一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
PRMU2024-6(2024-05)

# 大域・局所特徴統合埋め込みに基づくオープン語彙時系列行動検出

グエンチュン タイン[†,††] 川西　康友[††,†] 駒水　孝裕[†††,†] 井手　一郎[†,†††]

† 名古屋大 大学院情報学研究科　〒464–8601 名古屋市千種区不老町
†† 理化学研究所 ガーディアンロボットプロジェクト　〒619–0288 京都府相楽郡精華町光台 2–2–2
††† 名古屋大学 数理・データ科学教育研究センター　〒464–8601 名古屋市千種区不老町
E-mail: †nguyent@cs.is.i.nagoya-u.ac.jp, ††yasutomo.kawanishi@riken.jp, †††taka-coma@acm.org,
††††ide@i.nagoya-u.ac.jp

**あらまし**　オープン語彙時系列行動検出（オープン語彙 TAD）は，クローズド語彙時系列行動検出（クローズド語彙 TAD）の検出対象を拡張し，学習データに含まれない語彙で指定された未知行動クラスを検出することを目的とするタスクである．オープン語彙 TAD は行動区間の候補の提案とその候補における行動の識別の 2 段階手法にするのが一般的である．しかし前段での誤りが後段や最終結果に影響を及ぼす可能性がある．さらに，従来手法での時系列の文脈分析器は，局所的または大域的な文脈のいずれかに注目している．大域的な文脈のみに注目すると，瞬間的な詳細情報が不足し，各行動の識別が難しくなる．一方，局所的な文脈のみに注目すると，行動と非行動との区別が難しくなる．また近年，動画像特徴抽出の際に複雑な自己注意機構を用いることで生じるランク落ちによる各フレームでの識別力の低下が指摘されている．本研究では，これらの問題を解決するため，大域・局所特徴統合埋め込みを活用した 1 段階手法を提案する．この手法では，動画像の特徴から行動区間候補の提案と識別を同時にすることで，2 段階手法における誤差蓄積の問題を解決する．さらに，大域・局所特徴の統合的な埋め込みを動画像特徴抽出に導入することで，各フレームでの識別力を維持しつつ全体としての時系列における文脈を理解できるようになり，効果的な行動検出を実現する．先行研究と比較して，THUMOS14 データセットで最大 16.6 ポイント，ActivityNet-1.3 データセットで最大 8.3 ポイント，性能を向上した．

**キーワード**　時系列行動検出，オープン語彙，視覚と言語，映像解析

## 1. ま え が き

時系列行動検出（Temporal Action Detection; TAD）は，動画像理解における重要なタスクであり，動画中の行動区間の特定と行動認識の二つのサブタスクを含んでいる．TAD は，行動カテゴリが事前に定義されるクローズド語彙 TAD と，事前に定義されないオープン語彙 TAD に大別される．クローズド語彙 TAD では，動画像内の行動の区間を特定し，それに対応する行動のカテゴリを認識することが目的である．本研究では，訓練セットに存在しないカテゴリに属する行動に対しても区間特定とカテゴリ認識を要求する，より挑戦的なオープン語彙 TAD に焦点を当てる．これは，動画像から未知の行動を正確に検出し認識可能とすることで，動画像理解の能力を拡張し，実世界の複雑で多様なシナリオにおいても有効なシステムを構築する可能性を探るためである．

人間が未知の事象を単純な説明に基づいて認識する能力は，視覚と言語に関しての蓄積された知識を活用する能力に由来する．近年，大規模な視覚言語モデル [1] の開発により，コンピュータがこの能力を模倣できるようになってきた．この能力は，ゼロショット画像分類などのタスクで有効であることが証明され，物体検出，行動認識，時間的行動検出を含む様々なタスクに拡張されている．オープン語彙 TAD でも同様に，視覚言語モデルを活用することが画期的なアプローチとして登場した．このアプローチは，事前に定義された行動カテゴリの制約を超えて行動を認識し，行動区間を特定するという革新的な課題に対処する．既存のオープン語彙 TAD 手法 [2], [3] では，2 段階手法を採用している．1 段階目では，動画像から行動区間の候補を抽出する．これに続き，2 段階目では，前段階で生成された行動区間候補に対して行動を認識する．しかし，このような 2 段階手法では，前段での誤りが後段の行動認識の精度に大きな影響を与える可能性がある．これに対して，本研究では，先行研究とは異なり，近年の 1 段階クローズド語彙 TAD [4], [5] からの着想を得て，2 段階手法の誤差蓄積を減少させるために，オープン語彙 TAD に対して 1 段階手法を提案する．

さらに，最近の研究では，行動の区間特徴を抽出するためのさまざまな手法が提案されており，特にコンピュータビジョン分野で盛んに研究されている物体検出における進歩に触発された時間的行動分析が注目されている．ただし，物体検出で対象とされている多くの物体は明確な境界を持っており，検出が比較的容易であるのに対して，人の行動は開始・終了の境界が曖昧なため，より挑戦的な課題である．Anchor-Free Saliency-based Detector (AFSD) [6] と TemporalMaxer [7] は，それぞれ 1D 畳み

