Multimodal Sport Summarization of Football Games

Dimitrios Roidis
Christina Tzortzaki

July 1, 2025

Introduction

- Huge growth in **multimedia content** and demand for concise **sports highlights**
- Challenge: Detect emotionally and semantically meaningful moments
- Modern methods use **multiple modalities**:
  - **Audio** (emotions, intensity)
  - **Text** (commentary events)
  - **Visual** (actions, replays)

Dataset Overview

- 6 Football Matches in MP4 format
- English-language commentary for consistency

Selected for:

- Global popularity
- High viewer engagement
- Rich emotional moments (goals, fouls, crowd reactions)

- Broadcast-quality audio ensures reliable emotional signal extraction
- Matches chosen with accessible official highlights for evaluation
- 10.5 hours of gameplay analyzed
- Final test: Portugal vs Spain (3–3) – a goal-rich, high-intensity match

Methodology – Audio Modality

Used Behavioral Signals API for **emotion recognition** and **feature extraction**:

- Emotions: happiness, sadness, anger
- Positivity/negativity, vocal strength, speaker traits

Designed a custom saliency score:

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

Retained top 5–8% emotionally intense segments ($\geq$ 4 seconds)
Added 10-second padding before each highlight to capture buildup
Result: Emotion-rich audio segments for highlight candidate generation

Methodology – Pipeline of Audio Modality

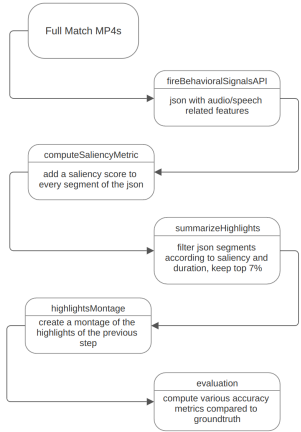

Figure: Flow diagram of audio modality pipeline

Methodology – Audio Modality

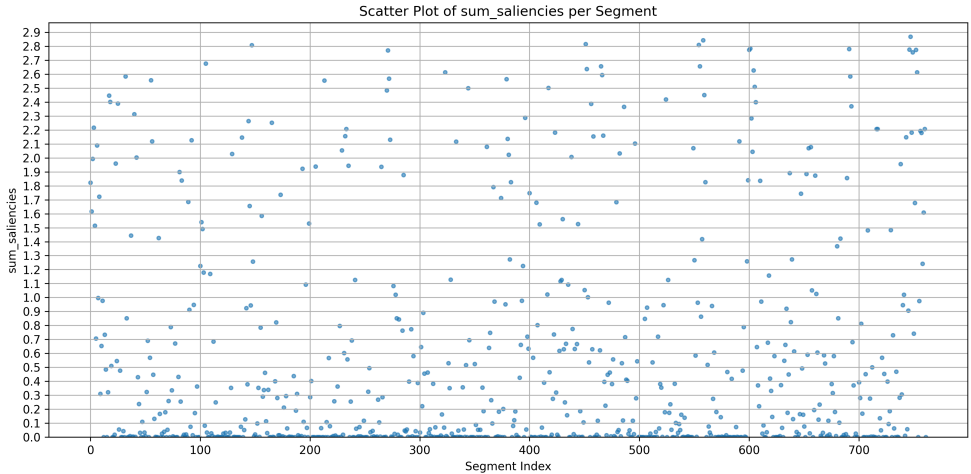

Figure: Scatter plot of our saliency metric for Portugal 3 - 3 Spain

Methodology – Text Modality

Used Whisper ASR to generate **timestamped transcripts**

- Chunks created to fit within transformer model limits
- Used Mistral-7B-Instruct for two-stage highlight extraction:
  - Pass 1: Extract **all** notable events (goals, red cards, etc.)
  - Pass 2: Select **top 6%** most critical moments

Output: JSON of timestamped highlight segments

Evaluation: How Did We Measure Performance?

How Did We Measure Performance?

- Ground Truth: Extracted from official highlight videos
  → Manually annotated JSON with start and end timestamps
- Predicted Highlights: Generated from our multimodal pipeline
  → Same format, converted to seconds for matching
- Custom Python Evaluation Script computed:
  - Precision: % of predicted highlight time that matches ground truth
  - Recall: % of actual highlights correctly predicted
  - F1 Score: Harmonic mean of precision and recall
  - Global IoU: Overlap / Union across all segments
  - AUC: Treating second-by-second prediction as binary classification
  - TP / FP / FN counts for second-level analysis

Evaluation Modes

- Strict Mode:
  - → Only counts exact second-by-second overlap
  - → High precision, low tolerance
- Flexible Mode (75%):
  - → Counts if 75% of ground-truth segment is covered
  - → Tolerates slight misalignments
- Superflexible Mode (20%):
  - → Very lenient, accepts partial overlap
  - → Reflects the subjectivity of highlight timing

Audio Modality: Evaluation Results

Evaluated under 3 modes:

- Flexible ($\geq$75%): allows partial coverage
- Super Flexible ($\geq$20%): most lenient
- Metrics computed: Precision, Recall, F1, IoU, AUC

Table: 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**

Text Modality: Challenges & Performance Issues

- Despite structured pipeline, LLM-generated highlights underperformed
- Key limitations:
  - ASR lacks temporal precision (crowd noise, overlapping speech)
  - LLMs sensitive to prompt engineering and token chunking
  - Text-only highlights ignore emotion and visuals (tone, intensity, player movement)
- Conclusion:
  - Text modality alone is not sufficient for reliable highlight detection
  - Needs to be fused with audio and visual modalities to enhance robustness

Conclusions & Future Work

- Developed a multimodal pipeline (audio, text, visual) for football highlight summarization
- Audio-based saliency proved to be the most effective single modality
  - Emotion recognition + custom saliency metric captured high-intensity segments
  - Precision held up well across strict and relaxed evaluation schemes
- LLM-based text pipeline showed limited reliability
  - Errors from ASR noise, vague or fast-paced commentary, and LLM hallucinations
  - Lacked alignment with emotional or visual context
- Key Insights:
  - Emotionally rich audio cues outperform pure transcript-based approaches
  - Subjective human highlight editing introduces evaluation challenges
  - Loose matching criteria (flexible modes) better reflect real-world variability

Thank You!

Questions?