LSTM Blockを用いたサッカー動画における動作分類

篠田　拓樹†　　青野　雅樹††

† 豊橋技術科学大学 情報・知能工学課程　〒 441–8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1
†† 豊橋技術科学大学 情報・知能工学系　〒 441–8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1
E-mail: †shinoda.hiroki.vo@tut.jp, ††masaki.aono.ss@tut.jp

あらまし　近年，スポーツ分野では，審判の補助や選手の分析等に映像を用いた技術が使われている．しかし，サッカーにおける映像の編集作業は手作業で行われている．例として，ハイライトの作成においては，編集者が 90 分の試合からゴールシーン等の注目するべき箇所を抜き出し，それらをつなぎ合わせることによって作成される．この作業において，注目するべき箇所を自動で検出することが出来れば編集にかかる時間が大きく減少する．本研究では，少ない特徴量で局所的な時間関係を取り入れるために LSTM Block(前向き LSTM，後ろ向き LSTM，双方向 LSTM) を用い，マルチクラスの動作分類問題に取り組む．ベンチマークとの比較実験では，全体の評価値は既存手法より低い結果となった．しかし，特定のクラスにおいては既存手法に近い結果を示した．ベンチマークの再現実装モデルと提案手法に同じ特徴量を入力として比較を行った実験では，“Shot on target”，“Substitution” において再現実装モデルよりも提案手法が高い結果となった．

キーワード　動作分類，マルチクラス問題，動画，サッカー，スポーツ，LSTM

1 はじめに

1.1 研究背景

近年，スポーツ分野では，審判の補助や選手の分析等に映像を用いた技術が使われている．しかし，サッカーの放送における編集は，人の手で行われていることが多い．例として，試合のハイライトの作成が挙げられる．ハイライトは，編集者が試合中からゴールシーン等の注目するべき箇所を切り抜き，つなぎ合わせることによって作成される．この作業は，1 試合を見て，編集作業を行わなければならないため，多くの時間がかかる．このような問題に取り組むため，サッカー映像に対するデータセット [1], [2] が提供されている. [1], [2] を用いた [3] では，時間的関係性を考慮したモデルを提案し，ハイライトの自動生成を目的とした研究が行われている．また, [4] では，アクション前後の関係性からクラスタリングを応用した手法が提案されている. [6] では，アクションの不均一性を改善するためにサンプリングとマスキングの手法を取り入れている．

これらの研究では，単一の特徴量を用いたモデルとして 1 次元畳み込み [5], [6] やクラスタリング [4] が用いられている. [7] では，Transformer [9] が用いられているが複数の特徴量を入力としている．また，これらのモデルの入力は短い時間となっている．このことから，本研究では，少ない特徴量で局所的な時間関係を取り入れるために LSTM を組み込んだ手法を提案し，アクションの特定・分類精度の向上を目的とする．

1.2 本論文の構成

2 章では，関連研究について述べる．3 章では，提案手法について述べる．4 章では，提案手法とベンチマークとの比較実験について述べる．5 章では，結論と今後の課題について述べる．

2 関連研究

2.1 サッカー動画に関するデータセット

サッカー動画に関する動作分類に用いられるデータセットとして [1], [2] が存在する. [1] は，トレーニングデータ 300 試合，バリデーションデータ 100 試合，テストデータ 100 試合からなる計 500 試合のサッカー映像を提供するデータセットである．また，映像だけでなく [8], [14], [15] を用いて抽出した特徴量を提供している．クラス数は，Goal，Subsituation，Foul の 3 種類であり，6,637 アクションのフレームレベルアノテーションを付与している. [2] は, [1] の拡張データセットである. [1] の試合に加えてチャレンジデータ 50 試合を含んでいる．また，クラス数は [1] の 3 クラスを含む 17 クラスに拡張されており，110,458 アクションのアノテーションが付与されている．そして，新しく複数のタスクである Action Spotting，Camera shot segmentation，Replay grounding の 3 つのタスクが設定されている，Action Spotting は，17 クラスのアクションの種類とフレームの位置を特定するタスクである．Camera shot segmentation は，カメラの切り替えを検出するタスクである．Replay grounding は，リプレイと映像を結びつけるタスクである．

