時系列行動セグメンテーションを用いた自動車組立映像の解析

久保　莞太[1,a)]　日下部　尊[1]　永井　裕也[1]　濱田　悠樹[1]　廣瀬　雄大[1]　森本　文哉[1]
宮田　将光[1]　玉城　大生[1]　久冨　あすか[2]　伊藤　浩隆[2]　東園　雄太[2]　小野　智司[1,b)]

概要：近年，自動車組立作業をはじめとする生産現場において，作業員の行動解析の要望が高まっており，作業手順の把握と作業時間の計測を自動化する手段として，時系列行動セグメンテーションの応用が期待されている．しかし，本研究で対象とする自動車組立作業映像は，一般的なデータセットと比較して，撮影映像内において注目すべき作業員よりも自動車車体が大きな領域を占めており，さらにそれが常に移動している点に難しさがある．本研究では，既存の時系列行動セグメンテーション手法を本問題に適用し，その有効性を検証する．加えて，単視点と多視点の映像による比較実験を行い，注目領域の情報量の変化による性能への影響を検証する．

Analysis of Automobile Assembly Work Videos by Temporal Action Segmentation

Abstract: In recent years, there has been a growing demand for analysis of worker behavior in factories such as automobile assembly. Temporal action segmentation is expected as a solution that can confirm work procedures and the time required for each task. However, compared to general datasets, automobile assembly work videos targeted in this study are difficult because automobiles occupy larger areas than the worker of interest in the video, and the automobiles are constantly moving. Therefore, this study verifies the effectiveness of several existing temporal action segmentation methods on this problem. This study also attempts to use multiview videos to improve temporal segmentation accuracy.

1. はじめに

近年，製造業における人手不足解消と作業効率向上の観点から，組立作業の行動解析の要望が高まっている．組立作業における行動解析を行うことにより，各作業に要する時間の計測の自動化が可能となるほか，作業員がマニュアルと同様の手順で作業を行っているかの確認を行うことが可能となる．このような需要の高まりから，新たな行動解析技術として深層ニューラルネットワーク（Deep Neural Network: DNN）を活用した時系列行動セグメンテーション，すなわち，図 1 に示すような，映像を構成するフレーム単位で行動クラスの認識を行う技術が広く研究されている [1–15].

本研究では，自動車組立工場における作業員の行動解析に着目する．一般に，時系列行動セグメンテーションタスクとして適用・公開されているデータセットは，料理を行う人物 [15–18] やおもちゃを組み立て・分解する人物 [1] など，行動する人物や行動に伴う物体が映像の画角内において主体となることが多い．料理を対象としたデータセットを例とすると，ほとんどの行動は手がどのように動いているか，どのような食材や道具を使用しているかということのみに着目している．これに対して，自動車組立における映像は，作業員よりも自動車車体が大きく写り，低速ではあるが，画角内を車体が移動しているため，DNN が作業員に注目することを妨げる恐れがある．また，多くの時系列行動セグメンテーション手法では画角全体を特徴抽出するため，異なる行動クラス間で同じ値を持つ特徴量を含む場合があり，行動クラスの判別がより困難となることが想

[1] 鹿児島大学
Korimoto, Kagoshima, Kagoshima 8900065, Japan
[2] トヨタ車体研究所
Kokubu, Uenodan, Kirishima, Kagoshima 8994461, Japan
[a)] k0243350@kadai.jp
[b)] ono@ibe.kagoshima-u.ac.jp

