DEIM2022 D23-1(day2 p18)

動画境界検出タスクにおける評価方法の検証

行武　俊秀†,††　　水船　公輔††　　松林　達史††

† 明治大学大学院　〒 164-8525 東京都中野区中野４丁目２１－１
†† 株式会社 ALBERT　〒 169-0074　東京都新宿区北新宿 2-21-1　新宿フロントタワー 15F
E-mail: †cs213021@meiji.ac.jp, ††{kosuke_*mizufune*, *tatsushi_matsubayashi*}@*albert*2005.*co.jp*

あらまし　動画内の境界検出は，映画やスポーツなどを対象としてシーンやアクションの変化点を検出する技術であり，動画検索や要約のための詳細かつ多様な境界を検出することが求められている．このような境界検出タスクの評価方法には F 値を用いることが多いが，境界検出タスクにおける F 値の妥当性は明らかになっていない．本稿では Kinetics-GEBD データセットを用いて，F 値を用いた評価方法に対する妥当性を検証した．実験の結果，均一な間隔で境界を予測するだけであっても，教師あり学習手法の F 値に相当することが観察された．さらに，境界予測の誤差に一定の許容範囲が与えられる検出タスクでは，重複を避けて密に一定の間隔を持つ予測境界を与えた方が偶発的に正しい予測を取得しやすいが，F 値による評価では正解の偶然性を考慮できないため，予測モデルを正しく評価できないことを示した．これらの結果に基づき，境界検出モデルを評価するための適切なアプローチを提案し，その妥当性を確かめた．

キーワード　評価方法，評価指標，動画認識，動画境界検出，Generic Event Boundary Detection

1 はじめに

近年の動画コンテンツの増加を背景に，動画を時間方向にセマンティックな単位で分割することは学術的にも産業的にも広く注目されており，動画像の要約や検索といった幅広い応用が期待される．古典的な動画内の境界検出タスクとして，Action Localization [1–6] や Action Segmentation [7–9] の研究が進んでいる．しかしながら，これらのタスクは事前にアクションラベルを定義する必要があり，定義されたアクションしか境界を検出することができない．一方で，定義されたアクションのシーンを抽出するのではなく，映像作品のシーン境界 [10] や，sub-action の境界 [11]，より汎用的で一般的なイベント境界 [12] を検出するという，境界検出タスクが近年注目されている．Action Localization ではアクションの始点と終点，及びそのラベルを予測するのに対し，境界検出タスクでは特定の変化がいつ発生するのかのみを予測し，そのラベルは予測しない．

境界検出タスクは，あるフレームが境界であるかどうかの二値分類問題として定義できるため，二値分類問題の評価指標として一般に使用される F 値で評価することが多い [11, 12]．しかしながら，F 値による評価は不適切な評価指標となる場合があることが動画要約研究で指摘されており [13]，境界検出タスクにおいても不適切な可能性がある．さらに Powers [14] は，F 値による評価では予測が偶然正解となる確率が考慮されないと指摘している．したがって，境界検出タスクにおいて F 値が妥当な評価指標であるかの検証を行うことは非常に重要である．

本稿では，Kinetics-GEBD データセット [12] を用い，境界検出タスクにおいて F 値を評価指標に用いることが妥当かどうかを検証する．まず，動画の内容に基づかないランダムな境界検出と，動画の内容に基づく教師あり，教師なしのベースライン手法を F 値で評価した．実験の結果，単純な F 値を用いた評価では，均一な間隔で多数の境界を作成するだけでも評価値が高くなることが判明した．さらに詳細な分析を行なった結果，予測境界間隔と境界数に関して適当に調整を行うと正しい予測を取得しやすいが，F 値による評価では予測境界が偶然正解となる確率が考慮できないことが示された．したがって，境界検出タスクにおいて F 値による評価は適切ではない場合があるため，より適切な評価指標で評価を行う必要がある．

本稿における貢献は以下のようになる：

（1） 境界検出タスクにおける評価指標の妥当性を検証し，均一な間隔で境界を予測するだけであっても，教師あり学習手法の F 値に相当することを示した．

（2） F 値による評価では正解の偶然性を考慮できず，重複を避けて密に一定の間隔を持つ予測境界を与えた方が良い F 値になることを示した．

（3） 境界検出タスクにおいて適切な評価指標を提案し，その妥当性を検証した．

2 関連研究

動画認識の研究では，様々な行動認識モデル [15–21] が提案されている．しかしながら，これらの手法は動画 1 つに対して単一のイベント・アクションを想定しているため，複数のイベントを含む長時間の動画を扱うことはできない．一方，長時間の動画認識はさまざまな枠組みで研究が進められている．Action Localization [1–6] は長時間の動画に対して，各アクションラベルの始点と終点を予測する課題となっている．Action Segmentation [7–9] では各フレームごとにアクションラベルを予測する．これらの手法は長時間の動画を扱うものであるが，事前に定義されたアクション・イベントしか検出できない．

