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

信学技報
TECHNICAL REPORT OF IEICE.

Transformerを用いた人物行動検出

水野　颯介†　　柳井　啓司†

† 電気通信大学 大学院情報理工学研究科 情報学専攻
E-mail: †mizuno-s@mm.inf.uec.ac.jp, ††yanai@cs.uec.ac.jp

あらまし　ビデオ内の人間の行動を検出及び認識する行動検出のタスクにおいて，既存研究はCNNをベースにした手法が主流である．近年，自然言語処理で使用されるTransformerをコンピュータビジョンに活用したモデルが行動認識タスクにおいてCNNを用いた手法を上回っている．また，CNNとSelf-Attentionを組み合わせたモデルであるCoAtNetが画像認識タスクにおいて高い精度を達成した．しかし，Transformerをベースにした行動検出手法は少なく，CoAtNetをベースにした動画認識手法は存在しない．本論文では，Transformerをベースにした行動検出手法及び，CoAtNetを行動認識に拡張したVideo CoAtNetを提案する．実験の結果，Transformerベースの提案手法は，CNNベースの手法よりも高精度を達成し，あるクラスでは最大39%の精度向上を達成することを示した．

キーワード　行動認識，行動検出，Transformer

1. はじめに

これまでに, 深層学習を用いた行動認識の研究が広く行われてきた. そして近年では，行動認識技術の発展ともに，動画内における人間の検出及び行動の認識を同時に行う行動検出タスクが注目されている. 近年のコンピュータビジョンでは，自然言語処理モデルであるTransformerをベースにしたモデルや，CNNとSelf-Attentionを組み合わせたモデルCoAtNet [1] がCNNベースのモデルよりも高精度を達成している. しかし，Transformerをベースにした行動検出手法は少なく，CoAtNet [1] をベースにした行動認識手法は存在しない．本論文では，Transformerをベースにした1ステージ及び2ステージの行動検出手法と，CoAtNetを行動認識タスクに拡張したVideo CoAtNetの計3つの手法を提案する. 実験には，行動認識及び行動検出タスクのベンチマークデータセットであるUCF101-24 [2] とAVA [3] を使用した. 実験の結果，Transformerをベースにした2ステージの提案手法は，CNNベースの既存手法よりも高精度を達成した. 特に，人間と物体が相互作用するカテゴリで精度が向上し，あるクラスでは最大で39%の精度向上を達成した.

2. 関連研究

本論文は，人物検出と行動認識を同時に行う行動検出タスクの研究であり，物体検出，行動認識及び行動検出の研究に関連する．また，本研究で使用する行動認識モデルは，画像認識モデルを行動認識に拡張したモデルであるため，画像認識の研究についても紹介する．

2.1 物体検出

物体検出は，行動検出と非常に似たタスクであり，物体検出手法に着想を得た行動検出手法はとても多い．物体検出の手法は，物体の検出と分類を同時に行う1ステージ手法と，物体の検出後に分類を行う2ステージ手法に分けることができる．2ステージ手法ではFaster R-CNN [4] が代表例として挙げられる．当時の物体検出モデルでは，物体の領域候補の検出に画像処理手法であるSelective Search [5] を用いることが一般的であったが，Region Proposal Network(RPN) と呼ばれるCNNで構築することでEnd-to-Endな物体検出モデルを実現した．一方で，1ステージ手法の代表例としてYOLO [6] が挙げられる．Faster R-CNNと同時期に発表されたモデルであり，画像をグリッドに分割し，各グリッドセルに対して「物体が含まれる確率」と「物体が何であるか」の条件付き確率を計算することで検出と分類の同時実行を可能にした．1ステージ手法は，2ステージ手法に比べて高速に処理が可能である一方で，精度が劣ってしまったり，小さな物体の検出が困難であるという課題があった．しかし，近年ではTransformerベースの手法も数多く登場しており，中でもDETR [7] は，Transformerを初めて物体検出に応用したモデルで，1ステージ手法でありながら2ステージ手法と同等の性能を達成した．DETRでは，物体検出を直接集合予測問題として捉え，推論結果とground-truthの対応づけを二部マッチング問題として考えることでEnd-to-Endを実現した．本論文では，1ステージの行動検出手法にDETRを，2ステージの行動検出手法にFaster R-CNNを用いて，人物の検出を行う．

