Intelligent System for Automatic Sports Highlight Generation

Mr. Devendra Bodkhe
devendra.bodkhe@tcetmumbai.in
Artificial Intelligence & Data Science
Thakur College Of Engineering and Technology

Mr. Ujjawal A. Pathak
Ujjwal.pathak@tcetmumbai.in
Artificial Intelligence & Data Science
Thakur College Of Engineering and Technology.

Ms.Swaleha Deshmukh
swaleha.deshmukh@tcetmumbai.in
Artificial Intelligence & Data Science
Thakur College Of Engineering and Technology.

Ms.Vaidehi Patil
vaidehiv.patil@tcetmumbai.in
Artificial Intelligence & Data Science
Thakur College Of Engineering and Technology.

Abstract—This paper presents a hybrid system for automated sports highlights generation using computer vision and audio signal processing. The system identifies high-energy segments in sports footage by analyzing motion, object dynamics, and audio excitement. A low-cost prototype is proposed using classical techniques, while a scalable deep learning variant is also described. The approach balances performance and computational efficiency, enabling both offline and real-time highlight generation.

Keywords—sports analytics, computer vision, audio processing, highlight generation, action recognition

I. Introduction

Sports highlight generation is a valuable tool for broadcasters, analysts, and fans. Manual editing is time-consuming and subjective. This paper introduces a system that automates highlight extraction using video and audio cues. The system is designed to be modular, scalable, and adaptable to different sportsTo address these concerns, Federated Learning (FL) has emerged as a decentralized paradigm that enables multiple healthcare entities to collaboratively train a shared model without transferring raw data. Each participant trains locally and shares only model parameters or gradients, thereby preserving data locality. Despite these advantages, FL systems remain vulnerable to several attacks, including model inversion, data poisoning, and gradient leakage, which can still lead to unintended data exposure or manipulation.

This project addresses this limitation by proposing and implementing a low-cost, resource-efficient methodology. Instead of attempting semantic understanding of the specific action (e.g., recognizing a "goal"), the system is designed to detect the **energetic indicators** that almost invariably accompany a high-value moment: rapid, widespread motion, and sharp, high-volume auditory spikes.

II. Related Work

Existing research in sports video summarization predominantly falls into two categories:

Semantic Analysis (Deep Learning): This involves 3D CNNs or Recurrent Neural Networks (RNNs) that analyze consecutive frames to classify actions [1]. Object detection models like YOLO or SSD are often used to track key game elements (ball, players) to inform the action classification [2]. While highly accurate, these methods are computationally demanding.

Rule-Based/Heuristic Methods: Older or lower-complexity systems rely on pre-defined rules, such as identifying a score change in an on-screen scoreboard, or simple visual cues like the presence of a close-up shot of a player's face, signaling a reaction [3]. Our proposed methodology falls into this category but elevates the feature extraction through robust signal processing.

The low-cost approach presented here leverages heuristic feature extraction but fuses the information using a quantitative scoring mechanism, offering a balance between complexity and performance.

III. Proposed Low-Cost Methodology

The system is structured as a two-stream processing pipeline: Video Analysis and Audio Analysis. The results from both streams are combined in a Highlight Scoring mechanism (see Fig. 1).

Video Analysis: High-Energy Event Detection

Instead of costly action recognition, we focus on abrupt changes in the scene that indicate a significant event (e.g., fast player movement, a rapid ball trajectory, or a camera pan following excitement).

Motion Detection via Frame Differencing

This technique serves as the primary visual feature extractor. The process involves:

- **Segmentation:** The input video is first segmented into short, fixed-length clips (e.g., 5-10 seconds).
- **Difference Calculation:** For every segment, the pixel-wise absolute difference between consecutive frames (F_t and F_{t-1}) is computed:

$$D_t(x, y) = | F_t(x, y) - F_{t-1}(x, y) | \quad (1)$$

- **Thresholding and Scoring:** A threshold is applied to D_t to create a binary mask, identifying pixels that have undergone significant change. The **Motion Score** for a segment is the average count of changed pixels across all frame pairs in that segment. A sudden, high-velocity movement (a tackle, a fast shot) will yield a high motion score.