図 1　Kinetics-GEBD データセットの例．点線がアノテーションされた境界であり，動画によって多様な種類の境界が存在する．上：オブジェクトの変化．中：アクションの変化．下：環境 (明るさ) の変化．

事前の定義を必要としない認識モデルを構築するため，動画中の変化点のみを予測する境界検出タスクが提案されている．Temporal Action Parsing [11] は 1 アクションを構成する複数の sub-action の変化を認識するタスクである．人間にとって sub-action の全てを定義することは難しく，クラス分類問題としてタスクを定義することが困難である．しかしながら，sub-action の境界のみを認識するタスクを定義することで，アノテーションの一貫性という点で質の高いデータセットを構築した．

その他の境界検出タスクとしては，ショット境界検出 [22–25] やシーン境界検出タスク [10,26,27] が提案されている．ショットは同一のカメラから撮影された一連のフレームのことで，低レベルの視覚的特徴を用いることで容易に境界検出できることが知られている [28]．一方シーンは意味的に一続きのショットと定義されており，複雑な時間構造を持っているため，境界検出が難しい課題となっている [10,27]．

このような sub-action，ショットの変化や，明暗等の環境変化なども含めた一般的な動画境界を検出するタスクとして，Generic Event Boundary Detection (GEBD) [12] が近年提案された．この研究では行動認識研究で多用される Kinetics データセット [29] の一部が使用され，人手による境界のアノテーションが行われた．GEBD では動画のドメインがスポーツ [11,30] や映画等 [10] に限定されることがなく，検出する境界もアクションの変化だけではなく多様であるため，汎用性の高いタスクとなっている．

3　境界検出タスクと評価手法

境界検出タスクでよく使われる評価手法は F 値である [11,12]．F 値は Precision と Recall の調和平均で求まる：

$$\mathrm{F1} = \frac{2\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}, \tag{1}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \tag{2}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}. \tag{3}$$

ここで，TP，FP，FN はそれぞれ True Positive，False Positive，False Negative である．

境界検出では予測に時間方向の誤差が生じるため，各予測境界と真の境界との誤差が，許容誤差 d より小さければその予測を正解とする．なお，許容誤差には絶対誤差または動画全体の長さに比例した相対誤差どちらを設定しても良い．また，図 2 に示す通り，許容誤差の範囲が重複した場合，同じ境界に対し重複して予測を行うことはできず，2 つ目以降の予測境界は FP として扱う．

人手による境界アノテーションを用いたタスク評価では，曖昧性を排除するために複数人のアノテーションデータを利用することが多い．しかしながら，F 値による評価では最終的に参照するアノテーションを 1 つに決める必要がある．GEBD [12] では，複数人のアノテータ（アノテーションを行った評価者）によってアノテーションされた境界それぞれに対し，境界予測モデルによる境界との F 値を算出する．その F 値が最大となるアノテータのアノテーションを参照アノテーションとしたものを Max とする．一方，各アノテータの境界に対して，残りのアノテータによる境界との F 値を算出し，その F 値が最大となるアノテータのアノテーションを参照アノテーションとして使用したものを Confident として，本稿でも利用する．

4　実　　験

本稿では評価手法の妥当性を検証するため，Otani et al. [13] の研究を基に，[12] で提示された動画の内容に基づく教師あり・教師なしのベースライン手法と，動画の内容に基づかない境界生成手法を比較する．動画の内容に基づく境界検出手法として，[12] では教師ありの手法として Pairwise boundary Classifier

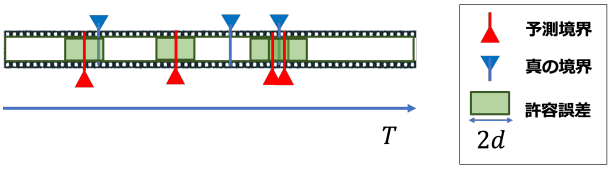

図 2　予測境界とその許容誤差．d は絶対誤差とする．１つ目の予測境界は許容誤差の範囲内に真の境界があるため，TP となる．一方，２つ目の予測境界は FP，２つ目の真の境界は FN となる．３つ目と４つ目の予測境界は重複した予測となるため，片方は TP，もう片方は FP となる．この時，TP は 2，FP は 2，FN は 1 となるため，Recall が 2/3，Precision が 1/2，F 値は 4/7 となる．