2.2 サッカー動画に関する動作分類

サッカー動画に関するデータセットのうち [1], [2] は，サッカー動画を扱う他のデータセットに比べて規模が大きく, アノテーションが正確に付与されていることから多くの研究に用いられている．A. Cioppa らは，アクション前後の時間的関係性を考慮し，Segmentation と Spotting の 2 つのモジュールからなるモデルを提案した [3]．Segmentation モジュールで

は，アクション前後の時間を 5 つの時間区域に分ける Time shift encoding を行い，時間関係を考慮するため提案された Temporal Segmentation Loss を損失関数として用いている．Spotting モジュールでは，アクションの種類とフレーム位置を出力とする YOLO like encoding を行い，アクションの種類とフレーム位置を考慮した Action Spotting Loss を損失関数として用いている．S. Giancola らは，アクションの前後関係に注目し，アクション前後のフレームに対してクラスタリングを応用した [4]．クラスタリング手法としてアクション前後のフレームに対して異なる [13] を用いることによって，アクションの前と後で異なる特徴量を抽出している．K. Vats らは，1 次元畳み込みのタワー構造を用いている [5]．イベント頻度が高いデータセットに対して精度を向上させることを目的として，ストライドサイズが異なる 1 次元畳み込み層を積み上げた 3 つのタワー構造を用いたモデルを提案した．B. Vanderplaetse らは，映像と音声を用いたマルチモーダルモデルを提案した [12]．映像では，各フレームに対して [8] を用いて特徴量を抽出している．音声では，音声スペクトログラムから [16] を用いて特徴量を抽出している．A. Cinppa らは，カメラ映像から 3 種類の特徴量を作成するモデルを提案した [11]．1 つ目は，フィールドを上部から見た 2 次元位置情報である．2 つ目は，2 次元位置情報を次元削減した特徴量ベクトルである．3 つ目は，2 次元位置情報をノードとしたプレイヤーグラフである．M. Tomei らは，データセットにおけるアクション数の不均一性を改善するためにサンプリングを行い，リプレイ等に隠れたアクションの認識精度を高めるためにマスキングを適応した [6]．サンプリングでは，アクション数に上限を設けることによって学習する際のアクション数の不均一性を改善している．マスキングでは，アクション前の数フレームを確率的に置き換えることによって汎化性能を向上させている．A. Vaswani らは，5 種類の特徴量抽出器によって得られた特徴量を用いて，時間的特徴を得るために [9] の構造を組み込んだ [7]．分類モデルでは，[9] の Encoder 構造を 3 層重ねた構造と線形層を用いている．

単一の特徴量を用いたモデルでは，1 次元畳み込み [5], [6] やクラスタリング [4] を用いた手法が提案されている．[7] では，モデルに Transformer [9] を用いているが複数の特徴量を入力としている．また，これらの研究では入力が短い時間となっている．このことから時間的局所性を捉えることが重要であると考え，本研究では，少ない特徴量で局所的な時間関係を捉えるために LSTM [17], [18] を用いたモデルを提案する．

3 提 案 手 法

3.1 概　　要

本節では，提案手法の概要について述べる．本研究では，LSTM [17], [18] を用いた 3 つのモデルを提案する．図 1 に提案手法の全体図を示す．図 1 は，2 つのモデルから構成されている．1 つ目が，特徴量抽出器である．特徴量抽出器では，[2] によって提案されたモデルを用いる．特徴量を抽出するモデルとして [10] によって事前学習された ResNet-152 [8] と PCA を用いる．特徴量抽出器では，フレーム数 T 枚の画像データ $(224, 224, 3)$ を入力として，各画像を ResNet-152 [8] を通し得られた 2048 次元の特徴量を PCA を用いて次元削減を行い，$(T, 512)$ の特徴量を出力する．2 つ目が，提案モデルである．提案モデルでは，特徴量抽出器によって得られた特徴量 $(T, 512)$ を LSTM Block に入力し，FC 層，Sigmoid 関数を通して 18 クラスの確率値を出力する．LSTM Block に関しては，提案手法ごとに構造が異なるため，3.2 節以降で詳細に説明を行う．

