Proceedings of the International Conference on Industrial Engineering and Operations Management

Publisher: IEOM Society International, USA DOI: 10.46254/GC03.20260370
Published: February 2, 2026

A Novel Trimmed Mean Absolute Deviated Local Ternary Patterns for Sports Video Summarization

Ahmad Alhabib
Electrical and Computer Engineering Department
Wayne State University
MI, USA
FJ8971@Wayne.edu

Muteb Aljasem
Electrical and Computer Engineering Department
Bowling Green State University
Ohio, USA
aljasem@bgsu

Auner Gregory
Surgery and Biomedical Engineering Department
Wayne State University
MI, USA
gauner@wayne.edu

Zaid Aldoulah
Electrical and Electronics Engineering Department
Alasala Colleges
Dammam, Saudi Arabia
Zaid.aldoulah@alasala.edu.sa

Abstract

The popularity of sports and the associated commercial benefits have encouraged broadcasters to generate and circulate an enormous volume of sports videos across online digital platforms. Effective management and processing of such extensive content is challenging. This brings a demand for the development of effective and efficient summarization techniques to handle large video collections while maintaining viewer engagement and optimizing storage and transmission requirements. This study introduces an automated system for summarizing videos of multiple sports genres through excitement-based event detection. The proposed approach analyzes the audio stream of sports videos to identify exciting events, which are then used to create concise video summaries. We propose a novel feature space, trimmed mean absolute deviated-local ternary patterns (TMAD-LTP), to better capture the distinctive traits of exciting segments within audio stream. We used our TMAD-LTP features to train a binary Support Vector Machine classifier to distinguish between excited and non-excited frames. The video frames corresponding to the exciting audio frames are identified as keyframes. The prior and later frames around these keyframes are appended to produce meaningful video skims that are then arranged sequentially to generate the highlights. Experimental evaluation on a diverse set of cricket and soccer videos demonstrates the impressive performance of the proposed method over contemporary approaches by achieving an average accuracy of 97.71%. These impressive results validate the efficacy of our approach for generating more useful sports highlights.

Keywords
Excitement detection, key-frame detection, local ternary patterns, trimmed mean absolute deviation, video summarization.

1. Introduction
The rapid expansion of multimedia content on the internet has made it extremely challenging to manage and analyze the vast amount of available data. Among this content, sports videos represent one of the largest and fastest-growing categories. Sports broadcasters produce massive volumes of video daily, and processing or analyzing this data is time-consuming for both humans and automated systems. In modern, fast-paced life, viewers also prefer highlights to watching full-length matches. This creates a strong motivation to develop intelligent and efficient techniques for sports video analytics, driven by commercial value and widespread audience demand.

Video summarization (VS) has become a popular solution for condensing lengthy videos into short, meaningful summaries. It has broad applications across domains such as sports (Javed et al. 2019; Antonio et al. 2017), surveillance (Veesam et al. 2025; Hossain et al. 2025), healthcare (Shoumi et al. 2025; Oliveira et al. 2025), and media (Paul et al. 2025; Apostolidis et al. 2025). However, designing a generic summarization framework that works across multiple sports is highly challenging due to differences in game rules, temporal dynamics, and event structures. Most existing approaches (Min et al. 2017; Jingjing et al. 2018) are sport-specific and fail to generalize across different games. This research aims to overcome that limitation by proposing a generalized framework capable of summarizing videos from multiple sports. Moreover, many existing methods (Shanmukhappa et al. 2014; Evlampios et al. 2014) are computationally intensive, making them less practical for large-scale, real-time applications. To address this, our approach proposes a novel and lightweight audio-based feature to detect exciting segments efficiently, which are then used to generate game highlights.

