G-014

サッカーにおける実況内容を含めたマルチモーダルな行動認識手法の検討

熊倉多香音† 折原 良平† 田原 康之† 大須賀昭彦† 清 雄一†

† 電気通信大学大学院情報理工学研究科情報学専攻 〒 182–8585 東京都調布市調布ヶ丘 1-5-1

E-mail: †kumakura.takane@ohsuga.lab.uec.ac.jp, ††orihara@acm.org, †††{tahara,ohsuga,seiuny}@uec.ac.jp

1 はじめに

サッカーは世界中で人気のスポーツであり, 135 カ国にフットボールクラブが存在している [4]. また, 2022 年のワールドカップでは約 163 万人もの観客を動員しており, Mordor Intelligence [10] によるとサッカーの市場規模は USD 7 億 4.145 万ドル (2024 年) と推定されている. このようなサッカーの人気ぶりから, 近年, 選手の動きや状況を把握して分析するスポーツ分析の分野が活発に研究されている. スポーツ分析は, チームでの戦略・プレイヤーのパフォーマンス判断・プレイヤーのスカウトや試合での判定・ハイライト生成など、多岐にわたる用途で活用されている. 例えば, 映像の要約を人手で作成するには, 前後半合わせて約 90 分のビデオを手動でトリミングして編集する必要があるため、多大な時間と労力が必要となる. よって, 自動生成が可能になることで, 時間や労力を削減できるだけでなく, 効率的な戦術レビューや即時的な映像提供も可能となる. このため, 放送動画から自動でプレイヤーの行動を認識させる技術は活発に取り組まれている研究分野である.

本研究において取り組んでいるのは, サッカー映像内で特定のアクションがいつ発生したかを識別する, 時間的行動検出 (Temporal Action Detection: TAD) タスクの一つである, Action Spotting [2] である. 図 1 のように, Goal や Corner, Yellow Card などの行動がどの瞬間に行われたか推測することを目的としている.

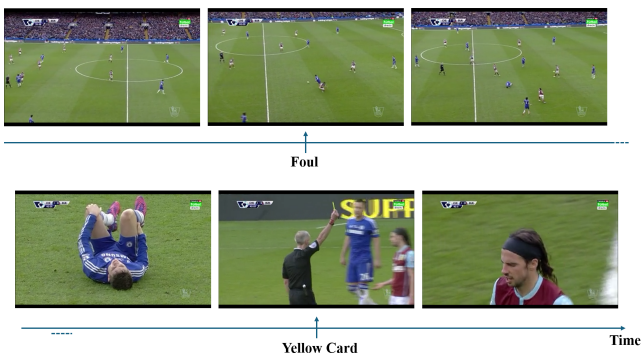

図 1: Action Spotting の例

既存研究では, 主にサッカーの放送映像から得た視覚的特徴を利用して, Action Spotting に取り組んでいるものが多く提案されている. しかし, リプレイが存在することやカメラアングルの問題により, 各行動全てを映像内に収めてはいない. 映像内には写されていないものの実際の試合では起こっている行動は不可視行動として定義されており, 可視行動に比べて精度が低いことが課題となっている [20]. そこで, 近年の研究ではグラフ [1] ・音声 [5] [15] [18] [20] のモダリティを活用することで不可視行動の精度向上を図っている.

例えば, Gan ら [5] は視覚と聴覚両方のモダリティに対する Transformer [19] ベースのマルチモーダルサッカーシーン認識手法を提案した. Visual Transformer にビデオフレームを入力し, Audio Transformer に音声スペクトログラムを入力し, 最終的にそれらの推定結果に対して Late Fusion を行うことで, 視覚・聴覚両方のモダリティを扱っている. また, VanderPlaetse ら [18] は ImageNet で事前学習した ResNet を Visual Stream に, AudioSet で事前学習した VGGish [9] を Audio Stream に設定し、各 Stream で特徴量を抽出した後に, 2 つのモーダルを 7 つの方法で融合させ, その結果を比較している. また Cartas ら [1] は, サッカー選手や審判をグラフのノードとして識別・表現し, それらの時間的相互作用をグラフのシーケンスとしてモデル化した. そして, Xarles ら [20] らは VGGish を用いて, 音声のログメルスペクトログラムから特徴量を抽出し, Transformer Encoder の入力で視覚的特徴量とマージすることで, 音声及び映像のモダリティを活用するマルチモーダルなアプローチを提案した.

