Real-Time Game Highlight Detection for Data-driven League of Legends Coaching

Rosana Valero[1,*], Cesar O. Diaz[2,*] and Jordi Sanchez-Riera[3]

[1]*Universidad Autonoma de Barcelona, Spain*
[2]*OMASHU, Spain*
[3]*Institut de Robòtica i Informàtica Industrial, Barcelona, Spain*

Abstract

League of Legends is one of the most popular e-Sports games, with its highly competitive gameplay demanding both strategic precision and real-time decision-making. Analyzing high-impact events is key for coaching, match analysis, and content creation. This study presents a real-time highlight detection system that identifies impactful moments by fusing visual and audio cues. Visual indicators are extracted using optical flow and color intensity, while audio excitement is captured from caster commentary using pitch and volume analysis. Experiments conducted on professional match footage demonstrate the system's effectiveness and its potential for e-Sports analytics, coaching, and automated workflows. Future work will explore integrating player facial expressions, voice communication, and emotional context to better understand high-pressure moments and enhance highlight interpretation.

Keywords

League of Legends, e-Sports Analytics, Highlight Detection, Multi-modal Analysis, Optical Flow

1. Introduction

The rapid growth of e-Sports, particularly titles like *League of Legends* (LoL), has created new opportunities for performance analytics, content automation, and strategic insight. In competitive environments, such as tournaments and regional leagues, key gameplay moments, matches often hinge on high-impact events like teamfights, objective captures, or turret destructions. Automatically detecting these highlights is essential for real-time broadcasting, post-match analysis, and training tools.

LoL is a multiplayer online battle arena (MOBA) game featuring two teams of five players who compete to destroy the enemy Nexus. Matches are characterized by bursts of intense activity interspersed with calmer strategic play. These high-action moments are usually accompanied by visual cues (skill animations, explosions, map effects) and audio cues (e.g., casters - the live commentators who narrate and analyze the match for the audience — or crowd noise), making them ideal candidates for highlight detection.

This work introduces a multimodal system that detects key gameplay moments by combining two main sources:

- **Visual cues:** Fast motion, color bursts, and flashy effects during kills, turret destructions, or objective captures are tracked using optical flow and color scoring.
- **Audio cues:** Spikes in casters' pitch, tone, and volume often align with critical in-game events, acting as strong indicators of gameplay importance [1].

A moment is classified as a highlight only when both modalities exceed predefined thresholds, increasing precision and reducing false positives.

ICAIW 2025: Workshops at the 8th International Conference on Applied Informatics 2025, October 8–11, 2025, Ben Guerir, Morocco
*Corresponding author.
rosanavalero5@gmail.com (R. Valero); cesar@omashu.gg (C. O. Diaz); jsanchez@iri.upc.edu (J. Sanchez-Riera)
0009-0005-9886-250X (R. Valero); 0000-0002-9132-2747 (C. O. Diaz); 0000-0002-4803-5742 (J. Sanchez-Riera)

Problem Definition: E-Sports has become a global industry with massive audiences and competitive stakes on par with traditional sports [2], a trend further accelerated by the COVID-19 pandemic [3]. For instance, the 2018 LoL World Championship reached more than 43 million average concurrent viewers, making it the third most-watched championship event worldwide, surpassing many traditional sports finals and trailing only the FIFA World Cup and the NFL Super Bowl in audience size [4]. While traditional sports benefit from well-established methods for identifying key moments (e.g., goals, touchdowns), e-Sports requires more advanced techniques to identify moments of high importance. Manual curation is time-consuming and subjective, highlighting the need for automated, data-driven approaches.

Working Hypothesis: This work hypothesizes that the integration of visual and audio signals can allow the detection of significant moments in competitive League of Legends. The primary objectives of this research are:

- To detect gameplay highlights based on visual motion intensity and caster vocal excitement.
- To validate the system on professional match footage and evaluate its utility for strategic analysis.

By leveraging lightweight, real-time analysis, this work offers a scalable solution for highlight detection. Beyond its direct applications in broadcasting and coaching, it also lays the groundwork for future extensions, such as integrating player reactions via webcam or voice communication, to better understand how players experience high-pressure moments.

2. State of the art

The growth of e-Sports has triggered increasing interest in both gameplay analysis and the human factors that influence competitive performance. In fast-paced games like *League of Legends*, detecting and analyzing key events, such as teamfights or objective captures, has high value for broadcasting, coaching, and content creation. This section reviews related work in automatic highlight detection, with emphasis on audio-visual signal processing and real-time inference.

