

#### 3.4 学習のための目的関数

#### 3.4.1 マルチスケール動画像解析の損失関数

マルチスケール動画像解析では,各時刻 t について  $(d_t^s, d_t^e, p(a_t))$ を決定する.ここで, $p(a_t)$  はその瞬間が行 動 a である確率を表し, $(d_t^s, d_t^e)$  はそれぞれ時刻 t から行 動開始点と終了点までの距離である.このモジュールで は,行動か否かの分類には Focal Loss [7],距離の回帰には DIoU Loss [19] の 2 つの損失関数を使用する.Focal Loss は背景と行動の標本数のバランスが取れていない状況を効 果的に扱う.一方,DIoU Loss は行動の境界までの距離を 正確に回帰するために使用され,行動の正確な位置決めを 目的とする.

$$L_{\rm MVA} = \sum_{t}^{T} \left( L_{\rm Focal}^{t} + \lambda L_{\rm DIoU}^{t} \right), \tag{1}$$

ここで, $\lambda$ はバランス係数である.

3.4.2 動画像-テキスト統合の損失関数

ここでは,テキスト特徴と動画像特徴(投影された特徴 量 Z<sup>0</sup> とマルチスケール層の特徴量 Z\*)のクロスモーダ ルな類似性を捉える.これを実現するために,Contrastive Loss を使用し,特徴空間で一致する組を求め,異なるクラ スの組を分離する.

$$L_{\rm TVA} = L_{\rm text \to projection} + \gamma L_{\rm text \to multi-scale}, \qquad (2)$$

ここで, $L_{\text{text}\rightarrow\text{projection}}$ はテキスト特徴と動画像の投影された特徴との関係を捉えるための Contrastive Loss を計算 する.同様に, $L_{\text{text}\rightarrow\text{multi-scale}}$ はテキスト特徴と動画像の マルチスケール特徴との関係を求めるための Contrastive Loss を計算する.これらの損失関数をバランスするために 係数  $\gamma$  を導入する.

提案手法の損失関数は,以下のように全ての合計として 定義する. *α* はバランス係数である.

$$L_{\rm Total} = L_{\rm MVA} + \alpha L_{\rm TVA} \tag{3}$$

## 4. 実験

提案手法による未知の行動検出能力を Open-vocab で評価した結果とその詳細について報告する.事前訓練 ViL モデルとしては, CLIP B/16 の埋め込みを主に使用した. また, ViL モデルからの知識の有用性を考察するために, Closed-set の設定 ( $D_{\text{train}} = D_{\text{test}}$ )で追加実験を行なう.

データセット:行動区間認識の評価に広く使用されて いる Thumos14 データセット [2] を用いて評価する.この データセットは 20 の行動ラベルを含む 413 の行動区間が 取り出されていない動画像からなる.

ランダム分類: Open-vocab シナリオのために無作為に2 つのサブセット, すなわち 75/25(75%の行動が訓練,25% の行動が検証)と50/50(50%の行動が訓練,50%の行動 が検証)に分割して評価した.文献[13]で示された方法に 倣い,75/25のサブセットは10セットに分割し,50/50の サブセットは5セットに分割した.これらの分割はデータ の多様性を保証し,モデルの汎用性とロバスト性を評価す るのに役立つ.

評価指標:推定した行動候補は,重複する可能性がある ため,Soft-NMS [1]を用いて重複を除去することで,最終 的な行動の出力を生成する.実験結果は,一般的に正解と の時間的な重複率(tIoU)に対してしきい値処理をするこ とで正しく推定できたか否かを判定し,その割合である平 均適合率(mAP)により評価した.

比較手法:提案手法は先行研究の1段階手法および2段 階手法 TAD/Open-vocab TAD と比較した.さらに,提案 モデルの構造を基にしたベースラインモデルも比較した. このベースラインモデルは,マルチスケールを1つに限定

エゴル	動画偽性ッ	テキスト	mAP	
	勤同 1844年143	特徴	75/25	50/50
OV-TAD [13]	CLIP $B/32$	CLIP $B/32$	12.0	8.0
	CLIP $L/14$	CLIP $L/14$	15.8	10.5
	ALIGN	ALIGN	8.7	6.5
	BASIC	BASIC	20.8	16.5
	I3D	CLIP $B/32$	14.1	11.5
	I3D	CLIP $L/14$	17.0	14.5
	I3D	ALIGN	10.4	8.3
	I3D	BASIC	22.9	19.5
OV-TAD [13]	CLIP B/16	CLIP B/16	12.9	9.1
OV-TAD [13]	I3D	CLIP $B/16$	15.6	12.9
ベースライン	I3D	CLIP $B/16$	15.8	9.6
提案手法	I3D	CLIP B/16	24.0	13.7

表 1 Open-vocab 設定での検出結果の平均適合率 [%] tIoU = [0.3:0.1:0.7]

し,テキスト特徴と総合する際にマルチスケール特徴を使用しない.

