Multimodal Sport Summarization of Football Games

Dimitrios Roidis
Christina Tzortzaki

June 2025

1 Introduction

The explosion of multimedia content and the growing demand for concise, informative sports highlights have motivated research into automatic summarization of football games. Modern approaches leverage multiple modalities, including audio, text, and visual streams, to identify and assemble the most important moments into coherent summaries. The challenge lies in detecting emotionally and semantically significant events from noisy, complex data sources, and fusing these heterogeneous signals to produce a high-quality highlight reel. This project explores a multimodal approach to football match summarization, combining audio emotion detection, text-based event extraction, and visual analysis to automatically generate highlight candidates.

2 Objective

The objective of this project is to develop a multimodal system capable of extracting and fusing salient information from football match videos to produce meaningful highlights. Specifically, we aim to:

- Extract emotionally intense audio segments using affective features derived from the match soundtrack.
- Identify semantically important textual events from the match commentary using large language models.
- Integrate these signals with visual cues to determine highlight-worthy segments.

- Evaluate the effectiveness of this approach in capturing key match moments and reducing irrelevant content.

Ultimately, our goal is to build a framework that enables scalable, automated summarization of football matches using state-of-the-art tools in speech processing, natural language understanding, and computer vision.

3 Data

To define and then evaluate our approach, we collected video recordings from six popular football matches, each provided in MP4 format. These matches were carefully selected based on their global popularity and high viewer engagement, ensuring that they featured a rich variety of emotionally significant moments such as goals, fouls, and crowd reactions. All matches included English-language commentary, which allowed us to maintain linguistic consistency across samples and ensured that our audio-based saliency metrics remained stable and interpretable.

Each MP4 file contained both the video and synchronized audio track. We prioritized official broadcast-quality footage, as it captures professional commentary, crowd noise, and match sounds in a balanced and realistic manner—elements essential for detecting nuanced emotional shifts through the Behavioral Signals API.

Matches with English commentary where chosen, due to their abundance and the robustness they offer in saliency metric engineering. For example Spanish commentary is more expressive which could potentially lead us to choose different saliencies or thresholds to extract the highlights.

Another key criterion in the selection of these matches was the ease of access to official highlight videos. Since highlights serve as our ground truth for evaluating the effectiveness of our highlight detection pipeline, having clear, publicly available reference highlight clips made it easier to compare our predicted segments with real-world highlights. This alignment enabled more meaningful validation and tuning of our saliency thresholds and feature fusion strategies.

In total, the dataset covers approximately 10.5 hours of gameplay, including pre- and post-match commentary. Each match was processed independently to extract both audio and visual features. Emotional saliency scores were computed per audio segment using our custom metric, and these scores were later integrated with other modalities to identify and rank highlight-worthy moments. Our data collection strategy aimed to balance emotional richness, consistency, and verifiability to support robust multi-modal analysis. Final tests were performed on Portugal - Spain (3-3), a well-known match for football fans, with intense moments and a lot of goals.

4 Methodology

4.1 Audio Modality

We used the **Behavioral Signals API** [1] to extract low-level audio features and perform emotion recognition on the full match audio. This API provides high-level affective and behavioral features, including probabilistic scores for emotions such as happiness, sadness, and anger; measures of positivity and negativity; vocal strength indicators (strong/weak); speaker attributes like gender and age; and behavioral signals such as speaking rate and hesitation. We selected this API for its robust documentation, integrated functionality, and consistent performance across varied audio content.

To detect emotionally salient moments within a match, we analyzed the API's timestamped scores across these features. Based on empirical observation of actual football highlights, we noticed that signals like happiness, sadness, positivity, negativity, and vocal strength tended to spike around key events. This insight led us to design a custom saliency metric that captures emotional intensity across segments, helping us isolate the most emotionally charged moments in the audio stream.

