タスク分割とサッカードメイン知識の統合による試合映像からのテキスト速報自動生成

田中智可良 [1,2]　永田亮 [3]　高村大也 [2]　市瀬龍太郎 [1]
[1] 東京科学大学　[2] 産業技術総合研究所　[3] 甲南大学
tanaka.c.e04e@m.isct.ac.jp　nagata-nlp2026 @ ml.hyogo-u.ac.jp.
takamura.hiroya@aist.go.jp　ichise@iee.e.titech.ac.jp

概要

スポーツのテキスト速報は，試合中に発生する重要イベントを時系列に沿って伝える情報サービスであり，放送映像の視聴が困難な状況における戦況把握に有用である．しかし，多人数が複雑に連動するサッカーでは，刻々と変化する戦況やプレーを逐次言語化する人的負荷が高く，提供が一部の注目試合に限られるという課題がある．本研究では，サッカー映像からアクションやセットプレー等のイベント，および選手座標等の試合状態を自動推定し，それらを根拠としてテキスト速報を生成する手法を提案する．さらに，サッカードメイン知識を統合することで，イベント種類・順序などの整合性を改善し，既存手法より事実整合性の高い記述を実現する．

1　はじめに

スポーツメディアにおいて，データ会社等によって配信されているライブテキスト（以下，テキスト速報）は，試合の動向をリアルタイムに言語化して伝える独自のメディア形式である．テキスト速報は，リアルタイムでの試合の放送映像の視聴が困難な状況においても，試合展開を把握できる点で大きな利点を持つほか，試合後の振り返りとしての活用にも適している．一方で，テキスト速報の作成は現在も人手で行われており，高い専門性と多大な工数を要することが課題となっている．そのため，配信対象は注目度の高い一部の試合に留まっており，他の多くの試合では利用することができない．

試合映像から自動的にテキスト速報を生成する代表的な既存手法 [1, 2] では，映像特徴を大規模言語モデル（LLM）へ入力し，テキストを直接生成するエンドツーエンド（E2E）型手法を採用している．しかし，スポーツ映像の認識においては，種目特有

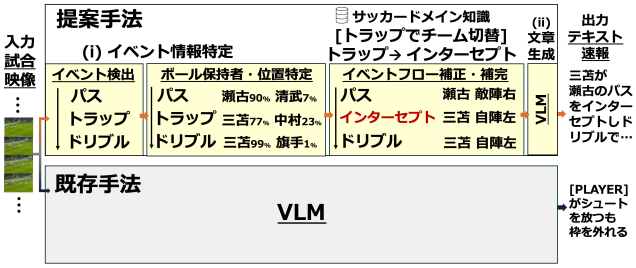

図 1　提案手法（TD-SDKs）と既存手法（E2E 型）の比較

のアクション認識に加え，プレーへの関与選手の同定や位置の把握など複合的な情報を整合的に扱う必要がある．E2E 型手法では，これらの根拠情報が明示的に保持・制御されない場合があり，生成文中で重要箇所の取り違えや事実誤りが生じ得る．とりわけサッカーは，選手数が多く関与選手が頻繁に切り替わるうえ，プレーが連続して進行し区切りが曖昧である．そのため，イベントの言語化において情報の漏れや誤認識（誰が・どこで・何を）が起きやすく，高い事実性を担保したテキスト速報生成は依然として困難である．

そこで本研究では，サッカーの試合映像から，事実性の高いテキスト速報を自動生成することを目的とし，Task Decomposition and Soccer Domain Knowledge-based System（TD-SDKs）を提案する．図 1 に示すように，TD-SDKs はテキスト速報生成を (i) イベント情報特定と (ii) 文章生成に分割し，段階的に実行する．特に (i) イベント情報特定は，(i-a) イベント検出，(i-b) 関与選手の同定，(i-c) 位置情報の推定の 3 つのタスクに分割し，生成文の構成要素となる情報（以下，根拠情報）を個別に抽出する．続いて，抽出されたこれらの根拠情報に対し，サッカー固有のドメイン知識に基づく整合性制約を用いて，イベントの流れにおける矛盾の補正や欠落した情報の補完を行う．(ii) 文章生成では，補正後のイベント情報列と試合映像に基づき視覚言語モデル

（VLM）を用いてテキスト速報の形式で文章化する．このように，イベントを構成する要素を明示的に特定し，ドメイン知識に基づきそれらの整合性を担保してから文章化することで，映像と矛盾する記述を抑制し，事実整合性の高い速報生成を実現する．

