Indexing and Summarization of Sports Videos Using Multi-modal Approach

KrupaShree M V
Dept. of Computer Science and Engineering
PES University
Bangalore, India
krupashreemv@gmail.com

Meenal Bagare
Dept. of Computer Science and Engineering
PES University
Bangalore, India
bagare.meenal@gmail.com

Melvin Jojee Joseph
Dept. of Computer Science and Engineering
PES University
Bangalore, India
melvinjjoseph2002@gmail.com

Naveen Kumar Reddy G
Dept. of Computer Science and Engineering
PES University
Bangalore, India
naveenreddy8825@gmail.com

Sandesh B J
Dept. of Computer Science and Engineering
PES University
Bangalore, India
sandesh_bj@pes.edu

***Abstract*—Traditional methods of sports summarization heavily depend on large teams performing manual editing, where enormous portions of game footage are picked over through by humans to select the most crucial moments for the compilation of highlights. This is, however, a time-consuming, resource-intensive process and leads to uneven-coverage of events. In this paper, we present a novel multi-modal approach to sports summarization, in which we combine multiple modalities: Twitter data, audio features, and video content to automate and enhance the entire process of sports summarization. This approach is based on the integration of diverse modalities, which should streamline the summarization process and hence enhance efficiency in covering sports events. This methodological novelty has the potential to transform sports summarization into a scalable and efficient solution, delivering engaging and informative highlights to sports enthusiasts from all over the globe.**

***Index Terms*—Highlight generation, LLMs, multimodal analysis, sports analytics, deep learning, text analysis, audio analysis, video analysis**

I. Introduction

Sports Video summarization is the condensing of lengthy sports videos into shorter summaries that capture the important events of a game. In an age where people do not have the time to sit for hours together to watch their favorite game, this can help viewers efficiently enjoy the entire game in very little time.

Traditional approaches to summarizing sports videos rely heavily on manual efforts by video editors who sift hundreds of hours of video footage from multiple camera angles and positions in order to identify the significant events and then manually edit this into a highlights video. This process is not only expensive but also time consuming. It also lacks an element of audience engagement. However, this approach can be made better by leveraging the advancements in deep learning, computer vision and natural language processing. Furthermore, the audience reactions can be captured from various social media platforms to help create a better experience for the sports viewers.

This research paper aims to design and build a system that can efficiently summarize sports videos by using the advancements in deep learning, computer vision and natural language processing. Further, this project also aims to capture user sentiments towards significant events in the sports videos by analyzing the tweets related to the game, thereby enabling us to effectively capture audience reactions and insights to create more engaging highlights for the viewer.

The core of this paper involves a multi-modal approach that seamlessly integrates multiple modals of data - visual, audio and textual in addition to the twitter data of that game. This ensures that our system is accurate and effective in generating summaries of sports videos and also adds audience engagement and insights into our video summarization process.

Audio Analysis is crucial for our system in order to help identify events based on valuable information obtained from the audio track such as crowd intensity, commentator fervor, and player reactions, adding the summaries with a deeper layer of context and emotion.

Visual Analysis plays a key role in ensuring that no event goes unnoticed. Our system will analyze the visual element which is scoreboard to identify pivotal moments and highlights within the videos.

Textual Analysis is performed both on the commentary data of the game as well as on the twitter feed provided. This will ensure that there is a human element in summaries generated. The commentary is analyzed using a large language model to classify the event as significant or not significant.

By incorporating various modalities, our aim was to create a superior system than the traditional systems being used to generate highlights. Additionally by adding audience reactions and insights from twitter we aim to generate a more enriching summary of the sports video.

II. Literature Survey

The literature survey carried out for Sports Video Summarization and Event Detection demonstrates a variety of approaches leveraging audio, visual, and textual modalities.

Paper [1] proposes a multi-modal architecture for cricket highlight generation using YOLO for segmenting video based on the bowler's position and a fine-tuned BERT model to classify events from commentary. Their approach achieves a very high accuracy (97%) in classifying events due to a comprehensive list of event-specific word corpus. The temporal misalignment between commentary and video introduces errors in classification.

In paper [2], a weighted dynamic heartbeat graph is used for extracting events from a stream of tweets. A temporal graph is created from the content of the tweet, and a rule-based classifier selects event candidates using graph properties. This method has been validated on benchmarks such as the FA Cup, US Election and Super Tuesday, the results showcases its reliability in detecting events.