The proposed saliency score is a weighted sum of the API's affective indicators, defined as:

$$\mathbf{Saliency} = (\mathbf{happy} + \mathbf{positive}) \cdot \mathbf{strong} + (\mathbf{angry} + \mathbf{negative}) \cdot \mathbf{strong}$$

This formulation emphasizes emotionally charged speech, whether positive or negative, when coupled with vocal strength. After scoring all segments using this metric, we applied a threshold-based filtering process to retain only a top percentage (5-8%) of segments with the highest saliency scores which at the same time have a duration bigger or equal than 4 seconds. This helped reduce noise and focus on the most emotionally intense audio moments, which were later fused with other modalities to detect highlight-worthy content. Finally those extracted highlight segments where used to create a video montage of them combined along with a padding of 10 seconds before each segment. That is to ensure that we do not lose important moments due to unresponsive commentator during the buildup of goal scoring chances.

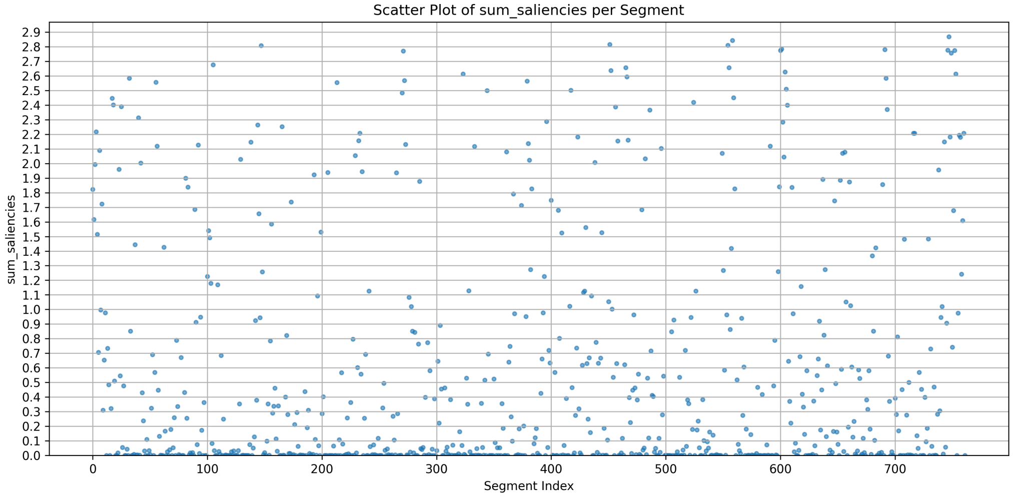

Figure 1: Scatter plot of our saliency metric for Portugal 3 - 3 Spain

4.2 Text Modality

The text modality component of our pipeline was designed to extract semantically meaningful events directly from the match commentary. We began by transcribing the full audio of the match using the **Whisper** automatic speech recognition (ASR) model. This model outputs timestamped text segments, providing both the spoken content and its corresponding temporal boundaries.

Once the full transcript was generated, it was segmented into smaller chunks to satisfy the token input limits of transformer-based models. We used a custom chunking function that preserved line and sentence structure while ensuring that each chunk remained under approximately 800 tokens when tokenized with the **Mistral-7B-Instruct** tokenizer from Hugging Face.

Each chunk was processed in the first pass through the language model with a prompt instructing it to extract all notable match events, such as goals, red cards, penalty kicks, and missed chances. The output consisted of a detailed list of timestamped candidate highlights covering the full match.

In the second pass, the combined list of candidate highlights was reformatted and re-chunked for a new LLM prompt. This prompt instructed the model to select only the most critical 6% of moments from the full list—roughly equivalent to 5–6 minutes in a standard match. The model returned a filtered list of top highlights in the same timestamped format, which we then parsed using regular expressions and saved in JSON format. These final timestamped segments were used in the evaluation and for video clip extraction.