2 関連研究

2.1 サッカー映像でのテキスト速報生成

サッカー映像からテキスト速報を自動生成する手法として，映像を構成するフレームから空間的・時間的特徴を抽出し，これを LLM へ入力してテキストを生成する E2E 型手法がある [1, 2]．データセット [1, 2] として，実際の試合映像に対して，人手で作成されたテキスト速報が付与されたものが公開されており，映像とテキストの対応付けに基づく E2E 型モデルの学習に使用されている．Rao らは，サッカー映像からの特徴抽出に特化した視覚エンコーダ MatchVision を提案し，テキスト速報生成とイベント分類等を併用した事前学習により生成性能を向上させた [2]．一方で，この手法は生成過程で参照すべき根拠情報（イベント種別・関与選手・位置など）を明示的な中間表現として保持しないため，生成文における「誰が・どこで・何を」といった要素の取り違えや事実誤りが生じうる．

これに対し本研究は，根拠情報を明示的に推定・保持し，さらにサッカードメイン知識に基づく整合性制約でイベント列を補正・補完してから文章化する点で既存の E2E 型手法 [2] と異なる．このアプローチにより，E2E 型手法で懸念されるハルシネーションを抑制し，事実整合性を担保できる点が既存手法に対する優位性である．

2.2 サッカー映像でのアクション認識

サッカー映像におけるアクション認識は，試合中の出来事（イベント）を映像から同定する動画理解タスクであり，SoccerNet [3] および SoccerNet-v2 [4] を契機に発展してきた．特に，イベントをキックやファウルの瞬間のように 1 点の時刻で検出する Action Spotting が中核として位置づけられている．2023 年からはボールに直接関与するアクションに特化した Ball Action Spotting（BAS）が導入され，パスやシュートなど 12 種類の主要なアクションの検出が求められる．代表的手法の T-DEED [5] は，隣接フレーム間の微小な特徴の違いを強調するように設計されたモデル構成と，推定時刻のずれを損失に含める学習方法により，高い時間分解能でイベント時刻を推定する．本研究では T-DEED を，イベント種別推定の基盤技術として用いる．

2.3 サッカー映像での試合状態推定

サッカーの試合映像から試合状態を再構成する枠組みとして，Game State Reconstruction（GSR）[6] が提案されている．GSR では，映像中の選手を検出して選手領域（BBox）を取得し，フレーム間で同一選手を対応付ける追跡処理によって一貫した選手追跡 ID を付与する．さらに，ピッチラインの推定に基づくカメラキャリブレーションを行い，画像上の選手位置をピッチ平面上の 2 次元座標へ射影することで，試合状態を 2 次元トップビューとして再構成するタスクである．また，各選手領域を入力として，外観特徴に基づきチームや背番号といった選手属性を推定・付与する．本研究では，GSR の枠組みをイベント検出，イベント関与選手の同定および位置情報の推定に適用する．

3 提案手法：TD-SDKs

サッカーの試合映像から事実性の高いテキスト速報を生成する TD-SDKs を提案する．本手法は，試合映像から構造化されたイベント情報を抽出し，それに基づき VLM がテキストを生成する 4 つのステップで構成される（図 2）．

(1) イベント検出 図 2(1) に示す本ステップでは，フレーム数 N の試合映像を入力とし，イベント種別 c_i と発生時刻 t_i からなるイベント列 $E = [(t_i, c_i)]_{i=1}^{M}$ を抽出する（M は検出されたイベント数）．まず，パスやシュート等のアクションイベントについては，T-DEED [5] を用いて各フレーム（$1 \ldots N$）に対し 13 クラスのイベント生起確率を推定する．次に，各クラスの生起確率の時系列を固定長 W の非重複時間窓に分割し，各窓内で生起確率が最大となるフレームが所定の閾値を超える場合に，そのフレームに対応する時刻をイベント発生時刻 t_i として採択する．並行して，GSR [6] により全フレームの選手・ボール座標等の試合状態を取得する．さらに，図 2(1) 下部に示す T-DEED の検出対象外（セットプレー等）のイベントは，GSR から得られるボール・選手位置と，特定領域内におけるボール滞在時間に基づくルールベース判定により補完する．以上により得られた「時刻とクラスのみ」からなるイベント