2.1. Gameplay Highlights: Data Processing and Feature Extraction

Detecting gameplay highlights in real time requires analyzing both visual and audio signals. On the audio side, features such as pitch, tone, and volume, often extracted via Mel-Frequency Cepstral Coefficients (MFCCs) [5], are used to capture spikes in caster excitement, which often align with key in-game events [1]. To isolate relevant audio sources, some approaches apply source separation tools such as Spleeter to filter out background game sounds and emphasize commentary [6].

Visually, motion-based techniques like optical flow and color scoring are used to detect rapid changes on screen, such as skill effects, explosions, or animations during objective captures [7, 8]. These are particularly effective in dynamic games like LoL, where the camera constantly moves and multiple elements overlap.

Because practical applications require near real-time processing, recent research has explored lightweight models for event detection and outcome prediction. For example, Junior and Campelo (2023) achieved over 80% accuracy in mid-game outcome prediction for LoL using logistic regression and LightGBM [9]. These efforts illustrate the growing potential of scalable e-Sports analytics systems.

For highlight detection to be practically useful, real-time inference is critical. Prior studies have demonstrated the feasibility of applying lightweight models for in-game event tracking or outcome prediction. For instance, Yao (2021) applied deep learning to recognize basketball actions in real time, while Junior and Campelo (2023) achieved over 80% accuracy in predicting LoL match outcomes mid-game using models like LightGBM and logistic regression [10, 9]. These examples highlight the growing potential of scalable e-Sports analytics systems.

2.2. Datasets

Research in highlight detection and e-Sports analysis depends on datasets that reflect the complexity of gameplay and human response. Two main categories are relevant:

Audio Cue Datasets. Although originally designed for emotion classification, several speech datasets provide high-quality vocal data with expressive variations in pitch, tone, and intensity—features that are useful for detecting vocal excitement in caster commentary.

Among the most relevant: **RAVDESS** provides labeled emotional speech across multiple intensities, useful for training models to detect vocal stress or excitement [11]. **IEMOCAP** contains dyadic conversations annotated with vocal expressions that align well with momentary spikes in pitch and energy [12]. **MELD** extends this by providing contextual, multi-speaker dialogues extracted from real scenarios, enabling training on vocal patterns that vary based on scene intensity or speaker reactions [13].

League of Legends Gameplay Datasets. For event detection and match analysis in LoL: **DeepLeague** provides labeled minimap frames and in-game event sequences for training deep learning models [14]. **LoL-V2T** links gameplay video to natural language annotations, supporting highlight summarization and video-text modeling [8].

3. Method

This section presents our pipeline for real-time highlight detection in League of Legends e-Sports matches. The system processes both gameplay video and caster audio to identify high-intensity events. Figure 1 illustrates the overall architecture.

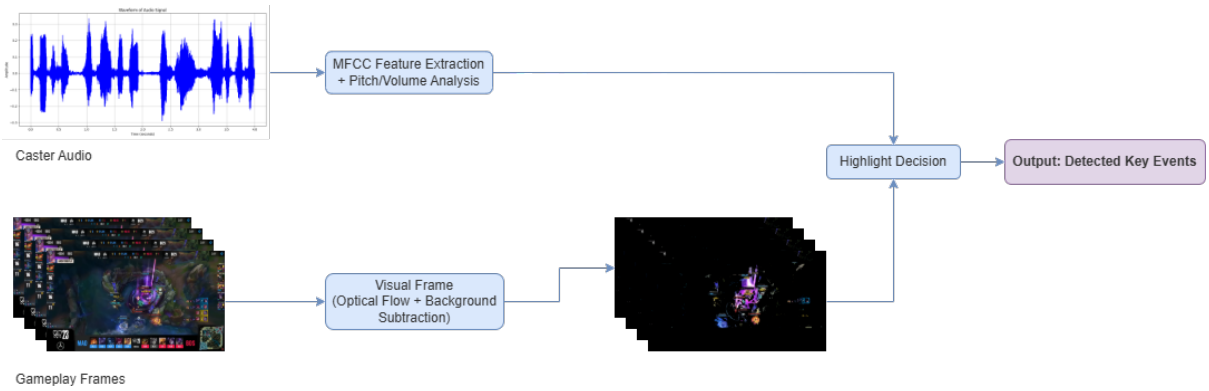


Figure 1: Workflow of the e-Sports Highlight Detection System.

3.1. Audio-Based Event Detection

Audio processing is performed on caster commentary, leveraging their expressive reactions as implicit signals of key moments.