2.2 画像認識

画像認識では，CNNをベースにしたモデルが長年使われてきた．しかし，近年では自然言語処理モデルであるTransformerをベースにしたモデルや，CNNとTransformerのSelf-Attentionを組み合わせたモデルが最先端の性能を達成している．

TransformerをベースにしたモデルにVision Transformer [8] やSwin Transformer [9] が挙げられる．Vision Transformerは，画像パッチを単語のように扱い，パッチ全体に対してSelf-Attentionを計算してモデリングする．しかし，Transformerを言語タスクから画像タスクに適用する際に，入力トークン数が多いことや，Self-Attentionの計算コストが画像サイズの二乗分必要に

なることが課題としてあった．そこで，画像サイズに対してスケーラブルな汎用バックボーンとして提案されたのが Swin Transformer である．Swin Transformer では，パッチの集合であるウィンドウを定義し，ウィンドウ内で Self-Attention を計算することで局所的な関係をモデリングする．また，レイヤー毎にウィンドウをシフトさせることでウィンドウ間の関係をモデリングすることで，大域的な関係をモデリングする．このようにすることで，Swin Transformer は画像サイズに対してスケーラブルに Self-Attention を計算することが可能になった．

近年のコンピュータビジョンでは，主流のモデルが CNN ベースから Transformer ベースへと遷移してきている．しかし，CNN と Self-Attention には互いに長所がある．CNN は，少ないパラメータ数で効率良く学習ができることや，局所性を保持できるなどの利点がある．一方で Self-Attention は，入力全体を一度に処理できたり，入力によって重みが決まるという利点がある．CoAtNet [1] は，これらの CNN と Self-Attention の長所を組み込んだモデルである．全体の構造は，5 つのステージから構成されており，ステージ毎に画像解像度を落としながらチャネル数を増やして受容野を拡大していく．本論文では，CoAtNet を行動認識タスクへ拡張したモデルである Video CoAtNet を提案する．

2.3 行動認識

行動認識は，画像認識と同様に Transformer ベースのモデルが CNN ベースのモデルの性能を上回っている．CNN ベースのモデルとして SlowFast [10] が挙げられる．SlowFast は，低フレームレートで動作し空間特徴をモデリングする Slow pathway と，高フレームレートで動作し時間特徴をモデリングする Fast pathway で構成されるモデルである．CNN ベース手法の中でも性能が良く，ビデオタスクにおいて多く用いられている．一方で，Transformer ベースのモデルとして Video Swin Transformer [11] が挙げられる．Video Swin Transformer は，前述した画像認識モデル Swin Transformer のパッチ分割及びウィンドウの定義を時間方向に拡張したモデルであり，最先端の性能を達成している．本論文では，入力動画からコンテキスト特徴を抽出する際に Video Swin Transformer を用いる．

2.4 行動検出

近年，行動認識技術の発展と共に，行動検出のタスクが注目されている．行動検出とは，行動の認識と行動主 (アクター) の検出を同時に行うタスクのことである．Sun ら [12] は，アクターとコンテキストの関係をモデリングする Actor-Centric Relation Network(ACRN) を提案した．ACRN は，アクターの検出後に行動を認識するという 2 ステージの行動検出手法である．一方で，Köpüklü ら [13] はアクターの検出と行動の認識を同時に行う 1 ステージ手法である You Only Watch Onece(YOWO) を提案している．YOWO は，2DCNN と 3DCNN を並列にしたモデルであり，2DCNN と 3DCNN の特徴マップを効果的に融合する Channel Fusion and Attention Module(CFAM) が提案されている．本論文では，アクターとコンテキストの関係をモデリングするために ACRN を，チャネル間の関係をモデリングするために CFAM を用いる．

3. 手　　法

本論文では，行動検出手法を 2 つと行動認識を 1 つの計 3 つの手法を提案する．1 つ目は，行動検出手法 YOWO と物体検出手法 DETR を組み合わせた 1 ステージの行動検出手法である．2 つ目は，最先端の行動認識モデルである Video Swin Transformer と，既存の行動検出手法の ACRN や YOWO の CFAM を組み合わせた 2 ステージの行動検出手法である．3 つ目は，最先端の画像認識モデルである CoAtNet を行動認識に拡張した Video CoAtNet である．各手法について詳しく説明していく．