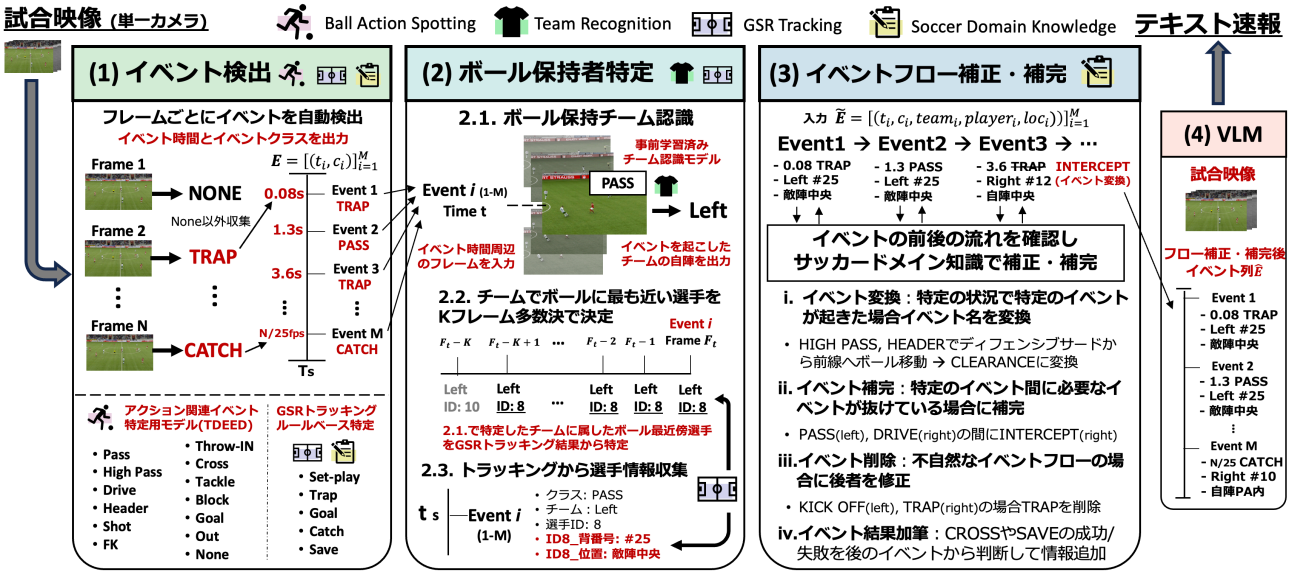

図 2　TD-SDKs の概観図

列 E を次ステップの入力とする．

(2) ボール保持者特定　本ステップでは，(1) で得たイベント列 E と GSR による試合状態を入力とし，図 2(2) に示す通り，各イベントにボール保持チーム，保持選手，および位置情報を付与した構造化イベント列 $\tilde{E} = \{(t_i, c_i, \text{team}_i, \text{player}_i, \text{loc}_i)\}_{i=1}^{M}$ を出力する．まず，BAS データセットに付随するイベントごとの保持チーム（left/right）および時間情報を用いて，ボール保持側のサイドを特定するように学習したチーム認識モデルを用い，イベント時刻 t_i 周辺の映像から保持チーム team_i を推定する（図 2(2.1)）．次に，同チーム内でボールに最も近い選手を保持者 player_i として特定する．図 2(2.2) に示すように，前後 K フレーム（K は奇数）における保持者推定結果の多数決を用いることで，選手遮蔽による誤判定を抑制する．距離計算については，足元のプレーではピッチ上の座標距離を，ヘディング等の空中プレーでは画像上の BBox 間距離を用いることで，浮いているボールのピッチ平面への投影に伴う座標の歪みに対応する．最後に，ピッチ座標を「敵陣中央」等の領域名 loc_i へ写像し，背番号等の属性情報とともに統合する．

(3) イベントフロー補正・補完　本ステップでは，前段で生成された構造化イベント列 $\tilde{E}$ に対し，サッカードメイン知識に基づく遷移整合性制約を適用し，補正・補完を行う．少数のフレーム特徴のみから個別に検出されたイベントには，誤検出や欠落，あるいはサッカーの競技特性上不自然な遷移が含まれ得る．そこで本研究では，イベントフローの整合性を担保するため，以下の 4 つの処理を定義し適用する．(i) イベント変換：特定の状況下で発生したイベントを，文脈に即したより適切なイベント名へと置換する（例：自陣ボックス内でのハイパスをクリアに変換）．(ii) イベント補完：特定のイベント間に介在すべきイベントを補完する（例：相手のパスと味方のドリブルの間にインターセプトを補完）．(iii) イベント削除：サッカーの流れにおいて不整合なイベントを排除する（例：ボールアウト後はセットプレーのみ）．(iv) イベント結果加筆：後続イベントのボール保持チームや位置情報を参照し，アクションの成否情報を付与する．これらの処理を通じて，イベント列 $\hat{E}$ を，最終的な文章生成における根拠情報として用いる．