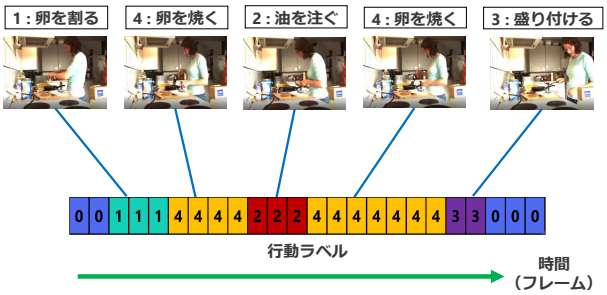


図 1: 時系列行動セグメンテーションの例
（目玉焼き料理映像）

定される．

本研究では，自動車組立映像における新たなデータセットを作成し，既存の複数の時系列行動セグメンテーション手法を自動車組立映像に適用することで，その有効性を検証する．作成したデータセットは一般的なデータセットと比較して，人や物体が画角内を占める割合が大きく異なるため，既存手法が本問題においても有効であるかどうかは不明である．よって，複数の既存手法を 5 分割交差検証を行うことで汎化性能を評価し，本データセットに対して最も有効な手法を模索する．

また，本研究は，入力として単視点映像のみを用いる場合と，作業員を多視点から撮影した映像を用いる場合との比較実験を行う．単視点映像では作業員が遮蔽物に隠れる恐れや，画角によって作業員の手元や移動の把握が困難となる場合がある．このため，別視点から作業員の特徴量を抽出することで，注目領域の情報量を増やすことによる認識精度の向上を期待し，注目領域の情報量の変化による性能への影響を検証する．

2. 関連研究

2.1 時系列行動セグメンテーションの概要

時系列行動セグメンテーションは，映像を時間によって各行動区間に分割（セグメント化）することを目的とした映像認識タスクである [12]．各セグメントは事前に定義された行動ラベルの 1 つが割り当てられる．すなわち，映像を構成するフレーム単位で行動クラスの認識を行うことを目的とする．長さ T の映像 $x = (x_1, x_2, \ldots, x_T)$ が与えられたとき，時系列行動セグメンテーションモデルを形式的に表すと次のような出力となる．

$$s_{1:N} = (s_1, s_2, \ldots, s_N) \tag{1}$$

ここで，N は行動セグメント数，$s_n = (c_n, l_n)$ は行動ラベル c_n が割り当てられた長さ l_n の連続した行動セグメントを表す．また，式 (1) はフレーム単位に行動ラベルを出力する [3] とみなすこともでき，その場合以下の式で表される．

$$y_{1:T} = (y_1, y_2, \ldots, y_T) \tag{2}$$

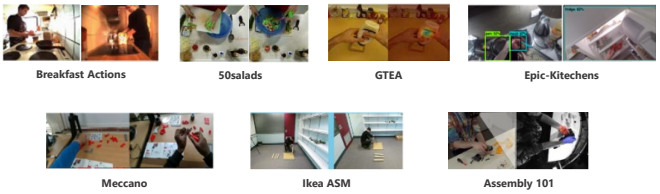


図 2: 時系列行動セグメンテーションにおける主なデータセットの一覧 [24]

ここで，y_t はフレーム単位の行動ラベルを表す．これら 2 つの表現は同一であるが，研究の目的に応じて使い分けられる．式 (1) や式 (2) で表されるような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．どのような行動がいつ開始され終了するのかを把握することができ，人間が次にどのような行動を取るのかを理解することが可能となる．このような特徴から，監視システムなどの映像におけるセキュリティ向上や，料理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が期待されている．

2.2 データセット

時系列行動セグメンテーションの主要なデータセットの例を図 2 に示す．図 2 で示したデータセットは，breakfast actions [15], GTEA [16], 50salads [18], Epic-Kitechens [19] などの料理映像が多く，ベンチマークとして頻繁に使用される．その他に，おもちゃや家具の組立映像 Meccano [20], Ikea ASM [21]，Assembly101 [1] も存在する．時系列行動セグメンテーションのデータセットを作成する際，映像を構成する全てのフレーム単位で正しい行動クラスをアノテーションする必要があり，時間的コストが高い．このため，ImageNet [22] や Kinetics [23] のような大規模なデータセットが存在せず，タスク毎にデータセットを作成する必要がある．