(PC)，教師なしの手法として PredictAbility (PA) という手法を提案している．それぞれ以下のアルゴリズムで境界を検出する．

- PC: ある時刻 t の前後のフレームをそれぞれ特徴ベクトルに変換し，連結された前後特徴を 2 値分類モデルの入力として，時刻 t が境界であるかを学習する．特徴変換にはバックボーンモデルは ResNet50 [31] の ImageNet 事前学習済モデルを使用し，ファインチューニングを行う．推論時は各フレーム t ごとに境界である確率を計算し，閾値を超えるような連続したフレームの中心を予測境界とする
- PA: ある時刻 t の前後のフレームをそれぞれ特徴ベクトルに変換し，その距離の局所最大を境界として定義する．特徴抽出は PC と同様，ResNet50 [31] の ImageNet 事前学習済モデルを使用する．局所最大の検出には，画像上の塊検知（blob detection）で使われている古典的な Laplacian of Gaussian (LoG) [32] を利用する．具体的には，特徴ベクトルの距離の符号を負にした後に LoG フィルタを施し，その勾配が負の値から正の値に変化する点を境界とする．

一方，動画の内容に基づかない境界生成手法として，以下の 2 手法を定義する．

- Random: 動画フレーム $\{1,\ldots,T\}$ の中から，ランダムで M 個の境界を生成する．
- Uniform: 動画を $M+1$ 個に均等に分割するような M 個の境界を生成する．

図 4 にそれぞれ境界を 3 個生成した時の Random，Uniform の予測誤差を示す．Random による境界はそれぞれ独立に予測されるため，予測境界からの許容誤差の範囲は重なる可能性がある．一方 Uniform による予測境界は，境界数 M が $2d(M+1) \leqq T$ を満たす場合に許容誤差が重なることがない．また，Random による境界は動画ごとに異なるが，Uniform による境界の相対位置は全ての動画で同じとなる．

Random や Uniform は動画の内容に基づかない境界の予測方法であるため，偶発的に得られる評価値としての良い指標となることが期待される．妥当な評価指標では，PC や PA のような動画に基づいた手法による境界検出の評価値が，Random や Uniform と比較して高くなることが期待される．さらに人間によって付与された境界が最も高い評価値となると考えられる．

4.1　データセット

本稿では，評価データとして Kinetics-GEBD データセット [12] を使用する．このデータセットは Kinetics-400 [33] の内 55,351 個の動画から構成され，約 149 万個の境界アノテーションが 3~5 人のアノテーターによって施されている．トレーニング，バリデーション，テストデータセットの比率は約 1:1:1 となっているが，テストデータセットのアノテーションは公開されていないため，本稿ではバリデーションデータセットを評価に使用する．トレーニングデータは PC の学習のみに使用している．なお，動画は YouTube から取得する必要があり，一部削除や非公開となった動画も含むため，全ての動画は使用できない．本稿では，トレーニングデータとして 17,025 個，バリデーションデータとして 16,884 個の動画を使用する．

4.2　実験設定

PC と PA は画像サイズを 224×224 とした RGB 画像を使用し，あるフレーム t に関して前後 5 フレームを入力としている．なお，3 フレームごとに 1 枚ずつサンプリングを行った．PC では，あるフレーム t の前後 0.15 秒以内に真の境界が存在すれば正例，そうでなければ負例とし，Binary Cross Entropy Loss を用いて学習する．PA で使用する LoG フィルタの標準偏差の値は 15 としている．PC と PA は再現実験を行い，[12] で報告されている値とほぼ同じ F 値を確認した．Random は動画ごとに異なる境界を予測するため，ランダム試験を 100 回繰り返し，その平均を表示している．また，Random と Uniform で予測する境界の数 M は 1 から 15 個までを検証した．人手によって付与された境界の評価は，アノテーターごとに評価を行い，その平均値を使用している．評価に使用する許容誤差には，閾値を 5%とした相対誤差を使用する．

4.3　Discussion

図 3 に PC，PA，Random，Uniform による境界と，人手によって付与された境界の Recall，Precision，F 値を示している．なお，人手によって付与された境界は Human として表記している．まず，Uniform の F 値は教師あり学習の手法である PC に相当し，特に境界数が 9 個の時に最も F 値が高くなった．さらに，Uniform では境界の数を増やしたときの Precision の低下に対して，Recall の上昇が非常に大きいことが観察された．一方，Random も教師あり学習である PC に近い F 値を出しており，教師なし手法の PA を上回っている．