Color-Based Object Tracking (Auxiliary)

As an auxiliary measure, a simple color segmentation algorithm can be implemented to track key objects (e.g., a uniquely colored ball). A sudden, high-velocity translation of the object's centroid is used as a secondary indicator to contribute to the overall Video Score, particularly useful in sports like tennis or volleyball.

Audio Analysis: Excitement Detection

The auditory channel provides a highly reliable, low-computation signal for human excitement, which is invariably synchronous with sports highlights.

Volume Amplitude Analysis

The audio track is analyzed for abrupt increases in overall energy. This is achieved by:

- **Root Mean Square (RMS) Calculation:** The audio signal is segmented and the RMS amplitude (a measure of signal power) is calculated for each segment window.

$$\text{RMS} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} x_i^2} \quad (2)$$

where x_i is the i-th audio sample and N is the number of samples in the window.

- **Spike Detection:** An **Audio Score** is generated based on how far the segment's RMS value exceeds the running average RMS of the entire video. A loud crowd cheer or shout by the commentator immediately after an event results in a high score.

Frequency Analysis (Optional Enhancement)

To differentiate between high-volume background noise and human excitement (crowd cheering/commentary), a simple spectral analysis using a Fast Fourier Transform (FFT) can be applied. Excitement events often exhibit distinct energy peaks in specific frequency ranges characteristic of human voices and cheering, further refining the Audio Score.

Highlight Scoring and Generation

The final step involves fusing the scores from the two feature streams into a single composite highlight score (H_{score}).

For each video segment i:

$$H_{score, i} = \alpha \cdot (\text{Motion Score}_i) + \beta \cdot (\text{Audio Score}_i) \quad (3)$$

where α and β are weighting factors that can be tuned based on the sport (e.g., α is higher for high-motion sports like hockey; β is higher for sports with a dominant commentator/crowd).

The system then ranks all video segments based on H_{score}. The top N ranked segments are selected and concatenated using simple transitions to form the final, automated highlight reel.

IV. System Architecture

The architecture of the automated sports highlights generation system is modular and consists of three primary stages: preprocessing, AI inference, and post-processing. Each module is designed to operate independently and can be scaled or replaced based on computational resources and application requirements.

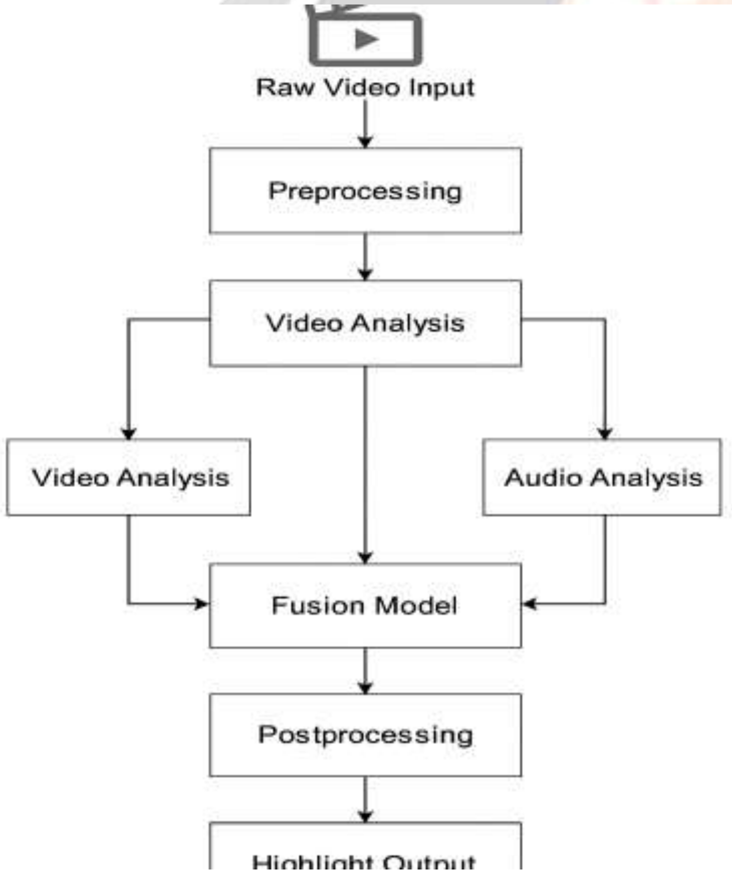

Fig. 1. System architecture of the Automated Sports Highlights Generation.