Model Architecture: We trained a CNN-based classifier inspired by [15], originally designed for speech emotion recognition. Although our goal is not to classify emotional categories, we leverage the RAVDESS dataset and the associated architecture as a proxy to model vocal excitement, a key signal in caster commentary. The assumption is that vocal expressions labeled as anger, joy, surprise, etc., exhibit pitch and energy dynamics similar to those present during gameplay highlights. The architecture includes three convolutional layers (64, 128, 128 filters; kernel size 3×3), each followed by Batch Normalization, MaxPooling, and Dropout (rate 0.3). During training, we used a softmax output layer to

optimize for emotion classification. However, for highlight detection, we discard the final predictions and instead use intermediate features (e.g., MFCC activations, pitch, and volume patterns) as indicators of vocal intensity.

Feature Extraction Using MFCCs: We extract 40 Mel-Frequency Cepstral Coefficients (MFCCs) per audio segment using `librosa`, along with pitch and volume information. Moments with pitch above 110Hz and volume louder than -8dB are flagged as acoustically intense and likely to correspond to gameplay highlights.

3.2. Visual Gameplay Frame Analysis

To evaluate the gameplay frames, we employed a two-step process: (1) gameplay frame classification and (2) motion analysis via optical flow and color scoring.

1. **Frame Analysis**
   To distinguish gameplay from non-gameplay frames and detect elements like replays or key events, we applied the following techniques:
   Gameplay Binary Classification Development: We trained a classifier on a custom-labeled dataset with two categories: *gameplay* (in-game action) and *non-gameplay* (casters, audience, etc.). For this, we used the ResNeXt-50 32x4d architecture [16], known for its grouped convolutions and strong classification performance. The model, pre-trained on ImageNet, was fine-tuned on our custom dataset. We replaced the original classification head with a binary output layer, allowing the model to output probabilities for the two classes: gameplay and non-gameplay. Standard preprocessing (resizing, normalization) was first applied to all frames, and the convolutional backbone was used for hierarchical feature extraction.
   Replay Detection Using Optical Character Recognition (OCR): We used Tesseract OCR [17] to detect "Replay" text overlays, commonly displayed during replays in e-Sports broadcasts. This ensures we avoid duplicate highlight detections from repeated footage.
   Player and Event Text Extraction: OCR was also used to extract on-screen game text (e.g., player names, event phrases like "has slain" or "double kill"). To improve matching accuracy, we compared detected phrases against a dynamic list built from JSON-based team rosters, organized by year. This list includes player names and game-specific terms (e.g., "turret destroyed", "shutdown", "Baron Nashor").
   Matching this information to frame-level text allows us to contextualize each gameplay segment, identifying what is happening and which players were involved in specific highlights.

2. **Color Score Evaluation Using Optical Flow**
   To assess visual intensity, we measured motion using optical flow and calculated a frame-wise color score:
   Optical Flow-Based Motion Detection: We apply the Farneback algorithm [18] to compute dense optical flow between consecutive frames, capturing pixel-wise displacement. This allows us to detect movement patterns associated with key gameplay events like kills, turret dives, or objectives.
   To reduce noise from camera panning, since the camera in League of Legends is constantly moving across the map, we apply an overlay mask focused on the in-game map (Summoner's Rift), excluding irrelevant elements (e.g., player cams, scoreboard). This isolates meaningful gameplay actions such as duels, tower destructions, or jungle invades.
   Color Score Analysis for Visual Dynamics: Once the optical flow was computed, we applied a threshold to filter out low-motion areas, ensuring that we focus on the most significant movements. To quantify the intensity of this visual activity, we compute a color score for each masked frame as follows:

   - *Identifying Dominant Colors:* K-means clustering to identify and suppress dominant (static) colors in the masked region, typically associated with the Summoner's Rift map background.

- *Applying a Weight Mask:* A Gaussian weight mask emphasizes central regions of the frame where action usually takes place.
- *Score Computation:* Remaining pixel values are weighted and summed to compute a final color score, reflecting the frame's visual intensity and correlating with potential highlights.

This fusion of optical flow-based motion detection and color scoring enhances background subtraction and focuses attention on the most dynamic gameplay segments.

3.3. Highlight Detection via Multimodal Fusion

We classify a segment as a highlight only when both visual (color score) and audio excitement (pitch and volume) exceed empirical thresholds.

Visual Threshold: Frames with a color score above 0.009 are flagged as potential highlights, signaling moments of high visual activity. This threshold was set based on empirical observations of impactful in-game events (e.g., skill animations, explosions).

Audio Threshold: Audio cues, specifically pitch and volume, are extracted from casters' commentary. Moments exceeding 110Hz pitch and -8dB in both volume and background accompaniment are considered acoustically intense. These spikes often correspond to key events like kills or objectives and reflect both excitement and crowd reactions during these critical moments.