込み層と最大値または平均値プーリングを利用して局所的な時間的文脈を捉える畳み込みニューラルネットワークによる方法である．これらは特に行動境界の開始時刻と終了時刻近くで，一つの行動を別の行動や非行動と区別する能力が欠ける可能性がある．一方，G-TAD [8] は長期的な時間的文脈を捉えるためにグラフを利用しているが，各瞬間の詳細な情報を見逃すことがあり，行動と非行動との区別が困難になることがある．最近では，ActionFormer [4] などの Transformer による方法が有望な結果を示している．しかし，Transformer の自己注意機構には，計算コストと行動区間の検出における限界がある．自己注意機構では動画像内の各フレームと他のフレームとの関係を評価し，フレーム対の注意スコアを保持する．特に長い動画像を処理する場合，これは計算とメモリ使用量の大幅な増加を意味する．また，時系列行動検出においては，行動の正確な開始と終了の境界を特定することが極めて重要である．しかし，自己注意機構が全ての位置間の関係を計算することで，モデルの層が深くなるにつれて，各フレームの特徴がより類似し，行動区間が検出しづらくなるという，ランク落ち問題 [9] が知られている．本研究では，これらの問題に対処するため，局所的および大域的な特徴を組み合わせる統合埋め込みによって行動区間特定能力を強化し，TAD の性能を向上できる手法を提案する．

本研究の主な貢献は以下の通りである．

- **1 段階 TAD 手法**：オープン語彙 TAD において，2 段階手法により生じる問題を解決することを目指した 1 段階 TAD 手法を採用する．提案手法のネットワークは，時間的文脈分析（Temporal Context Analysis; TCA）と動画像・テキスト整合（Video-Text Alignment; VTA）の 2 つの分岐を持つ．前者は時間的動画像特徴の抽出を目的とし，後者は動画像特徴とテキスト特徴の整合を目的とする．
- **大域・局所エンコーダ**：TCA 分岐において，全体的な動画像文脈を分析するため，大域・局所エンコーダを提案する．大域特徴は行動と非行動を区別し，局所特徴は各行動を識別するための詳細な時間的行動特徴を抽出する．

本報告の残りの部分は以下のように構成される：2. では関連研究について紹介する．その後，3. で提案手法を説明する．次に，4. で実験結果を提示する．最後に，5. でまとめる．

## 2. 関 連 研 究

### 2.1 クローズド語彙時系列行動検出

クローズド語彙 TAD は，動画像理解の重要なタスクである．これは，事前に定義された行動を含む未編集の動画像において，時間的行動を特定し，認識するタスクである．クローズド語彙 TAD の特徴は，テストセットのラベル集合が訓練セットのラベル集合の部分集合となっていることである．近年の TAD 手法は大まかに 1 段階と 2 段階の手法に分かれている．2 段階手法では，行動空間候補を生成し，その後これらの行動区間候補を特定の行動カテゴリに分類する．先行研究 [8], [10], [11] では，この 2 段階手法に従い，行動空間候補生成ネットワークと認識ネットワークの 2 つの独立したネットワークにより TAD を実現している．しかし，2 段階手法は End-to-End で訓練できないため，計算が複雑であり，処理時間も大きい．2 段階手法とは異なり，1 段階手法は，単一のネットワークで行動を特定し，認識することによって，行動検出処理を合理化する大きな進歩を遂げている．いくつかの研究 [12]〜[14] は，畳み込みニューラルネットワーク，グラフ畳み込みネットワーク，および Transformer によるネットワークを構築し，その有効性を示している．

### 2.2 オープン語彙時系列行動検出

オープン語彙 TAD はクローズド語彙 TAD を拡張した新たな研究分野として登場し，訓練データセットに存在しない行動区間の特定と認識に焦点を当てている．オープン語彙 TAD の前には，オープンセット TAD が提案された．その代表的な手法 [15], [16] では，学習していないカテゴリを扱うために「Unknown」クラスを導入することでオープンセット TAD を実現している．それとは対照的に，オープン語彙 TAD は学習していない個々の行動カテゴリを詳細に識別することを目指している．その代表的な手法 [2], [3] では，オープン語彙 TAD に対処するために，事前学習済みの画像・言語モデルの能力を利用する 2 段階手法を採用している．本研究では，先行研究とは異なり，1 段階手法を採用する．さらに，効率的に行動区間を分析するために，階層的な特徴ブロック内での大域・局所エンコーダによる大域・局所統合埋め埋め込みを提案する．