3.1 手法 1：YOWO ＋ DETR

1 つ目の提案手法は，YOWO [13] と DETR [7] を組み合わせた 1 ステージの行動検出手法である．図 1 に手法の全体像を示す．この手法では，2DCNN 特徴と 3DCNN 特徴を YOWO の CFAM で効果的に融合した特徴量を DETR に入力し，アクターの検出と行動の認識を同時に行う．YOWO では，Non Maximum Suppression(NMS) やアンカーの数・アスペクト比などを人手でチューニングする必要があるのに対し，手法 1 では DETR の二部マッチングを用いることでそれらを省くことができる．処理の流れを以下に示す．

1. 入力クリップを 3DCNN に，キーフレームを 2DCNN に入力して特徴マップを得る．
2. CFAM で 2DCNN 特徴マップと 3DCNN 特徴マップを融合し，新たな特徴マップ $F \in \mathbb{R}^{C\times H\times W}$ を得る．
3. 特徴マップ $F \in \mathbb{R}^{C\times H\times W}$ を 1×1Conv で $d(=512)$ チャネルに圧縮する．
4. サイズ $H\times W$ の特徴マップ d 個を，$H\times W$ 個の d 次元特徴として Transformer encoder へ入力する．
5. Transformer encoder は，$H\times W$ 個の中間行動特徴 (各 d 次元) を出力する．
6. $H\times W$ 個の中間行動特徴と N(=100) 個の行動クエリを Transformer decoder へ入力する．
7. Transformer decoder は，N 個の行動特徴 (各 d 次元) を出力する．
8. N 個の行動特徴をそれぞれ Feed Forward Network(FFN) に送り，N 個の推論結果 (バウンディングボックスとクラス) を出力する．

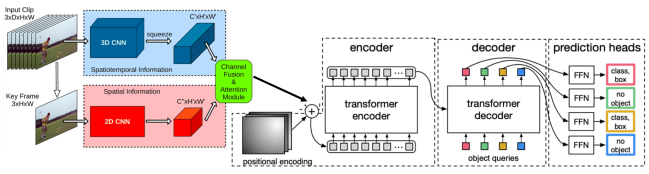

図 1　手法 1：YOWO+DETR

3.2 手法 2：Video Swin Transformer ＋ ACRN/CFAM

1 つ目の提案手法は，行動認識モデル Video Swin Transformer をベースに，ACRN や CFAM を組み合わせた 2 ステージの行動検出手法である．図 2 に手法のアーキテクチャを示す．

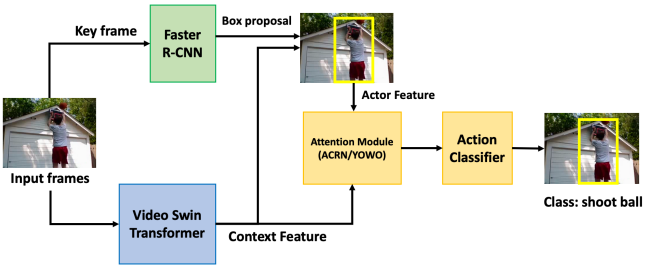

図 2　手法 2：Video Swin Transformer ＋ ACRN/CFAM

処理の流れを以下に示す．

1. 特徴マップと領域候補の抽出
   (a).入力クリップ $X \in \mathbb{R}^{C\times T\times H\times W}$ を Video Swin Transformer に入力し，コンテキスト特徴を抽出する．
   (b).キーフレーム $X_{t=\lfloor T/2\rfloor} \in \mathbb{R}^{C\times H\times W}$ を Faster R-CNN [4] に入力し，人物の領域候補を取得する．
2. RoIAlign を用いて，コンテキスト特徴から各領域候補のアクター特徴を抽出する．
3. ACRN 及び CFAM を用いて，アクターとコンテキスト間やチャネル間の Attention を計算する．
4. FC 層を適用し，分類スコアを得る．

3.2.1　Video Swin Transformer

入力クリップ $X \in \mathbb{R}^{3\times T\times H\times W}$ を Video Swin Transformer に入力し，コンテキスト特徴 $F \in \mathbb{R}^{8C\times \frac{T}{2}\times \frac{H}{32}\times \frac{W}{32}}$ を抽出する．Video Swin Transformer と同様に，パッチサイズを 2×4×4，ウィンドウサイズを 8×7×7 と定義する．また，$T(=32)$ は入力フレーム長，H, W は入力解像度，C は最初のステージの隠れ層のチャネル数を示し，Swin-T の時は $C = 96$，Swin-B の時は $C = 128$ である．

