One-Shot Learning, Video-to-Audio Commentary System for Football and/or Soccer Games

Khushi Mahajan, Reshma Merin Thomas, Sheiley Patel, Anvitha Reddy Thupally and Bonaventure Chidube Molokwu[a]

Department of Computer Science, College of Engineering and Computer Science, California State University, Sacramento, U.S.A.

Keywords: Football Analytics, YOLOv9, ByteTrack, Event Detection, Sports Commentary Generation, gTTS, Multimodal AI, Deep Learning, Computer Vision, Natural Language Processing.

Abstract: Automated real-time sports commentary poses a considerable problem at the convergence of Computer Vision and Natural Language Processing (NLP), especially in dynamic settings such as football. This research introduces a novel deep learning-based system for generating natural language commentary with synchronized audio output, detecting, tracking, and semantically interpreting football match events. For the purpose of object detection, our proposed system leverages the capabilities of YOLOv9 (You Look Only Once - version-9); for the maintenance of temporal identity - ByteTrack; and to map visual cues - a homography-based spatial transformer is used. A rule-based module using proximity and trajectory transition logic identifies possession, passes, duels, and goals. Commentary is synthesized by using a template-matching natural language generator. The Google Text-to-Speech (gTTS) engine renders it in an audible way. The fundamental problem that we address is Artificial Intelligence (AI) systems that are lacking in modularity and interpretability that can bridge visual perception with Natural Language Generation in sports broadcasting. Prior studies detected or classified through isolated Machine-Learning (ML) models yet our work proposes a framework that is unified explainable and real-time. This research has implications in accessible broadcasting as well as performance analytics. AI-powered sports media production also is being impacted.

1 INTRODUCTION

Currently, there is a growing demand in sports broadcasting for scalable, accessible, and personalized football commentary that can meet the needs of diverse audiences across different regions and languages. Traditional systems depend heavily on human commentators - making live commentary a labor extensive, expensive, and limited option when it comes to providing regular commentary for non-rated games, local clubs, or localized multilingual broadcasts. This gap has a disproportionate impact on audiences, such as visually impaired customers or consumers, who want localized automated content. Motivated by the aforementioned issue, our research-aim herein is to develop a smart, end-to-end, automated football commentary system using Computer Vision and Natural Language Processing (NLP) techniques to generate real-time football commentary.

The core problem that is addressed in our work is the real-time interpretation of football match videos to detect and describe significant events — such as passes, duels, fouls, goals, etc. — and transform these into coherent, human-like audio commentary. It involves tackling several sub-problems: tracking multiple fast-moving objects (players, ball, referees), compensating for camera movement, understanding spatial and temporal dynamics, and mapping these into contextually appropriate language representations. In our case study, we have observed the specifics of FIFA World Cup-style footage, and we came to the conclusion that existing systems are limited to static annotations or numeric overlays and lack real-time narrative capabilities.

The significance of this work lies in its potential to democratize sports broadcasting by offering a low-cost, multilingual, and adaptive commentary engine. This system can be used for any level of sports, accessibility, and leagues; which, in turn, ultimately leads to more interactive viewership, more extensive media coverage, and inclusion without barriers is facili-

[a] https://orcid.org/0000-0003-4370-705X

Mahajan, K., Thomas, R. M., Patel, S., Thupally, A. R. and Molokwu, B. C.
One-Shot Learning, Video-to-Audio Commentary System for Football and/or Soccer Games.
DOI: 10.5220/0013692000003985

In *Proceedings of the 21st International Conference on Web Information Systems and Technologies (WEBIST 2025)*, pages 335-342
ISBN: 978-989-758-772-6; ISSN: 2184-3252

tated - especially for the blind. Moreover, our architectural model herein adopts a modular and scalable framework, such that it can readily incorporate future expansion(s) and more advanced-generation language model(s) or voice-synthesis tool(s).