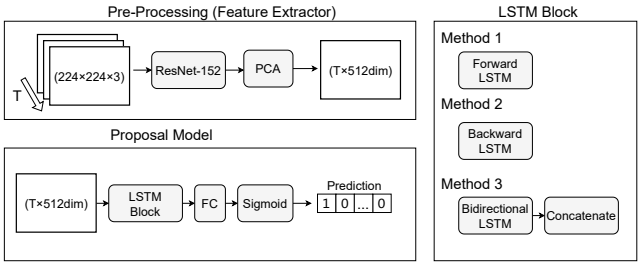

図 1: 提案手法

3.2 提案手法 1

提案手法 1 では，図 1 の LSTM Block に前向き LSTM を用いる．図 2 にモデルの詳細を示す．前向き LSTM では，3.1 節で述べた特徴量抽出器から得られた特徴量 $(T, 512)$ を入力とし，$(T, 512)$ の出力が得られる．この出力のうち時系列の最後の出力 $(1, 512)$ を取得し，FC 層と Sigmoid 関数を通して 18 クラスの確率値を出力する．

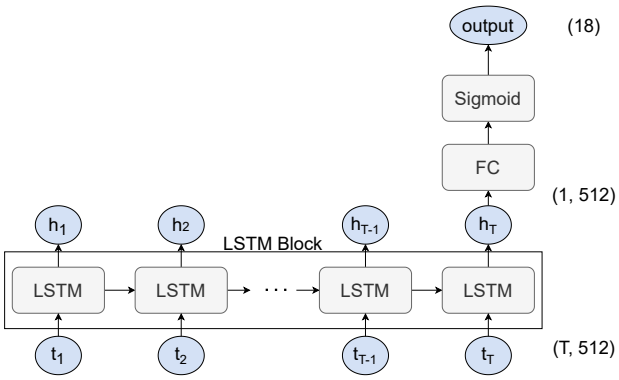

図 2: 提案手法 1

3.3 提案手法 2

提案手法 2 では，図 1 の LSTM Block に後ろ向き LSTM を用いる．図 3 にモデルの詳細を示す．後ろ向き LSTM は，3.1 節で述べた特徴量抽出器から得られた特徴量 $(T, 512)$ を時系列を反転させて，前向き LSTM に入力することによって実現する．時系列を反転させた特徴量 $(T, 512)$ を入力とし，後ろ向き LSTM から $(T, 512)$ の出力が得られる．この出力のうち時系列の最後の出力 $(1, 512)$ を取得し，FC 層と Sigmoid 関数を通して 18 クラスの確率値を出力する．

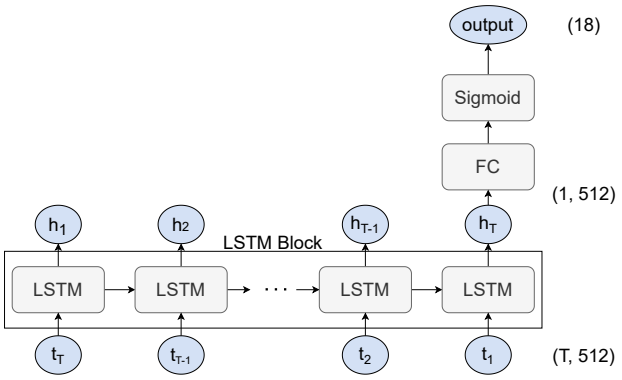

図 3: 提案手法 2

3.4 提案手法 3

提案手法 3 では，図 1 の LSTM Block に双方向 LSTM を用いる．図 4 にモデルの詳細を示す．双方向 LSTM は，3.1 節で述べた特徴量抽出器から得られた特徴量 $(T, 512)$ を入力する．出力として前向き・後ろ向き LSTM の出力 $(T, 1024)$ が得られる．このうち，前向き LSTM の最終出力 $(1, 512)$ と後ろ向き LSTM の最終出力 $(1, 512)$ を結合する．得られた出力 $(1, 1024)$ を入力とし，FC 層と Sigmoid 関数を通して 18 クラスの確率値を出力する．