しかしながら，動画を均一に区切るような境界は動画の内容に基づいておらず，F 値が高くなることは直感に反する結果となっている．さらに，Random，Uniform は両者ともに動画内容によらずに境界を予測する手法であるにも関わらず，大きな F 値の差が確認される．これらの結果から，検出境界が特定の条件を満たすときに F 値による評価が高くなると考えられる．F 値による評価は，不適切な評価指標となる場合があることが指摘されており [13]，境界検出タスクにおいても不適切な可能性がある．さらに Powers [14] は，F 値による評価では予測が偶然正解となる確率が考慮されないと指摘している．そこで，予測された境界が偶然正解となる確率を考察することにより，どのような場合に F 値が高くなるのかを検証する．

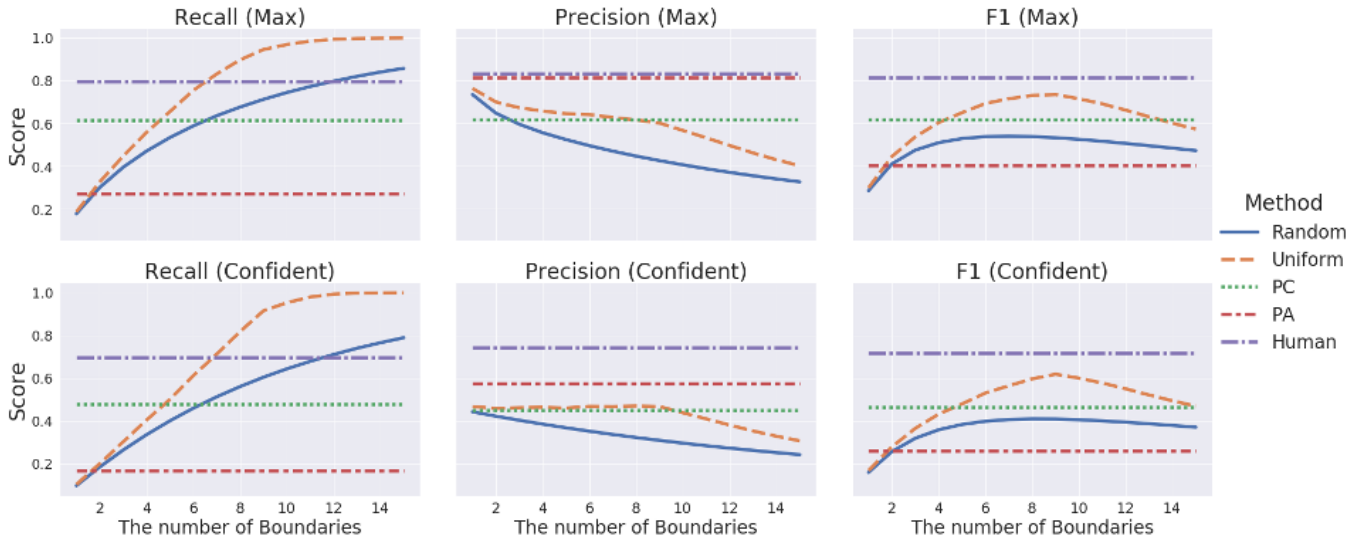

図 3　Kinetics-GEBD データセットにおける，異なる境界検出手法の Recall，Precision，及び F 値．１行目は Max による評価設定で，２行目が Confident による評価設定となっている．

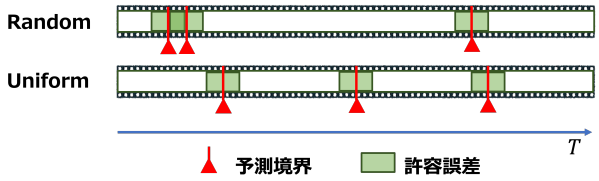

図 4　Random と Uniform による予測境界．それぞれ境界の数は 3 個としている．また，許容誤差が被っている領域は深緑で示している．

まず，予測境界と真の境界に関して，それぞれ正例の割合 Prevalence と Bias を下記に定義する：

$$\text{Prevalence} = \frac{\text{RP}}{T}, \tag{4}$$

$$\text{Bias} = \frac{\text{PP}}{T}, \tag{5}$$

ここで，RP は真の境界とその許容誤差を含めた領域，PP は予測境界とその許容誤差を含めた領域である．Prevalence は真の境界の数に依存するため，予測の精度や予測境界数に関係なく一定値を取る．また Prevalence が高い時，ランダムに境界を予測した時にその境界が偶然正解となる確率が高くなる．一方，Bias は予測された境界の数に依存するため，予測モデルのアルゴリズムや閾値の調整によって値が変わる．Bias が高い時，真の境界が偶然予測される確率が高くなる．