2.3 時系列行動セグメンテーションモデル

時系列行動セグメンテーションに関する研究は広く行われており，主に深層学習ベースの手法が多く提案されている．近年では，ネットワークアーキテクチャとして時間畳み込みネットワーク Temporal Convolutional Network（TCN），Recurrent Neural Network（RNN），Transformer が組み込まれることが多い．

Lea らは時間的に広い範囲の特徴を捉えるための TCN をベースとした手法を提案した [11]．特にエンコーダ・デコーダ型の TCN は，プーリングとアップサンプリングを用いて効率的に広い範囲の時間的特徴を捉えることを可能にした．Huang らは Gated Recurrent Units（GRU）と Graph Convolution Network（GCN）をベースとした Graph-based Temporal Reasoning Module（GTRM）を提

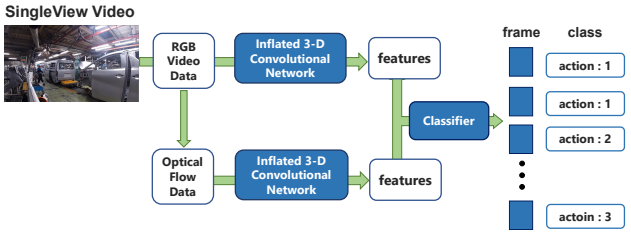


図 3: 提案方式の構成と処理手順

案した [10]．この手法は，GRU によりフレーム入力を時系列順に取り込み，GCN により隣接するノード（行動）との関係に基づいて特徴を捉えることを可能にした．また，Yi らはエンコーダと複数のデコーダで構成される Transformer をベースとした ASFormer を提案し，フレームの時間的関係を捉えることを可能にした [5]．

しかし，これらの時系列行動セグメンテーション手法は，「料理」 [15,16,18] や「おもちゃの組立・分解」 [1]，「ネジ締め動作」 [2] などの手元の細かい作業動作解析に適用した程度である．また，自動車組立における作業員の行動解析では，注目すべき作業員よりもラインを流れる自動車車体の方が映像の画角内を占める面積が大きいため，時系列行動セグメンテーション手法が作業員の領域を注視することを妨げる恐れがある．

3. 研究方法

3.1 方針および特徴

本研究では，3.4 節にて後述する 6 つの既存の時系列行動セグメンテーション手法を適用することで，自動車組立作業における行動解析を試み，最適な時系列行動セグメンテーション手法を模索する．しかし，単視点映像のみを用いて DNN を学習させる場合，カメラから遠方の位置で作業をする人物の画像領域が極めて小さくなり，作業員の手元や移動の把握が困難となる場合や，作業員がドアなどの遮蔽物に隠れる場合があるため，識別が困難となる恐れがある．このため，既存手法が自動車組立作業映像においても有効であるかどうかは不明である．

よって本研究では，単視点映像のみを用いた検証に加え，多視点映像を用いた検証を行う．多視点映像を用いることで，別視点から作業員の特徴量を抽出し，DNN に注目させることで認識精度の向上を期待する．また，多視点映像における検証は 3.5 節にて後述する 2 つの方法で行う．

3.2 提案方式の構成と処理手順

時系列行動セグメンテーション手法の多くは，図 3 に示すように RGB 映像特徴とオプティカルフロー特徴とを入力とする．すなわち，特徴抽出器として Inflated 3D ConvNet（I3D）[25] を用いることを前提としている．本研究では図 3 における行動分類モデルとして，3.4 節にて後述する 6 手法を用いて比較を行う．

3.3 Two-Stream Inflated 3D ConvNet（I3D）

I3D は，Two-Stream ConvNet [26] を応用した行動認識のための深層学習モデルであり，近年の時系列行動セグメンテーションにおいて，多くの手法の映像用特徴抽出器として用いられる．I3D は，図 3 に示すように，映像の動きを含めた特徴を捉えるために，それぞれ RGB 画像とオプティカルフローを入力とする 2 つのストリームを有している．RGB 画像をもとにフレーム内の物体やその位置を空間的特徴として学習し，オプティカルフローをもとに物体やシーンの動きなどの時系列的な特徴を学習する．また，ネットワーク部分を 2D ConvNet から 3D ConvNet に膨張（Inflated）させ，空間要素に時間要素を合わせた 3 次元で畳み込みを行い，複数フレーム画像から動きを解析する．