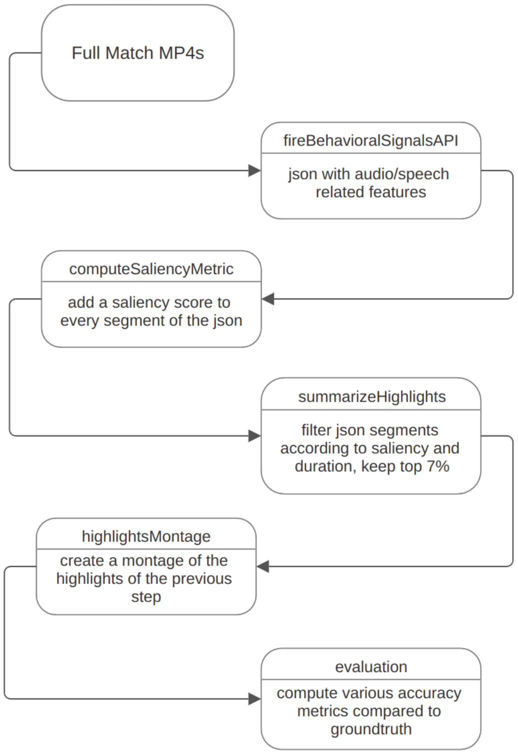
Full Match MP4s
fireBehavioralSignalsAPI
json with audio/speech related features
computeSaliencyMetric
add a saliency score to every segment of the json
summarizeHighlights
filter json segments according to saliency and duration, keep top 7%
highlightsMontage
create a montage of the highlights of the previous step
evaluation
compute various accuracy metrics compared to groundtruth

Figure 2: Flow diagram of audio modality pipeline

5 Evaluation

To assess the effectiveness of our highlight detection pipeline, we compared our predicted segments against manually curated ground-truth highlights. The evaluation was performed using both strict and flexible matching strategies to account for potential temporal misalignments between predicted and true highlight boundaries.

5.1 Ground Truth and Predictions

For each match, the official highlight clips were used to construct ground-truth annotation files in JSON format, where each entry marked a start and end timestamp of a highlight segment. Similarly, the predicted highlight segments—generated by our saliency-based multimodal pipeline—were stored in the same format. Timestamps were originally in `hh:mm:ss` format and converted to seconds for uniform processing.

5.2 Evaluation Metrics

We implemented a custom evaluation script in Python to compute the following metrics:

- **Precision**: Proportion of predicted seconds that are correctly aligned with ground-truth highlight seconds.
- **Recall**: Proportion of ground-truth highlight seconds that are correctly covered by predictions.
- **F1 Score**: Harmonic mean of precision and recall.
- **Global IoU (Intersection over Union)**: Measures the total overlap duration divided by the total union duration across all predicted and ground-truth segments.
- **AUC (Area Under Curve)**: Computed by treating the second-wise prediction and ground-truth masks as binary classifiers. This is useful for understanding the discriminative capacity of the predicted highlight mask.
- **TP, FP, FN**: Second-level counts of true positives, false positives, and false negatives.

5.3 Evaluation Modes

We introduced two modes of evaluation to accommodate different levels of strictness:

- **Strict Mode**: Requires second-by-second overlap between predicted and ground-truth segments. A prediction only contributes positively if it directly matches the annotated time range.
- **Flexible Mode**: Allows a prediction to count as a true positive if it covers at least 75% of a ground-truth segment's duration. This tolerance accounts for the natural variability in highlight timing and improves robustness.
- **Superflexible Mode**: Allows a prediction to count as a true positive if it covers at least 20% of a ground-truth segment's duration. A more relaxed and forgiving mode to compensate for the objectivity of a commentator's temperament.

6 Results

In this section, we present the evaluation results of our audio-based highlight detection method using the ground-truth annotations for the test match(Portugal 3 - 3 Spain). We tested our system under three evaluation modes: *Strict*, *Flexible* ($\geq$ 75%overlap), and *Super Flexible* ($\geq$ 20% overlap). Table **??** summarizes the key metrics across these modes.