## 3. 提 案 手 法

ここでは，オープン語彙 TAD を定義する．オープン語彙 TAD とは，入力動画像フレーム $\mathcal{V} = \{v_i\}_{i=1}^{T}$ と行動カテゴリのプロンプトテキスト $\mathcal{A} = \{a_t\}_{i=1}^{M}$ が与えられた場合に，動画像内に含まれる各行動の行動区間とカテゴリ $Y_i = \{s_i, e_i, a_i\}_{n=1}^{N}$ を特定することである．ここで $s_i$ は開始時刻を，$e_i$ は終了時刻を，$a_i$ は行動カテゴリラベルを表す．$T$ と $N$ はそれぞれ動画像 $\mathcal{V}$ における長さと検出された行動区間の数を，$M$ は訓練セットまたはテストセットにおける行動の数を表す．オープン語彙 TAD の難しい点は，訓練用のラベルセット $D_{\text{train\_label}}$ とテスト用のラベルセット $D_{\text{test\_label}}$ の非重複性である．これは，$D_{\text{train\_label}} \cap D_{\text{test\_label}} = \emptyset$ と書け，テスト用のラベルセットで評価される行動が訓練フェーズ中に完全に未知であることを保証する．

提案手法を明確にするために，まず 3.1 節で概要を述べ，その後，3.2 節と 3.3 節で各構成要素を詳細に説明する．最後に，3.4 節では学習時の目的関数とテスト時の処理の概要を説明する．

### 3.1 手法の概要

提案手法はアンカーフリーの 1 段階 TAD 手法を採用する（図 1）．提案手法のネットワークは，時間的文脈分析（Temporal Context Analysis; TCA）と動画・テキスト整合（Video-Text Alignment; VTA）の 2 つの分岐を持つ．TCA は，行動区間（各行動の開始・終了時刻）を特定し，各フレームを行動または非行動に分類することを目的としている．一方，VTA は動画像特

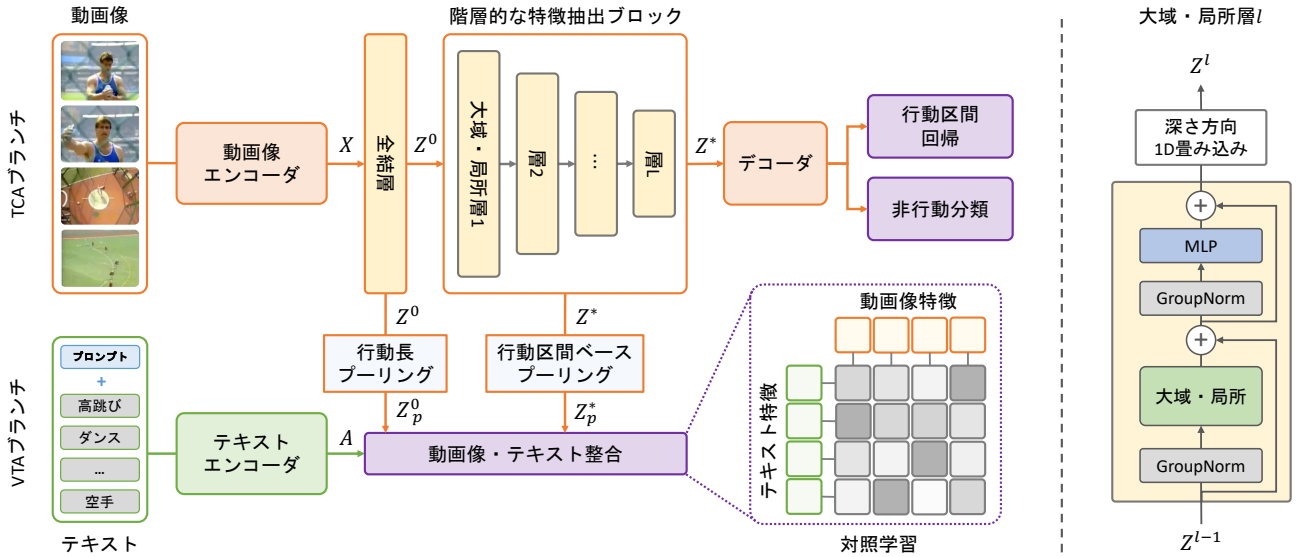


図 1: 提案手法のネットワークは，時間的文脈分析（TCA）と動画像・テキスト整合（VTA）という 2 つの分岐を持つ．TCA では，各動画像フレームが事前訓練済みの動画像エンコーダに入力され，その後，階層的なエンコーダとデコーダによって行動区間が判定される．VTA では，事前訓練済みのテキストエンコーダを使用して行動テキストラベルが埋め込まれ，動画像特徴と整合が取られる．これにより，動画像特徴と行動テキスト特徴が統合され，未学習の行動であっても認識可能になる．

徴とテキスト特徴の相関を理解し，行動ラベルを割り当てることを目的としている．

TCA は事前訓練済みの動画像エンコーダを使用して，入力動画像からフレーム単位の特徴を抽出する．その後，様々な時間的持続時間を持つ行動に対処するために，階層的な特徴ブロックを導入する．提案手法では，階層的な特徴ブロック内の大域・局所エンコーダを用いた大域・局所統合埋め込みにより，局所的および大域的な時間的文脈の両方を分析する．階層的な大域・局所エンコーダにより，時間的パターンの包括的な分析が可能となり，デコーダが行動区間を正確に特定し，各フレームが行動か非行動かを判断できるようになる．