The paper [3], develops automatic detects and summarization of important events in cricket match.This model uses techniques like optical character recognition(OCR),sound detection, and replay detection to extract important events such as boundaries and wicket.This method combine both CNN and OCR for event detection and it gives higher accuracy compared to other techniques.The system splits up video into separate shots,detects key frames using audio cues,recognizes scoreboard information using OCR, and generates highlights based on detected events.

This paper [4] addresses a simple, near real-time performance system for events detection in soccer game retransmissions and video summaries generation. Two acoustic features, namely, block energy and the acoustic repetition index, have been used to identify important events of the game, such as goals. The objective is to enhance event detection and summary generation in the context of sports broadcasts.

While these approaches offer valuable insights, our work aims to address the gaps in multimodal integration for football video summarization. Unlike [1] and [3], which is specific to cricket, we aim to incorporate techniques specific to football, such as integrating commentary, scoreboard detection, and Twitter data to enhance summaries. In order to increase detection reliability, we will also get around the drawbacks in [4] by integrating both visual and audio modalities.

III. Proposed Methodology

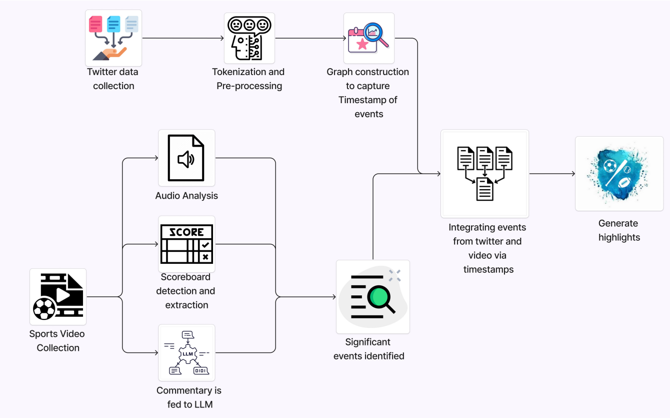
Twitter data collection
Tokenization and Pre-processing
Graph construction to capture Timestamp of events
Audio Analysis
SCORE
Scoreboard detection and extraction
LLM
Commentary is fed to LLM
Sports Video Collection
Significant events identified
Integrating events from twitter and video via timestamps
Generate highlights

Fig. 1. Proposed architecture for the Multi-modal Sports Summarization.

A. Twitter Analysis

The rapid growth of social media platforms, especially Twitter, has opened up access to a large amount of real-time information during important events. Unfortunately, this information is usually noisy, diverse, and dynamic, making it very difficult to accurately detect real and significant events. Oftentimes, such traditional methods do not scale or adapt well in environments with rapidly changing trends. The main objective is to develop a robust system to support very high data spikes while also identifying the important events and adapting to the change in trends, accuracy being mandatory while redundancy is minimum.

A hybrid approach that uses graph-based event detection and goal keyword analysis is what has been put in place to ensure this.

1) Graph-Based Event Detection: Cleaning the tweets consists of removing unnecessary elements such as URLs, mentions, and hashtags, and filtering for the English tweets only. Further, preprocessing is text normalization, punctuation removal, tokenization, and stop word removal, which hence yield relevant text. Process the tweets into time buckets to capture the temporal structures that would enable an aggregation analysis over the preferred time intervals.

The construction of a co-occurrence graph for each time bucket is intended to explore the relationships between words [2]. The centrality measures for the graph nodes are used to calculate the significance of words [2]. Two major components are

- **Growth Factor (GF):** The change in the significance of word relationships over time is calculated as:

$$GF_t = \text{Weight}_{t+1} - \text{Weight}_t \tag{1}$$

- **Aggregated Centrality (AC):** Calculated based on degree centrality, measuring relative node importance:

$$C(v) = \frac{\deg(v)}{n-1} \tag{2}$$

$$AC = \sum_{v \in G} C(v) \tag{3}$$

These two measures—growth factor and aggregated centrality—are derived from the methodology proposed in [2] and are combined to calculate the heartbeat score (HS), which identifies significant events:

$$HS_t = GF_t \times AC_{t+1} \tag{4}$$

2) Goal Keyword Analysis: This includes the analysis of tweets according to the keyword ”goal” to listen for joint references to goals that occurred during a match. The collected tweets are grouped within five-minute ranges and are based on the counts of the keyword goal for each five-minute interval, allowing us to analyze time-dependent frequencies of mentions of the goal.