3.2.2　Faster R-CNN

キーフレーム $X_{t=\lfloor T/2\rfloor} \in \mathbb{R}^{C\times T\times H\times W}$ を Faster R-CNN [4] に入力し，人物の領域候補 $B = (b_i, ..., b_R)(b_i = (x_1^i, y_1^i, x_2^i, y_2^i))$ を取得する．本論文では，既存研究と同様に COCO データセット [14] で事前学習し，AVA データセット [3] で fine-tune したモデルを使用する．

3.2.3　Actor-Centric Relation Network(ACRN)

ACRN では，アクターとコンテキスト間の関係をモデリングする．まず，Video Swin Transformer から得られたコンテキスト特徴 $F \in \mathbb{R}^{C'\times T'\times H'\times W'}$ に時間プーリングを適用し，$F' \in \mathbb{R}^{C'\times H'\times W'}$ を得る．次に，Faster R-CNN から得られたアクターの領域候補 $B = (b_i, ..., b_R)$ に該当するコンテキスト特徴 F' の部分をクロップしたアクター特徴 $A_i \in \mathbb{R}^{C'\times h\times w}$ を得る．このアクター特徴 A_i に Average Pooling を適用し，アクター特徴を 1 次元ベクトル $a_i \in \mathbb{R}^{C'}$ に変換した後，コンテキスト特徴 F' の各位置に連結した $F_i'' \in \mathbb{R}^{2C'\times H'\times W'}$ を得る．その後，特徴マップ F_i'' に 1×1Conv と 3×3Conv を適用して最終的な特徴マップを得る．

3.2.4　Channel Fusion and Attention Module(CFAM)

CFAM は，画像変換や領域分割のタスクで用いられていたグラム行列に基づいた Attention メカニズムであり，チャネル間の関係をモデリングする．Video Swin Transformer から得られたコンテキスト特徴 $F \in \mathbb{R}^{C'\times T'\times H'\times W'}$ に時間プーリングを適用し，$F' \in \mathbb{R}^{C'\times H'\times W'}$ を得る．次に，各チャネルを 1 次元ベクトルに変換した $F'' \in \mathbb{R}^{C'\times N}(N = H'\times W')$ を得る．そして，式 1 でチャネル間の特徴相関を表すグラム行列 $G \in \mathbb{R}^{C\times C}$ を求め，softmax を適用することでチャネル間の Attention マップ $M \in \mathbb{R}^{C\times C}$ を得ることができる．

$$G = F \cdot F^T, G_{ij} = \sum_{k=1}^{N} F_{ik} \cdot F_{jk} \tag{1}$$

元の特徴マップ F'' に Attention マップ M の影響を適用するために F'' と M を乗算し，元の入力特徴 F' と同じ形状に変形した $\bar{F''} \in \mathbb{R}^{C\times H\times W}$ を得る．そして，式 2 のように $\bar{F''}$ を元の入力特徴マップ F' に要素和演算を用いて学習可能なパラメータ α で結合した特徴マップ $C \in \mathbb{R}^{C\times H'\times W'}$ を得る．α は，初期値が 0 で徐々に重みを学習していく．

$$C = \alpha \cdot \bar{F''} + F' \tag{2}$$

3.3　手法 3：Video CoAtNet

2 つ目の提案手法は，最先端の画像認識モデルである CoAtNet [1] を行動認識タスクへ拡張した Video CoAtNet である．図 3 に全体のアーキテクチャを示す．CoAtNet と同様に 5 つのステージ (S0，S1，S2，S3，S4) から構成されており，最初のステージ S0 は 2 層の Conv 層から成り，以降の S1 から S4 は Mobile Inverted Bottleneck(MB) ブロックか Transformer ブロックで構成される．ステージ S1-S2-S3-S4 の組み合わせは，MB ブロックを C，Transformer ブロックを T とすると，CoAtNet で最も精度が高かった C-C-T-T の組み合わせを用いた．

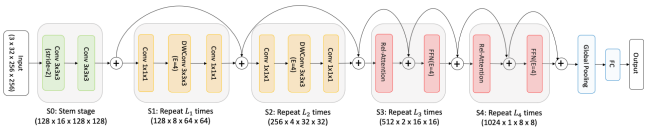

図 3　手法 3：Video CoAtNet

図 3 からも分かるように，Video CoAtNet ではステージ毎に時空間解像度を小さくしていく一方で，チャネル数 D は大きしモデルの受容野を拡大していく．各ステージの繰り返し回数 L とチャネル数 D をまとめたものを表 1 に示す．