Video summarization approaches can be categorized into static and dynamic techniques. Several studies (Issa and Shanableh, 2022a; Pan et al. 2022) have focused on extracting keyframes from sports videos to generate static summaries. Such techniques are storage-efficient as they only retain selected frames, but they fail to capture important cues such as audio, motion, and temporal continuity, making them less informative and engaging. On the contrary, studies (Javed et al. 2019; Javed et al. 2016; Banjar et al. 2024), which generate dynamic summaries, are capable of preserving audio and motion elements by selecting key video segments rather than static frames, resulting in a richer and more engaging viewing experience. However, such approaches demand more storage and resources. From a commercial perspective, user satisfaction takes precedence over computational efficiency, making dynamic summarization more suitable for sports applications, where excitement, continuity, and engagement are essential.

Effective key-event detection is crucial for the generation of a sports video summary. Prior research has explored the use of audio, visual, and combined audio-visual features for event detection and summarization. Existing audio-processing-based VS methods have used either time domain features (Shiqi et al. 2015), spectral features (Anant et al. 2015), with conventional machine learning (ML) or deep learning (DL), including end-2-end DL-based approaches. Visual modality-based approaches have used non-learning (Javed et al. 2016) as well as learning-based methods (Javed et al. 2019; Banjar et al. 2024) for summary generation. Learning-based VS approaches have either used handcrafted feature-based approaches (Javed et al. 2019; Javed et al. 2016;) or pure DL-based approaches (Hossain et al. 2025).

Summarizing sports videos poses significant challenges due to variations in game rules, lighting conditions, camera angles, shot types, event dynamics, and game structure. Further, most existing video summarization techniques (Min et al. 2017; Jingjing et al. 2018) are sport-specific, which limits their adaptability across different game types. The proposed framework addresses these challenges by introducing a method that can generate summaries for various sports in a consistent and generalized manner. Additionally, many current summarization approaches tend to be computationally intensive. The proposed method contributes to this direction by focusing on audio stream analysis to efficiently identify exciting segments within sports videos, thereby producing concise and informative highlights.

In this paper, we introduce an efficient and effective audio-based framework for sports video summarization. The method detects excitement levels in the audio track using the novel trimmed mean absolute deviation-local ternary patterns (TMAD-LTP) features. These features are used to train a binary SVM classifier to distinguish between excited and non-excited segments. The identified excited frames are mapped to their corresponding video frames, which are then combined with neighboring frames based on the desired summary length to form concise video skims. These skims are chronologically ordered to generate coherent highlights. Experimental evaluations on diverse cricket and soccer datasets demonstrate the effectiveness and efficiency of the proposed approach in

detecting key events and producing high-quality summaries. The major contributions of this research can be summarized as follows:

- We present a lightweight and effective audio-driven summarization framework to generate highlights of different sports.
- We propose a novel audio feature descriptor, TMAD-LTP, for extracting the salient attributes in exciting and non-exciting parts of an audio stream.
- Our method is robust to sports genre, game structure, background noise, broadcasters, editing effects, variations in camera angle, field colors, illumination conditions, and shot types.
- Rigorous experimentation on a diverse collection of Soccer and Cricket videos was performed to assess the efficacy of our VS framework.

2. Literature Review

This section critically investigates the state-of-the-art (SOTA) sports VS methods. Existing works have presented the audio-based, video-based, and audio-visual modalities-based VS methods using handcrafted features with conventional classifiers and end-to-end DL approaches.

Several studies relied on audio features to identify key events in sports videos. For example, (Baijal et al. 2015) used Mel-Frequency Cepstral Coefficients (MFCC) and delta-MFCC to detect excitement in rugby matches using a multi-stage classification approach. Shiqi et al. (2015) employed statistical rules to select key frames in tennis videos, ranking them based on audio energy. Kolekar et al. (2010) applied sequential association mining to link detected events with semantic concepts, while (Kolekar et al. 2011) used Bayesian Belief Networks for automatic excitement indexing. Islam et al. (2019) proposed a non-learning method using empirical mode decomposition for soccer event detection.