本実験では，Kinetics データセット [23] で事前学習させた I3D モデルを用い，各ストリームで抽出された特徴量を連結させ，これを行動分類モデルへの入力とする．

3.4 行動分類モデル

本研究では，6 種類の手法を行動分類器として用いる．

- TCN を複数積み重ねた手法である Multi-Stage TCN（MS-TCN） [3]
- MS-TCN を改良した MS-TCN++ [4]
- エンコーダと複数のデコーダで構成される Transformer をベースとした手法である ASFormer [5]
- ASFormer と同様のエンコーダを用い，デコーダ部分をフレーム単位の行動ラベルの代わりに行動セグメントを出力させるように改良した Unified Video Action Segmentation model via Transformers（UVAST） [6]

上記の 4 つのモデルに加えて，既存のモデルの性能を向上させる手法として下記の 2 種類を用いる．

- ドメインの特徴空間のズレを減らすことでソースモデルのターゲットデータに対する性能を向上させる手法である Self-Supervised Temporal Domain Adaptation（SSTDA） [7]
- モデルが予測したセグメントに対して後処理として修正を加える手法である Action Segment Refinement Framework（ASRF） [9]

上記 2 手法において，ベースとなる分類モデルには MS-TCN を使用する．

3.5 多視点映像を用いた検証

本研究では，単視点映像のみを用いた検証に加え，多視点映像を用いた検証を行う．多視点映像の利用は，以下の 2 通りの方法を検証する．

(1) **映像連結**: 図 4 に示すように，異なる 3 つの単視点映像を連結し，1 本の映像として特徴抽出を行い，行動分類モデルの学習に用いて評価を行う．映像を連結し特徴抽出する利点として，3 つの映像の特徴量を得る

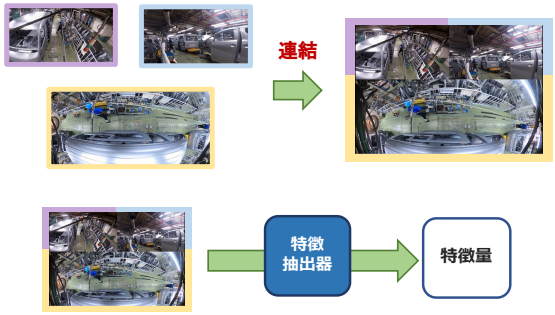


図 4: 多視点映像（映像連結）の処理手順

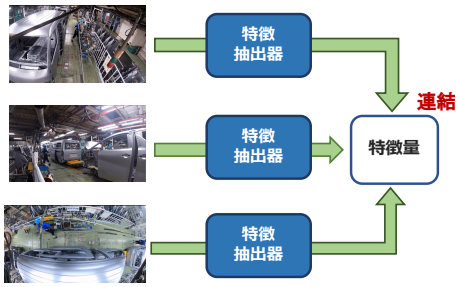


図 5: 多視点映像（特徴量連結）の処理手順

ことができ，かつ，単視点映像を個別に特徴抽出する場合と比べて学習に必要な時間を抑えることができる点が挙げられる．

(2) **特徴量連結**: 図 5 に示すように，異なる 3 つの単視点映像を個別に特徴抽出した後に，特徴量を連結して行動分類モデルの学習に用いて評価を行う．単視点映像を個別に特徴抽出した後に特徴量を連結する利点として，行動分類モデルへの入力となる特徴量の次元が 3 倍となり，学習に用いられる情報量が増加する点が挙げられる．