表 1　各ステージの詳細

Stages	Size	L	D
S0-Conv	1/2	2	128
S1-MBConv	1/4	2	128
S2-MBConv	1/8	6	256
S3-Transformer	1/16	14	512
S4-Transformer	1/32	2	1024

3.3.1　MB ブロック

MB ブロックは，MobileNetv2 [15] で提案されており，Depth-wise 畳み込みと Pointwise 畳み込み [16] から構成される．通

常の畳み込みは，空間方向とチャネル方向を同時に畳み込むのに対して，Depthwise 畳み込みは空間方向のみを，Pointwise 畳み込みはチャネル方向のみを畳み込む．MB ブロックでは，Pointwise 畳み込みでチャネル数を 4 倍に増やし，Depthwise 畳み込んだ後に Pointwise 畳み込みで元のチャネル数に戻すという流れになる．

3.3.2 Transformer ブロック

Transformer ブロックでは，通常の Self-Attention に加えて，Video Swin Transformer [11] の 3D Shifted Window を用いた Self-Attention の場合でも実験を行った．

4. 実　　験

4.1 実 験 設 定

データセット　本実験では，行動認識及び行動検出のタスクで広く用いられているビデオデータセットである UCF101-24 [2] と AVA [3] を使用する．UCF101-24 は提案手法 1 で，AVA は提案手法 2 と 3 で使用する．

UCF101-24 は，行動認識タスクのベンチマークデータセットである UCF101 のサブセットである．UCF101 とは異なり，動画の全フレームに対して行動クラスとバウンディングボックスが 1 つ付与されており，クラス数は 24 個である．

AVA は，ビデオの各クリップは YouTube から収集されており，総ビデオクリップ数は約 5 万 7 千個，総ラベル数は約 21 万個，クラス数は 80 個と大規模なデータセットである．図 4 に AVA の例を示す．図 4 のように，キーれフレームに行動ラベルとバウンディンボックスが複数付与されている．評価時には，サンプル数が少なくとも 25 個以上存在する計 60 クラスを使用する．

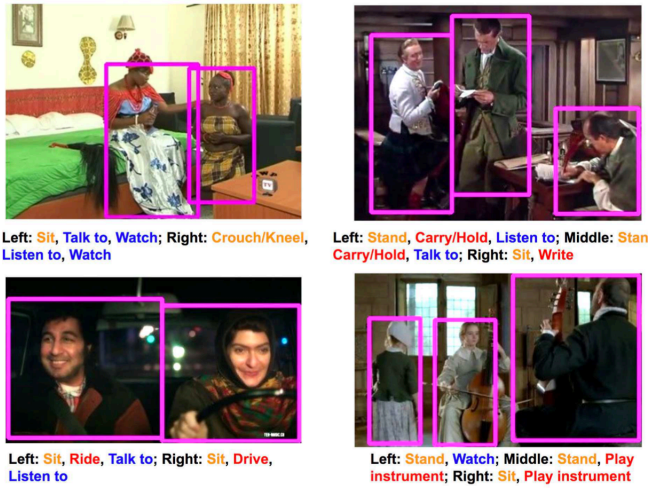

図 4　AVA データセットの例 ([3] から引用)

評価指標　提案手法 1 では，各クリップのキーフレームの行動インスタンスを評価し，定位精度と分類精度を使用する．定位精度とは，予測されたバウンディングボックスと ground-truth の IoU が 0.5 より高い行動インスタンスを True Positive とした時の recall のことである．分類精度とは，バウンディングボックスと ground-truth の IoU が 0.5 より高く，行動クラスが正しく分類された割合のことである．提案手法 2 と 3 では，各クリップのキーフレームの行動インスタンスを評価し，mAP(mean Average Precious) を使用する．True Positive の行動インスタンスでは，予測された人間のバウンディンボックスと ground-truth の間の IoU が 0.5 よりも高く，予測クラスと ground-truth が同じである．