しかし, これらの手法は観客や実況における盛り上がりや雰囲気・声には注目しているものの, 実況の内容には焦点を当てられていない. その理由は, Action Spotting の多くの既存研究で用いられている SoccerNet-v2 [2] データセットは, 表 1 のように, 試合映像ごとに言語が大きく異なり, 映像によっては言語が入っていないものも存在するからである. 言語が異なると音声特徴量は全く異なるものになるため, コメントの実況内容は精度向上には寄与していない.

また, 視覚的には Yellow Card は 1 回しか掲げていないように見える可視行動であっても, 実際には両チームの選手に対して出されるケースや, 視覚的には Red Card しか掲げていないものの, 実際は Yellow Card が 2 枚出されて Red Card に至るケース (YC → RC) も存在した. これらの状況において実況では, “Yellow Card to ○○, and to ○○”などの表現を用いて両選手にカードが出されていることが分かる場合や, “Second

表 1: 本研究にて FasterWhisper を用いて特定した, SoccerNet-v2 データセットにおける言語の内訳

言語	English	Spanish	Russian	German	French	Norwegian Nynorsk	Italian	Turkish
train	185	153	112	73	60	7	3	2
valid	71	46	38	18	20	0	0	2
test	59	55	42	30	8	2	2	0
challenge	26	14	26	16	14	0	0	0
total	341	268	218	137	102	9	5	4

言語	Korean	Polish	Latin	Welsh	Māori	Croatian	Hungarian	None
train	0	1	1	2	0	0	0	1
valid	1	0	0	1	1	0	0	2
test	0	0	0	0	0	0	0	2
challenge	0	0	0	0	0	2	2	0
total	1	1	1	3	1	2	2	5

Yellow Card, so Red!”などの表現を用いて YC → RC であることが分かる場合がある．また, 視覚的にゴールが確認できる場合でも, “Goal!!”と発言している場合も確認された．

これらの課題及び試合の様子に関する観察を踏まえ, 本研究では映像情報・音声情報に加え, コメンテーターの実況内容というテキスト情報の 3 つのモーダルに対して, Transformer ベースの ASTRA(Action Spotting TRAnsformer for Soccer Videos) モデルを活用する, マルチモーダルなサッカー行動認識手法を提案する．コメンテーターの実況内容を情報量として加えることで, 映像でも音声でも認識できていなかった不可視行動が認識できるようになることや, 可視行動の詳細な状況の理解により, データの少ないアクションに対する認識精度の向上を目指す．

2 関連研究

2.1 ASTRA

Xarles ら [20] らは, Action Spotting のために設計された Transformer ベースのモデルである ASTRA を提案した．ASTRA は SoccerNet 2023 Action Spotting Challenge の Challenge Set にて Average-mAP で 3 位の精度となったモデルである．まずビデオモーダルとして, Zhou ら [22] によって作成された Baidu Soccer Embeddings に対し, PFFN(Position-wise Feed-Forward Network) を通した．ここで PFFN では, 位置単位に順伝播ネットワークを形成することで、並列処理を可能としている．そして音声モーダルとして, 音声のログメルスペクトログラムを入力とし, AudioSet dataset [6] で事前学習された VGGish [9] を用いて Audio Embedding を得た．そして Video Embedding と Audio Embedding の特徴次元を揃え, Transformer Encoder, Decoder を介して結合されて処理した．そして得られた Embedding は, 時間的位置分類のための分類ヘッド (Λ_s) と予測をさらに洗練するための変位ヘッド (Λ_d) によって利用される．ASTRA は不可視行動の精度が可視行動のそれに比べて低いという課題に対し, 音声モーダルを活用することで対処した．他にも, Xarles らはタスク及びデータセット内に内在する, ラベルの変動性やロングテール分布に対して対処した．