次に，正例だけでなく負例についても着目した Informedness と Markedness を下記に定義する：

$$\text{Informedness} = \text{Recall} - \frac{\text{FP}}{\text{FP} + \text{TN}}, \tag{6}$$

$$\text{Markedness} = \text{Precision} - \frac{\text{FN}}{\text{FN} + \text{TN}}, \tag{7}$$

ここで TN は True Negative である．Informedness と Markedness はそれぞれ Recall，Precision に対し，負例の未検出，過検出の割合をペナルティとして与えたものと見なすことができる．これらの定義より，Precision，Recall は以下のように書き直すことができる [14]：

$$\text{Precision} = \text{Markedness}(1 - \text{Bias}) + \text{Prevalence}, \tag{8}$$

$$\text{Recall} = \text{Informedness}(1 - \text{Prevalence}) + \text{Bias}. \tag{9}$$

式 (8)，(9) より，Precision，Recall 及びその派生系である F 値は Bias 及び Prevalence の値に大きく影響されることがわかる．具体的には，Prevalence が高くなるにつれ Precision は線形に高くなり，Informedness の値に応じて Recall が減少する．一方，Bias が高くなるにつれ Recall が線形に高くなり，Markedness の値に応じて Precision が減少する．なお，ランダムな予測を行なった時，Informedness と Markedness は共に 0 となる．したがって Bias の値を大きくする境界検出モデルを作成すれば，Recall が上昇する一方，Precision は Prevalence の値から変わらない事になる．

境界検出タスクでは，図 2 に示すように，予測境界からの許容誤差の範囲によって Bias が定まる．Uniform では予測境界の数 M が $2d(M+1) \leqq T$ を満たす時に許容誤差の重なりが生じないため，Bias の値は $2dM/T$ となる．なお，本稿では許容誤差の値として，動画の長さに比例する相対誤差 $d/T = 0.05$ を設定しているため，Bias の値は $0.1M$ となる．さらに，図 5 右に PC と PA の Bias を示す．PC の Bias は 0.4 ~ 0.5 を中心に分布する一方，PA の Bias は 0.1 ~ 0.2 を中心に分布している．Uniform，PC，PA の Bias の値は図 3 下の Recall とほぼ同じ値になっており，境界検出モデルの Bias が増加するにつれ Recall が上昇することが確認される．

一方で，Prevalence は真の境界からの許容誤差の範囲によって定まる．図 5 左に真の境界の Prevalence の分布を示しているが，Prevalence は 0.4 ~ 0.5 を中心に分布していることが観察される．また，図 3 下の Precision を確認すると，Uniform の境界数が 9 個まで，すなわち境界数 M が $2d(M+1) \leqq T$ を満たす時，Precision の値が Prevalence の平均値とほぼ同じ値になっている．したがって，許容誤差の重複がないように予測すれば，

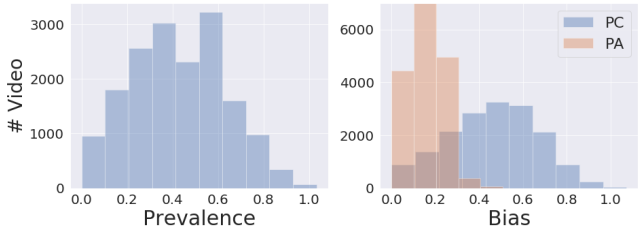

図 5　左図：真の境界における Prevalence のヒストグラム．右図：PC と PA における Bias のヒストグラム．

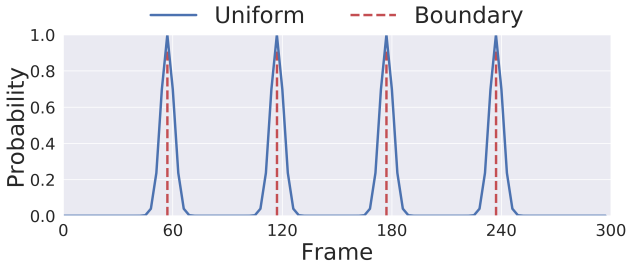

図 6　Uniform による境界に対して生成した予測確率．境界数は 4 としている．横軸はフレームのインデックスを，縦軸はそれぞれのモデルの予測確率（Probability）を示している．

予測境界の数に関わらず Precision の値は変化しないことがわかる．また，Random の Precision が下がり続けるのは，予測境界の許容誤差の重複によって False Positive が増大したためであると考えられる．