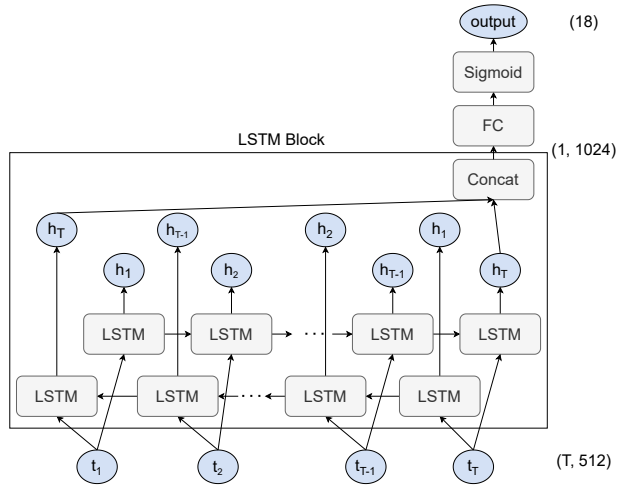

図 4: 提案手法 3

4 比較実験

4.1 実験内容

4.1.1 データセット

本研究では, [2] で提供された SoccerNet-V2 データセットを用いる．このデータセットは，ラ・リーガ (スペイン)，セリエ A(イタリア)，ブンデスリーガ (ドイツ)，リーグ 1(フランス)，プレミアリーグ (イングランド)，UEFA チャンピオンズリーグから 550 試合を抽出したサッカー映像のデータセットである．また，Action spotting, Camera shot segmentation, Replay grounding の 3 つのタスクが設定されており，本研究では Action Spotting をタスクとする．Action spotting におけるラベルでは，17 クラスのアクションがフレームレベルで付与されている．提供データは，トレーニングデータ 300 試合，バリデーションデータ 100 試合，テストデータ 100 試合，チャレンジデータ 50 試合 (本研究では未使用) に分けられている．

4.1.2 サンプリング

本研究では，先行研究で述べられたアクションの不均一性を改善するために [6] で提案されたサンプリング手法を用いる．この手法では，各アクションに上限を設けることによって，学習時における各エポックで用いるデータ内のアクション数を均一に近づける．

4.1.3 Non Maximum Suppression

テストデータにおける予測の出力には，Non Maximum Suppression(以下，NMS と示す) を適応する．NMS では，フレーム数 T_{NMS}, 閾値 η を設定し，各クラスの予測系列に対して T_{NMS} 範囲内で η 以上の確率を持つ最大の出力のみを予測結果として残す処理を行う．

4.1.4 評価指標

評価指標は, [1], [2] において考案された Average-mAP を用いる．Acerage-mAP を計算する際に時間的許容域 δ を考える．δ は，Ground Truth(以下，GT と示す) からの許容範囲を示し，その中にアクションの予測が入っていれば True Positive(以下，TP と示す) と判定する．また，GT 以外をアクションでないと判定した場合は，False Positive(以下，FP と示す) として，GT をファールではないと判定した場合を False Negative(以下，FN と示す) とする．

ここで，Average-mAP の計算手順を以下に示す．

（1） 閾値 η 以上の予測値における TP，FP，FN を数える．
（2） 閾値 η における Precision，Recall を計算する．
（3） 1，2 を閾値 $\eta = 0.0, 0.1, \cdots, 1.0$ に対して行う．
（4） 各閾値の Precision，Recall から PR 曲線を描画し，Area Under the Curve によって Average Precision(以下，AP と示す) を求める．
（5） AP を時間的許容域 $\delta = 5, 10, \cdots, 60$ に対して求める．
（6） 各時間的許容域 δ における AP を描画し，台形公式によって mean AP(以下，mAP と示す) を求める．
（7） mAP を各アクション毎に求め，平均した結果を Average-mAP とする．

4.1.5 実験条件

本実験では，ベンチマークとして [3], [6], [7] を用いる．データには，4.1.1 節で示したデータセットを用いた．データの分割は，データセットで提供されている分割方法を用い，トレーニングデータ 300 試合，バリデーションデータ 100 試合，テストデータ 100 試合とした．サンプリングにおける各アクションの上限値を 1000 として設定した．トレーニングデータにおける各アクションの数を表 1 に示す．モデルの入力フレーム数は $T = 30$ とした．損失関数は，Binary Cross Entropy を用いる．学習時には，エポック数を 400 として，バリデーションのスコアが最も低い重みを用いた．また，バッチサイズは 32 とした．最適化手法には SGD を用い，学習率 (lr) は 0.001, momentum は 0.8 として設定した．評価指標は，4.1.4 節で示