これら 2 つの検証を行うことで単視点映像と多視点映像との比較を行い，注目領域の情報量を増やすことにより性能が向上するか検証する．

4. 評価実験

4.1 解析対象となる映像の特徴

3 章で述べた方法に従い，自動車組立作業を行う作業員を撮影し，時系列行動セグメンテーションを試みる．対象とする作業映像は，ライン上を常に移動する車体に対してフロントドアとその付近の部品を取り付ける工程を担当する作業員を撮影したものである．2 台のカメラおよび 1 台の 360 度カメラで撮影を行った．各カメラの視点を図 6 に示す．このとき，360 度カメラの視点は「視点 3」に相当する．

撮影された 1 本の映像は，1 名の作業員が 1 台に対して作業を行う様子が含まれる．作業員が行動する範囲は，ライン上の前後約 7 メートル程度と幅広いため，画角の広いカメラで撮影を行った．このため，カメラから遠方の位置

(a) 視点 1　(b) 視点 2　(c) 視点 3

図 6: 3 台のカメラ視点

図 7: 車体映像の例

で作業員が作業を行う際は，作業員の画像領域が画角全体に対して極めて小さくなった．また，解析対象とは異なる作業員や，作業対象となる車体の前後の車体が画角内に入ることがあった．

撮影された映像は 20 本であり，映像の長さは各々 1 分程度であった．本映像に含まれる組立作業は，図 7 に示すような「フェンダーを取り出す」，「歩行」，「車体にセットする」，「ボルトを締める」など，31 クラスの行動で構成される．作業を行わずに作業員が移動するのみとなる「歩行」に相当する行動は各作業の間で行われるが，本研究では歩行に相当する行動を統一せずに各作業段階における歩行を異なる行動，すなわち別クラスとする点に留意されたい．また，「スタート位置へ移動」など一部の映像にのみ含まれる行動が 2 クラスほど含まれる．さらに，20 本の映像のうち，映像の開始から終了まで自動車車体が停止している中で作業を行っている映像が 1 本，映像の終盤の 5 クラス分の行動を行わず，作業の途中で終了する映像が 6 本存在する．自動車車体が停止している映像の中盤は，作業員が何も作業を行っていない区間が存在するため，その部分はカットした．

4.2 実験設定

本研究では下記の 2 点について検証を行った．まず実験 1 として，単視点映像を I3D 特徴抽出器で特徴抽出を行い，抽出した特徴量を 6 種類の行動分類モデルへ入力し，フレーム毎に作業動作の識別を行った．このとき，図 6 で示した「視点 2」を用いることとした．次に，実験 2 として，視点の異なる 3 本の映像の併用を試みた．このとき，3.5 節で述べた映像連結および特徴量連結の 2 種類の方法を適用した．また，得られる特徴量に対して，実験 1 において最も正解率が高かった行動分類モデルを適用することとした．

どちらの実験においても，汎化性能を評価するために 20 本の映像に対して 5 分割交差検証を行い，各モデルの正解率（Acc），編集スコア（Edit），F1 スコアの平均を算出した．正解率 Acc は，時系列行動セグメンテーションの評価

表 1: 単視点映像における比較実験結果

Model	Acc	Edit	F1@10%	F1@25%	F1@50%
MS-TCN	87.39	94.86	94.64	92.53	85.27
MS-TCN++	86.46	94.52	93.57	91.62	84.24
ASRF+MS-TCN	86.19	90.73	94.07	92.12	86.28
SSTDA+MS-TCN	85.81	91.60	94.49	93.02	86.27
ASFormer	86.09	92.09	93.28	91.18	84.52
UVAST	86.96	96.29	93.57	92.70	86.51

に最も広く用いられる指標であり，フレーム単位で予測行動ラベルが正解と合致する割合として定義される．編集スコア Edit [14] 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．F1 スコア（$F1@\tau$）[11] 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 $\frac{\tau}{100}$ 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）と再現度（Recall）の調和平均を表す．