Timestamp with the highest number of hits associated with the keyword goal is identified and ranked to extract top hits from significant occurrences. Unlike many works that examine goal-related keywords [2], this analysis is entirely focused on the one keyword goal for clear and effective analysis of goal indications in tweets.

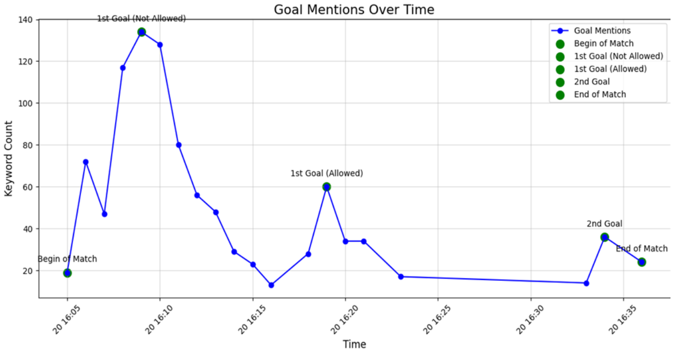

Fig. 2. Tweet Traffic During the FIFA World Cup 2022 Match Between Qatar and Ecuador.

3) Combined Methodology: Timestamps derived from heartbeat score (HS) [2] obtained from graphs along with a relevant keyword produce better performance in event detection. Generalized events then use the analysis of the structures and dynamics of the tweet stream, while currently specific events such as ”goal” engage keyword analysis. Merging these two has the most accurate precision and recall for high-performance adaptability to data trend changes for effective critical moment capturing.

B. Scoreboard detection and extraction

In sports video broadcasts, a superimposed scoreboard typically displays game details, such as the current score, game time and team names to improve the viewer’s game progression. As the scoreboard updates live for each major event (for example, goals; or other timekeeping, points scored), the detection and recognition of the scoreboard plays a key role in automated sports video analysis, serving as a primary source of information for detecting game events and extracting relevant statistics.

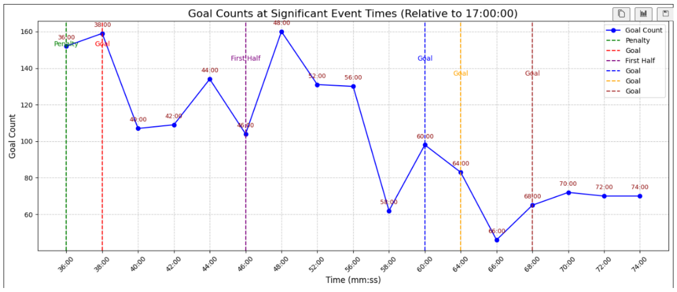

Fig. 3. Significant timestamps from twitter modal for France vs Croatia match

Our approach to this involves several principal stages, the first stage being frame extraction. It involves segmenting the video into individual frames. This segmentation allows for precise localization and analysis of the scoreboard across different frames.

Following frame extraction, in order to have a dynamic scoreboard detection model we employed YOLOv4, a deep learning object detection framework. This model has been trained to recognize the scoreboard dynamically within each frame thus adapting to diverse scoreboard placements. This detection model ensures flexibility and robust performance across various broadcast layouts without depending on fixed coordinates, making it highly suitable for general sports video analysis.

After locating the scoreboard, we made use of Optical Character Recognition (OCR) to extract the displayed scores, the game time and the team names. The OCR process focuses on the stable regions of the detected scoreboard, ensuring that extracted text remains consistent, capturing any updates in score or team information accurately across the video. As the scoreboard position remains fixed, the OCR can reliably interpret the changing details without interference from other screen elements. Then, a mechanism was implemented to capture the team names, score and game time from the scoreboard only post a significant event (e.g., goals or wickets) giving us accurate time stamp for the event.

It was observed that OCR extraction achieved 89.06% accuracy when processing videos with a resolution of 1280×720 pixels. For videos with resolutions below 854×480 pixels, the accuracy was negatively impacted in few of the frames.