5　Bias の影響を受けない評価指標

境界検出問題では，分類問題と異なり，負例の識別が重要ではないため，True Negative を用いない F 値による評価が用いられてきた．しかしながら，F 値は正例の割合に依存するため，モデルの比較を行うには Bias の影響を排除するような評価が必要である．この問題を解決するため，F 値に代替する評価指標の提案を行う．

情報検索の分野において使用される指標のひとつに Average Precision(AP) が挙げられる．AP では信頼度スコアによって予測をソートし，Precision-Recall Curve の下部分の面積をスコアとする：

$$\mathrm{AP} = \int_0^1 \mathrm{Precision}(\mathrm{Recall})\, d\mathrm{Recall}\,. \tag{10}$$

AP によって負例を考慮せずに評価を行うことが可能であり，さらに Precision を Recall の関数とみなして積分することで Bias の影響を排除している．

本稿では，境界の予測確率を信頼度スコアとして AP による精度評価を試みる．なお，PA はある時刻 t における前後のフレームの特徴ベクトルに対する距離を使用するため，信頼度スコアが正規化されていない．そこで本稿では，特徴ベクトルのコサイン類似度を使用する．また，Uniform，Random，及び Human に対しては，予測確率を疑似的に与える必要があるため，t 番目のフレームが境界である予測確率 $f(t)$ を，以下のように定義する：

表 1　Average Precision による各手法の評価結果

手法	AP	F 値 (Max)	F 値 (Confidence)
Human	**0.708**	**0.814**	**0.717**
PC	0.495	0.616	0.462
PA	**0.564**	0.405	0.259
Uniform(M=9)	0.474	**0.734**	**0.617**
Random(M=9)	0.452	0.530	0.408

$$f(t) = \sum_{\tau \in \mathcal{T}} K(\tau, t)\,, \tag{11}$$

$$K(\tau, t) = \exp\left(-\frac{|\tau - t|^2}{\sigma^2}\right), \tag{12}$$

ここで，動画内に含まれる予測境界のフレームの集合を $\mathcal{T}$ とし，その要素を $\tau \in \mathcal{T}$ と表記する．パラメータに関しては $\sigma = 5$ で固定とした．生成された予測確率の例として，M=4 における Uniform の予測確率を図 6 に示す．この疑似的な予測確率は，境界を中心として高くなることが確認される．

評価時は真の境界の周辺も正例とすることで，許容誤差を考慮した評価が可能である．以下の評価実験では，上記の F 値による実験と同様に，許容誤差には閾値を 5%とした相対誤差を使用した．

表 1 に Kinetics-GEBD における，Average Precision による評価実験の結果をまとめた．人間による境界予測 (Human) や予測モデルによる境界予測 (PC，PA) と比較して，Uniform および Random による予測は低い AP を示していることがわかる．また Uniform や Random は，その予測境界の数に依らずほとんど一定の Precision-Recall Curve(図 7) となっており，Recall に依らずほぼ一定の Precision となっている．これらの結果から，AP による評価によって，F 値の課題であった Bias に対するロバスト性が獲得できたと考えられる．

また，教師あり学習 (PC) と教師なし学習 (PA) の比較を行うと，PA のほうが高い AP となっている．ここで図 8 に，PC と PA による予測境界確率の例を示す．PC による境界は FP が多いのに対し，PA は FP の数が比較的少ないことが観察される．図 8 上を見ると，PA では大きな変化が起こった一点で大きな極大値をとっている．ここから，PA が明らかな境界のみを予測しており，FN が大きいモデルであることがいえる．対して，PC は，PA では予測できなかった境界も一部は予測しているが，FP も同時に大きい．図 8 下を見ると，PA は画像の変化に対して的確に極大値を取るように動いているのに対し，PC は境界にあたらない部分でも極大値を取っており，PC が多く誤検出を行っていることが読み取れる．

F 値による評価では，PC のように予測境界数が多い，すなわち Bias の高い予測モデルを高く評価してしまう．一方で，AP による評価は Bias に影響されないため，PC の過検出をより適切に考慮した上での評価ができていると考えられる．