Other studies (Javed et al. 2019; Wang et al. 2016; Mendi et al. 2013; Nguyen et al. 2014) have relied primarily on visual cues for summarization. Javed et al. (2019) proposed a replay detection and key-event identification method based on gradient transitions and motion history images. Wang et al. (2016) designed a soccer annotation method that synchronized textual and visual events using field zone detection. Mendi et al. (2013) leveraged optical flow to compute motion features, while (Javed et al. 2016a) developed a replay-based summarization approach that identified replays using score caption detection. Nguyen et al. (2014) focused on emotionally intense scenes, such as crowd reactions and player expressions, to generate summaries of soccer videos.

A combination of audio and visual features has also been widely explored, as it generally improves accuracy at the expense of computational cost. Javed et al. (2016) combined audio and visual cues for cricket summarization, first identifying excited audio segments and then classifying the corresponding frames. Merler et al. (2017) analyzed golf videos using player reactions and commentator tone. Hasan et al. (2013) and (Raventós et al. 2015) proposed generic and soccer-specific methods, respectively, using low- and mid-level audio-visual descriptors. Tomoki et al. (2018) applied convolutional neural networks for soccer highlight generation, while (Merler et al. 2018) curated golf and tennis highlights using both modalities. Javed et al. (2019a) introduced an approach using Acoustic Local Binary Patterns and SVM classification, and (Khan et al. 2020) proposed a deep neural network framework for multi-sport summarization.

Other research explored event modeling techniques. Tavassolipour et al. (2014) utilized a Bayesian network with Markov-based segmentation for soccer summarization. Nguyen et al. (2013) employed logo transitions to identify replay segments, while (Vinay et al. 2016) leveraged contextual cues in basketball environments. Dongmahn et al. (2014) integrated live text and social media data for basketball highlights. Tomet al. (2013) used spatio-temporal event streams, (Trinh et al. 2016) combined Short-Time Fourier Transform and Gaussian Mixture Models for excitement detection, and (Pushkar et al. 2018) merged event-based and excitement-based cues for cricket video clipping.

3. Proposed Methodology

This section provides the details of the proposed audio-driven VS framework. Section 3.1 provides the detailed computation method of our novel TMAD-LTP features, followed by classification and summary generation mechanisms in Section 3.2.

3.1 Feature Extraction using novel TMAD-LTP

To develop a robust audio-driven sports VS method, we require audio features that can better extract the characteristics of exciting and non-excited audio chunks from the input stream. For this purpose, we introduced an acoustic TMAD-LTP feature to reliably capture the distinctive attributes of excited and non-excited frames in

outdoor environments with massive background noise. Shown in Figure 1 is the computation of TMAD-LTP features.

Let $A[f]$ represent the input audio stream consisting of F frames, where we formed a disjoint frame window comprising 9 frames. Since we proposed an improved local ternary pattern (LTP) feature, which combine 9 frames in one window, inspired by the two-dimensional LTP features computed for 2D images [25], thus, we also created a frame window in our TMAD-LTP descriptor using the nine frames, with the middle one "*m"* surrounded by four neighboring frames each in left and right side. This makes each frame window of length 9. Next, the MLTP approach generates the ternary codes for each of the 8 neighboring frames corresponding to the central frame in the frame window. We compare the value of neighboring frames at a time with the central frame adjusted by the threshold "μ_2". More precisely, if the given neighboring frame value is greater than the aggregate of the central frame and threshold, then the ternary code of 1 is assigned at that location. Next, given neighboring frame value is less than the difference between the central frame and the threshold, then the ternary code of -1 is assigned at that location. Finally, if the given frame value lies between the sum of the central frame and threshold and the difference of the central frame and threshold, then zero is assigned as a ternary code. We compute these ternary codes as follows:

$$C(f^i, m, \mu_2) = \begin{cases} +1, & f^i \geq m + \mu_2 \\ 0, & m - \mu_2 < f^i < m + \mu_2 \\ -1, & f^i \leq (m - \mu_2) \end{cases} \quad (1)$$