C. Audio analysis

In this paper, this approach is a multi-modal framework that aids in the detection and annotation of key events in sports

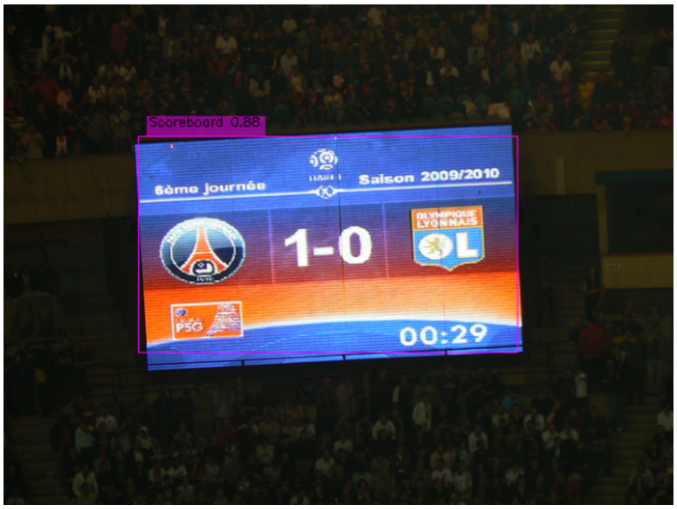

Fig. 4. Scoreboard Detection using YOLOv4 - 1

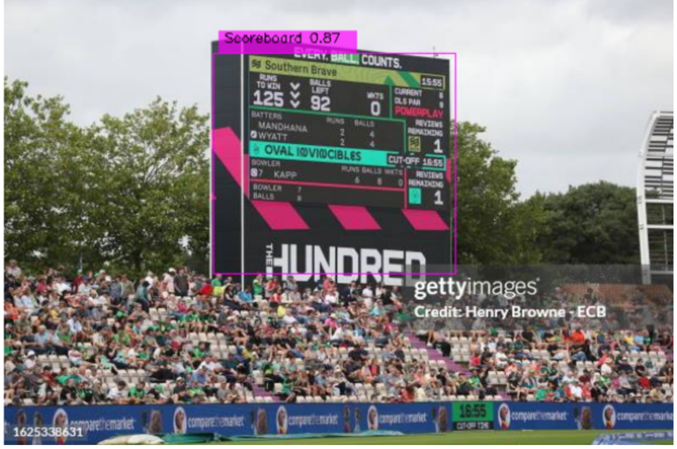

Fig. 5. Scoreboard Detection using YOLOv4 - 2

videos through audio analysis. The process can be broken down into four main steps: audio extraction, audio transcription, audio peak detection, and peak annotation with context-specific keywords. First, a video file undergoes a process wherein the audio is first extracted in WAV format and then used as an input for audio analysis and transcription.

After audio extraction, it is transcribed using an effective ASR model that enables English transcription. The outcome of the transcription is a full textual representation of the audio commentary describing the events as they appear in the video. It is further processed to include pre-defined sports-related keywords such as ”goal,” ”foul,” and ”penalty,” which are indicators of key events.

The audio signal contains peaks of interest that are analyzed through the help of both RMS energy and spectral flux as indicators of sudden change in intensity that are often represented with an event. These peaks are calculated within the audio signal by computing segments of higher energy or flux than average using thresholding to keep the highest peaks. This enables the system to identify the essential moments correlated with events of large intensity.

In order to annotate, peak times are aligned with transcriptions corresponding to relevant keywords; therefore, there will be contextual insights on keywords. A fuzzy matching algorithm compares words in the transcript around each peak with a set of predefined sports-related keywords, applying a high match threshold to actually capture key event mentions; thus this step allows for precise association between audio peaks and event keywords.

Finally, with the aim of indexing annotated events over their timestamps, it forms another index which further describes what happens in the video for the occurrences of each event. The found events are then saved in a structured format for easy retrieval and summarizing so that users can access important moments in quick time, review them as well, and generate summaries easily. This multi-modal method uses both audio signal analysis and natural language processing to realize efficient yet accurate sports video summarization.