2.2 Baidu Soccer Embeddings

Zhou ら [22] らは, サッカー放送映像に対するイベント検出のための 2 段階のアプローチを提案した．最初のステージでは, TPN [21], GTA [8], VTN [11], irCSN [17], I3D-Slow [3] という複数のアクション認識モデルをサッカーデータに Fine-Tuning して高次のセマンティック特徴を抽出した．次に Action Spotting の時間検出モジュールとして, NetVLAD++ [7] 及び Transformer を利用した．

2.3 Whisper 及び FasterWhisper

Radford ら [14] は, インターネット上の大量の音声データを用いた弱教師あり学習に基づく音声処理システムを提案した．本モデルは Sequence-to-Sequence Transformer モデルを用いて多言語音声認識・音声翻訳・話者識別・音声活動検出などの複数の音声処理タスクを学習させている．また, FasterWhisper [16] は Whisper モデルの改良版となるモデルであり, Ctranslate2 エンジンという高速な推論エンジンの使用, 8 ビット量子化の導入, 最適化処理の導入により, 高速かつ効率化を実現した音声認識モデルである．GPU を使用した際の Whisper と FasterWhisper を比較すると, Whisper に比べて最大 4 倍の処理速度向上を実現しながらも, GPU メモリや CPU メモリ使用量の大幅な削減を可能としている．

3 アプローチ

本研究では映像情報・音声情報・コメンテーターの実況内容という実況テキスト情報の 3 つのモーダルに対して, Transformer ベースのマルチモーダルなサッカー行動認識手法を提案する．SoccerNet の放送音声から FasterWhisper を用いて実況テキストの書き起こしを行い, 英語でないものに関しては GPT-4o [12] を用いてテキスト翻訳を行うことで, データセットの全ての試合映像に関して英語の実況テキストを得た．そして実況テキストに関して Text Embedding Large 3 [13] を用いて得た Embedding を, 既存の ASTRA モデル [20] に加えて学習させた．

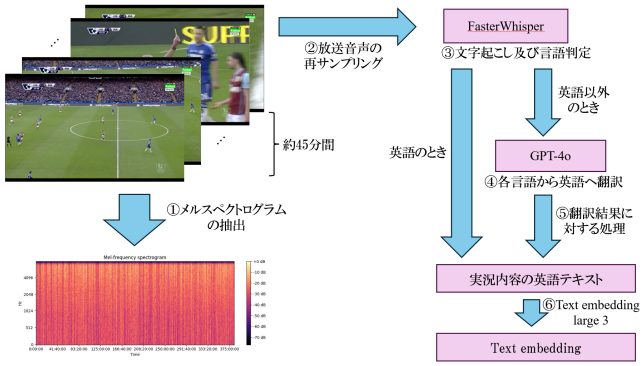

図 2: 音声モーダルとテキストモーダルのデータセット作成の流れ

まず, メルスペクトログラム・Text Embedding のデータセット作成の流れを図 2 に示す．そして ASTRA をベースとした本研究の提案モデルを図 3 に示す．

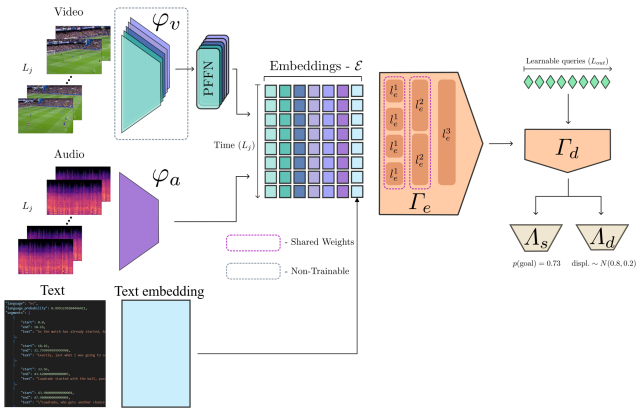

図 3: ASTRA モデルに対してテキストモダリティを追加した，本研究における提案モデル．図は Xarles ら [20] の図をアレンジしたものである．

3.1 音声モーダルのデータセット作成

音声情報を用いた学習を行うため，SoccerNet-v2 の放送音声からメルスペクトログラムを作成した．具体的にはまず，オリジナルの音声ファイルのサンプリングレートを 16,000Hz に設定し，オーディオチャネルをモノラルに設定することで，音声データのサイズを小さくし，処理を効率化した．