4.3 実験 1：単視点映像を用いた実験

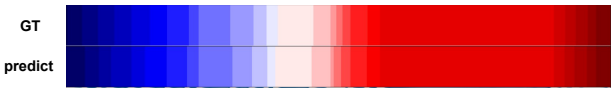
表 1 に各手法の Acc，Edit，F1 スコアの値を示す．結果から，全ての手法において 85%以上の正解率で行動解析を行えることが確認できた．これは，本実験において単一の車種かつ単一の工程を撮影したデータのみで学習を行っており，20 本の映像の間で各行動の分布の偏りが少なかったためと考える．

各手法のうち，最も正解率が高かった MS-TCN の結果に着目し，5 Fold の中で最も正解率が高かった Fold におけるテスト映像の識別結果を図 8(a) に示す．図の上部は正解，下部は予測結果を示しており，時系列に沿って左から右へと各行動の区間を色ごとに表している．各クラスは正解において青から赤に徐々に変化するように色分けされており，滑らかに変化していない箇所がエラーとなる．図 8(a) より，ほぼすべての行動セグメントが正解に類似しており，詳細な行動を高精度に認識できたことが確認できる．

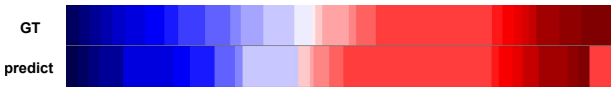
一方，図 8(b) に，MS-TCN の正解率が最も低かった Fold におけるテスト結果の識別結果を示す．結果から，中盤の行動において誤認識が生じており，また，正解と予測した行動セグメントとのズレが大きいことが確認できる．これは，図 8(b) に示すテストデータが，4.1 節で述べた，映像の開始から終了まで自動車車体が停止しており，作業員が何も作業を行っていない中盤の区間をカットした映像に相当しているため，中盤の行動において認識が困難となったことが原因であると考える．

4.4 実験 2：多視点映像を用いた実験

多視点映像を対象とした実験 2 では，映像連結と特徴量



（a） 最も Accuracy が高かった Fold におけるテスト結果



（b）最も Accuracy が低かった Fold におけるテスト結果

図 8: MS-TCN によるテスト結果の分析

表 2: 多視点映像における比較実験結果

Model	Input/feature	Acc	Edit	F1@10%	F1@25%	F1@50%
MS-TCN	単視点	87.39	94.86	94.64	92.53	85.27
MS-TCN	映像連結	86.91	94.43	94.16	92.57	84.58
MS-TCN	特徴量連結	87.64	94.20	94.35	92.77	87.44

表 3: 自動車が停止した映像におけるテスト結果の比較

Model	Input/feature	Acc	Edit	F1@10%	F1@25%	F1@50%
MS-TCN	単視点	48.82	75.86	52.83	41.51	30.19
MS-TCN	映像連結	40.83	65.52	44.00	44.00	28.00
MS-TCN	特徴量連結	41.47	58.62	44.00	40.00	36.00

連結の 2 つの方法について検証を行った．実験 1 において正解率が最も高いモデルであった MS-TCN を採用し，実験 1 と同様に 5 分割交差検証を実施した．表 2 に単視点，多視点（映像連結・特徴量連結）で学習を行った結果を示す．結果から特徴量を連結し学習することにより，単視点映像と比較して正解率と F1 スコアが向上することを確認した．これは，単視点映像を個別に解析して得られた特徴量を連結して学習することで，分類機へ入力される特徴量の次元が 3 倍に増加し，情報量が増加したためと考える．

一方，映像連結は単視点と比較して正解率などの低下がみられた．これは，表 3 に示すように，自動車車体が停止した映像におけるテスト結果が，単視点映像と比較して悪かったためである．これは，作業の対象となる自動車車体が作業の開始付近で停止しており，図 6 で示した「視点 1」，「視点 3」の画角に作業者が写ることがほぼなかったためである．

なお，特徴量連結においても同様に，自動車車体が停止した映像では性能の低下がみられたが，情報量が増加したことで他の映像で性能が改善したために，表 2 に示すように全体的に性能が向上した．