した Average-mAP を用いた．出力に用いる NMS のパラメータは，$T_{NMS} = 30, \eta = 0.5$ とした．

4.2 実験結果

表 2 に各モデルの評価値を示す．本実験で提案した手法では提案手法 3 の双方向 LSTM を用いたモデルが最も Average-mAP が高い結果となった．提案手法でクラスごとの評価値を比較すると提案手法 1，2 では，“Ball out” や “Clearance” のように他の提案手法よりも評価値が高くなっているクラスがある一方で，“Dir. free-kick” や “Offside” のように著しく低くなっているクラスが存在する．これに比べて提案手法 3 では，著しく低くなっているクラスが少なく，平均的な評価値となっている．提案手法 1，2，3 をベンチマークと比較すると，全体の評価値では低い結果となった．提案手法 3 とベンチマークのクラスごとの評価値を比較すると，“Corner” では提案手法 3 が [3] よりも 0.83 高い結果となった．しかし，“Shot on Target” では, [3] よりも 10.18 低い結果となった．

4.3 クラスごとの分析

4.3.1 Corner クラス

提案手法 3 において，高い評価値となった “Corner” について述べる．このクラスは，コーナーキックをアクションとするクラスである．1 試合 (2015/8/23 15:30 West Brom 2 - 3 Chelsea) の後半における予測とアクション箇所のグラフを図 5 に示す．1 試合 (2015/8/23 15:30 West Brom 2 - 3 Chelsea) における混合行列の結果を表 3 に示す．特定のフレームから前後 15 フレーム以内に予測値があるかを基準とした．図 5 からアクション箇所に対して予測ができている箇所が複数あることが分かる．一方で，アクション箇所以外で予測が出ている箇所が存在する．また，表 3 から正解率が 99.93%，再現率が 84.52%，適合率が 64.71%，F 値が 73.30% となっていることがわかる．このことから，アクション箇所に対する間違いは少なく，アクション以外の箇所での誤った予測が一定数起こっていると考えられる．

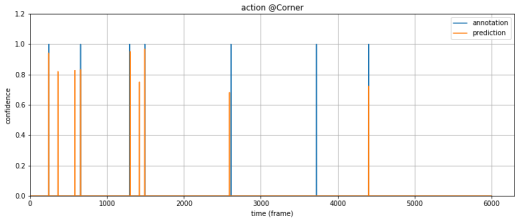

図 5: Corner クラスの予測とアクション箇所

4.3.2 Shot on target クラス

提案手法 3 において，低い評価値となった “Shot on target” クラスについて述べる．このクラスは，枠内へのシュートをアクションとするクラスである．1 試合 (2015/8/23 15:30 West Brom 2 - 3 Chelsea) の前半における予測とアクション箇所のグラフを図 6 に示す．1 試合 (2015/8/23 15:30 West Brom 2 - 3 Chelsea) における混合行列の結果を表 4 に示す．特定のフレームから前後 15 フレーム以内に予測値があるかを基準とした．図 6 からアクション付近で予測できている箇所が少なく，アクションのない箇所に関して予測が多く出ていることがわかる．また，表 4 から正解率が 99.84%，再現率が 14.29%，適合率が 7.69%，F 値が 10.00% となっていることがわかる．

アクション箇所に対して正しく予測した例とアクション箇所以外を誤ってアクションとして予測した例を図 7 に示す．図 7(a) では，中央から枠内へシュートを行うシーンであり，正しく予測できている．それに対して，図 7(b) では，サイドから中央へのパスをキーパーがキャッチしたシーンを枠内へのシュートシーンとして予測している．この例から Shot on target クラスでは，枠内へのシュートに対して，細かなシーンの違いを判別できていないことが考えられる．

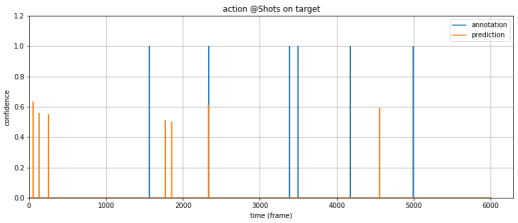

図 6: Shot on target クラスの予測とアクション箇所

4.4 同特徴量を入力としたモデル比較

4.4.1 実験内容

本研究における提案手法と [7] で提案された Transformer [9] を用いた手法の同特徴量を入力した際の比較実験を行う．特徴量には，本研究で用いた，ResNet-152 [8] と PCA によって得られた $(T, 512)$ 次元の特徴量を用いる．提案手法は，4.1.5 節に示した条件と同条件で実験を行った．本実験では, [7] のモデルの入力を変更し，再現実装を行った．図 8 に本実験で用いた [7] の再現実装モデルを示す．再現実装モデルは，特徴量抽出器から得られた $(T, 512)$ 次元の特徴量と Positional Encodeing を入力として，Transformer [9] の Encoding layer を 3 層，Sigmoid 関数を通して 18 クラスの確率値を出力する．また，提案手法と同様に 4.1.5 節に示した条件と同条件で実験を行った．