次に，メルスペクトログラムのシーケンス長と放送映像のシーケンス長を揃えるため，その音声ファイルをサンプリングレートを 100 で読み込み，短時間フーリエ変換 n_fft を 512，ホップサイズは次式 1 で定め，メルフィルタバンク n_mels を 128 と設定してメルスペクトログラムを計算した．ここで，len(y) は放送映像のシーケンス長である．そしてパワースペクトログラムを dB 単位に変換することで，数値処理を扱いやすくした．ここで，音声が含まれていない又は音声が途中で途切れている放送音声に対しては音声ファイルに 0 の値を追加して処理を行った．

$$\text{hop\_length} = \left\lfloor \frac{\text{len}(y) - n\_fft}{\text{len}(y) - 1} \right\rfloor + 1 \qquad (1)$$

これらの処理を SoccerNet-v2 データセットにおける全ての試合映像に対して行い，メルスペクトログラムを抽出することで，音声モーダルのデータセットを作成した．

3.2 テキストモーダルのデータセット作成

ASTRA モデルを活用して実況内容を含めた学習を行うため，音声から実況内容を Text Embedding として得た．具体的にはまず，音声モーダルのデータセット作成と同様の処理を行うことで，音声データのサイズを小さくし，処理を効率化した．

次に，FasterWhisper を用いて音声の文字起こしを行うとともに言語判定を行った．特に文字起こしを行う際に，発言の時系列情報を保持することに留意した．具体的には，各発言が何秒から何秒までの間に行われたかという時間情報を含む形式で文字起こしデータを保存した．この方法により，発言の正確なタイミングを確実に把握することが可能となる．

次に，FasterWhisper を用いた言語判定の結果，英語以外の言語であると判定された試合に対しては，GPT-4o [12] の API を使用して各言語を英語へ翻訳した．特に，テキストと時間の依存関係を維持するために，各試合を一括して翻訳するのではなく，各秒数における書き起こしテキストの生成結果を逐次翻訳した．ここで，GPT-4o に翻訳させたときのプロンプトは，“Translate the following text from (放送音声の言語) to English: (翻訳したい文章)”とした．放送音声の言語には，FasterWhisper の言語判定で出力された言語を指定した．

またここで，各秒数における書き起こしテキストが各言語における大文字で始まる場合，GPT-4o が固有名詞と判断し，翻訳が行われていなかったため元々の書き起こしテキストを利用した．また，書き起こしテキストが記号のみで構成されている場合についても翻訳を除外した．加えて，各書き起こしテキストにおいて，前のテキスト生成結果と同一の結果が生成されるケースが存在したため，そのケースに関しても翻訳を行わなかった．

ここで，GPT-4o の各出力に対して，以下の処理方法で翻訳結果を精査した．また各処理について，GPT-4o への入力とその出力，及び各処理結果の例を示す．下線部は FasterWhisper で書き起こされたテキストである．

（1） GPT-4o の出力には必ず文字列として認識させるための 1 対の引用符が存在する．そしてその内側にもう一対の引用符が存在する場合，その 1 対の引用符の内側を最終翻訳結果として採用した．

入力: “Translate the following text from Spanish to English: “Ahí vemos en el fondo a Remi, a Cahill.””

出力: “Sure, here is the translation:**\n\n**“There we see in the background Remi, Cahill...\””

処理結果: “There we see in the background Remi, Cahill...”

これにより，例のように GPT-4o の出力から “Sure, here is the translation:**\n\n**”の部分を省くことができる．

（2） GPT-4o の出力に必ず存在する 1 対の引用符の内側に，2 対以上の引用符が存在する場合，その複数の対となっている引用符の内側のうち，英語である確率が最も高いものを最終翻訳結果として採用した．

入力: “Translate the following text from Spanish to English: “colegiados que””

出力: “\“colegiados que\” can be translated to \“referees who\” in English.”