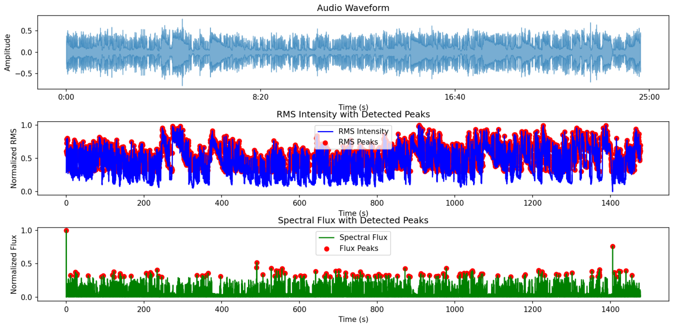

Fig. 6. Audio waves RMS intensity and Spectral flux

D. *Event Classification using commentary*

In the proposed multi-modal summarization approach, we fine tune a BERT (Bidirectional Encoder Representations from Transformers) model to classify football commentary text into specific events in football. This modality makes use of BERT’s natural language understanding to recognise and classify specific patterns in football commentary, and map each commentary segment to a specific event category such as ”free kick,” ”foul,” or ”attempt.”

The event classification process begins with us making use of the BERT model to obtain dense vector textual representation for the commentary text. Each commentary segment is cut off to 512 tokens because that is the maximum context window size in BERT and it is also long enough such that we do not lose relevant context around the specific event. This ensures that the model processes the entire necessary input without any loss due to truncation, therefore capturing unique patterns in commentary that will help in effectively classifying football events.

During training, we trained the BERT model for multi-class classification with 5 training epochs, achieving a final classification accuracy of 98%. This commentary-based classification helps our system to accurately recognize and prioritize significant moments by using commentary (text modality), enhancing the multi-modal summarization framework. This approach shows how LLM based fine-tuning can be used to achieve high precision for real-world sports event classification tasks.

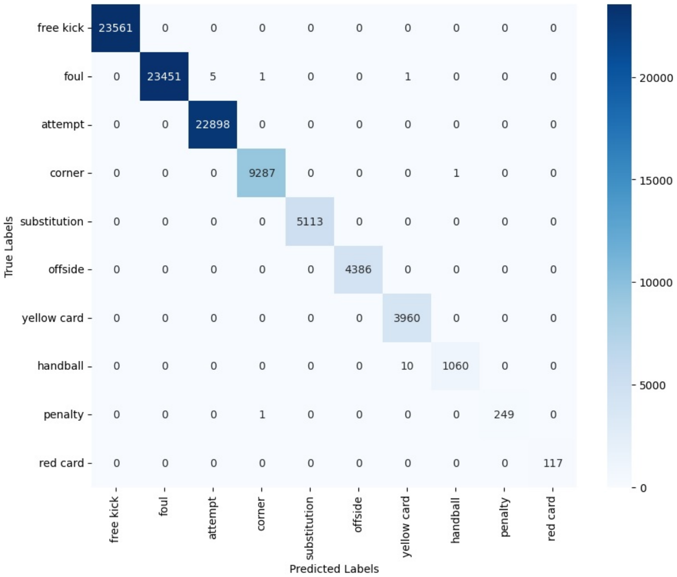

Fig. 7. Confusion matrix of football event classification using BERT.

IV. Multi-modal Integration Algorithm

In the proposed framework for multi-modal sports video summarization, integrating data from various different sources is of utmost importance to accurately identify and prioritize significant events. For this, we propose a timestamp clustering and weighting algorithm that synthesizes temporal data from all the modalities (commentary, crowd noise, scoreboard).

Let $T_{i,j}$ represent the j-th timestamp of the i-th modality, where $i \in \{1, 2, \ldots, M\}$ for each of the M modalities, and $j \in \{1, 2, \ldots, N_i\}$ for N_i timestamps from the i-th modality.

To determine overlapping timestamps across multiple modalities, we define a clustering window of Δt seconds that groups timestamps into clusters if they fall within Δt seconds of each other. Let C_k denote the k-th cluster, which includes timestamps $\{T_{i,j}\}$ such that:

$$|T_{i,j} - T_{i',j'}| \leq \Delta t, \tag{5}$$

$$\forall\, i, i' \in \{1, 2, \ldots, M\}, \quad j, j' \in \{1, 2, \ldots, N_i\}. \tag{6}$$

Each cluster C_k is calculated a weight w_k based on the number of modalities contributing to that cluster:

$$w_k = |\{i \mid T_{i,j} \in C_k\}| \,.$$

Thus, the weight w_k signifies the significance of the cluster, with higher values indicating more importance due to multi-modal consensus on the event's importance.