(4) VLM によるテキスト速報生成　図 2(4) に示す最終ステップでは，(3) で精緻化された構造化イベント列 $\hat{E}$ と，元の試合映像を VLM（Gemini 3.0 Pro Preview）への入力とする．プロンプトでは，$\hat{E}$ に含まれる「誰が・どこで・何をしたか」という根拠情報に基づき，得点につながる可能性が高い重要シーンを優先的に記述するよう指示する．また，ファウル，交代，オフサイドについては，VLM の映像理解能力を活かして補完的に記述するよう誘導することで，時間的文脈の理解を要するイベントも反映した，事実整合性の高いテキスト速報を出力する．

4　実験

目的と設定　TD-SDKs が事実性において既存の E2E 型手法より向上するかを，自動評価により検証

表 1　自動評価によるシステム比較結果（B-1: BLEU-1, R-1: ROUGE-1, BS-F1: BERTScore F1, QE: Qwen3-emb, LM-F1: Label Match F1）

Method	B-1	R-1	BS-F1	QE	LM-F1
MatchVision	0.23	0.23	0.84	0.64	0.57
Gemini 3.0 Pro Prev	0.26	0.31	0.87	0.73	0.78
TD-SDKs w/o 映像	**0.31**	0.358	0.882	0.768	0.67
TD-SDKs	0.29	**0.362**	**0.884**	**0.773**	**0.79**

表 2　3 人の評価者による一対比較結果．勝率はイベントフロー補正・補完ありが多数決で勝った割合を示す．

VLM 入力	#	Fleiss' κ	勝率	95% CI	p 値
イベント列のみ	40	0.207	0.40	0.26–0.55	0.27
イベント列と映像	40	0.194	0.58	0.42–0.71	0.43

する．加えて，イベントフロー補正・補完および映像入力の寄与を，人手によるアブレーション評価により確認する．評価には，2019–20 スイスリーグのスタンド俯瞰映像（テロップなし）から切り出した 30 秒クリップ（25 FPS，164 本）と，GSR データ [6] のフレーム単位トラッキングアノテーションを用いた．対応する既存テキスト速報がないため，サッカー経験者が参照テキスト速報を新規作成した（英語 1–3 文，重要イベント優先，チームは色で言及，背番号は #12 形式，中立表現）．初期検証として，画像認識で得られる情報に基づきテキスト速報が生成可能かを評価するため，根拠情報には GSR のアノテーションを入力に用いた．比較対象は，Gemini 3.0 Pro Preview，MatchVision [2]，および映像を VLM に入力しない TD-SDKs w/o 映像である．自動評価には，BLEU-1 [7]，ROUGE-1 [8]，BERTScore F1 [9]，Qwen3-emb [10]（文章埋め込み類似度），Label Match F1 を用いた．Label Match F1 では参照文および生成文から抽出されたイベント種別集合の一致度を F1 として算出した．一方，人手によるアブレーション評価では，イベントフロー補正・補完の寄与を検証するため，TD-SDKs の補正・補完あり／なしの 2 設定を比較し，VLM 入力を「イベント列のみ」または「イベント列と映像」とする 2 条件で，各条件につき同一クリップに対する 2 出力を提示して 3 名が一対比較した．評価者には映像を参照して，どちらがより妥当（事実性が高い）かを選択する．

結果　表 1 に自動評価結果を示す．TD-SDKs は多くの指標で E2E 型手法を上回り，タスク分割に基づく段階的生成手法が事実性向上に寄与する可能性が示唆される．また，映像情報を統合した設定では Label Match F1 が大きく向上した．この結果は，構