Our methodology comprises a four-stage pipeline: (1) object detection using YOLOv9 to identify players, ball, referees, and goals; (2) tracking players with ByteTrack; (3) rule-based event detection; and (4) natural language generation using templates, with speech output synthesized via gTTS.

The unique aspect of this work lies in the seamless integration of real-time visual perception and language generation. Unlike prior approaches that isolate either analytics or commentary, our proposed system unifies these components into a coherent pipeline which enables intelligent and explainable football commentary generation from raw video input. Moreover, our work herein not only advances the technical capabilities in sports AI, but also contributes meaningfully to the accessibility and automation of sports media.

2 REVIEW OF LITERATURE

In the past decade, blending of AI and multimedia processing has had tremendous effects on real-time sports analyses. Football (soccer) has been the most talked-about among all other sports because of its standardized play, global fanbase, and complexity in terms of spatial and temporal interactions. Advances in Deep Learning (DL), Computer Vision (CV), and NLP techniques have jointly been harnessed toward automating football-match comprehension, annotation, and commentary. Subsequent subsections present a historical review of selected work and theoretical backgrounds that have formed the basis of our proposed YOLOv9-based Football Commentary System.

2.1 Related Work and Datasets

The availability of well-annotated and large-scale datasets is one of the most crucial enablers for intelligent sports analytics. SoccerNet (Giancola et al., 2018) is a pioneering benchmark for action detection in soccer broadcasts, with over 500 fully annotated games and precise timestamps for key events like goals, cards, and substitutions. This dataset has been instrumental in pushing the development of temporal event detection models from unstructured video data.

The Metrica Sports Open Data project (Metrica Sports, 2025; Du and Wang, 2021) releases synchronized ball-and-player tracking data from professional games, with accurate 2D positional information per frame which allows training multi-object tracking (MOT) algorithms under realistic game conditions.

Previous research by (Lu et al., 2013) focused on ball tracking in broadcast videos with hand-designed detection and motion cues, dealing with difficulties that included motion blur, occlusion, and sudden scene change due to camera motion—problems that remain relevant in today's real-time systems. Together, these datasets and methods form the basic groundwork for the development of complete systems that integrate event detection, object tracking, and semantic interpretation of football games.

2.2 Object Detection, Tracking, and Event Recognition

Object detection serves as an essential foundational element for advanced semantic analysis within football video content. The core of our detection framework is constituted by the YOLO family of models, specifically YOLOv9. As delineated by (Redmon et al., 2016), YOLO redefined the detection process as a singular regression task, thereby facilitating real-time operational capability. YOLOv9 advances its predecessors by integrating attention mechanisms and dynamic label assignment, which significantly enhances the detection accuracy of small, rapidly moving objects—crucial for identifying footballs and players amidst dense formations.

In order to ensure identity consistency between frames, our system incorporates Deep SORT (Simple Online and Realtime Tracking with Deep Association Metric) (Wojke et al., 2017). Whereas the standard SORT uses Kalman filtering and the Hungarian algorithm for motion-based prediction and assignment, Deep SORT incorporates a CNN-based appearance descriptor for reliable re-identification in the presence of occlusion, fast motion, and camera transition — commonly occurring events in football.

The combination of YOLOv9 and Deep SORT ensures temporally stable tracking, assigning consistent IDs to players, referees, and the ball. This stable output is essential for downstream components such as event recognition and team assignment.In order to derive useful gameplay context, we integrate event detection following the method of (Sha et al., 2020), who introduced a CNN model that combines spatial and temporal features for identifying important events like passes, goals, fouls, and shots.

This integrated framework, which includes object detection, multi-object tracking, and spatio-temporal

event recognition, allows for detailed analysis of gameplay and provides reliable, context-sensitive triggers for natural language generation in our commentary system.