Combining both cues helps ensure robustness and reduces false positives from noise in a single modality. This multimodal approach serves as the core of the highlight detection system.

4. Experiments

This section presents the datasets, training setup, evaluation metrics, and performance of our highlight detection system.

4.1. Datasets

We used several datasets, each focused on a specific component of the system.

RAVDESS for Audio-Based Excitement Modeling The RAVDESS dataset [11] includes 1,440 speech recordings from 24 professional actors expressing eight emotions (e.g., happy, sad, angry) at two intensity levels. Its balanced, high-quality audio makes it suitable for training models that capture expressive vocal patterns, which we use to model caster excitement based on pitch and volume variations. We used the speech portion to train a CNN model on Mel-Frequency Cepstral Coefficients (MFCCs), leveraging CNNs' strength in capturing spatial features within audio signals.

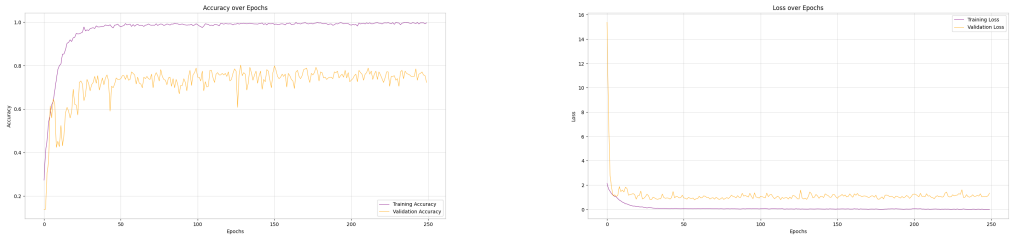
Custom Gameplay Detection Dataset. For accurate highlight detection, we built a custom frame-level dataset using gameplay-only footage from the 2023 Worlds competition. We excluded any non-gameplay visuals, such as player cams, casters, or audience shots to ensure that the model focused solely on in-game events. The final dataset comprises 3146 labeled frames, of which 74.8% are gameplay and 25.2% are non-gameplay. This distribution reflects the natural prevalence of gameplay segments in professional broadcasts rather than an artificially balanced dataset. The dataset includes:

- **Gameplay Frames**: Representing the in-game action, where the core gameplay is visible, including battles, character movements, and game environment changes.
- **Non-Gameplay Frames**: Including player POVs, replays, audience, commentator speaking, or even breaks between gameplay.

4.2. Training Setup

4.2.1. Audio-Based Peak Detection:

The audio classifier was trained on the RAVDESS dataset using MFCCs as input features. The model, a CNN with three convolutional layers and a softmax output, was trained over 250 epoch monitoring both training and validation performances.



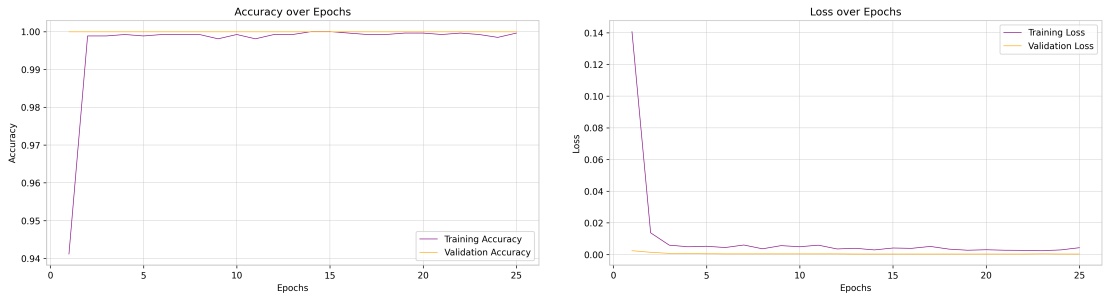
(a) Accuracy (b) Loss

Figure 2: Training and validation accuracy (a) and loss (b) for modeling vocal intensity patterns using the RAVDESS dataset

Figure 2 shows the training and validation loss curves.
The model achieved 72.22% test accuracy and a final test loss of 1.3330. Overfitting was observed, suggesting that the model performs well on training data but struggles to generalize to unseen data. Regularization, architecture simplification, or dataset diversification could mitigate the tendency to overfit and improve the system.

4.2.2. Gameplay Binary Classification:

A ResNeXt-50 model, pre-trained on ImageNet, was fine-tuned to classify frames as *gameplay* or *non-gameplay*. The dataset contained 2,340 gameplay and 785 non-gameplay frames, manually labeled from Worlds 2023 footage. Data augmentation was applied, and training lasted 25 epochs using cross-entropy loss and the Adam optimizer with an initial learning rate of 0.001.