#### 4.1 Open-vocab 設定での結果

表1は異なる動画像特徴とテキスト特徴の組み合わせを 使用した,提案手法とOV-TAD 手法による Open-vocab 設 定での検出の比較結果を示す.OV-TAD 手法では動画像 の I3D 特徴とテキストの BASIC 特徴の組み合わせが最も 高い精度を達成した.提案手法は動画像の I3D 特徴とテキ ストの CLIP B/16 特徴を使用したが, 75/25 の分割にお いて最も高い精度である 24.0% を達成した . CLIP B/16 モデルは BASIC モデルに比べて学習データ数が 1/10 で あるが,提案手法では,1.1%高い結果となった.さらに, 同じ CLIP B/16 特徴を使用した他のモデルと比較すると, 提案手法は 8.4% 上回った . 50/50 の分割では,検出がよ り難しくなり, BASIC 特徴の方が CLIP B/16 特徴より優 れた結果を示した.ただし,CLIP B/16 特徴を用いる他の モデルと比較すると,0.8%の改善があった.さらに,提案 手法はベースラインを上回り, Open-vocab 行動検出シナ リオでマルチスケール特徴の有効性を確認した.

## 4.2 Closed-set 設定での結果

表2は動画像のみを使用したモデルと動画像とテキスト を組み合わせたモデルによるClosed-set設定での結果を示 す.提案手法は動画像のみを用いるほとんどの手法に比べ て優れた性能を示し,動画像とテキストを用いる手法にお いて最も高精度であった.具体的には,マルチモーダルの 設定で,提案手法はmAPで55.5%の精度を達成し,さら に訓練する際にネガティブサンプリング(NS)戦略を使用

表 2 Closed-set 設定での検出結果の平均適合率 [%] tIoU = [0.3:0.1:0.7]

モデル	年	入力	動画像特徴	mAP
BMN [6]	2019		TSN	38.5
TAL-MR [18]	2020		I3D	43.3
VSGN [17]	2021		TSN	50.2
AFSD [5]	2021	Video	I3D	52.0
TadTR $[8]$	2021		I3D	46.6
ActionFormer [16]	2022		I3D	66.8
TriDet [14]	2023		I3D	69.3
EffPrompt [4]	2021		CLIP $B/16$	34.5
STALE [10]	2022		I3D	52.9
OV-TAD [13]	2022		CLIP $B/16$	29.0
OV-TAD [13]	2022	Video	ALIGN	32.7
OV-TAD [13]	2022	Text	BASIC	45.4
ベースライン	2023		I3D	39.9
提案手法	2023		I3D	55.5
提案手法(NS)	2023		I3D	59.5

した場合には 59.5% の精度を達成した.同じ I3D 特徴を 使用した STALE モデルと比較すると,提案手法は 2.6% から 6.6% の向上を示した.また,マルチスケールを1つ に限定したベースラインモデルは精度が著しく低下するこ とが確認された.これは,マルチスケールを使用すること の重要性を示す.

## 5. むすび

本研究は Open-vocab TAD のタスクにおいて,テキス ト特徴とマルチスケールの動画像特徴を導入し,1段階の 検出による手法を提案した.実験結果により,行動検出に おけるマルチスケール特徴の貢献を示した.

今後の方向性として,大規模なデータセットを用いた評価により,提案手法の汎用性を検証する予定である.また,現在のモデルはTransformerを用いているが,計算速度が遅いという課題を抱えている.そのため,より高速な計算を可能とする新たな手法を検討している.

## 謝辞

本研究の一部は JSPS 科研費 JP21H03519 の助成を受け たものである.また,本研究は名古屋大学のスーパーコン ピュータ「不老」の一般利用を利用して実施した.