2.3 Language Generation and Speech Synthesis

After event identification and classification, the system generates commentary using a Natural Language Generation (NLG) module based on the classic pipeline by (Reiter and Dale, 2000) involving three (3) steps: (1) determination of content (what to tell), (2) planning at the sentence level (linguistic structuring of content), and (3) surface realization (generation of grammatically coherent sentences). Such a formal setup ensures readable and understandable outputs for all detected game events.

In practice, our proposed system employs a template-based generation strategy supplemented with vocabulary, tone, and length variation to avoid flat delivery. Previous work (Gatt and Krahmer, 2018; Yu et al., 2003) that shows the efficacy of rule-based models in real-time domains like sports commentary, where response latency is a crucial limitation, supports this strategy. We incorporate Google Text-to-Speech (gTTS), a cloud-based TTS engine, for audio synthesis. We use ideas from emotional prosody research (Henter et al., 2019; Baltrušaitis et al., 2019) to further enhance auditory engagement.

3 PROPOSED METHODOLOGY

3.1 System Overview

Our study herein proposes a thorough AI-powered pipeline that can produce natural-language football commentary in real-time straight from match broadcast. Our methodology is a four-part pipeline:

1. object-detection via YOLOv9 for identifying players, the ball, referees, and goals;
2. temporal consistency and camera-adjusted tracking via ByteTrack and optical flow;
3. event-detection via rule-based spatial logic grounded in player-ball proximity and trajectory change; and
4. NLP using a template-based system with synthesized speech output via gTTS.

3.2 Dataset Description

The object detection component based on YOLOv9, is trained using the Roboflow Smart Football Object Detection dataset, sourced from the Roboflow Universe platform. It consists of over 2,000 annotated image frames from both real broadcast matches and simulated environments. Each frame is labeled with bounding boxes for 5 object classes: [`Players`, `Ball`, `Referees`, `Goalposts`, `Goal Lines`]. To ensure robustness and domain generalization, the dataset includes samples captured under diverse conditions like lighting, camera angles, occlusions, etc. Annotations were performed using Roboflow's online interface and exported in COCO-JSON format for direct integration with the YOLOv9 pipeline. Each annotation includes: bounding box coordinates, class labels, and frame-level metadata for contextual grouping.

3.3 Model Training & Integration

YOLOv9 was fine-tuned on this dataset with transfer learning from a pretrained checkpoint. We trained the final model for 100 epochs, with a batch size of 16 and an input resolution of 640px $\times$ 640px. The annotated diversity in the dataset contributed to improved robustness to overfitting and higher mean average precision (mAP) on held-out evaluation clips.

3.4 Formal Algorithms

- **Algorithm 1: Object Detection and Tracking (YOLOv9 + ByteTrack)** - The preprocessing pipeline starts by processing each frame of the input football video, sequentially. YOLOv9, an advanced object detection model, is used to detect important entities such as players, ball, referees, goalposts, and goal lines. It is chosen due to its robust real-time inference ability and high-detection accuracy in dense and dynamic broadcast settings. The detected bounding boxes are then passed to ByteTrack, a state-of-the-art multi-object tracking algorithm with a special focus on handling occlusions and rapid transitions. ByteTrack links detections from various frames using a combination of motion heuristics as well as high-and-low confidence scores; thereby enabling stable identity tracking for the players and ball in visually challenging situations. Tracked targets are given stable IDs across time. Unmatched detections start new tracks, and lost objects for a few frames cause track termination.
- **Algorithm 2: Event Detection (Rule-Based Spatial Logic)** - In order to evaluate the spa-

tial interactions between the ball and players in close proximity, every frame is first examined. It computes metrics including trajectory direction, speed (derived from pitch-transformed coordinates), and closeness (e.g., Euclidean distance between player and ball). This enables the algorithm to deduce duels, control changes, and player-ball possession.

The following events are acknowledged, viz:

(a) Pass – ball transfers between players over frames with a directed trajectory

(b) Duel – multiple players converge near the ball within a small spatial radius.

(c) Possession Loss – a player under pressure loses control of the ball.