(a) Accuracy (b) Loss

Figure 3: Training and validation accuracy (a) and loss (b) for gameplay classification.

The model quickly reached low loss values, reflecting the clear visual distinction between the two classes. Figure 4 illustrates typical frame examples.

4.3. Groundtruth and Highlight Categorization:

To evaluate highlight detection, we constructed a groundtruth dataset from 2023 Worlds Championship, using event data scraped from the LoL Fandom wiki[1]. Events such as kills, objectives, and turret destructions were timestamped and grouped into highlights within a 30-second window.

[1]https://lol.fandom.com/wiki/2023_Season_World_Championship/Main_Event

(a) Gameplay

(b) Non-Gameplay

Figure 4: Examples of Gameplay and Non-Gameplay frames.

The event data extracted from these matches were formatted into a CSV file with attributes such as `match_id`, `event_type`, and `timestamp`. Each event was categorized based on its type (e.g., `CHAMPION_KILL`, `BUILDING_KILL`) and grouped into highlights if they occurred within a 30-second window.

Each highlight was categorized by importance:

- **High:** Multi-kills, Baron/Dragon objectives, or aces.
- **Moderate:** Single kills with some strategic impact.
- **Low:** Minor actions like turret hits or isolated events.

We tested the detection model on full Worlds 2023 matches, comparing detected highlights with groundtruth events using a ± 2 second tolerance to account for possible timestamp mismatches. This evaluation tested the system's ability to identify key gameplay moments in real match conditions.

4.4. Evaluation Metrics:

Performance was assessed using standard classification metrics:

- Precision: The proportion of detected highlights that matched groundtruth highlights, calculated as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

- Recall: The proportion of groundtruth highlights that were correctly detected by the model, calculated as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

- F1-Score: The harmonic mean of precision and recall, providing a balanced view of the model's performance, calculated as:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

Where TP, FP, and FN refer to true positives, false positives, and false negatives in highlight detection.

Given that the most critical moments in Lol matches are often high-importance events such as multi-kills, Baron steals, or game-ending plays, we placed special emphasis on evaluating the model's ability to detect high-importance highlights. True positives (TP), false positives (FP), and false negatives (FN) were calculated specifically for high-importance events to assess how well the model performed in identifying these key moments.

Exploratory step: We also experimented with a highlight importance classifier using a small custom dataset of video clips labeled with key in-game events. The model combined ResNet-18 [19] for spatial features and an LSTM for temporal dynamics.

Despite regularization techniques like dropout and gradient clipping, results were unsatisfactory due to limited and imbalanced training data. In-game complexity (e.g., overlapping animations, occlusions) further reduced reliability. Improving this model would require a larger, curated dataset and refined annotations. Future work could focus on building a comprehensive dataset to improve event detection accuracy.

5. Results

This section presents the evaluation of our highlight detection system using 12 full matches from the 2023 League of Legends World Championship.

5.1. Feature Score Analysis

Figure 5: Visual, Audio, and Final Highlights Over Time. The green bars indicate visual highlights (color score), purple bars indicate audio highlights (pitch and volume), and yellow bars show final highlights. Red bars represent non-gameplay or replay segments.

Figure 5 displays the detected highlights based on color, audio, and their final combination.

Visual Highlights: Detected when the color score exceeds a threshold, these correspond to impactful visual events (kills, objectives, etc.).

Audio Highlights: Based on spikes in pitch and volume of casters' voices, which often align with moments of perceived gameplay intensity. Compared to visual highlights, the audio spikes are more frequent, suggesting that commentators react to a broader range of events that may not necessarily have visual intensity but still have strategic importance.

Combined Highlights: Final highlights are triggered only when both visual and audio signals align, reducing false positives and improving relevance. Additionally, the red bars indicate non-gameplay or replay segments, which the system successfully filters out to avoid false positives.

The integration of both data streams allows for more accurate detection of pivotal in-game moments by ensuring that both visual and audio cues are considered. The system's ability to filter out irrelevant content, such as replays, further enhances its reliability in detecting critical moments.