処理結果: “referees who”

これにより，例のように出力から英語の翻訳結果のみを抽出することができる．

（3） GPT-4o の出力に “translat”が含まれているにも関わらず，GPT-4o の出力に必ず存在する 1 対の引用符以外引用符が存在しない場合，元々の書き起こしテキストを翻訳結果として採用した．“translat”としたのは，“translation”及び “translate”の両方を含ませるためである．

入力: “Translate the following text from Spanish to English: “ 5 minutes.””

出力: “Sure, please provide the text you would like to have translated from French to English.”

処理結果: “ 5 minutes.”

これにより, 翻訳結果が存在しない出力に対して元の書き起こしテキストを採用した.

このようにして得られた、英語以外の言語の実況内容を書き起こしたテキストを翻訳し処理を行ったテキストと、英語の書き起こしテキスト全てに対して、Text Embedding を生成した. この Text Embedding を得るために、Text Embedding Large 3 [13] モデルを使用した. Text Embedding Large 3 は、テキストデータをエンベディングに変換するためのモデルの一種であり、汎用的に NLP タスクに応用することが可能である. ここで, 各書き起こしテキストは発言の開始時刻と終了時刻を保持しているため, その間の時刻に各書き起こしテキストのテキストエンベディングを挿入した. そして最終的に, 1 秒につき d 次元のテキストエンベディングを得て結合することで, 各試合に対して約 2700 × d のテキストエンベディングを得た.

3.3 ASTRA を用いた学習

本研究で提案するモデルは, ASTRA をベースとしてテキストモダリティを追加したモデルである. つまり, 主に ASTRA モデルに対して変更した点は, Transformer Encoder の前に Fusion している各モダリティに, Text Embedding を追加したことである. アーキテクチャを図 3 に示す. Baidu Soccer Embeddings を特徴次元方向に 5 つに分割し, PFFN に通して得られた Video Embeddings と, 3.1 章で得られたスペクトログラムをログメルスペクトログラムに変換し, VGGish モデルを通して得られた Audio Embedding と, 3.2 章で得られた Text Embedding を, Hierarchical Transformer Encoder に入力する前に結合した. これにより, 異なるモーダル間の依存関係を学習させる.

また, Text Embedding を結合する前に, Text Embedding に対して学習可能な時間的位置埋め込み及び特徴量に対する位置埋め込みを加えた. これにより, 本モデルを時間的・特徴量的な位置情報を考慮に入れて学習させる.

4 評　　価

本研究の提案モデルとして, コメンテーターの実況内容を情報量として加えることで, 映像でも音声でも認識できていなかった不可視行動が認識できるようになることや, 可視行動の詳細な状況の理解により, データの少ないアクションに対する認識精度の向上を目指した.

4.1 使用したデータセット

本研究では, サッカーの Action Spotting に対するデータセットとして, SoccerNet-v2 [2] を用いる. SoccerNet-v2 は, プレミアリーグ, UEFA チャンピオンズリーグ, リーグ・アン, ブンデスリーガ, セリエA, ラ・リーガにおいて 2014 年から 2017 年に開催されたサッカー試合 550 試合からなる, 多様なタスクに応じて提供されているデータセットである.

Action Spotting に対しては, 500 試合は 17 種類のアクションに対するアノテーションが公開されており, 残りの 50 試合は Challenge データセットとして主催者のみがアクセスできるアノテーションデータとなっている. そして 500 試合のうち, 300 試合を Train データ, 100 試合を Validation データ, 100 試合を Test データとして設定されており, 本研究においては Test データに対して 4.3 章に従って評価を行う.

4.2 実装の詳細

モデルのハイパーパラメータは ASTRA [20] と同様の設定にしている. 具体的にはモデル実装には Pytorch を採用し, Adam Optimizer を用いている. 初期学習率は 5×10^{-5}, 初期ウォームアップは 3 エポック, コサイン減衰による学習率ウォームアップは 50 エポックとした. このモデルには, 50 秒のクリップが $d = 512$ の埋め込み次元を用いて入力した.