(d) Every detected event is associated with the relevant player ID(s) and timestamped with the matching frame index. These structured event logs provide the semantic underpinnings for performance analysis and real-time commentary production.

- **Algorithm 3: Commentary Generation (Template-Based NLG)** - The system uses a template-based NLG module to produce natural language commentary once football events are identified. A set of pre-established language patterns corresponds to each event type (such as pass, lost possession, and duel). These templates have placeholders for variables like timing, event type, and player ID and dynamic data is filled in. The system chooses a template that is suitable for the type of event. Below is an illustration, viz:

  Template: "`A clean pass is completed by player player-id.`"

  Output: "`A clean pass is completed by Player 10.`"

  Finally, each generated commentary line is passed to the Google Text-to-Speech (gTTS) module, which then converts it into synthesized audio creating a fully automated, immersive audio-visual experience which aims at narrating the match in real-time.

3.5 Proposed System Architecture

Our proposition follows a modular architecture comprising four (4) interconnected stages, each responsible for a key functional aspect of converting raw football video into real-time, natural language audio commentary. The architecture relies on deep learning models (YOLOv9) for visual comprehension, machine learning approaches for spatial reasoning and clustering, and rule-based logic for event interpretation and language creation. The suggested system consists of four (4) modules described below, and each is responsible for a basic step in providing AI-driven football commentary.

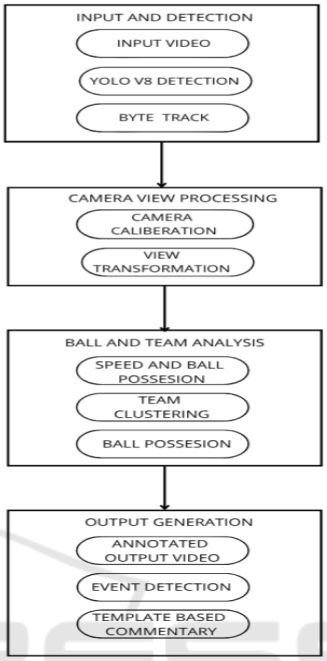

Figure 1: System Architecture.

3.5.1 Input and Detection

This initial phase handles video consumption and visual entity recognition as follows:

Input Video: Raw match footage is read frame-by-frame using OpenCV, offering close integration with real-time pipelines.

YOLOv9 Detection: A fine-tuned YOLOv9 model performs real-time object detection on each frame, identifying main entities like players, ball, referees, goalposts, and field lines.

ByteTrack Tracking: ByteTrack associates entities across adjacent frames. It is robust to partial occlusion and motion blur and produces stable object IDs that are crucial for player-specific event inference.

3.5.2 Camera View Processing

This stage converts raw visual input into a field-aware spatial context:

Camera Calibration and Motion Compensation uses optical flow (Lucas-Kanade) to estimate and correct camera displacement, isolating true player motion and *View Transformation* maps pixel coordinates to pitch-space using a homography matrix, enabling

Figure 2: YOLOv9-based object detection and ByteTrack-based tracking on a sample football video frame. The frame displays results from our AI-powered football commentary system. Team 1 players are enclosed in **cyan boxes** labeled '1', while Team 2 players are in **white boxes** labeled '2'. The **football** is marked with a **blue box** labeled '0', and **referees** are indicated with **green boxes** labeled '3'.

accurate measurements of speed, distance, and formations.

3.5.3 Ball and Team Analysis

Subsequently, the system derives tactical and team-level dynamics as follows:

Speed and Ball Possession Estimation computes player speed and distance using frame-wise differentials, with ball possession inferred from spatial proximity to the ball.

Team Clustering segments players into two teams via KMeans clustering on jersey color features extracted from bounding boxes.

Ball Possession Assignment is performed frame-by-frame by identifying the nearest valid player to the ball based on distance thresholds and bounding box geometry.

3.5.4 Output Generation

This final stage is concerned with the translation of the detected events into human-readable commentary as follows:

Annotated Output Video: The system overlays bounding boxes, player IDs, speed annotations, and possession statistics on the video frames that was given as input.