For customizable durations of summaries, clusters C_k are sorted in descending order by weight w_k, giving higher priority to events that have a multi-modal consensus. Based on the required length of the summary video, the top clusters are selected to fit the target duration. Thus, the number of clusters selected N_{selected} varies to produce a summary of the desired length.

$$N_{\text{selected}} = \text{Top clusters to get target summary duration}$$

This selection mechanism allows the algorithm to effectively integrate the timestamps generated by each modality into a single summary of the desired length.

V. Results

This section showcases the performance of our proposed multi-modal system for sports video summarization, demonstrating how it incorporates information from Tweets, scoreboard detection, and audio analysis.

A. Events Detection from Twitter Data Result

The hybrid Twitter-based event detection system effectively combines graph-based event detection with goal keyword analysis to identify key moments during the event. By analyzing the structure of tweet streams and the frequency of the keyword 'goal' within specific time intervals, the system detects important events with high accuracy—95% of the time. While occasional spikes and dips in tweet activity sometimes caused missed event detections, the system quickly adapted to these trends, ensuring that significant events were detected in a timely manner.

TABLE I
Goal Detection By Twitter

Match	Detected	Total	Goal Ratio (%)
France vs. Croatia	5	6	83.33
Qatar vs. Ecuador	2	2	100

B. Scoreboard Extraction and Detection Result

YOLOv4 successfully detected scoreboards in different videos and extracted essential information (team names, scores, time) from high-resolution frames (1280×720, 30 frames per second) with an accuracy of 89.06%. Low-resolution and noisy frames sometimes negatively impacted the algorithm's accuracy.

C. Audio Analysis and Commentary Classification Result

The audio analysis module, using Whisper for transcription, achieved 95% transcription accuracy. Event detection based on RMS intensity and peak finding was generally accurate in detecting key events. However, noisy timestamps occurred occasionally due to suboptimal audio quality. Using BERT for commentary classification, the model achieved over 95% accuracy. Some misclassifications occurred due to differences between commentary and actual events.

We chose the Goal & Scoring Ratio as the primary metric for evaluating the audio modality, defined as:

$$\text{Goal Ratio} = \frac{\text{Goals and Scoring Opportunities Detected}}{\text{Total Goals and Scoring Opportunities}} \tag{7}$$

Accuracy was not selected as it could be misleading due to the sparse nature of key events in a football match. Instead, precision and recall were used to evaluate the model's capability of correctly identifying relevant key events.

TABLE II
GOAL RATIO FOR AUDIO MODALITY

Match	Detected	Total	Goal Ratio (%)
France vs. Croatia	10	15	66.6
Portugal vs. Spain	8	12	66.6
Belgium vs. Brazil	10	16	62.5

The audio modality's Goal Ratio demonstrates its ability to capture goals based on sound intensity. While some goals were missed, the system consistently detected key moments, highlighting its importance in event detection.

D. Multi-modal Integration Result and Comparative Analysis

To evaluate our overall model, we compared its output with popular highlights available on YouTube for each match. These highlights serve as widely accepted benchmarks for key event detection quality.

Our system-generated summaries captured all events included in popular YouTube highlights and additional near-goal events not present in the highlights. While these extra events could be considered false positives, they represent key moments that enhance the viewer's experience, making our summaries more enriching.

Metric Definitions:

- **Precision**: $\frac{\text{Relevant Events Captured by Model}}{\text{Total Events Captured by Model}}$
- **Recall**: $\frac{\text{Relevant Events Captured by Model}}{\text{Total Relevant Events in YouTube Highlights}}$

TABLE III
PRECISION METRICS FOR MULTI-MODAL INTEGRATION

Match	YouTube Events	Model Events	Precision
France vs. Croatia	7	10	70.0
Portugal vs. Spain	8	11	72.0
Belgium vs. Brazil	6	9	66.6

TABLE IV
RECALL METRICS FOR MULTI-MODAL INTEGRATION

Match	Generated Events	Relevant Events	Recall
France vs. Croatia	10	9	90.0
Portugal vs. Spain	11	9	81.8
Belgium vs. Brazil	9	7	77.0

Goal Detection in the Overall System: The scoreboard detection modality ensured all goals were successfully identified, yielding a 100% Goal Ratio in a few matches. This result highlights the crucial role of scoreboard detection in augmenting the system's accuracy for key event identification.