Here, $C(f^i, m, \mu_2)$represents the ternary codes, f^idenotes the neighboring frames, with *i* being the index of the corresponding neighbor, μ_2 is the threshold, and *m* is the central frame value within the given frame window.
To generate the MLTP features, we computed the difference of magnitude between *m* and f^i. In the baseline LTP features (Adnan et al., 2018), a hard-coded fixed value is used as a threshold, which makes LTP features unstable in the presence of background noise in the audio stream. Since field sports videos are played outdoors and contain much environmental noise and other distortions, thus, presents a demand to generate effective feature schemes more robust to background noise. Considering this limitation, we propose an adaptive threshold strategy, named trimmed mean absolute deviation (TMAD), in our feature descriptor to counter the aforementioned limitation. Specifically, our adaptive threshold computation mechanism for each frame window is shown in Eqs. (2)-(4).

$$\mu_1 = mean[w(z_1, z_2, z_3, \ldots\ldots, z_n)] \quad (2)$$

$$D_i[z_1, z_2, z_3, \ldots\ldots, z_n] = [z_1, z_2, z_3, \ldots\ldots, z_n] - \mu 1 \quad (3)$$

$$\mu_2 = mean\ (D_i[z_1, z_2, z_3, \ldots\ldots, z_n]) \quad (4)$$

First, we calculated the mean of all values, μ_1, in a frame window using Eq. (2). Next, we computed the difference of each frame with $\mu_1$as shown in Eq. (3). Finally, we computed the threshold using the average of all values obtained in Eq. (3). Thus, $\mu_2$represents the threshold used in Eq. (1) to generate our TMAD-based LTP codes. This TMAD adaptive threshold scheme enables our TMAD-LTP features to be robust against background noise and other distortions without any significant computational cost.

Next, we split our adaptive ternary codes into positive and negative binary patterns, C^+ and negative C^-. For the positive patterns, we set the value of 1 if the code is +1 and the rest values to zero. For the negative patterns, we set the value of 1 if the code is -1 and the rest values to zero as mentioned in Eqs. (5) and (6).

$$C^+(f^i, m, \mu_2) = \begin{cases} 1, & if\ C(f^i, m, \mu_2) = +1 \\ 0, & Otherwise \end{cases} \quad (5)$$

$$C^-(f^i, m, \mu_2) = \begin{cases} 1, & if\ C(f^i, m, \mu_2) = -1 \\ 0, & Otherwise \end{cases} \quad (6)$$

It is to be noted that uniform patterns can capture more salient attributes than non-uniform patterns in the input signal (Adnan et al., 2018); therefore, we created two uniform patterns, namely C_u^+ and C_u^-, and transformed them into decimal forms using Eqs. (7) and (8).

$$C_u^+(f^i, m, \mu_2) = \sum_{j=0}^{7} 2^j \times C^+(f^i, m, \mu_2) \quad (7)$$

$$C_u^-(f^i, m, \mu_2) = \sum_{j=0}^{7} 2^j \times C^-(f^i, m, \mu_2) \quad (8)$$

Next, we calculated the histograms for both positive and negative uniform patterns, namely, H_n^+ and H_n^-, as follows:

$$H_n^+(C_u^+, n) = \sum_{k=1}^{K} (C_k^+, n\ ) \tag{9}$$

$$H_n^-(C_u^-, n) = \sum_{k=1}^{K} (C_k^-, n\ ) \tag{10}$$

Where n represents the histogram bins. We conducted a thorough experiment to identify the number of bins suitable to capture all salient attributes. Specifically, our analysis concluded that the starting 20 uniform pattern bins of both H_n^+ and H_n^- combined are sufficient to gather maximum significant information from the audio. Thus, we collected the starting 10 bins each from $H_n^+(C_u^+, n)$ and $H_n^-(C_u^-, n)$, and concatenated them to form a 20-dimensional novel descriptor as:

$$TMAD - LTP = \left[H_n^+(C_u^+, n) \,||\, H_n^-(C_u^-, n)\right] \tag{11}$$

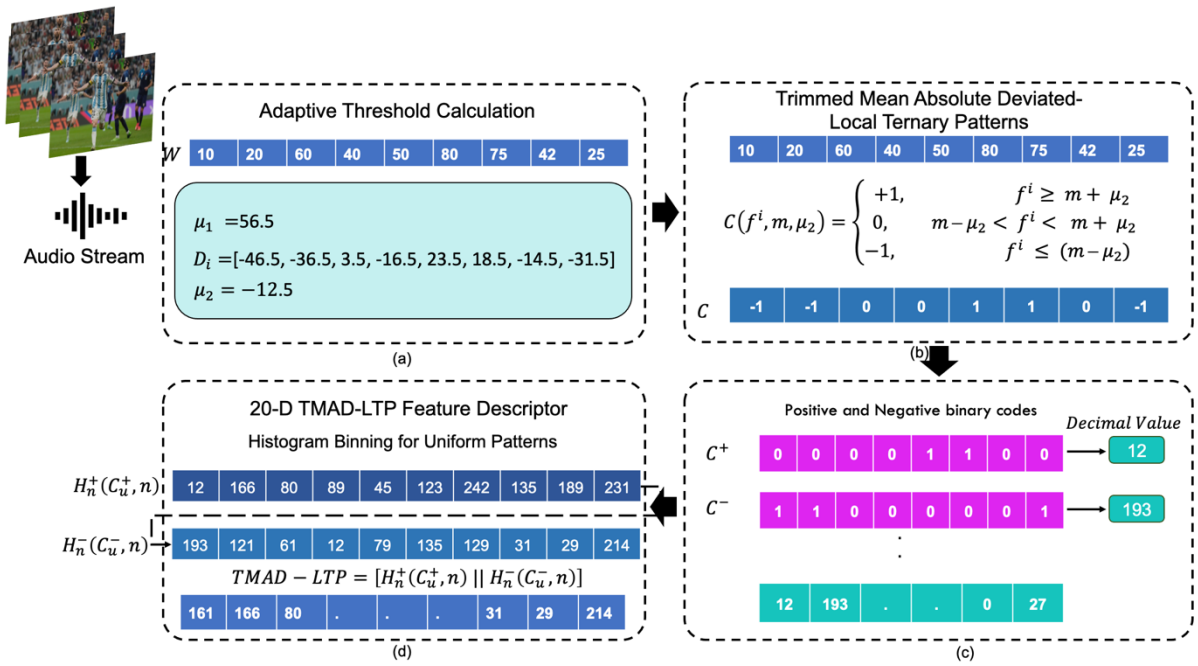

Figure 1. TMAD-LTP feature computation.

3.2 Classification and summary generation

Since our classification problem is binary, therefore, we used a Support Vector Machine (SVM) classifier due to its simplicity, effectiveness, strong generalization ability, robustness against overfitting, and consistent performance. We used our novel TMAD-LTP features to represent the audio stream. These features reliably extract distinct acoustic traits available in audience reactions, commentator tone, and excitement levels during significant moments in the game. SVM employs kernel functions to project the training feature vectors into a higher-dimensional space, where it determines the optimal separating hyperplane that maximizes class distinction. The model achieves this by minimizing a cost function, which, for excitement detection, is defined as follows:

$$min_{y,s,\xi} \frac{1}{2}\|y\|^2 + P\sum_{t=1}^{T} \xi_q \tag{12}$$

w.r.t

$$a_i(y.\phi(v_i) + s) \geq 1 - \xi_q, \qquad \xi_q \geq 0 \tag{13}$$

where y and s represent the hyperplane parameters, ξ_q are slack variables allowing soft margin classification, P is the penalty term, and $\phi(v_i)$ denotes the kernel mapping of the input feature vector v_i.

Once trained, the SVM classifier predicts the excitement label for each audio frame. Consecutive frames classified as excited are grouped with their corresponding prior and later frames chronologically to form an excitement segment, which serves as an indicator of potential key events. The proposed framework uses the detected key events to generate concise video chunks, which are arranged in sequential order to generate the final summary of the input sports video.