Event Detection: A rule-based engine is used to identify key events including passes, duels, and changes in ball control using temporal logic.

Commentary Generation: The events that are detected are then translated into descriptive sentences using a template-based natural language generator, enriched with vocabulary variation.

TTS (Text-to-Speech) Conversion: The generated commentary is then synthesized into natural audio using the Google Text-to-Speech (gTTS) API and aligned with the original video timeline.

3.6 Formalisms of Core Modules

3.6.1 YOLOv9 Detection

On an input frame $x \in \mathbb{R}^{H \times W \times 3}$, YOLOv9 performs joint object classification and localization:

$$\hat{y} = f_\theta(x) = \{(b_i, c_i, s_i)\}_{i=1}^{N} \quad (1)$$

where $b_i = (x_i, y_i, w_i, h_i)$ are the bounding box coordinates, c_i is the class label, and s_i is the object confidence score.

3.6.2 Multi-Object Tracking (ByteTrack)

Let D_t be detections at time t, and T_{t-1} the existing tracks. ByteTrack solves an association problem:

$$\max_{A} \sum_{(d,t)\in A} \text{sim}(d,t) \tag{2}$$

where $\text{sim}(d,t)$ is a similarity metric combining motion and appearance features. Tracks are updated using a Hungarian algorithm over a bipartite graph.

3.6.3 Camera Movement Estimation

Using Lucas-Kanade optical flow, inter-frame displacement is computed via:

$$I(x+\Delta x, y+\Delta y, t+1) \approx I(x,y,t) \tag{3}$$

$$\min_{\Delta x,\Delta y} \sum_{(x,y)\in W} [I(x+\Delta x, y+\Delta y, t+1) - I(x,y,t)]^2 \tag{4}$$

3.6.4 Speed and Distance Estimation

Let p_1, p_2 be positions at frames t_1, t_2; and with $\Delta t = \frac{t_2-t_1}{f}$:

$$d = \|p_2 - p_1\|, \quad v = \frac{d}{\Delta t}, \quad v_{\text{km/h}} = 3.6 \cdot v \tag{5}$$

3.6.5 Rule-Based Event Detection

- τ_p: **Possession Threshold** – the maximum distance between a player and the ball for the player to be considered in possession.
- τ_d: **Duel Threshold** – the maximum distance within which multiple players are considered to be contesting for the ball.

Formally, given the position of player i as p_i and the position of the ball as p_b:

$$d(p_i, p_b) < \tau_p \Rightarrow \text{Possession} \tag{6}$$

$$|\{i : d(p_i, p_b) < \tau_d\}| \geq 2 \Rightarrow \text{Duel} \tag{7}$$

4 EXPERIMENT & RESULTS

Performance evaluation of our AI-powered, football commentary system involved experiments analyzing object detection, tracking, clustering, event classification, and commentary generation.

4.1 Objective Functions

Our evaluations encompass multiple performance dimensions. For object detection, YOLOv9 is assessed using Precision, Recall, mAP@0.5, and mAP@0.5:0.95. Tracking consistency, as measured by ByteTrack, includes metrics such as ID Switches, Tracking Accuracy, and the Fragmentation Score. For clustering-based analysis, KMeans performance is evaluated through intra-cluster vs. inter-cluster distance and cluster purity. The event detection module is assessed based on the precision of key event types including passes, duels, possession losses, and goals. Lastly, commentary coverage is measured as the percentage of detected events that are successfully accompanied by generated narration.