一方，VTA は事前訓練済みのテキストエンコーダを通じて，行動テキストプロンプトを処理する．次に，対照損失を使用して，動画像特徴とテキスト特徴の間の整合を取る．最終的に，各行動区間に対応する行動カテゴリのラベルが出力される．

### 3.2 時間的文脈分析（TCA）

TCA は，入力動画像 $\mathcal{V}$ を事前訓練済みの動画像エンコーダ [17] を使用してエンコードすることから始まり，特徴列 $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T) \in \mathbb{R}^{T \times D}$ を得る．次に，スキップ接続をもつ全結合層 $E : \mathbb{R}^D \to \mathbb{R}^{D'}$ を利用し，各入力特徴 $\mathbf{x}_t$ を $D'$ 次元空間に埋め込み，階層的な特徴抽出ブロックの入力となる $\mathbf{Z}^0 = (E(\mathbf{x}_1), E(\mathbf{x}_2), \ldots, E(\mathbf{x}_T)) \in \mathbb{R}^{T \times D'}$ を得る．階層的な特徴ブロックでは，$L$ 層の特徴抽出層を通して，階層的な特徴量集合 $\mathbf{Z}^* = \{\mathbf{Z}^1, \mathbf{Z}^2, \ldots, \mathbf{Z}^L\}$ が得られる．これらは様々な区間長の行動に対処するように設計されている．$\mathbf{Z}^*$ 内の各特徴量を処理する $l$ （$1 \le l \le L$）番目の層では，提案する大域・局所エンコーダによる処理を経て，$\mathbf{Z}^l = \text{Encoder}_{l-1}^l(\mathbf{Z}^{l-1})$ を得る．最終的に，$\mathbf{Z}^*$ は，行動の開始と終了の時刻を特定し，行動か非行動かを識別するためにデコードされる．

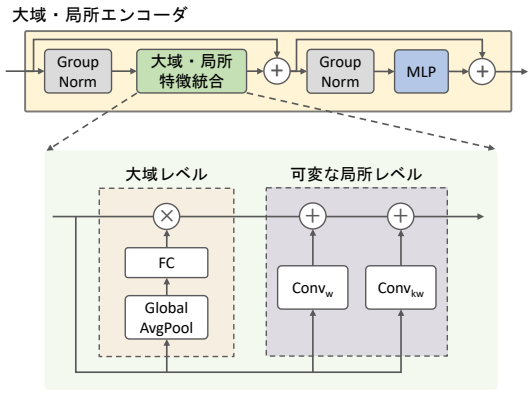


図 2: 大域・局所特徴統合部は，行動と非行動を区別することを目的とした大域レベル部と，特定の行動を識別するための詳細な時間的行動特徴を抽出するために設計された可変な局所レベル部から構成される．

#### 3.2.1 階層的な特徴ブロック内の大域・局所エンコーダ

ActionFormer [4] の Transformer エンコーダに触発された，時間的特徴を処理するための大域・局所エンコーダによる大域・局所統合埋め込みを提案する．図 2 に示すように，提案するエンコーダでは，Transformer の Multi-head Attention と Layer Norm に相当する部分を，それぞれ大域・局所特徴統合部と Group Norm に置き換える．大域・局所特徴統合部は 2 つのレベルで設計されている．

- 大域レベル：行動と非行動の区別を目的とした特徴を抽出する．動画像の全体的なコンテキストを捉えるために，各瞬間の特徴を動画全体の平均特徴と比較する．
- 可変な局所レベル：特定の行動を識別するための詳細な時間的行動特徴を抽出する．意味的なコンテキストを効果的

に捉えるために，可変な畳込みウィンドウサイズを使用する．

また，Layer Norm は全体的な動画像の正規化をする．動画像内では各種の行動と背景が交錯ため，Group Norm [18] を用いて，異なるフレーム間での特徴の一貫性を保ちつつ，動画内の各グループごとに適切な正規化をする．これにより，行動や非行動の特徴をより効果的に区別する．大域・局所エンコーダの後，層間のダウンサンプリングのためにストライド付きの 1D 深さ方向畳み込みを適用し，出力系列長を半分にする．