図 3　アブレーションスタディの質的結果例と修正例．

造化データのみでは捉えにくい試合状況の手掛かりを映像入力が補完し得ることを示唆している．表 2 にアブレーション結果を示す．「イベント列と映像」を入力とした条件では，補正・補完ありの手法が勝率 0.58 と高い傾向を示した．さらに，図 3 に示す例では，補正・補完によりイベント列がサッカーとして自然な流れへと整合化され，それに伴い出力文章もより妥当な内容となる様子が観察される．一方で，表 2 の p 値は有意水準（0.05）を下回っておらず，評価者間一致も Fleiss' $\kappa \approx 0.20$ と低い．したがって，本アブレーションの結果は統計的に有意な差としては結論づけられず，傾向として解釈するにとどめる必要がある．今後，イベント検出や選手識別の精度向上により，補正・補完が安定的に機能すれば，効果がより明確に確認できる可能性がある．

5　おわりに

本研究では，サッカー映像から事実性の高いテキスト速報を自動生成するという課題に対し，タスク分割とドメイン知識を統合した手法 TD-SDKs を提案した．GSR アノテーションを根拠情報とした初期検証の結果，自動評価において既存手法を上回り，タスク分割のアプローチが事実整合性の向上に寄与することを示した．一方で，人手評価における補正・補完の効果は統計的な有意差には至らず，傾向の提示にとどまった．今後は，選手・位置情報を実際の認識モデルによる推定結果に置き換えた条件下での性能検証が重要な課題である．

謝辞

この成果の一部は，産総研政策予算プロジェクト「フィジカル領域の生成 AI 基盤モデルに関する研究開発」の結果得られたものである．

参考文献

[1] Hassan Mkhallati, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. SoccerNet-caption: Dense video captioning for soccer broadcasts commentaries. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)**, pp. 5074–5085, 2023.

[2] Jiayuan Rao, Haoning Wu, Hao Jiang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards universal soccer video understanding. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 8384–8394, 2025.

[3] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Marc Van Droogenbroeck. SoccerNet: A scalable dataset for action spotting in soccer videos. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)**, pp. 1824–1834, 2018.

[4] Adrien Deliege, 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 **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 4508–4519, 2021.

[5] Artur Xarles, Sergio Escalera, Thomas B. Moeslund, and Albert Clapés. T-DEED: Temporal-discriminability enhancer encoder-decoder for precise event spotting in sports videos. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)**, pp. 3410–3419, 2024.

[6] Vladimir Somers, Victor Joos, Silvio Giancola, Anthony Cioppa, Seyed Abolfazl Ghasemzadeh, Floriane Magera, Baptiste Standaert, Amir Mohammad Mansourian, Xin Zhou, Shohreh Kasaei, Bernard Ghanem, Alexandre Alahi, Marc Van Droogenbroeck, and Christophe De Vleeschouwer. SoccerNet game state reconstruction: End-to-end athlete tracking and identification on a minimap. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)**, pp. 3293–3305, 2024.

[7] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)**, pp. 311–318, 2002.

[8] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, pp. 74–81. Association for Computational Linguistics, 2004.

[9] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In **Proceedings of the International Conference on Learning Representations (ICLR)**, 2020.

[10] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. **arXiv preprint arXiv:2506.05176**, 2025.

A　イベントフロー補正・補完の詳細

表 3　イベントフロー補正・補完ルール（変換／補完・除去／加筆）

変換前／操作	変換後／対象	条件
(i) 状況に応じたイベント表記の変換		
DRIVE	TRAP	直前が同一チームのボール移動系統イベント
TRAP	INTERCEPT	直前が別チームのボール移動系統イベント
TRAP	CUT	直前が別チームかつ非ボール移動系統イベント
PASS	SHOT	敵 PA 内の PASS の直後に別チームの {GOAL, BLOCK, TACKLE, CATCH, SAVE, SAVE(fail), HEADER, AERIAL DUEL} が続く
HIGHPASS HEADER	CLEARANCE	自陣ディフェンシブサードで実行（ただし直前が GOAL KICK なら除外）
HIGHPASS	SHOT	敵 PA 内で HIGHPASS を実行
HEADER	SAVE	GK が自陣 PA 内で実行
INTERCEPT	CATCH	GK が実行
BLOCK	CATCH	相手 SHOT 直後に GK が実行
CROSS	IN/OUTSWING CROSS	CORNER KICK 後の CROSS に対し，選手位置から IN/OUTSWING を判定
(ii) 欠落補完・誤検出除去		
追加	INTERCEPT	PASS 系統の後に別チームのイベントが続く（チーム切替）
追加	CUT	{TRAP, DRIVE, INTERCEPT, BLOCK, TACKLE} 後に別チームのイベントが続く
追加	PASS	{TRAP, DRIVE, INTERCEPT, BLOCK, TACKLE} 後に同一チームだが別選手のイベントが続く
削除	PASS 系統	直後に同一チーム・同一選手のイベントが続く（重複除去）
追加・置換	SHOT	GOAL 前に SHOT が無い場合，直前の PASS/HIGHPASS を SHOT へ置換または SHOT を追加
追加	OUT	GOAL 直後に OUT が存在しない
削除	OUT	前後イベント間隔 ≤2 秒，または直後にセットプレーが無い
(iii) 特定イベントの結果情報の加筆		
CROSS	追記: *out of the box after the cross*	クロス後に PA 外へ出るイベントがある場合：決定機を逃したニュアンスを付与
SAVE	SAVE(fail)	直後に GOAL が来る場合：反応したが止められなかったニュアンスを付与