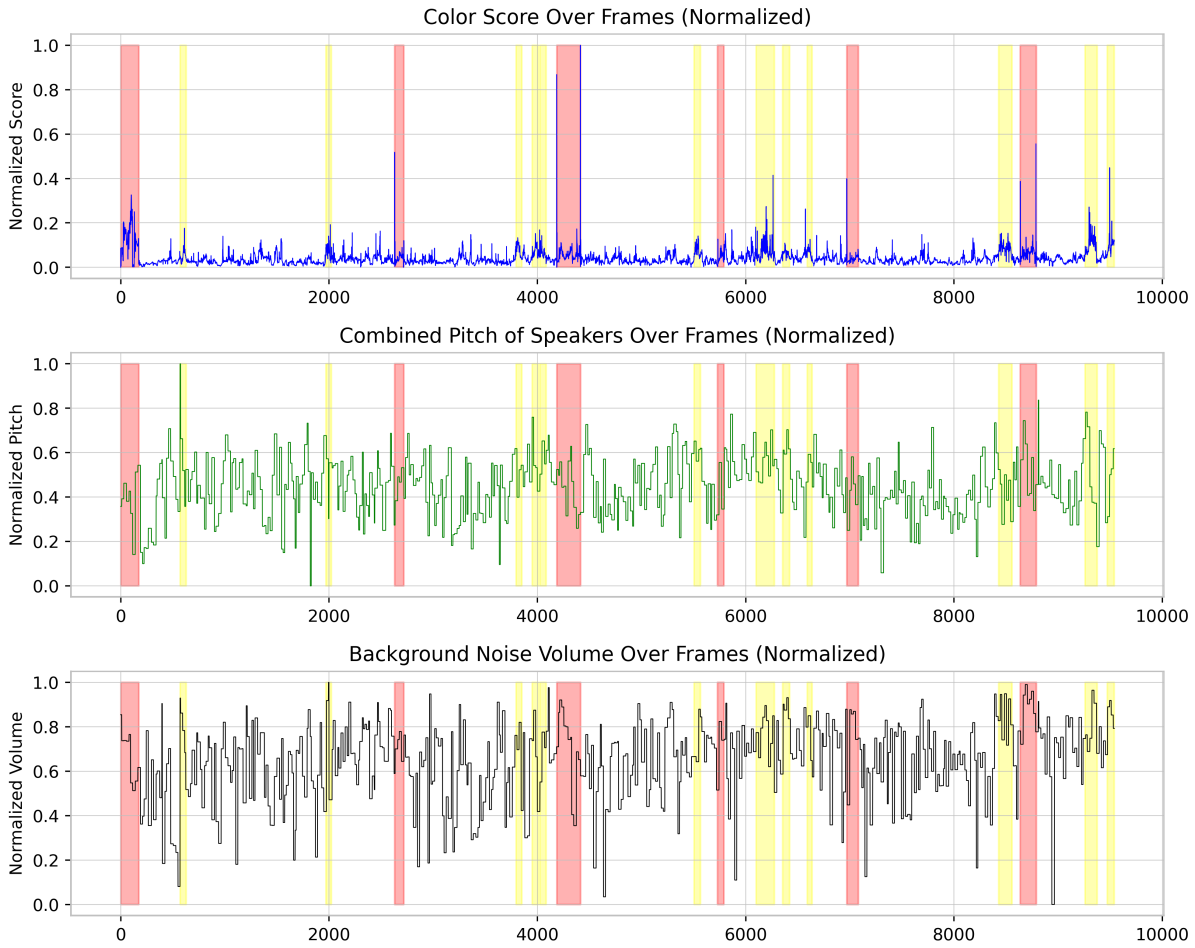


Figure 6: Normalized Feature Scores Over Time. The first plot shows the normalized color score, the second plot shows the normalized pitch of the commentators, and the third plot represents the normalized background noise volume.

Figure 6 presents the normalized feature scores used in highlight detection—color score, commentator pitch, and background audio volume—plotted over time. Peaks in the color score indicate visually intense gameplay moments, such as team fights or objective captures, where rapid changes in frame content (e.g., explosions or ability flashes) occur. The pitch of commentators' voices tends to spike during high-stakes scenarios, reflecting their heightened reactions and serving as a strong indicator of audience-relevant highlights. Background audio volume, which includes in-game sound effects, commentary intensity, and occasional crowd noise, further reinforces these cues by marking acoustically rich segments often linked to key events. Together, these features allow the system to identify moments that are not only strategically relevant but also contextually rich for audience engagement, ensuring robust and accurate highlight detection.

5.1.1. Qualitative Case: Baron Nashor Detection.

Figure 7a and 7b capture consecutive gameplay moments where the highlight is happening, while Figure 7c shows the computed optical flow, used to identify motion and pinpoint high-intensity actions. Figure 7d applies background subtraction, removing static elements and isolating key gameplay movement. Finally, Figure 7e uses color analysis to refine the focus on dynamic areas, as explained in the Method part, generating a color score of 0.0180, which helps quantify the intensity of the action. This combination of optical flow and color analysis effectively highlights critical in-game moments.