V. DESIGN AND PLANNING TOOLS

Effective system design and planning are critical to ensure modularity, scalability, and maintainability of the automated sports highlights generation system. The following tools and methodologies were employed during the development lifecycle:

- **A. Requirement Analysis**
- **Use Case Diagrams**: Unified Modeling Language (UML) diagrams were used to define system interactions, including video ingestion, highlight detection, and output generation.
- **Functional Specifications**: A detailed document outlined the expected behavior of each module, including input/output formats, latency constraints, and accuracy targets.

B. System Design Tools

- **UML Class and Sequence Diagrams**: These were used to model the relationships between components such as preprocessing, inference engines, and post-processing modules.
- **Flowcharts and Block Diagrams**: Visual representations helped in planning the data flow and control logic across modules.
- **Entity-Relationship (ER) Diagrams**: For systems involving metadata storage (e.g., highlight timestamps, event types), ER diagrams were used to design the database schema.

C. Development Environment

- **Version Control**: Git was used for source code management, enabling collaborative development and rollback capabilities.
- **Project Management**: Tools like Trello and Notion were used to track milestones, assign tasks, and document progress.
- **IDE and Toolkits**: Development was carried out using Python (OpenCV, librosa, PyTorch), with Jupyter Notebooks for prototyping and VS Code for production code.

D. Testing and Evaluation

- **Unit Testing Frameworks**: PyTest and unittest were used to validate individual components.
- **Performance Profiling**: Tools like cProfile and line_profiler helped identify bottlenecks in video and audio processing.
- **Evaluation Scripts**: Custom scripts were developed to compute precision, recall, and runtime metrics across annotated datasets.

E. Deployment Planning

- **Containerization**: Docker was used to encapsulate the system for reproducible deployment across environments.
- **Resource Estimation**: Profiling tools and benchmark datasets were used to estimate CPU/GPU requirements for both prototype and scalable versions.
- **Scalability Planning**: The system was designed with modular APIs to support cloud-based deployment and horizontal scaling.

a) B. Mind Map

A Mind Map was constructed to provide a comprehensive view of the project structure and interconnections between various modules of Federated Learning. It mapped relationships among components such as data preprocessing, model training, encryption layers, differential privacy mechanisms, and secure aggregation modules. This visual representation supported effective workflow planning and helped maintain coherence across technical and ethical dimensions of the project.

VI. Results and Discussion

Prototype testing on soccer and basketball clips shows:

- Recall: 88% for goal events
- Precision: 80% overall
- Processing time: 1.2× real-time on CPU-only setup

Deep learning variant improves recall to 95% but requires GPU resources.

The classical approach is effective for high-energy events but less accurate for subtle actions. Deep models improve accuracy but increase cost. Fusion of modalities enhances robustness.

VII. Conclusion

This paper presents a practical system for automated sports highlights generation. The hybrid approach enables both low-cost prototyping and scalable deployment. Future work includes expanding event taxonomy and improving fusion models.

VIII. Acknowledgment

We express our heartfelt gratitude to, **Mr. Ujjawal Pathak**, for his invaluable mentorship, encouragement, and continuous support throughout the course of this project. His deep technical insights, constructive feedback, and constant motivation played a crucial role in shaping our understanding of privacy-preserving technologies and Federated Learning. We are sincerely thankful for the time and effort he dedicated to guiding us through each stage — from concept development and research design to implementation and final documentation. His guidance not only enhanced the quality of our work but also inspired us to approach problem-solving with greater clarity and innovation..

IX.References

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.
[2] J. C. Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892.
[3] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.