TABLE V
GOAL DETECTION BY SCOREBOARD MODALITY

Match	Detected Goals	Total Goals	Goal Ratio (%)
France vs. Croatia	6	6	100.0
Portugal vs. Spain	5	6	83.3
Belgium vs. Brazil	3	3	100.0

By integrating scoreboard detection, our system ensures complete coverage of goals, addressing the limitations of audio-based detection. This proposed multi-modal approach provides a more reliable summarization of critical moments in football matches.

VI. CONCLUSION AND FUTURE WORK

The proposed multi-modal integration framework successfully summarizes sports events by merging data from multiple modalities, such as audio commentary, scoreboard detection, and social media analysis. By using the state of the art technologies in transcription, object detection and event classification (BERT), this system creates game highlights, significantly reducing the need for intensive manual editing. By integrating data from multiple different modalities, the system is able to detect key events accurately.

While the framework is able to achieve positive results, transitions between clips at times seem abrupt, sometimes cutting into the middle of an event. To enhance the viewing experience, future work needs to be carried out in order to improve the current switch from segment to segment by predictive models that capture the entire context of the highlight.

Additionally, exploring how this framework could integrate with AR/VR could enable fans to experience highlights in a more immersive and interactive way, potentially creating new avenues for engagement and fan retention.

Overall, the system gives promising results for automated sports summarization. As techniques in multimodal integration grow, this framework can be extended to provide more personalized, fluid, and engaging highlight presentation methodologies that enhance relations between fans and sports content.

ACKNOWLEDGMENT

We would like to express out gratitude to Dr. Sandesh B J, Professor and Chairperson, Department of Computer Science and Engineering, PES University, for his continuous guidance, assistance, and encouragement throughout the development of this UE21CS461A Capstone Project Phase-2.

We are grateful to the Capstone Project Coordinators, Dr. Farida Begam, Professor and Prof. Vandana M. Ladwani, Associate Professor, Department of Computer Science and Engineering, PES University for organizing, managing, and helping with the entire process.

We take this opportunity to thank Dr. Sandesh B J, Chairperson, Department of Computer Science and Engineering, PES University, for all the knowledge and support we have received from the department. We would like to thank Dr. B.K. Keshavan, Dean of Faculty, PES University for his help.

We are deeply grateful to Dr. M R Doreswamy, Chancellor, PES University, Prof. Jawahar Doreswamy, Pro-Chancellor, PES University, Dr. J Suryaprasad, Vice Chancellor, PES University and Prof. Nagarjuna Sadineni, Pro-Vice Chancellor, PES University for providing to us various opportunities and enlightenment every step of the way. Finally, this project could not have been completed without the continual support and encouragement we have received from our family and friends.

References

[1] H. Sattar, M. S. Umar, E. Ijaz and M. U. Arshad, ”Multi-Modal Architecture for Cricket Highlights Generation: Using Computer Vision and Large Language Model,” 2023 17th International Conference on Open Source Systems and Technologies (ICOSST), 2023, pp. 1-6, doi: 10.1109/ICOSST60641.2023.10414235

[2] Z. Saeed, R. Ayaz Abbasi, M. I. Razzak and G. Xu, ”Event Detection in Twitter Stream Using Weighted Dynamic Heartbeat Graph Approach [Application Notes],” in IEEE Computational Intelligence Magazine, vol. 14, no. 3, pp. 29-38, Aug. 2019, doi: 10.1109/MCI.2019.2919395

[3] A. Bhalla, A. Ahuja, P. Pant and A. Mittal, ”A Multimodal Approach for Automatic Cricket Video Summarization,” 2019 6th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2019, pp. 146-150, doi: 10.1109/SPIN.2019.8711625

[4] Helenca Duxans, Xavier Anguera and David Conejero Telefnica Investigacin y Desarrollo,Audio based Soccer Game Summarization,June 2009, Broadband Multimedia Systems and Broadcasting, 2009. BMSB ’09. IEEE International Symposium on, DOI: 10.1109/ISBMSB.2009.5133759