学習設定　Video Swin Transformer は，ImageNet [17] と Kinetics [18] [19] で事前に学習させる．具体的には，Swin-Tiny は ImageNet-1K と Kinetics400 [18] で，Swin-Base は ImageNet-22K と Kinetics600 [19] で事前学習する．一方 Video CoAtNet の実験では，Kinetics [18]〜[20] で事前学習を行わずに AVA で学習する．Kinetics は，公開されているビデオデータセットの中で最も巨大であり，学習に膨大な計算コストを必要とするため，今回の実験では Kinetics での事前学習を行わずに AVA で学習を行う．それに伴い，Video Swin Transformer における Kinetics での事前学習なしの場合でも実験を行う．オプティマイザは，全ての手法で AdamW [21] を使用する．提案手法 1 には，初期学習率が 3×10^{-4}，重み減衰が 5×10^{-2} の AdamW を使用する．提案手法 2 には，初期学習率が 10^{-4}，重み減衰が 10^{-3} の AdamW を使用する．スケジューラは，linear warmup と cosine decay を設定した．バッチサイズは 64(1GPU あたり 8 動画，8GPU) で 20 エポックの学習を行った．

4.2 手法 1 の実験結果

YOWO と DETR を組み合わせた手法 1 と YOWO の精度を比較したものを表 2 に示す．手法 1 は定位精度が 66.9%，分類精度が 78.9% と YOWO を下回る結果となった．しかし，表 2 内の DETR の部分と手法 2 を比較すると，2DCNN 特徴のみ 3DCNN 特徴のみを利用する場合よりも精度が向上していることが分かる．これは，CFAM が効果的に働いていることを示す．また，2DCNN 特徴と 3DCNN 特徴の場合を比較すると，2 つのことが分かる．1 つ目は，2DCNN 特徴のみの方が分類精度が高くなっている点である．これは，UCF101-24 データセットは空間情報がとても重要であることが既存研究でも報告されていることから妥当な結果であると言える．2 つ目は，3DCNN 特徴のみの方が定位精度が高くなっている点である．これは，時間情報の恩恵が得られていると考えられる．

表 2　手法 1 の実験結果

手法		定位精度	分類精度
YOWO	2D	91.7	85.9
	3D	90.8	**92.9**
	2D ＋ 3D ＋ CFAM	**92.7**	92.3
DETR	2D	58.6	**75.5**
	3D	**75**	72.8
YOWO+DETR(手法 1)		66.9	**78.9**

4.3 手法 2 の実験結果

4.3.1 最先端手法との比較

Video Swin Transformer と ACRN 及び CFAM を組み合わせた提案手法 1 と最先端手法の精度の比較を表 3 に示す．表 3 より，提案手法 1 は 28.3mAP を達成し，ACAR [22] に次いで 2

番目に高い精度を達成していることが分かる．しかし，ACAR は推論時に特徴バンク ACFB を用いて学習時の入力フレーム数よりも多くのフレームを使用している．そこで，特徴バンク ACFB を使用しない場合で比較したところ，提案手法 2 の方が 0.5mAP 高い精度を達成した．

表 3　手法 2 と最先端手法の精度比較

model	Reference	pre	mAP
AVA baseline [3]	CVPR18	K400	15.6
ACRN [12]	ECCV18	K400	17.4
YOWO [13]	arXiv19	K400	19.2
LFB [23]	CVPR19	K400	27.6
SlowFast, R50 [10]	ICCV19	K400	24.7
SlowFast, R101 [10]	ICCV19	K600	27.3
Context-Aware [24]	ECCV20	K400	28.0
X3D-XL [25]	CVPR20	K400	26.1
ACAR [22], R50	CVPR21	K400	**28.8**
ACAR w/o ACFB [22]	CVPR21	K400	27.8
WOO, SFR50 [26]	ICCV21	K400	25.2
WOO, SFR101 [26]	ICCV21	K600	28.0
Swin-T	-	K400	21.4
Swin-B	-	K600	26.5
Swin-B+ACRN	-	K600	**28.3**
Swin-B+CFAM	-	K600	26.5

4.3.2　クラス/カテゴリ毎の比較

AVA データセットの全 80 クラスは，大きく「Person Movement」,「Object Manipulation」,「Person Interaction」の 3 カテゴリに分けることができる．「Person Movement」カテゴリは「jump」などの人間のみに関係する行動クラス，「Object Manipulation」は「answer phone」などの人間が物体を操作する行動クラス，「Person Interaction」は「hand shake」などの人間同士の行動クラスである．ここでは，SlowFast [10] を基準にした場合の提案手法 1 の精度改善をクラス毎に示したものを図 5 に，カテゴリ毎に示したものを図 6 に示す．また図中では，「Person Movment」を赤色，「Object Manipulation」を緑色，「Person Interaction」を青色で示している．図 5 より，「Object Manipulation」カテゴリに属するクラスの精度がよく改善されていることが分かる．最も精度が向上したクラスは，「play musical instrument」で約 39% 向上した．一方で，「Person Movement」カテゴリに属するクラスの中には，精度が低下しているものも多いことが分かる．