#### 参考文献

 Bodla, N., Singh, B., Chellappa, R. and Davis, L. S.: Soft-NMS-Improving object detection with one line of code, *Proceedings of the 2017 IEEE International Conference on Computer Vision*, pp. 5561–5569 (2017).

- [2] Idrees, H., Zamir, A. R., Jiang, Y.-G., Gorban, A., Laptev, I., Sukthankar, R. and Shah, M.: The THUMOS challenge on action recognition for videos " in the wild ", *Computer Vision and Image Understanding*, Vol. 155, pp. 1–23 (2017).
- [3] Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z. and Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision, *Proceedings of the* 2021 International Conference on Machine Learning, pp. 4904–4916 (2021).
- [4] Ju, C., Han, T., Zheng, K., Zhang, Y. and Xie, W.: Prompting visual-language models for efficient video understanding, *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27,* 2022, Proceedings, Part XXXV, Springer, pp. 105–124 (2022).
- [5] Lin, C., Xu, C., Luo, D., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F. and Fu, Y.: Learning salient boundary feature for anchor-free temporal action localization, *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3320–3329 (2021).
- [6] Lin, T., Liu, X., Li, X., Ding, E. and Wen, S.: BMN: Boundary-Matching Network for temporal action proposal generation, *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision*, pp. 3889–3898 (2019).
- [7] Lin, T.-Y., Goyal, P., Girshick, R., He, K. and Dollár, P.: Focal loss for dense object detection, *Proceedings of the 2017 IEEE International Conference on Computer Vision*, pp. 2980–2988 (2017).
- [8] Liu, X., Wang, Q., Hu, Y., Tang, X., Zhang, S., Bai, S. and Bai, X.: End-to-end temporal action detection with transformer, *IEEE Transactions on Image Processing*, Vol. 31, pp. 5427–5441 (2022).
- [9] Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N. and Li, T.: CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning, *Neuro*computing, Vol. 508, pp. 293–304 (2022).
- [10] Nag, S., Zhu, X., Song, Y.-Z. and Xiang, T.: Zero-shot temporal action detection via vision-language prompting, Computer Vision-ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III, Springer, pp. 681–697 (2022).
- [11] Pham, H., Dai, Z., Ghiasi, G., Kawaguchi, K., Liu, H., Yu, A. W., Yu, J., Chen, Y.-T., Luong, M.-T., Wu, Y., Tan, M. and Le, Q. V.: Combined scaling for zero-shot transfer learning, *Computing Research Repository arXiv Preprints*, arXiv:2111.10050 (2021).
- [12] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I.: Learning transferable visual models from natural language supervision.
- [13] Rathod, V., Seybold, B., Vijayanarasimhan, S., Myers, A., Gu, X., Birodkar, V. and Ross, D. A.: Open-Vocabulary Temporal Action Detection with Off-the-Shelf Image-Text Features, *Computing Research Repository arXiv Preprints*, arXiv:2212.10596 (2022).
- [14] Shi, D., Zhong, Y., Cao, Q., Ma, L., Li, J. and Tao, D.: TriDet: Temporal action detection with relative boundary modeling, *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18857–18866 (2023).
- [15] Wang, M., Xing, J. and Liu, Y.: ActionClip: A new

paradigm for video action recognition, *Computing Research Repository arXiv Preprints*, arXiv:2109.08472 (2021).

- [16] Zhang, C.-L., Wu, J. and Li, Y.: ActionFormer: Localizing moments of actions with transformers, Computer Vision-ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IV, Springer, pp. 492–510 (2022).
- [17] Zhao, C., Thabet, A. K. and Ghanem, B.: Video selfstitching graph network for temporal action localization, *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision*, pp. 13658–13667 (2021).
- [18] Zhao, P., Xie, L., Ju, C., Zhang, Y., Wang, Y. and Tian, Q.: Bottom-up temporal action localization with mutual regularization, *Computer Vision—ECCV 2020:* 16th European Conference, Glasgow, UK, August 23– 28, 2020, Proceedings, Part VIII, Springer, pp. 539–555 (2020).
- [19] Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R. and Ren, D.: Distance-IoU loss: Faster and better learning for bounding box regression, *Proceedings of the AAAI Conference* on Artificial Intelligence, Vol. 34, No. 07, pp. 12993– 13000 (2020).