6　結　　論

本稿では，動画境界検出に広く用いられている F 値による評

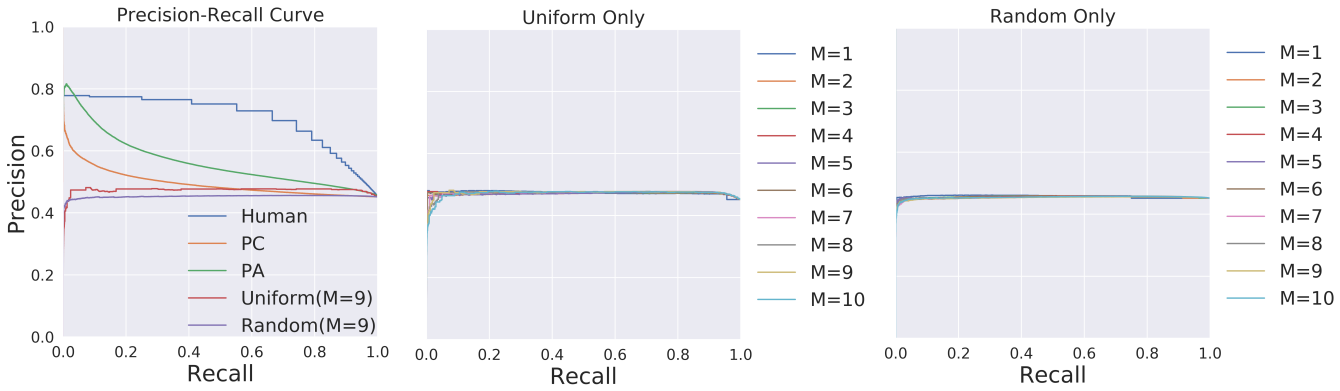

図 7　左：今回使用したすべての手法の PRCurve での比較．中央：予測数 M を変化させた場合の Uniform の PRCurve．右：予測数 M を変化させた場合の Random の PRCurve．Random と Uniform での予測の個数を M としている．

価の妥当性を検証した．その結果，F 値による評価には問題があることが明らかになった．F 値による評価は予測境界が偶然正解となる確率を考慮できないため，許容誤差の重複を避けるように一定間隔の境界を密に予測すれば，F 値が高くなることが観察された．さらに，この原因が Bias の増加によるものであることが示された．そのため，F 値による評価方法ではモデルの比較が困難であると判明した．

したがって，本稿では Average Precision を用いて Bias の変化にロバストな評価を行うことを提案した．Average Precision による評価では，動画の内容と無関係な予測はその予測の数に依らず低い値となり，Bias の影響を排除できた．また，AP による PA と PC の比較によって，F 値による評価では考慮できなかった過検出を，AP によって適切に考慮できることを示した．

本稿ではフレーム単位の境界予測確率に対して評価を行なった．しかしながら，予測確率のみを評価対象とした時，重複した予測に対し，F 値のように罰則を与えることができない．したがって，フレーム単位の予測確率のみではなく，検出された境界単位での新しい評価も検討し，本稿で得られた結果と同様の傾向が見られるかどうかを検証することが今後の課題である．また，Kinetics-GEBD 以外のデータセットに対する評価や，複数人のアノテーターを考慮した評価指標についても検討していきたい．

文　献

[1] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1130–1139, 2018.

[2] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19, 2018.

[3] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3889–3898, 2019.

[4] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang. Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5734–5743, 2017.

[5] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In Proceedings of the IEEE International Conference on Computer Vision, pp. 2914–2923, 2017.

[6] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10156–10165, 2020.

[7] 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, pp. 3575–3584, 2019.

[8] Peng Lei and Sinisa Todorovic. Temporal deformable residual networks for action segmentation in videos. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6742–6751, 2018.

[9] 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, pp. 14024–14034, 2020.

[10] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16, pp. 709–727. Springer, 2020.

[11] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Intra-and inter-action understanding via temporal action parsing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 730–739, 2020.

[12] Mike Zheng Shou, Stan Weixian Lei, Weiyao Wang, Deepti Ghadiyaram, and Matt Feiszli. Generic event boundary detection: A benchmark for event segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8075–8084, 2021.

[13] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkila. Rethinking the evaluation of video summaries. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7596–7604, 2019.

[14] David Powers. Evaluation: From precision, recall and f-measure to roc, informedness, markedness & correlation. Journal of Machine Learning Technologies, Vol. 2, No. 1, pp. 37–63, 2011.