実際に，「Object Manipulation」カテゴリは約 4.6%，「Person Interaction」カテゴリは約 1.2% 精度が向上していた．これは，Video Swin Transformer を用いることでより局所的な関係をモデリングでき，人間と物体が関係する「Object Manipulation」カテゴリの精度が向上したと考えられる．一方で，「Person Movement」カテゴリは精度が約 1.1% 低下した．これは，「Person Movement」カテゴリは 1 人の人間のみに着目するため，Video Swin Transformer の強みである局所的なモデリングを活かせていないことと，SlowFast の Fast pathway が時間情報を良く捉えていることが原因であると考えられる．

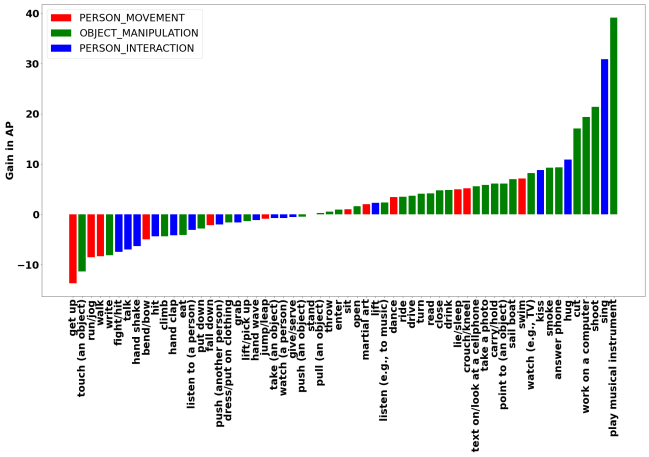

図 5　クラス毎の精度改善

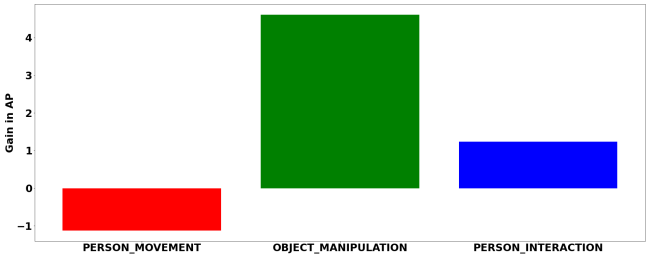

図 6　カテゴリ毎の精度改善

4.4　手法 3 の実験結果

2 つ目の提案手法である Video CoAtNet と最先端の行動認識モデルである Video Swin Transformer との精度比較を表 4 に示す．表 4 より，Video CoAtNet は Video Swin Transformer の精度を下回る結果となった．精度が低下した原因は，2 つ考えられる．1 つ目は，MB ブロックが行動認識にあまり適していなかった可能性があることである．実際に，MB ブロックは画像認識モデルで多く用いられるが，行動認識モデルで用いられることはあまりない．精度向上のためには，行動認識用に最適な CNN ブロックを模索する必要がある．2 つ目は，本来の性能を発揮できていない可能性があることである．CoAtNet は，画像認識の大規模データセットである ImageNet-1k(画像数：約 130 万枚) と JFT-300M(約 3 億枚) で最先端の精度を達成している．今回の実験に使用した AVA は，約 5.7 万本の動画から構成されており，行動認識の大規模データセットである Kinetics-400 [18](動画数：約 30 万本)，Kinetics-600 [19](約 50 万本)，Kinetics-700 [20](約 65 万本) に比べると小さなデータセットと言える．よって，本来の性能を確認するためにも Kinetics での実験は必要である．

表 4　Video Swin Transformer vs. Video CoAtNet

model	mAP
Video Swin Transformer	**8.77**
Video CoAtNet(Self-Attention)	1.83
Video CoAtNet(Window-Attention)	1.86

