動作クラスのクエリに対応した SST による動作区間検出

ER16060 畑中和輝　　指導教授：藤吉弘亘

1. はじめに

スマートフォンや動画サイトの普及によって，動画コンテンツがより身近になっている．それに伴い，動画の内容を短時間で把握できる要約動画の需要が増加している．動画要約を行うためには特定の動作区間を検出する必要がある．そこで本研究では，動作クラスをクエリとして入力し，対応するクラスに特化した動作区間検出を目的とする．そのために，動作区間検出が可能なネットワークである Single-Stream Temporal Action Proposals(SST)[1] を特定クラスの動作区間検出に拡張する．

2. Single-Stream Temporal Action Proposals

SST は，動作区間を検出可能な手法である．SST は図 1 のように Convolutional 3D Network(C3D)[2] により動画像の特徴を抽出する Visual Encoder と，Gated Recurrent Unite(GRU) により時間的特徴を抽出する Sequence Encoder の 2 つのモジュールから構成される．SST は，数フレームの動画像を入力し，その時刻及びそれまでの時系列が動作区間であるかどうかの確率 $\boldsymbol{c_t}$ を出力し，上位 N 個を検出区間とする．

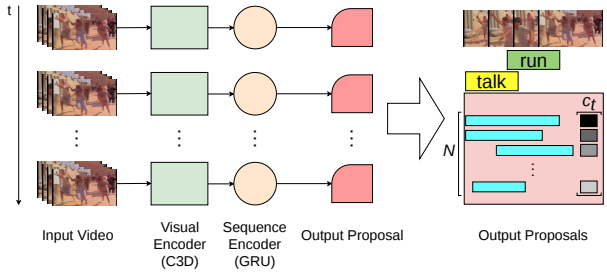

図 1: SST のネットワーク構造

3. 提案手法

本研究では，検出したい特定クラスの情報をクエリとして，ネットワークの中間層に入力することで，対応するクラスの動作区間検出する手法を提案する．図 2 に提案手法の構成を示す．提案手法は，C3D の出力にクラス ID に対応した one-hot vector をクエリとして結合する．その後，全結合層に入力することで，SST と同様に動作区間である確率 $\boldsymbol{c_t}$ を出力する．クエリを結合したことで特定クラスの特徴が強調され，特定クラスを含む候補区間の確率が高くなる．出力された区間のうち，最も確率の高い区間を検出区間とする．

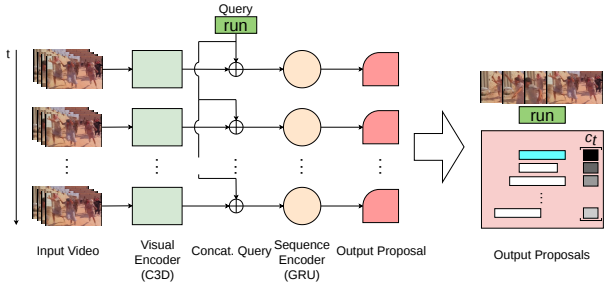

図 2: 提案手法

4. 評価実験

提案手法の有効性を評価実験によって検証する．

4.1. 実験概要

本実験では，AVA データセット [3] を用いる．AVA データセットは，stand，carry などの一般動作を含んだ動画のデータセットである．各動画は 22.4 秒であり，学習データに 8,192 本，評価用データに 2,450 本用いる．学習時のバッチサイズを 32，更新回数を 50epoch とし，最適化関数には Adam を用いる．GRU の隠れ層のユニット数は 128 である．定量的評価として，最も確率が高い検出区間を tIoU を閾値とした precision で評価する．クラス毎での精度では，タスクが異なり，従来手法と比較できない．そのため，各クラスの区間を集約し，SST と同様に複数の動作を含んだ区間として精度を評価する．集約後，1 つの動画に対して検出区間が複数現れた場合，検出区間の中で最も確率が高い区間を採用する．

4.2. 実験結果

定量的評価の結果を図 3 に示す．提案手法の presicion は，SST よりわずかに低下しているが，ほぼ同等である．

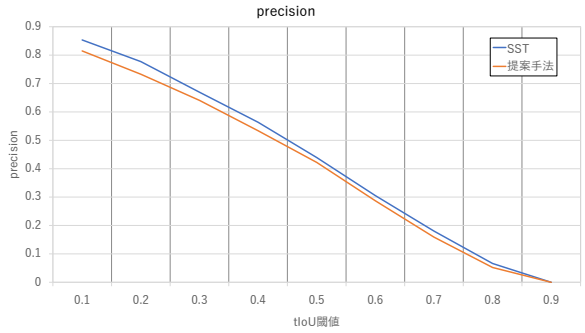

図 3: precision による精度比較

4.3. 検出区間の可視化

図 4 に検出区間の可視化結果を示す．SST では複数の候補区間を 1 つの区間として検出している．一方で，listen to のクエリとして与えた提案手法は，対応する区間が検出されている．しかし，talk to をクエリとして与えた結果のように，ほかのクラスの動作区間まで検出する結果も存在する．データセットには，同じ区間に複数のクラスの動作が含まれている場合があるが，単一のクラスとして学習させている．そのため，重複区間による影響によって他の動作クラスの区間も検出していると考えられる．また，実際の動画では，GT に含まれていないが話している場面が検出されており，GT には存在しないが実際は動作を行っている区間の検出もしている．

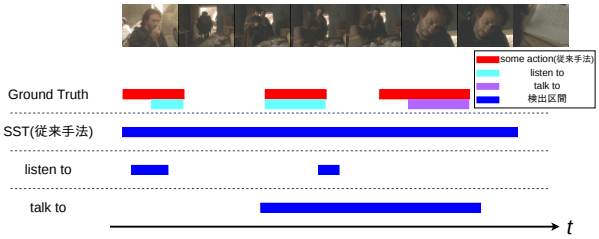

図 4: 従来手法及びクエリによる検出区間の例

5. おわりに

本研究では，クエリを追加することで，動作クラスに対応した動作区間検出を行う手法を提案した．提案手法を用いることにより，クエリによる検出区間の変化を実現した．今後は，複数クラスのクエリを使用することで，複数クラスの動作区間検出の対応を目指す．

参考文献

[1] S. Buch, *et al.*, “SST: Single-Stream Temporal Action Proposals”, CVPR, 2017.

[2] D. Tran, *et al.*, “Learning Spatiotemporal Features with 3D Convolutional Networks”, ICCV, 2015.

[3] Chunhui Gu, *et al.*, “AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions”, CoRR, 2017.