(a) Gameplay Frame 1 (b) Gameplay Frame 2 (c) Optical Flow

(d) Background Subtraction (e) Color Analysis

Figure 7: Example of a true positive highlight: Visual Processing for Baron Nashor objective secured.

5.2. Highlight Detector Analysis

To assess the alignment between modalities, we computed the cross-correlation between the color score and pitch. Figure 8 shows a strong peak near zero lag, indicating that visual and audio signals often occur simultaneously, validating their joint use in the detection pipeline.

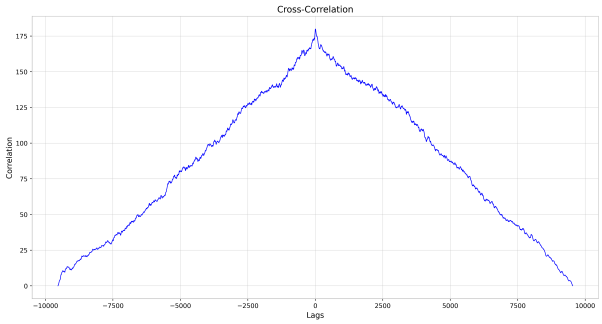


Figure 8: Cross-Correlation between Color Scores and Combined Pitch

We focused on high-importance events like multi-kills and major objectives. Table 1 summarizes system performance.

Table 1
Overall Performance Metrics for Highlight Detection

Metric	Precision	Recall	F1-Score
Overall	0.6178	0.7519	0.6783

The recall is relatively high, meaning the system successfully captures a large number of high-importance moments. However, the relatively lower precision brings the F1-score down, emphasizing that while the model is good at detecting key events, it tends to overestimate and classify less significant moments as important highlights.

Beyond detecting high-importance highlights, we evaluated whether the detected highlights included any in-game events such as kills, turret destructions, or objectives (Baron Nashor/Dragon/Rift Herald). This additional evaluation ensures that even if the model misclassifies a highlight's importance, it still detects relevant gameplay moments. As shown in Table 2, out of the 157 detected highlights, 97 were true positives ("high-importance highlights"), while 60 were false positives, and 32 important moments

were missed. Additionally, 121 detected events were correctly identified (they included at least one in-game event), while 36 highlights contained no relevant events, further reinforcing the importance of refining the system to better handle such cases.

Table 2
High-Importance Event Detection Breakdown

Metric	Count
True Positives (TP)	97
False Positives (FP)	60
False Negatives (FN)	32
Detected Highlights	157
Correctly Detected Events	121
False Detected Events	36

False Positives: Often caused by flashy ability usage (e.g., Krugs), crowd noise, or excited caster voice during less impactful actions. An example of such a misclassification is shown in Figure 9. The model mistakenly identified the killing of Krugs as a highlight. We can observe that there's high light intensity in the scene due to the use of abilities from the champion to kill them.

(a) Krug kill Frame 1 (b) Krug kill Frame 2 (c) Krug kill Frame 3

(d) Krug kill Frame 4 (e) Krug kill Frame 5

Figure 9: Example of a false positive highlight: Krug kill considered as a highlight.

Despite the promising results, some false positives were caused by moments with crowd noise or rapid camera movements, which the model mistakenly interpreted as high-importance highlights. Additionally, large quantities of light effects in the game also disturbed the model, leading to further misclassifications.

However, the model consistently captured high-importance moments, such as multi-kills, Baron steals, and team fights, confirming its effectiveness in detecting the most critical moments of a match.

By refining the model to better differentiate between subtle in-game events and non-relevant moments, and incorporating additional data to enhance its understanding of various highlight categories, future iterations could further improve the detection of these high-importance events.

6. Conclusions

This work presented a multi-modal highlight detection system for *League of Legends*, combining visual dynamics and caster audio to identify and segment key in-game moments. By leveraging video cues

such as optical flow and color intensity, alongside peaks in pitch and volume from commentary, the system effectively pinpointed events of strategic significance.

The system showed strong performance, particularly in recall (75.19%), capturing most high-impact events such as team fights and objective captures. However, its lower precision (61.78%) revealed a tendency to flag minor moments as highlights. With an F1-score of 67.83%, the system strikes a balance between detecting key events and the need for further refinement to reduce false positives. This balance demonstrates the system's ability to reliably identify key events while highlighting the need for refinement in filtering less relevant events.