#### 3.2.2 出　　力

大域・局所エンコーダによって階層的な特徴量集合 $\mathbf{Z}^*$ にエンコードされた特徴量は，軽量な畳み込みネットワークからなるデコーダを用いてデコードされ，行動区間回帰と非行動分類の 2 つのヘッドへ入力される．前者は，フレーム $t$ において，フレーム $t$ から各行動の開始時刻と終了時刻までの距離 $(d_t^s, d_t^e)$ を推定する．ただし，$d_t^s = s_t - t$ はフレーム $t$ から行動の開始時刻 $s_t$，$d_t^e = e_t - t$ はフレーム $t$ から行動の終了時刻 $e_t$ までの時間差を表す．一方，後者はフレーム $t$ における非行動ではない確率 $p(a_t)$ を予測する．したがって，入力動画像 $\mathcal{V}$ が与えられた場合，TCA は複数の行動 $\widehat{Y} = \{\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_N\}$ を出力する．ここで，$\widehat{y}_t = (d_t^s, d_t^e, p(a_t))$ である．

### 3.3 動画像・テキスト整合（VTA）

VTA は，検出対象となる各行動に対して，「An action of」等のプロンプトと行動テキストラベルを連結したテキストを入力として受け取り，事前訓練済みのテキストエンコーダ [1] によってエンコードして，テキスト特徴集合 $\mathbf{A} = \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_M\}$ を得る．ここで $M$ は行動の数を表す．その後，動画像の行動特徴とテキスト特徴の整合を取る．これには，動画像特徴 $\mathbf{Z}^0$ に対して行動長プーリングして得た特徴 $\mathbf{Z}_p^0$ と，$\mathbf{Z}^*$ に対して適切な層から回帰によって得られる特徴に対する長さ方向のプーリングを用いて得た特徴量 $\mathbf{Z}_p^*$ を用いる．具体的には，各行動区間から抽出されたフレームごとの動画像特徴に対して時間方向の平均プーリングを使用する．このプーリング処理は，区間内の特徴を集約し，代表的な特徴 $\mathbf{Z}_p$ を生成する．これらの特徴と，テキスト特徴 $\mathbf{A}$ の整合を取る．

### 3.4 目的関数と推論過程

#### 3.4.1 TCA の目的関数

動画像の各フレーム $t$（$1 \le t \le T$）に対して，区間回帰ヘッドと非行動分類ヘッドの出力に対してそれぞれ異なる損失関数 $L_{\mathrm{BR}}$ と $L_{\mathrm{BC}}$ を利用する．前者の $L_{\mathrm{BR}}$ は Distance Intersection over Union (DIoU) 損失 [19] を使用して，行動区間の境界までの距離を正確に回帰し，行動の正確な推定を目指す．後者の $L_{\mathrm{BC}}$ は Focal 損失 [20] を使用して，背景と行動間の不均衡なサンプルを効果的に扱う．TCA の全体的な損失関数 $L_{\mathrm{TCA}}$ は，各時間ステップ $t$ に対する上記の損失 $L_{\mathrm{BR}}(t)$ と $L_{\mathrm{BC}}(t)$ の重み付き和として定義される．

#### 3.4.2 VTA の目的関数

VTA は，動画像特徴とテキスト特徴という 2 つのモダリティ間の関係をモデル化することが目的である．そこで，対応する動画像とテキストペアの表現間の距離を最小化する．一方で，非対応ペアの距離を最大化する．損失関数 $L_{\mathrm{VTA}}$ は，2 つの対照項 $L_{\mathbf{Z}_p^0 \to \mathbf{A}}$ と $L_{\mathbf{Z}_p^* \to \mathbf{A}}$ で構成される．前者は動画像特徴 $Z_0^p$ とテキスト特徴 $A$ とを，後者は階層的な処理後の動画像特徴 $Z_*^p$ をテキスト特徴 $A$ と整合する．

#### 3.4.3 テスト時の処理

テスト時には, Soft Non-Maximum Suppression (Soft-NMS) [21] を使用して，重複する行動区間の提案を軽減する．TAD のための Soft-NMS は，行動区間候補の確率点数を，より高い点数の提案との temporal Intersection over Union (tIoU) に基づいて微調整する．既存の NMS は，重複する予測行動区間の中からスコアが最も高いものを選択し，他の行動区間を削除する．しかし，この方法では，複数の密接に関連する行動が同一画像内で同時に発生した場合，点数が低い行動区間をすべて削除してしまう．それに対して，Soft-NMS は重複する行動区間を完全に排除するのではなく，重複の度合いに基づいてそれらの点数を動的に調整する．これにより，動画像全体での行動の正確で包括的な検出を実現する．

## 4. 実　　験

ここでは，実験を通じて提案手法を評価する．さらに，提案手法の主要な特性について深く分析するために，Ablation Study を実施する．

### 4.1 実 験 設 定

**データセット**：TAD の研究で広く使用されている THUMOS14 [22] と ActivityNet-1.3 [23] データセットで評価した．THUMOS14 データセットには，20 の行動カテゴリを含む 413 本の動画像からなる．ActivityNet-1.3 データセットは，200 の行動カテゴリを含む約 20,000 本の動画像からなる大規模な行動データセットである．これらのデータセットをオープン語彙データ分割戦略 [3] に従って訓練セットとテストセットに分けた．この戦略は，行動カテゴリを特定の比率で無作為に分割し，各行動カテゴリを含む動画像を対応するセットへ分割する．分割割合として，「75/25 分割」と「50/50 分割」の 2 つの比率を用いた．前者は行動カテゴリと対応する動画像の 75% を訓練セットとして選択し，残りをテストセットとした．前者では無作為に 10 回，後者では無作為に 5 回の分割をした．後者は 50% を訓練セットとして選択し，残りをテストセットとした．さらに，Activity-Net1.3 データセットでは，既存手法 [3] で説明されている「スマート分割」戦略を採用し，行動カテゴリの階層を活用し，ラベルの 25% をテストセットに割り当てた．