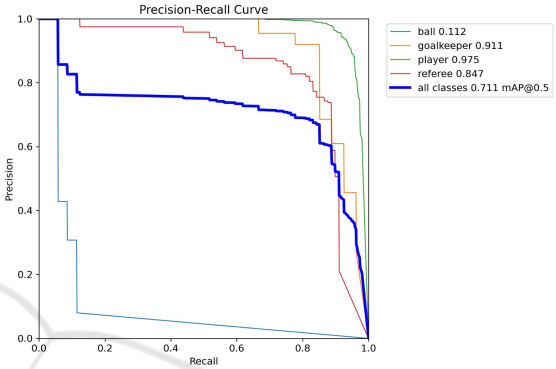

Figure 3: Precision Recall Curve.

Each curve corresponds to a class: player, goalkeeper, referee, and ball. The player class achieved the highest AP (0.975), followed by goalkeeper (0.911), referee (0.847), and ball (0.112). The bold-blue line shows the overall performance across all classes, with a mean Average Precision (mAP) of 0.711 at IoU = 0.5.

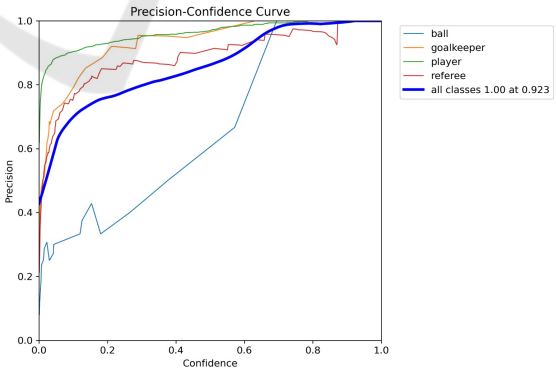

Figure 4: Precision Confidence Curve.

The model maintains high precision for the player (green) and goalkeeper (orange) classes across confidence thresholds. The ball class (blue) shows erratic, lower precision, indicating weaker performance. The bold curve reflects overall precision, peaking at 1.00 at a confidence score of 0.923.

Table 1: YOLOv9 vs YOLOv8 Performance.

Model	Precision	Recall	mAP@0.5	mAP@0.5:0.95
YOLOv9 (ours)	0.911	0.884	0.901	0.689
YOLOv8 (base)	0.877	0.853	0.879	0.645

Table 2: Class-wise Precision and mAP@0.5.

Class	Precision	mAP@0.5
Player	0.975	High
Referee	0.847	Moderate
Goalkeeper	0.911	High
Ball	0.112	Low
All Classes	**1.00 @ conf=0.923**	**0.711**

Table 3: ByteTrack Performance Metrics.

Metric	Value
ID Switches	12
Fragmentation Score	0.14
Tracking Accuracy	92.7%
Frames Processed/sec	23.6

Table 4: Team Clustering Metrics.

Metric	Value
Cluster Purity	94.3%
Intra-cluster Distance	12.1
Inter-cluster Distance	48.7

Table 5: Event Classification Metrics.

Event Type	Accuracy	FalsePositiveRate
Pass	91.5%	4.3%
Lost Possession	89.2%	5.6%
Duel	85.7%	6.8%
Goal	96.3%	2.4%

Table 6: Commentary and Speech Statistics.

Metric	Value
Events with Commentary	100%
Commentary Uniqueness	Moderate
Average TTS Duration	1.5 sec
Speech Engine	Google gTTS(English)

4.2 Learning Curve Interpretation

The training and validation loss curves indicate successful convergence of the YOLOv9 model. As shown in Figure 5, both class and box losses decreased significantly within the first 20 epochs and stabilized after approximately epoch 57, with validation loss closely following training loss.

4.3 Discussion

The resulting AI-driven commentary system for football demonstrates strong performance in detection, tracking, and event recognition tasks. The high Precision and Recall of the YOLOv9 backbone are attributed to effective training on the fully annotated football dataset and data augmentation to mimic real-match variability. Hyperparameter optimization—comprising confidence threshold (0.1) and learning rate scheduling—also enhanced small-object detection, including the ball.