またモデルにおける ASTRA と異なる箇所は, Embeddings の数 $\mathcal{E}$, Text Embedding の埋め込み次元, 及び Text Embedding に対する位置埋め込みである. まず, 視覚的データに対応する Baidu Soccer Embeddings が 5, 音声データに対応するログメルスペクトログラムを VGGish に通して得た Audio Embedding が 1, テキストデータに対応する Text Embedding が 1 の計 $|\mathcal{E}| = 7$ の Embeddings を利用した. また, Text Embedding の埋め込み次元は d=512 とした.

4.3 評 価 指 標

本手法に対する評価指標として, Average-mAP を利用した. この指標は, 異なる許容差 δ において, mAP の AUC を定量化したものである. mAP は, 全てのアクションクラス間の平均適合率 (AP) を平均化した値である. AP は, 縦軸に適合率 (Precision), 横軸に再現率 (Recall) を取ってプロットした PR 曲線を要約した値である. Action Spotting においては, 検出結果と正解値が一定の時間範囲で一致している必要があるため, 異なる許容差 δ を設定している.

そして SoccerNet では Average-mAP に対して tight Average-mAP 及び loose Average mAP の指標を採用している. tight Average-mAP は δ を 1 秒から 5 秒と設定しているのに対し, loose Average-mAP は δ を 5 秒から 60 秒と設定している. 本研究においても, tight Average-mAP 及び loose Average-mAP を用いて評価した. また各アクションクラスに対し, tight Average-AP 及び loose Average-AP を用いて評価した. ここで, 結果として示す全ての指標は各モデルについてランダムシードで 5 回学習を行い, 平均を取った値である.

4.4 本提案モデルに対する評価

音声スペクトログラム, 実況テキスト, 及び Text Embeddings は本研究で作成したものであるため, (i) 映像モダリティのみ, (ii) 映像モダリティ及び音声モダリティ, (iii) 映像モダリティ, 音声モダリティ, 及び実況テキストモダリティ (提案モデル) の

表 2: 全行動、可視行動、不可視行動それぞれに対する Average-mAP

使用したモダリティ	全行動		可視行動		不可視行動	
	tight	loose	tight	loose	tight	loose
映像	65.62	77.92	70.87	81.95	36.55	52.29
映像+音声	66.19	77.98	71.61	82.03	**37.28**	**52.59**
映像+音声+実況テキスト (提案手法)	**66.21**	**78.06**	**71.65**	**82.35**	36.72	52.23

表 3: 各アクションクラスにおける全行動, 可視行動, 不可視行動それぞれに対する Average-AP. 上表が tight Average-AP, 下表が loose Average-AP の結果である.

行動クラス	Shots on Target			Clearance			Indirect Free-Kick			Direct Free-Kick			Red Card			YC → RC		
	全	可視	不可視	全	可視	不可視	全	可視	不可視	全	可視	不可視	全	可視	不可視	全	可視	不可視
映像	61.11	61.22	12.74	65.45	78.98	48.93	**55.53**	**63.88**	43.23	73.78	78.46	**14.75**	40.41	65.11	0.00	**28.75**	**34.49**	0.00
映像+音声	60.77	60.87	**17.29**	**66.04**	**80.37**	48.36	55.45	63.55	43.44	73.43	78.45	13.87	38.77	63.92	0.00	24.75	28.71	0.00
映像+音声+実況テキスト	**61.21**	**61.33**	16.00	65.87	79.20	**50.03**	55.26	63.37	**43.97**	**73.79**	**78.71**	14.17	**41.45**	**67.41**	0.00	28.28	34.18	0.00

行動クラス	Shots on Target			Clearance			Indirect Free-Kick			Direct Free-Kick			Red Card			YC → RC		
	全	可視	不可視	全	可視	不可視	全	可視	不可視	全	可視	不可視	全	可視	不可視	全	可視	不可視
映像	68.12	68.25	9.97	81.59	88.21	72.75	**75.76**	**78.49**	72.64	78.23	82.83	**20.45**	44.00	73.18	0.00	48.14	54.75	0.00
映像+音声	68.12	68.25	**15.53**	81.34	**88.25**	71.53	75.13	78.31	71.49	**78.27**	83.33	18.09	**46.13**	72.30	0.00	44.78	49.59	0.00
映像+音声+実況テキスト	**68.14**	**68.26**	10.41	**81.71**	88.13	**73.16**	75.66	78.26	**72.76**	78.07	**83.46**	18.42	45.11	**74.03**	0.00	**50.67**	**57.76**	0.00

計 3 ケースについて学習を行い, 評価した. その結果を表 2, 3 に示す. ここで表において, tight A-mAP を “tight”, loose A-mAP を “loose”として省略している.