4.4.2 実験結果

実験結果を表 5 に示す．結果から同特徴量を入力とした際には，全体の評価値において本研究の提案手法よりも再現実装モデル [7] の方が 4.16 高い結果となった．多くのクラスでは，再現実装モデル [7] が最も高い評価値となったが，“Shot on target” や “Substitution” においては，提案手法 3 が最も高い評価値となった．

5 おわりに

本研究では，サッカー動画における動作分類を行う手法を提案した．具体的には，少ない特徴量で局所的な時間関係を取り入れるため，LSTM Block を組み込んだモデルを提案した．

表 1: トレーニングデータにおけるアクション数

アクション種類	Ballout	Throw-in	Foul	Ind. free-kick	Clearance	Shots on tar.	Shots off tar.	Corner	Substitution	Kick-off	Yellow card	Offside	Dir. free-kick	Goal	Penalty	Yel. → Red	Redcard	Background
サンプリング前	19097	11391	7084	6331	4749	3463	3214	2884	1700	1516	1238	1265	1379	995	96	24	34	3234692
サンプリング後	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	995	96	24	34	1000

表 2: ベンチマークと提案手法における Average-mAP とクラスごとの mAP

Method	Average-mAP	Shown	UnShown	Ball out	Throw-in	Foul	Ind. free-kick	Clearance	Shots on tar.	Shots off tar.	Corner	Substitution	Kick-off	Yellow card	Offside	Dir. free-kick	Goal	Penalty	Yel. → Red	Red card
X. Zhou, et al. [7]	**73.77**	**79.28**	**47.84**	**87.87**	**87.81**	**84.23**	**71.22**	**76.50**	**66.28**	**65.69**	**88.92**	**78.98**	**82.83**	**79.11**	**67.97**	**75.05**	**86.81**	**93.63**	**35.60**	25.62
M. Tomei et al. [6]	63.49	68.88	38.03	58.67	67.99	69.27	64.29	76.24	37.94	50.12	88.80	71.28	65.19	76.76	58.32	66.83	77.60	89.20	30.83	**30.00**
A. Cioppa et al. [3]	41.61	43.54	28.88	66.40	59.76	54.64	40.91	51.90	23.77	26.71	72.60	49.86	36.07	40.95	29.16	44.04	70.16	39.60	0.00	0.78
提案手法 1	27.49	32.48	18.03	38.54	34.12	44.22	9.63	30.40	11.24	25.78	64.88	55.30	30.31	27.21	22.30	8.76	64.69	0.00	0.00	0.00
提案手法 2	20.84	22.88	13.94	9.09	40.97	13.10	19.11	35.28	9.09	4.55	69.82	46.44	24.60	20.75	3.03	40.99	11.25	6.20	0.00	0.00
提案手法 3	30.63	32.75	17.97	34.69	36.25	43.78	21.75	34.46	13.59	23.71	73.43	58.61	26.47	28.72	21.40	39.38	64.41	0.00	0.00	0.00

(a) アクション箇所に対して正しく予測した例

(b) アクション箇所以外を誤って予測した例

図 7: 正しい予測例と誤った予測例

表 3: Corner クラスの混合行列

		Prediction		
		Positive	Negative	Total
Annnotation	Positive	11	2	13
	Negative	6	11305	11311
	Total	17	11307	11324

表 4: Shot on target クラスの混合行列

		Prediction		
		Positive	Negative	Total
Annnotation	Positive	1	6	7
	Negative	12	11305	11317
	Total	13	11311	11324

LSTM Block には，前向き LSTM，後ろ向き LSTM，双方向 LSTM の 3 手法を適用した．提案手法は，全体の評価値では既存手法よりも低くなったが，特定のクラスにおいては既存手法に近い結果となった．評価値の低くなったクラスでは，シーンの細かな違いを認識できていないと推察される．また，同特徴量を入力とする実験では，特定のクラスにおいて再現実装モデルよりも提案手法の方が高い結果となった．今後の課題としては，細かなシーン違いにおける認識精度の向上が挙げられる．