略記: ボール移動系統は PASS/HIGHPASS/CROSS/HEADER/SHOT/THROWIN を含む．PA: ペナルティエリア，GK: ゴールキーパー．

表 3 に，本手法で用いるイベントフロー補正・補完ルールの詳細を示す．映像から推定されたイベント列に対し，サッカーのドメイン知識に基づくルールを適用することで，プレーの文脈として不自然な箇所を修正し，高精度な解析を実現する．以下に，各ルールの設計指針を述べる．

状況に応じたイベント表記変換　学習データの不均衡性などに起因し，特定のイベント（DRIVE や PASS 等）が本来とは異なる意味で誤検出されるケースがある．例えば，DRIVE が同一選手によって連続して検出される場合には始めの DRIVE を TRAP へ変換したり，敵陣ペナルティエリア（PA）内での PASS を後のイベント（GK のセーブやゴール等）との関係性から SHOT へ変換したりするなど，プレーの前後関係に基づいた適正化を行っている．

イベント欠落補完と誤検出除去　イベントの検出漏れや重複によって生じる，サッカーとして明らかに矛盾したイベントのシーケンスを修正する．具体的には「GOAL の直前に SHOT が存在しない」「同一選手が連続して PASS を行う」といった，サッカーの競技特性上起こり得ないデータの不整合を抑止する．これにより，パス回しや攻守交代の文脈を正しく保持することが可能となる．

特定イベントの結果情報の加筆　テキスト速報の生成における正確性を向上させるため，特定のイベントに対し結果のニュアンスを付与する．例えば，GK がシュートに触れたもののそのままゴールとなった場合，単なる「SAVE」の検出では「シュートを防いだ」という誤った解釈を招く恐れがある．そこで，直後の GOAL イベントの有無を確認し，表記を「SAVE (fail)」へ変更することで，「セーブを試みたが失点した」という事実関係を正確に反映させている．

B　ボール・選手間距離算出の詳細

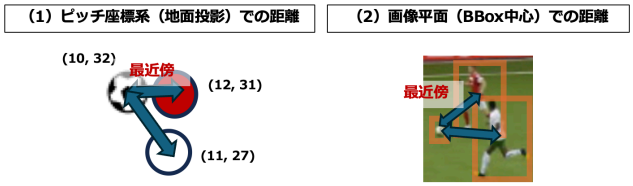

図 4　GSR データにおけるボール―選手距離の算出方法（ピッチ距離／ BBox 距離の切替）

図 4 に，ボール保持者特定に用いる距離算出手法の使い分けを示す．ボールの高さに起因する特定誤りを防ぐため，以下の 2 手法をイベント種別に応じて切り替える．

(1) ピッチ座標系距離　選手・ボール間のピッチ座標の距離を用いる．カメラ画角に依存せず正規化された尺度で計算可能だが，空中にあるボールに対しては投影誤差が大きくなる欠点がある．

(2) 画像平面距離　BBox 中心間の距離を用いる．実装が簡便でボールが空中にある際でも一貫した 2D 観測が可能だが，遠近（スケール変化）によって距離が歪む課題がある．

切替ルールとして，イベント直後にボールが空中にあると想定される特定のイベント（HIGHPASS, SHOT, HEADER, CROSS）の発生直後から次イベントまでは手法 (2) を適用し，それ以外は手法 (1) を用いる．これにより，浮いているボールのピッチ平面への射影により座標が歪む問題に対応する．