まず, (i) 映像モダリティのみ, (ii) 映像モダリティ及び音声モダリティ, (iii) 映像モダリティ, 音声モダリティ, 及び実況モダリティの 3 ケースについて, tight Average-mAP, loose Average-mAP を表 2 に示す.

表 2 から, 本提案モデルは全行動・可視行動における tight A-mAP 及び loose A-mAP に対して最も精度が高くなった. また, 不可視行動における tight A-mAP についても次点の精度となった. このことから, Action Spotting における, 映像情報・音声情報・実況テキスト情報の 3 つのモーダルを利用したマルチモーダル学習の有効性を確認できた.

次に, 各アクションに対する本モデルの有効性を確認する. (i) 映像モダリティのみ, (ii) 映像モダリティ及び音声モダリティ, (iii) 映像モダリティ, 音声モダリティ, 及び実況モダリティの 3 ケースについて, 一部のアクションクラスにおける全行動・可視行動・不可視行動のそれぞれについての Average-AP の結果を表 3 に示す. 全てのアクションクラスにおける Average-AP の結果は http://www.ohsuga.lab.uec.ac.jp/information/Average-AP.pdf に掲載した.

図 3 のうち最も注目すべき点は, Red Card や YC → RC, Clearance, Yellow Card, Indirect Free-Kick, Direct Free-Kick における精度である. まず, Red Card や YC → RC については映像のみのモデルに音声を加えることで精度を落としていたが, 提案モデルではこれらのアクションクラスに対して精度を上げることとなった. 次に Clearance について, 可視行動については精度の差が出ていなかったが, 提案モデルを用いることで, 不可視行動については精度が大きく改善していた. また Yellow Card について, 全体行動や可視行動については映像のみのモデルが最も精度が高かったが, 不可視行動については提案モデルや映像+音声のモデルの精度が高かった. そして Direct Free-Kick について, 可視行動については提案モデルの精度が高かったが, 不可視行動については映像のみのモデルが最も精度が高くなった. 同様に, Shots on Target について, 可視行動については提案モデルの精度が最も高かったが, 不可視行動については映像+音声のモデルが最も精度が高くなった. 対して Indirect Free-Kick について, 全行動や可視行動については映像のみのモデルが最も精度が高かったが, 不可視行動については提案モデルの精度が最も高かった.

5 おわりに

本研究では映像情報・音声情報・コメンテーターの実況内容という実況テキスト情報の 3 つのモーダルに対して, Transformer ベースのアーキテクチャである ASTRA を応用して, マルチモーダルなサッカー行動認識手法を提案した. まず映像情報については, Baidu Soccer Embeddings を PFFN に通して Video Embeddings を得た. 次に音声情報については, SoccerNet-v2 の放送音声からメルスペクトログラムデータセットを作成し, ログメルスペクトログラムに変換した上で VGGish モデルを通し, Audio Embedding を得た. そして実況テキスト情報については, 放送音声から FasterWhisper を用いて文字起こしを行い, 全ての書き起こしテキストを英語に統一した上で, Text Embedding Large 3 を用いて Text Embedding を得た. これ

らの映像・音声・実況テキストの 3 つのモーダルの Embeddings を活用して学習した提案モデルは, 映像及び音声モーダルを活用したモデルと比べて, tight A-mAP が 0.26, loose A-mAP が 0.80 改善した. また, データ数の少ないアクションであるために精度が落ちていた Red Card や YC → RC についても, Red Card の tight A-mAP は映像のみのモデルと比べて 1.04, YC → RC の loose A-mAP は映像のみのモデルと比べて 2.53 改善した. これらの結果は, Action Spotting において, 映像・音声・実況テキストを活用したマルチモーダル学習の有効性を示唆している.