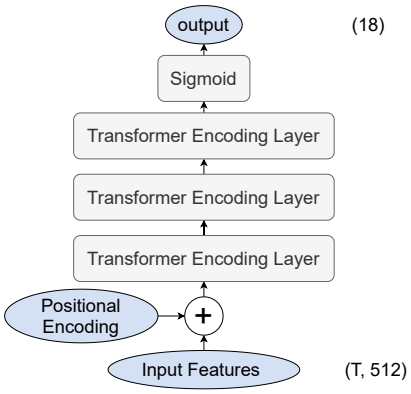

図 8: 再現実装モデル

文　　献

[1] S. Giancola, et al., "Soccernet: A scalable dataset for action spotting in soccer videos", in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 1711-1721.

[2] A. Deliege, et al., "Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4508-4519.

[3] A. Cioppa, et al, "A context-aware loss function for action spotting in soccer videos", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

表 5: 提案手法とベンチマークの同特徴量入力時の実験結果

Method	Average-mAP	Shown	UnShown	Ball out	Throw-in	Foul	Ind. free-kick	Clearance	Shots on tar.	Shots off tar.	Corner	Substitution	Kick-off	Yellow card	Offside	Dir. free-kick	Goal	Penalty	Yel. → Red	Red card
再現実装モデル [7]	**34.79**	**54.14**	**52.29**	**54.14**	**52.29**	**45.59**	**28.70**	**40.51**	9.68	20.80	**76.23**	55.52	30.17	**34.19**	**24.17**	**42.12**	63.45	**13.84**	0.00	0.00
提案手法 1	27.49	32.48	18.03	38.54	34.12	44.22	9.63	30.40	11.24	**25.78**	64.88	55.30	**30.31**	27.21	22.30	8.76	**64.69**	0.00	0.00	0.00
提案手法 2	20.84	22.88	13.94	9.09	40.97	13.10	19.11	35.28	9.09	4.55	69.82	46.44	24.60	20.75	3.03	40.99	11.25	6.20	0.00	0.00
提案手法 3	30.63	32.75	17.97	34.69	36.25	43.78	21.75	34.46	**13.59**	23.71	73.43	**58.61**	26.47	28.72	21.40	39.38	64.41	0.00	0.00	0.00

2020, pp. 13126-13136.

[4] S. Giancola, et al., "Temporally-Aware Feature Pooling for Action Spotting in Soccer Broadcasts.", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4490-4499.

[5] K. Vats, et al., "Event detection in coarsely annotated sports videos via parallel multi-receptive field 1d convolutions.", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 882-883.

[6] M. Tomei, et al., "Rms-net: Regression and masking for soccer event spotting.", in 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021, pp. 7699-7706.

[7] X. Zhou, et al., "Feature Combination Meets Attention: Baidu Soccer Embeddings and Transformer based Temporal Detection.", arXiv preprint arXiv:2106.14447, 2021.

[8] K. He, et al., "Deep residual learning for image recognition", in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770-778.

[9] A. Vaswani, et al., "Attention is all you need", in Proceedings of the Advances in neural information processing systems, 2017, pp. 5998-6008.

[10] J. Deng, et al., "ImageNet: A large-scale hierarchical image database.", in proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248-255.

[11] A. Cinppa, et al., "Camera Calibration and Player Localization in SoccerNet-v2 and Investigation of their Representations for Action Spotting.", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4537-4546.

[12] B. Vanderplaetse, et al.,"Improved soccer action spotting using both audio and video streams", in proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 896-897.

[13] R. Arandjelovic, et al., "NetVLAD: CNN architecture for weakly supervised place recognition.", in proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 5297-5307.

[14] D. Tran, et al., "Learning spatiotemporal features with 3d convolutional networks.", in proceedings of the IEEE international conference on computer vision, 2015, pp. 4489-4497.

[15] J. Carreira, et al., "action recognition? a new model and the kinetics dataset.", in proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299-6308.

[16] S. Hershey, et al., "CNN architectures for large-scale audio classification.", in proceedings of 2017 ieee international conference on acoustics, speech and signal processing (icassp), 2017, pp. 131-135.

[17] F A. Gers, et al., "Learning to forget: Continual prediction with LSTM.", in proceedings of Neural computation, 2000, 12.10: pp.2451-2471.

[18] H. Sak, et al., "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition.", arXiv preprint arXiv:1402.1128, 2014.