**評価指標**：結果は，平均精度（mean Average Precision; mAP）により評価する．この指標は，tIoU に対する閾値に応じて，正しく推定された行動の割合を測定するものである．

**比較手法**：提案手法を以下の方法と比較した．

- ベースライン I：階層的な特徴抽出ブロックの層を 1 層にしたモデルで，提案手法の特徴ピラミッドの有効性を評価する．
- ベースライン II [24]：階層的な特徴抽出ブロックの各層に，

表 1: THUMOS14 および ActivityNet-1.3 データセットにおける結果．異なる tIoU 閾値での mAP を報告する．THUMOS14 は [0.30:0.10:0.70] の範囲，ActivityNet-1.3 は [0.50:0.05:0.95] の範囲での平均 mAP である．テキスト特徴の各グループ内で最も良い結果を**太字**で強調し，全体で最も良い結果を下線で示す．

| モデル | 画像/動画像特徴 | テキスト特徴 | THUMOS14 [22] | | | | | | | | ActivityNet-1.3 [23] | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 75/25 | | | | 50/50 | | | | スマート | 75/25 | 50/50 |
| | | | 0.3 | 0.5 | 0.7 | 平均 | 0.3 | 0.5 | 0.7 | 平均 | 平均 | 平均 | 平均 |
| OV-TAD [3] | CLIP B/16 | CLIP B/16 | 21.8 | 13.2 | 3.7 | 12.9 | 18.0 | 8.9 | 2.2 | 9.5 | 23.4 | 21.4 | 19.5 |
| | I3D | CLIP B/16 | 27.8 | 15.3 | 4.3 | 15.6 | **23.6** | 12.8 | 3.2 | 12.9 | 24.1 | 22.1 | 20.1 |
| | CLIP B/32 | CLIP B/32 | 21.4 | 11.8 | 3.5 | 12.0 | 15.4 | 7.7 | 1.9 | 8.0 | 22.6 | 19.4 | 17.3 |
| | I3D | CLIP B/32 | 25.1 | 13.9 | 4.0 | 14.1 | 21.0 | 11.3 | 3.0 | 11.5 | 23.9 | 20.2 | 18.2 |
| ベースライン I | I3D | CLIP B/16 | 33.6 | 19.6 | 4.3 | 19.2 | 15.9 | 8.9 | 2.0 | 8.9 | 20.4 | 16.8 | 13.3 |
| ベースライン II [24] | I3D | CLIP B/16 | 32.1 | 25.3 | 13.6 | 24.0 | 18.6 | 15.2 | 9.3 | 14.6 | 28.2 | 22.6 | 20.8 |
| 提案手法 | I3D | CLIP B/16 | **38.4** | **29.9** | **16.0** | **28.6** | 22.2 | **17.7** | **10.1** | **16.9** | **30.9** | **25.3** | **22.3** |
| OV-TAD [3] | CLIP L/14 | CLIP L/14 | 28.6 | 15.4 | 4.2 | 15.8 | 21.0 | 9.8 | 2.0 | 10.5 | 28.7 | 24.6 | 22.4 |
| | I3D | CLIP L/14 | 30.1 | 16.8 | 4.7 | 17.0 | **26.1** | 14.3 | 3.6 | 14.5 | 28.1 | 24.8 | **22.8** |
| ベースライン I | I3D | CLIP L/14 | 32.0 | 18.7 | 4.6 | 18.6 | 17.5 | 9.1 | 1.8 | 9.4 | 19.2 | 15.6 | 13.4 |
| ベースライン II [24] | I3D | CLIP L/14 | 35.3 | 26.6 | 13.6 | 25.5 | 18.7 | 15.8 | 9.6 | 15.0 | 28.8 | **24.9** | 21.1 |
| 提案手法 | I3D | CLIP L/14 | **39.1** | **30.7** | **16.5** | **29.1** | 21.4 | **17.6** | **10.8** | **16.8** | **29.8** | **24.9** | 21.7 |