また残された課題として 4 点挙げられる. 第一に, モーダル間の重要度には注目されていないことである. 図 3 で示すように, 各アクションにおけるモーダルの重要度が異なることが考えられるため, 将来の展望として, 各モーダルの重要度を考慮に入れた際の性能評価が挙げられる. 第二に, 音声モーダルや実況テキストを学習に含めた結果, やや精度は上がったものの, 大幅な精度改善には至らなかった点である. これに対し, より高度なクロスモーダルな関係を捉えるモデルの導入が有効であると考えられる. 第三に, Text Embedding が Action Spotting に適した特徴量とはなっていない点である. 本研究では Text Embedding を直接 Transformer Encoder に入力したが, Text Embedding の特徴量をより大きくし, Transformer Encoder の前に別のモデルを通すことで, Action Spotting に適した特徴量を学習させることが可能であると考えられる. 第四に, loose A-mAP の改善に対して tight A-mAP の改善が小さい点である. これは, FasterWhisper で書き起こされたテキストが 5 秒以上継続する場合, その間同じ Embedding を保持するためであると考えた. そこで, より時系列を詳細に捉えた Text Embedding 表現が不可欠である.

提案手法は, 映像・音声のモダリティに対し, 実況テキストを含ませることの有効性を示唆したが, 今後は本モデルの改良を重ねることで, より高精度な Action Spotting が期待できる.

6 謝　　辞

本研究は JSPS 科研費 JP22K12157, JP23K28377, JP24H00714 の助成を受けたものです.

文　　献

[1] Alejandro Cartas, Coloma Ballester, and Gloria Haro. A graph-based method for soccer action spotting using unsupervised player classification. In *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports*, MM '22. ACM, October 2022.

[2] Adrien Deliège, Anthony Cioppa, Silvio Giancola, Meisam J. Seikavandi, Jacob V. Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B. Moeslund, and Marc Van Droogenbroeck. Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4503–4514, 2021.

[3] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition, 2019.

[4] INSIDE FIFA. Fifa publishes professional football report 2023. `https://inside.fifa.com/legal/news/fifa-publishes-professional-football-report-2023`, (Accessed on 04/06/2024).

[5] Yaozong Gan, Ren Togo, Takahiro Ogawa, and Miki Haseyama. Transformer based multimodal scene recognition in soccer videos. In *2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pp. 1–6, 2022.

[6] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780, 2017.

[7] Silvio Giancola and Bernard Ghanem. Temporally-aware feature pooling for action spotting in soccer broadcasts, 2021.

[8] Bo He, Xitong Yang, Zuxuan Wu, Hao Chen, Ser-Nam Lim, and Abhinav Shrivastava. Gta: Global temporal attention for video action understanding, 2021.

[9] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. Cnn architectures for large-scale audio classification, 2017.

[10] Mordor Intelligence. サッカー市場規模とサッカー市場株式分析 - 成長傾向と成長傾向予測 (2024 〜 2029 年). `https://www.mordorintelligence.com/ja/industry-reports/football-market`, (Accessed on 04/06/2024).

[11] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network, 2021.

[12] OpenAI. Gpt-4o, 2024.

[13] OpenAI. text embedding large 3, 2024.

[14] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[15] Muhammad Bilal Shaikh, Douglas Chai, Syed Mohammed Shamsul Islam, and Naveed Akhtar. Maivar: Multimodal audio-image and video action recognizer. In *2022 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pp. 1–5, 2022.

[16] SYSTRAN. Faster whisper. `https://github.com/SYSTRAN/faster-whisper`, 2024. Accessed: 2024-06-10.

[17] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks, 2019.

[18] Bastien Vanderplaetse and Stéphane Dupont. Improved soccer action spotting using both audio and video streams, 2020.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[20] Artur Xarles, Sergio Escalera, Thomas B. Moeslund, and Albert Clapés. Astra: An action spotting transformer for soccer videos, 2024.

[21] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition, 2020.

[22] Xin Zhou, Le Kang, Zhiyu Cheng, Bo He, and Jingyu Xin. Feature combination meets attention: Baidu soccer embeddings and transformer based temporal detection, 2021.