5. おわりに

本論文では，行動検出と行動認識のタスクに取り組み，3 つの手法を提案した．1 つ目は，行動検出手法である YOWO [13] と物体検出手法である DETR [7] を組み合わせた 1 ステージ手法である．実験の結果，通常の YOWO よりも精度が劣る結果となった．2 つ目は，行動認識モデルである Video Swin Transformer [11] と行動検出手法である ACRN [12] 及び CFAM [13] を組み合わせた手法である．ACRN を組み合わせた場合に，特徴バンクを使用しない手法の中で最も高い精度を達成することができた．また，SlowFast [10] をベースにしたモデルと比較して，人間と物体が相互作用する行動クラスの精度を向上することができた．3 つ目は，最先端の画像認識モデルである CoAtNet [1] を行動認識に拡張した Video CoAtNet である．Video Swin Transformer と比較した結果，精度が劣る結果となってしまった．

手法 2 は，AVA データセット [3] の行動クラスが人物と物体の相互作用を表すものが多いため，物体の検出を考慮すると更なる精度向上が期待できると考える．そのため，検出器で人間と物体の両方を検出し，人間と物体及びコンテキスト情報を考慮したモデルを提案し，精度改善を目指したい．また，AVA はロングテールなデータセットであるため，ロングテールを解消するような工夫を施すと更なる精度改善が期待できると考える．

手法 3 の Video CoAtNet は，MB ブロック [15] ではなく行動認識に最適な CNN ブロックを模索する必要がある．また，CoAtNet は大規模データセットで本来の性能を発揮するので，行動認識タスクの大規模データセットである Kinetics での実験も行う予定である．

文　献

[1] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *arXiv preprint arXiv:2106.04803*, 2021.

[2] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[3] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proc.of IEEE Computer Vision and Pattern Recognition*, pp. 6047–6056, 2018.

[4] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, Vol. 28, pp. 91–99, 2015.

[5] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, Vol. 104, No. 2, pp. 154–171, 2013.

[6] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proc.of IEEE Computer Vision and Pattern Recognition*, pp. 779–788, 2016.

[7] C. Nicolas, M. Francisco, S. Gabriel, U. Nicolas, K. Alexander, and Z. Sergey. End-to-end object detection with transformers. In *Proc.of European Conference on Computer Vision*, 2020.

[8] 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 *Proc.of International Conference on Learning Representations*, 2021.

[9] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *Proc.of IEEE International Conference on Computer Vision*, 2021.

[10] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proc.of IEEE International Conference on Computer Vision*, pp. 6202–6211, 2019.

[11] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv preprint arXiv:2106.13230*, 2021.

[12] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Kevin Murphy, Rahul Sukthankar, and Cordelia Schmid. Actor-centric relation network. In *Proc.of European Conference on Computer Vision*, pp. 318–334, 2018.

[13] Okan Köpüklü, Xiangyu Wei, and Gerhard Rigoll. You only watch once: A unified cnn architecture for real-time spatiotemporal action localization. In *arXiv preprint arXiv:1911.06644*, 2019.

[14] Tsung-Yi Lin, Michael Maire, Serge J Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc.of European Conference on Computer Vision*, 2014.

[15] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proc.of IEEE Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018.

[16] Bing Liu, Danyin Zou, Lei Feng, Shou Feng, Ping Fu, and Junbao Li. An fpga-based cnn accelerator integrating depthwise separable convolution. *Electronics*, Vol. 8, No. 3, p. 281, 2019.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc.of IEEE Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[18] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[19] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018.

[20] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc.of International Conference on Learning Representations*, 2018.

[22] Junting Pan, Siyu Chen, Mike Zheng Shou, Yu Liu, Jing Shao, and Hongsheng Li. Actor-context-actor relation network for spatio-temporal action localization. In *Proc.of IEEE Computer Vision and Pattern Recognition*, pp. 464–474, 2021.

[23] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proc.of IEEE Computer Vision and Pattern Recognition*, pp. 284–293, 2019.

[24] Jianchao Wu, Zhanghui Kuang, Limin Wang, Wayne Zhang, and Gangshan Wu. Context-aware rcnn: A baseline for action detection in videos. In *Proc.of European Conference on Computer Vision*, pp. 440–456. Springer, 2020.

[25] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proc.of IEEE Computer Vision and Pattern Recognition*, pp. 203–213, 2020.

[26] Shoufa Chen, Peize Sun, Enze Xie, Chongjian Ge, Jiannan Wu, Lan Ma, Jiajun Shen, and Ping Luo. Watch only once: An end-to-end video action detection framework. In *Proc.of IEEE International Conference on Computer Vision*, pp. 8178–8187, 2021.