一般的な Transformer エンコーダを組み込む 1 段階手法で，提案された大域・局所エンコーダを強調する．

- OV-TAD [3]：2 段階手法で，事前訓練済みの画像とテキストの共通埋め込みを使用するオープン語彙 TAD の既存手法である．

**実装の詳細**：提案手法を 3. の説明に従って実装し，TCA の階層的な特徴抽出ブロックは 6 層とした．動画像エンコーダには，Kinetics データセットで事前学習された 2 ストリームの Inflated 3D (I3D) ConvNet を使用して動画像特徴 [17] を抽出した．テキストエンコーダには，事前学習された Contrastive Language-Image Pre-training (CLIP) モデル [1] を使用してテキスト特徴を抽出した．CLIP モデルには，Large (L) や Base (B) など，複数の亜種がある．これらの名称は，モデルのアーキテクチャとサイズを示す．最適な結果を得るために，各データセットのモデルの複雑さと利用可能な訓練データを慎重に検討し，適切なハイパパラメータを選択した．

### 4.2 実験結果

表 1 に，提案手法と他の比較手法の性能を示す．THUMOS14 では mAP@[0.30:0.10:0.70] の範囲，ActivityNet-1.3 データセットでは mAP@[0.50:0.05:0.95] の範囲で平均された異なる tIoU 閾値での mAP を報告する．結果を CLIP B と CLIP L モデルのテキスト特徴を使用する 2 つのグループに分ける．

**THUMOS14 の結果**：提案手法は，表 1 で示される両グループにおいて，他の手法を大きな差で上回り，一貫して最高の結果を達成した．特に mAP@0.7 での結果は，ベースライン I と OV-TAD の手法をほぼ 3 倍の性能で上回り，提案手法の顕著な利点を示した．これは，階層的な特徴が未知行動検出をより正確に達成する上での大きな貢献を示す．Transformer エンコーダを使用するベースライン 2 と比較した場合でも，提案された大域・局所エンコーダの結果は優れた性能を示した．

表 2: TCA における時間的動画像抽出エンコーダを変更した比較．[0.30:0.10:0.70] の範囲での平均 mAP (↑) を示す．

| 動画像抽出エンコーダ | CLIP B/16 | | CLIP L/14 | |
|---|---|---|---|---|
| | 75/25 | 50/50 | 75/25 | 50/50 |
| 畳み込み [6] | 22.4 | 12.8 | 23.2 | 12.9 |
| 平均プーリング [7] | 21.8 | 12.0 | 24.3 | 13.1 |
| 最大プーリング [7] | 26.7 | 13.8 | 28.5 | 14.1 |
| Transformer [4] | 24.0 | 14.6 | 25.5 | 15.0 |
| 大域・局所（提案手法） | **28.6** | **16.9** | **29.1** | **16.8** |

**ActivityNet-1.3 の結果**：表 1 の両グループにおいて，提案手法の性能はほとんどの分割シナリオで他の手法を上回った．特にスマート分割では，提案手法は CLIP B グループで 10.5% まで，CLIP L グループで 10.6% の精度向上を示した．さらに，提案手法は 75/25 分割と 50/50 分割でも顕著な向上を示した．一方，ベースライン I はこのデータセットで最適ではなかった．これは，データセットの大規模で多様な性質，特に訓練中に十分に訓練されなかった多数のラベルの存在が原因と考えられる．これらの結果は，階層的な特徴抽出による提案手法がこれらの課題を克服し，優れた結果を達成する上での重要な貢献を示す．

### 4.3 Ablation Study

提案手法の効果を評価するために，THUMOS14 データセッ

トを使用し，「75/25 分割」と「50/50 分割」の評価設定において，エンコーダ部分を比較検討した結果を報告する．

比較検討の結果を表 2 に示す．提案した大域・局所エンコーダが他のエンコーダと比較して一貫して最高の mAP を達成していることがわかる．特に，75/25 と 50/50 の両方の比率で CLIP L/14 を使用した場合の精度が際立つ．異なるエンコーダの中では，最大値プーリングが平均プーリングや畳み込みよりも優れた性能を示している．また，Transformer と比較すると，提案されたエンコーダの優位性がわかる．この結果は，大域・局所エンコーダが時間的文脈の理解において重要な役割を果たすこと，および行動区間検出の精度と頑健性を向上させるために有効であることを示唆する．

## 5. むすび

本報告では，オープン語彙 TAD における大域・局所特徴の統合埋め込みによる効果的な時系列行動検出を提案した．提案した 1 段階手法は，時間的文脈分析 (TCA) とビデオ・テキスト統合 (VTA) で構成される．TCA は時間的行動特徴の抽出を目的とし，大域・局所統合埋め込みのための，大域・局所エンコーダを導入する．大域レベルでは行動と非行動を区別し，局所レベルでは特定の行動を識別するための詳細な時間的行動特徴を抽出する．VTA は動画像特徴と行動テキスト特徴のアラインメントを目的とする．THUMOS14 [22] および ActivityNet-1.3 [23] データセットを用いた実験により，提案手法の検出性能の優位性を確認した．

**謝辞** 本研究の一部は JSPS 科研費 JP21H03519 と JP24H00733 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．

### 文　献

[1] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” Proceedings of the 2021 International Conference on Machine Learning, pp.8748–8763, 2021.