[15] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference

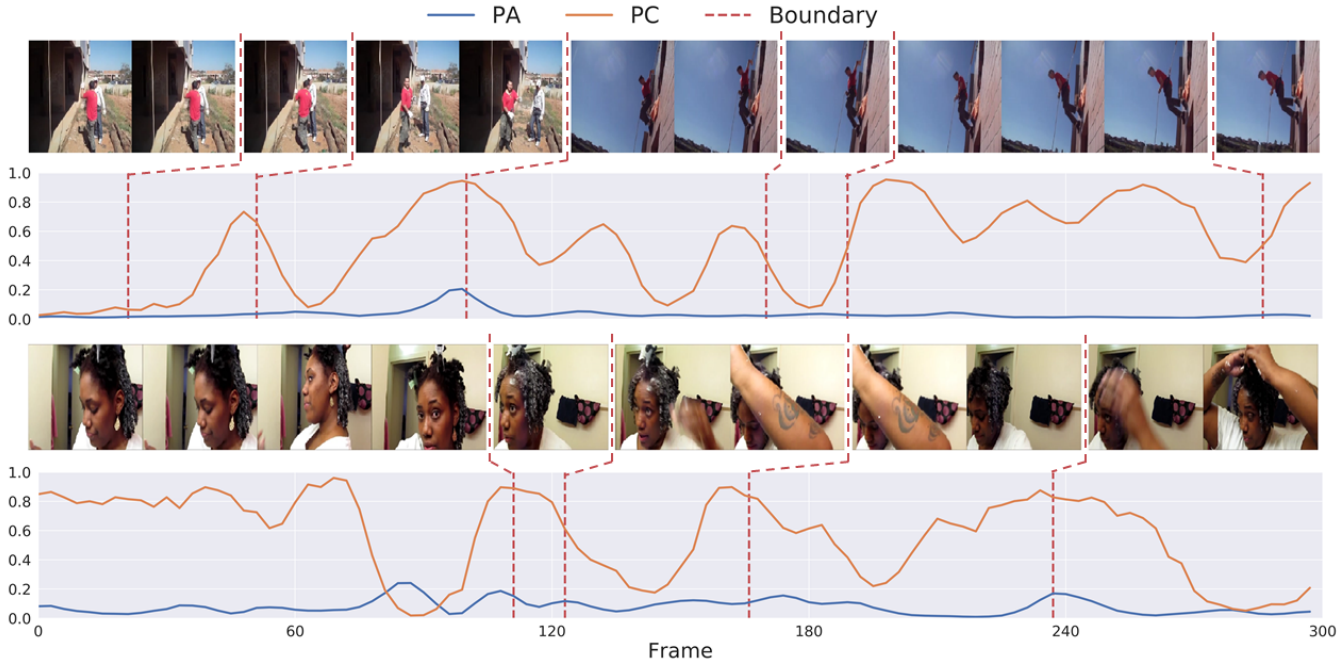

図 8　2 つの異なる動画における，PA（青）と PC（橙）の予測確率の例．真の境界は赤の波線（Boundary）で表示している．横軸はフレームのインデックスを，縦軸はそれぞれのモデルの予測確率（Probability）を示している．

on computer vision, pp. 4489–4497, 2015.

[16] 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.

[17] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 6450–6459, 2018.

[18] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 6202–6211, 2019.

[19] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7083–7093, 2019.

[20] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 203–213, 2020.

[21] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 6546–6555, 2018.

[22] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. Shot and scene detection via hierarchical clustering for re-using broadcast video. In International Conference on Computer Analysis of Images and Patterns, pp. 801–811. Springer, 2015.

[23] Hong Shao, Yang Qu, and Wencheng Cui. Shot boundary detection algorithm based on hsv histogram and hog feature. In 5th International Conference on Advanced Engineering Materials and Technology, pp. 951–957, 2015.

[24] Shitao Tang, Litong Feng, Zhanghui Kuang, Yimin Chen, and Wei Zhang. Fast video shot transition localization with deep structured models. In Asian Conference on Computer Vision, pp. 577–592. Springer, 2018.

[25] Tomáš Souček, Jaroslav Moravec, and Jakub Lokoč. Transnet: A deep network for fast detection of common shot transitions. arXiv preprint arXiv:1906.03363, 2019.

[26] Anyi Rao, Linning Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, and Dahua Lin. A local-to-global approach to multi-modal movie scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10146–10155, 2020.

[27] Shixing Chen, Xiaohan Nie, David Fan, Dongqing Zhang, Vimal Bhat, and Raffay Hamid. Shot contrastive self-supervised learning for scene boundary detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9796–9805, 2021.

[28] Costas Cotsaces, Nikos Nikolaidis, and Ioannis Pitas. Video shot detection and condensed representation. a review. IEEE signal processing magazine, Vol. 23, No. 2, pp. 28–37, 2006.

[29] 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.

[30] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. Soccernet: A scalable dataset for action spotting in soccer videos. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 1711–1721, 2018.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016.

[32] Tony Lindeberg. Feature detection with automatic scale selection. International journal of computer vision, Vol. 30, No. 2, pp. 79–116, 1998.

[33] 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.