Audio and visual cues complemented each other well, with cross-correlation analysis confirming their synchronization. Casters' pitch and volume provided emotional signals that helped identify moments of excitement, even when visual intensity was low. However, the reliance on casters introduced bias, as their reactions sometimes exaggerated the significance of events. For instance, casters might raise their voices simply to build excitement when two enemies head toward the same point, even if nothing significant happens. Tailoring audio models more specifically to e-Sports could mitigate this issue.

Overall, the system shows strong potential for advancing e-Sports analytics. With further improvements in precision, audio modeling, fine-tuning, and dataset expansion, it could support applications such as content creation, post-game review, and automated match summarization across competitive titles.

7. Future Work

While this work focuses on highlight detection using video and caster audio, future research could include information coming directly from the players, such as webcam footage or team voice chat, especially during high-pressure moments of a match.

Since highlights often represent the most intense and emotionally charged segments of a match, they offer a natural opportunity to analyze how players respond under pressure, handle stress, make decisions, and work as a team under pressure. Capturing facial expressions, tone of voice, or other behavioral cues during these moments could help reveal how individuals manage stress, make decisions, and collaborate as a team.

This kind of analysis could support coaching and performance optimization, helping teams better understand emotional resilience, stress responses, and player tendencies in critical scenarios. Although this study focused on League of Legends, the proposed multimodal approach could be generalized to other competitive video games where the combination of video and audio cues might enable automatic detection of crucial moments.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] Kempe-Cook, L., Sher, S., Su, N.: Behind the Voices: The Practice and Challenges of Esports Casters. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (2019)

[2] Hamari, J., Sjöblom, M.: What is eSports and Why Do People Watch It? Internet Research **27**(2), 211–232 (2017)

[3] Campbell, W., Goss, A., Trottier, K., Claypool, M.: Sports versus Esports - A Comparison of Industry Size, Viewer Friendliness, and Game Competitiveness. In: Global Esports: Transformation of Cultural Perceptions of Competitive Gaming, pp. 1–12. Bloomsbury Academic, London (2021)

[4] Fakazlı, A.: The Effect of Covid-19 Pandemic on Digital Games and eSports. In: International Journal of Science Culture and Sport, pp. 335–344 (2020).

[5] Ali, S., Tanweer, S., Khalid, S., Rao, N.: Mel Frequency Cepstral Coefficient: A Review. In: Proceedings of the 2021 EAI International Conference (2021).
[6] Hennequin, R., Khlif, A., Voituret, F., Moussallam, M.: Spleeter: A Fast and Efficient Music Source Separation Tool with Pre-trained Models. Journal of Open Source Software **5**, 2154 (2020).
[7] Zou, Y., Luo, Z., Huang, J.-B.: DF-Net: Unsupervised Joint Learning of Depth and Flow Using Cross-Task Consistency. arXiv preprint arXiv:1809.01649 (2018)
[8] Tanaka, M., Simo-Serra, E.: LoL-V2T: A Dataset for Esports Video and Text Processing. IEEE Transactions on Multimedia (2023)
[9] Junior, J.B.S., Campelo, C.E.C.: League of Legends Real-Time Result Prediction. In: XVI Brazilian Conference on Computational Intelligence (CBIC), Salvador, Brazil (2023)
[10] Yao, P.: Real-Time Analysis of Basketball Sports Data Based on Deep Learning. IEEE Transactions on Neural Networks and Learning Systems (2021).
[11] Livingstone, S.R., Russo, F.A.: The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A Dynamic, Multimodal Set of Facial and Vocal Expressions in North American English. PLoS ONE **13**(5), e0196391 (2018).
[12] Busso, C., Bulut, M., Lee, C., Narayanan, S.: IEMOCAP: Interactive Emotional Dyadic Motion Capture Database. Language Resources and Evaluation **42**(2), 335–359 (2008)
[13] Poria, S., Hazarika, D., Majumder, N., Cambria, E.: MELD: A Multimodal EmotionLines Dataset. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (2019)
[14] Cho, K., Lee, H.: DeepLeague: A Dataset for Esports Gameplay Analysis. In: Proceedings of the 32nd International Conference on Esports Data (2018)
[15] Chakraborty, S.: Speech-Emotion-Recognition. GitHub repository, https://github.com/Shreyasi2002/Speech-Emotion-Recognition--1, last accessed 2025/06/06
[16] Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated Residual Transformations for Deep Neural Networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5987–5995 (2017)
[17] Smith, R.: An Overview of the Tesseract OCR Engine. In: Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR), vol. 2, pp. 629–633 (2007)
[18] Farnebäck, G.: Two-Frame Motion Estimation Based on Polynomial Expansion. In: Scandinavian Conference on Image Analysis (SCIA), vol. 2749, pp. 363–370 (2003)
[19] He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016).