[2] C. Ju, T. Han, K. Zheng, Y. Zhang, and W. Xie, “Prompting visual-language models for efficient video understanding,” Proceedings of the 17th European Conference on Computer Vision, vol.35, pp.105–124, 2022.

[3] V. Rathod, B. Seybold, S. Vijayanarasimhan, A. Myers, X. Gu, V. Birodkar, and D.A. Ross, “Open-vocabulary temporal action detection with off-the-shelf image-text features,” Computing Research Repository arXiv Preprints, arXiv:2212.10596, pp.1–33, 2022.

[4] C.-L. Zhang, J. Wu, and Y. Li, “ActionFormer: Localizing moments of actions with transformers,” Proceedings of the 17th European Conference on Computer Vision, vol.4, pp.492–510, 2022.

[5] D. Shi, Y. Zhong, Q. Cao, L. Ma, J. Li, and D. Tao, “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.

[6] C. Lin, C. Xu, D. Luo, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, and Y. Fu, “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.

[7] T.N. Tang, K. Kim, and K. Sohn, “TemporalMaxer: Maximize temporal context with only max pooling for temporal action localization,” Computing Research Repository arXiv Preprints, arXiv:2303.09055, pages=1-11, year=2023, pp.••–••, ••.

[8] M. Xu, C. Zhao, D.S. Rojas, A. Thabet, and B. Ghanem, “G-TAD: Sub-graph localization for temporal action detection,” Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.10156–10165, 2020.

[9] Y. Dong, J.-B. Cordonnier, and A. Loukas, “Attention is not all you need: Pure attention loses rank doubly exponentially with depth,” In Proceedings of the 2021 International Conference on Machine Learning, pp.2793–2803, 2021.

[10] Z. Qing, H. Su, W. Gan, D. Wang, W. Wu, X. Wang, Y. Qiao, J. Yan, C. Gao, and N. Sang, “Temporal context aggregation network for temporal action proposal refinement,” Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.485–494, 2021.

[11] C. Zhao, A.K. Thabet, and B. Ghanem, “Video self-stitching graph network for temporal action localization,” Proceedings of the 18th IEEE/CVF International Conference on Computer Vision, pp.13658–13667, 2021.

[12] R. Zeng, W. Huang, M. Tan, Y. Rong, P. Zhao, J. Huang, and C. Gan, “Graph convolutional networks for temporal action localization,” Proceedings of the 17th IEEE/CVF International Conference on Computer Vision, pp.7094–7103, 2019.

[13] C. Wang, H. Cai, Y. Zou, and Y. Xiong, “RGB stream is enough for temporal action detection,” Computing Research Repository arXiv Preprints, arXiv:2107.04362, pp.1–11, 2021.

[14] X. Liu, Q. Wang, Y. Hu, X. Tang, S. Zhang, S. Bai, and X. Bai, “End-to-end temporal action detection with transformer,” IEEE Transactions on Image Processing, vol.31, pp.5427–5441, 2022.

[15] W. Bao, Q. Yu, and Y. Kong, “OpenTAL: Towards open set temporal action localization,” Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.2979–2989, 2022.

[16] M. Chen, J. Gao, and C. Xu, “Cascade evidential learning for open-world weakly-supervised temporal action localization,” Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.14741–14750, 2023.

[17] J. Carreira and A. Zisserman, “Quo vadis, action recognition? A new model and the kinetics dataset,” Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, pp.6299–6308, 2017.

[18] Y. Wu and K. He, “Group normalization,” Proceedings of the 2018 European Conference on Computer Vision, vol.128, pp.3–19, 2018.

[19] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, “Distance-IoU loss: Faster and better learning for bounding box regression,” Proceedings of the 34th AAAI Conference on Artificial Intelligence, no.07, pp.12993–13000, 2020.

[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” Proceedings of the 16th IEEE International Conference on Computer Vision, pp.2980–2988, 2017.

[21] N. Bodla, B. Singh, R. Chellappa, and L.S. Davis, “Soft-NMS—Improving object detection with one line of code,” Proceedings of the 16th IEEE International Conference on Computer Vision, pp.5561–5569, 2017.

[22] H. Idrees, A.R. Zamir, Y.-G. Jiang, A. Gorban, I. Laptev, R. Sukthankar, and M. Shah, “The THUMOS challenge on action recognition for videos ”in the wild”,” Computer Vision and Image Understanding, vol.155, pp.1–23, 2017.

[23] C.H. Fabian, E. Victor, G. Bernard, and C.N. Juan, “ActivityNet: A large-scale video benchmark for human activity understanding,” Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, pp.961–970, 2015.

[24] N. Trung Thanh, K. Yasutomo, K. Takahiro, and I. Ichiro, “One-stage open-vocabulary temporal action detection leveraging temporal multi-scale and action label features,” 18th International Conference on Automatic Face and Gesture Recognition, pp.1–10, 2024.