Table 1: Updated evaluation results for the audio modality across three evaluation modes.

Mode	**Precision**	**Recall**	**F1**	**Global IoU**	**AUC**	**TP**	**FP**
Strict	0.6606	0.3529	0.4601	0.2988	0.6730	72	37
Flexible	0.6818	0.3676	0.4777	0.2988	0.6805	75	35
Super Flexible	**0.7578**	**0.4755**	**0.5843**	**0.2988**	**0.7348**	**97**	**31**

Although the absolute scores—particularly recall and F1—may initially appear modest, it is important to interpret them in the context of audio-driven highlight detection. Unlike visual or score-based methods, the audio modality relies heavily on linguistic, prosodic, and emotional cues, which vary significantly between commentators and broadcasting styles.

Moreover, the subjectivity inherent in human-annotated highlights introduces ambiguity: one commentator may emphasize a tactical pass or a build-up, while another may highlight only goal-scoring moments. These differences affect emotional saliency patterns detected in the audio and make it difficult for algorithmically generated segments to exactly match manually selected ones.

The strict evaluation mode penalizes even slight mismatches in timing, which explains the relatively low recall. As the overlap threshold relaxes (moving to flexible and super flexible modes), we observe a clear increase in recall and F1 score, without a drastic loss in precision. This suggests that the system effectively captures emotionally significant moments, even if its predicted segment boundaries do not exactly align with the human-edited highlight timestamps.

These results show that our audio-based saliency model is capable of identifying key emotional moments in football matches, and serves as a reliable modality for multi-modal highlight detection when complemented by visual or textual signals.

On the case of text driven highlights, despite the clear structure of the pipeline and detailed instructions, the LLM-generated highlights performed poorly in terms of evaluation metrics. This underperformance can be attributed to several inherent limitations of this modality. First, automatic transcription often fails to preserve contextual clarity and temporal precision, especially when football commentary includes overlapping speech, crowd noise, and rapid transitions. Second, LLMs are sensitive to prompt phrasing and token limits, which can lead to hallucinations, missed events, or vague summaries. Finally, highlights derived purely from language ignore other salient emotional or visual signals (e.g., tone, volume, player movements), making it difficult to compete with multimodal approaches. This reinforces the notion that textual data, when used in isolation, may lack the granularity and reliability needed for accurate highlight detection in dynamic sports contexts.

7 Conclusions

In this project, we developed and evaluated a multimodal pipeline for automatic football highlight summarization, integrating audio, text, and visual cues. Our primary focus was on the audio and textual modalities, with audio-based saliency emerging as the most effective single modality for identifying emotionally charged match moments. By leveraging emotion recognition APIs and designing a custom saliency metric, we were able to isolate segments of heightened emotional intensity that often aligned with key match events.

Evaluation across strict and relaxed matching schemes demonstrated that while exact temporal alignment with ground-truth highlights is challenging, especially due to subjectivity in commentary and highlight editing, our approach was able to capture a substantial portion of meaningful content. Notably, the performance improved under more tolerant evaluation modes, reflecting the difficulty of matching algorithmically derived segments to manually edited human highlights.

On the other hand, the text-based pipeline using large language models (LLMs) proved less reliable. Despite employing advanced transcription and prompting techniques, the LLM struggled to consistently extract semantically important events due to transcription noise, ambiguity in commentary, and the limitations of text-only interpretation. This outcome reinforced the importance of using multimodal signals—especially affective and acoustic features—for robust highlight detection.

Overall, our findings highlight both the potential and limitations of automated summarization in sports media. Audio remains a powerful but underutilized source of information for highlight generation. Future work should explore more sophisticated fusion strategies across modalities, introduce vision-based event detection, and incorporate reinforcement learning or user feedback loops to further refine and personalize highlight reels.

References

[1] Behavioral signals api. https://behavioralsignals.com/. Accessed: June 2025.