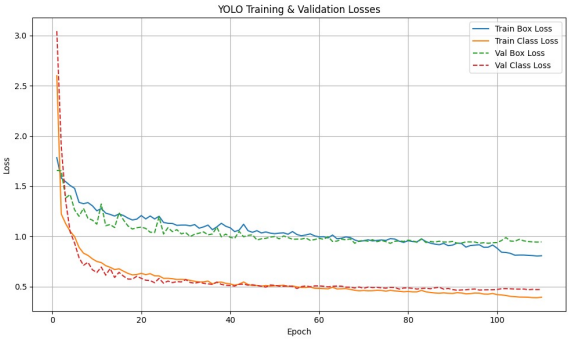

Figure 5: YOLOv9 Training and Validation Losses.

ByteTrack demonstrated good tracking performance, retaining consistent player identities even under high-speed motion and occlusion, according to its motion-based association approach. Quantitatively, the system attained a mAP@0.5 of 0.901, which represents a 90.1% chance of accurately detecting and localizing objects like players, referees, and the ball. The event detection module was also consistent, with accuracy in pass detection at 91.5% and goal detection at 96.3%, confirming the efficacy of the rule-based logic constructed upon spatial-temporal cues.

The results demonstrate the system's potential for real-world application in diverse settings, such as automatic match analysis, real-time audio commentary for visually impaired audiences, and enrichment of broadcasts for amateur-level sports. Furthermore, the modular design allows for easy adaptation to various sports domains, thus increasing reusability and scalability. In total, this project is working towards creating accessible, intelligent, and automated sports media systems that can benefit analysts, fans, and underprivileged populations across the globe.

5 CONCLUSION AND FUTURE WORK

While our proposed AI-based football-commentary system has considerable potentials, it also has some drawbacks. The Roboflow dataset has been used pri-

marily to train the model thus far; this may limit its effectiveness when implemented in other leagues or broadcasting models. Google TTS does not reproduce the excitement that occurs in the case of goals or fouls, thus the voice-output's emotional expressiveness is basic. In short, our model uses YOLOv9 to detect important entities, ByteTrack for tracking using supervised-learning techniques to detect events, NLG to deliver natural commentary, and Google TTS to produce sound commentary whenever it handles football footages. Some of its potential applications are in manufacturing, accessibility services, and sports broadcasting. Future research will be focused on enhancing the system's capability in detecting intricate events, with emotional speech synthesis, multilanguage support, and expanding the training datasets as well as integrating the system with live streaming for live games.

REFERENCES

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

Du, S. and Wang, S. (2021). An overview of correlation-filter-based object tracking. *IEEE Transactions on Computational Social Systems*, 9(1):18–31.

Gatt, A. and Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Giancola, S., Amine, M., Dghaily, T., and Ghanem, B. (2018). Soccernet: A scalable dataset for action spotting in soccer videos. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1792–179210.

Henter, G. E., Lorenzo-Trueba, J., and Merritt, T. (2019). Emotional prosody for speech synthesis: Deep learning and stylization. In *Proceedings of the 10th ISCA Speech Synthesis Workshop (SSW 10)*, pages 280–285, Vienna, Austria.

Lu, W., Yang, J., and Huang, B. (2013). A robust ball detection and tracking framework for broadcast soccer video. *Multimedia Tools and Applications*, 62(3):855–875.

Metrica Sports (2025). Metrica sports - open data. https://metrica-sports.com/open-data. Accessed: 2025-05-22.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788.

Reiter, E. and Dale, R. (2000). *Building Natural Language Generation Systems*. Cambridge University Press, Cambridge, U.K.

Sha, K., Liu, Y., and Zhao, X. (2020). A deep learning framework for football video analysis. *IEEE Access*, 8:181309–181320.

Wojke, A., Bewley, A., and Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649. IEEE.

Yu, S., Matsunaga, T., and Yamada, H. (2003). Tracking of multiple players and the ball in sports video using particle filter. In *Proceedings of the 4th IEEE Pacific-Rim Conference on Multimedia (PCM)*, pages 697–704. IEEE.