5. 結論

本研究では，自動車組立映像における行動解析を時系列行動セグメンテーション手法により試みた．実験により 87%の正解率，94%の編集スコア，85%の F1 スコアで行動解析が行えることを確認でき，自動車組立作業において，個々の作業に要した時間の解析や手順の正確さなど，定量的に評価を行える可能性が示唆された．また，多視点映像を用い，特徴量を連結することで性能が改善することを確

認した．

今後は，多視点を用いた学習における有意性の検証と異なる車種・異なる作業工程を対象とした実験を検討する．

参考文献

[1] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21096–21106, 2022.

[2] Takuya Kobayashi, Yoshimitsu Aoki, Shogo Shimizu, Katsuhiro Kusano, and Seiji Okumura. Fine-grained action recognition in assembly work scenes by drawing attention to the hands. In *2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, 2019.

[3] Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3575–3584, 2019.

[4] Shi-Jie Li, Yazan AbuFarha, 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*, pp. 1–1, 2020.

[5] Fangqiu Yi, Hongyu Wen, and Tingting Jiang. Asformer: Transformer for action segmentation. In *The British Machine Vision Conference (BMVC)*, 2021.

[6] Nadine Behrmann, S. Alireza Golestaneh, Zico Kolter, Juergen Gall, and Mehdi Noroozi. Unified fully and timestamp supervised temporal action segmentation via sequence to sequence translation. In *European Conference on Computer Vision (ECCV)*, pp. 52–68, 2022.

[7] Min-Hung Chen, Baopu Li, Yingze Bao, Ghassan Al-Regib, and Zsolt Kira. Action segmentation with joint self-supervised temporal domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[8] Min-Hung Chen, Baopu Li, Yingze Bao, and Ghassan AlRegib. Action segmentation with mixed temporal domain adaptation. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020.

[9] Yuchi Ishikawa, Seito Kasai, Yoshimitsu Aoki, and Hirokatsu Kataoka. Alleviating over-segmentation errors by detecting action boundaries. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2322–2331, 2021.

[10] Yifei Huang, Yusuke Sugano, and Yoichi Sato. Improving action segmentation via graph-based temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[11] Colin Lea, Michael D. Flynn, Rene Vidal, Austin Reiter, and Gregory D. Hager. Temporal convolutional networks for action segmentation and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[12] Li Ding and Chenliang Xu. Tricornet: A hybrid temporal convolutional and recurrent network for video action segmentation. *arXiv preprint arXiv:1705.07818*, 2017.

[13] Muheng Li, Lei Chen, Yueqi Duan, Zhilan Hu, Jianjiang Feng, Jie Zhou, and Jiwen Lu. Bridge-prompt: Towards ordinal action understanding in instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19880–19889, 2022.

[14] Colin Lea, Austin Reiter, René Vidal, and Gregory D Hager. Segmental spatiotemporal cnns for fine-grained action segmentation. In *European conference on computer vision*, pp. 36–52. Springer, 2016.

[15] 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 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 780–787, 2014.

[16] Alireza Fathi, Xiaofeng Ren, and James M. Rehg. Learning to recognize objects in egocentric activities. In *CVPR 2011*, pp. 3281–3288, 2011.

[17] Yin Li, Zhefan Ye, and James M. Rehg. Delving into egocentric actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[18] 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*, pp. 729–738, 2013.

[19] Dima Damen, Hazel Doughty, Giovanni Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pp. 1–23, 2022.

[20] Francesco Ragusa, Antonino Furnari, Salvatore Livatino, and Giovanni Maria Farinella. The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1569–1578, 2021.

[21] Yizhak Ben-Shabat, Xin Yu, Fatemeh Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 847–859, 2021.

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

[23] 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.

[24] Guodong Ding, Fadime Sener, and Angela Yao. Temporal action segmentation: An analysis of modern technique. *arXiv preprint arXiv:2210.10352*, 2022.

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

[26] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, Vol. 27, , 2014.