4. Dataset

This section provides the details of the data repository used to investigate the competence of the proposed method. We have used this dataset (Banjer et al. 2024), comprising YouTube videos of Cricket and Soccer. This collection is diverse in terms of genre, broadcasters, editing effects, illumination conditions, including daylight matches and night games, formats, duration of games, and tournaments, etc. It is a common practice in sports video summarization to use YouTube videos for performance assessment, as this practice is also adopted by comparative studies.

The collection of Cricket includes the audio streams of videos comprising Test, One-day, and Twenty20 matches. Moreover, these videos were collected from different tournaments, including bilateral series, tri-lateral series, Leagues, and multi-national series, including the World Cup. The collection of soccer includes matches from the 2014 and 2018 World Cups, the La Liga Cup, Euro Cups 2012 and 2016. The extracted audios of these videos, comprising the exciting and non-exciting clips, were used for assessment. All audio frames containing loud spectators and commentators' cheers were marked as exciting, and the rest as non-exciting. Both Cricket and Soccer have 1782 audio clips with a total of 3564. For both sports, we used 1142 audios for training and the rest 640 for evaluation. Both the training and testing collections contain an equal number of exciting and non-exciting samples.

5. Results and Discussion

This section has provided the details of different experiments conducted for performance assessment, evaluation metrics, and a discussion of the results.

5.1 Evaluation Metrics

We have used objective metrics to evaluate the performance of the proposed method. For this purpose, we used precision, recall, F-1 score, accuracy, and error rate computed through true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

In our case, precision (Pr) represents the capability of our framework to correctly identify the key-events from the total number of events and is computed as:

$$\boldsymbol{P_r} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \tag{14}$$

where TP are the true key-events detected, and FP are the non-key-events falsely classified as the key-events.

Recall (Re) represents the capability of our framework to correctly detect the key-events from the total key-events in the video. Recall is computed as follows:

$$\boldsymbol{R_e} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \tag{15}$$

where FN represents the true key-events that are misclassified as non-key-events by the classifier.

Accuracy (Ac) represents the correctly labelled key-events/non key-events by the classifier to the total number of events. We computed the accuracy as follows:

$$\boldsymbol{A_c} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{Total\ Samples}} \tag{16}$$

Error (Er) is the ratio of misclassified key-events/non key-events to the total number of events and computed as:

$$\boldsymbol{E_r} = \frac{\boldsymbol{FP+FN}}{\boldsymbol{Total\ Samples}} \tag{17}$$

F1-score is the measure of the weighted average of Pr and Re. It has been observed that some methods achieve better precision values, whereas some have better recall values. In these scenarios, the F1-score provides the true performance measure. F1-score of 1/0 illustrates the best/worst classification accuracy, and is computed as follows:

$$\boldsymbol{F1-score} = \boldsymbol{2\times\frac{(P_r \times R_e)}{P_r+R_e}} \tag{18}$$

5.2 Assessment of Proposed Method

To test the efficacy and generalizability of our approach for reliable VS generation, we assessed our method on the audio streams of two different genres of sports, i.e., Soccer and Cricket. We extracted the TMAD-LTP features from the audio stream and used them to train the SVM to determine the exciting audio segments. In the first stage, we used our method to train on the TMAD-LTP features computed from the 1142 audio segments of soccer videos and evaluated the performance on the remaining 640 audios. Our approach attained the average precision, recall, F1-score, accuracy, and error rate of 97.89%, 97.46%, 97.68%, 97.54%, and 2.46%, as shown in Table 1. Similarly, in the next stage, we examined our method's classification ability on the audio streams of Cricket videos. Again, 1142 audio segments were used for training, and the rest unseen 640 audio streams for testing. On Cricket, we obtained the average precision, recall, F1-score, accuracy, and error rate of 98.17%, 97.45%, 97.92%, 97.87%, and 2.13%. The outcomes of this experiment revealed that our TMAD-LTP features can better capture

salient attributes in the exciting and non-exciting audios, and with SVM, detect the key-events effectively. Moreover, these remarkable results on two entirely different genres of sports demonstrate the generalization power of our VS method.

Table 1. Assessment of the proposed method.

Sports Genre	Precision (%)	Recall (%)	Accuracy (%)	Error (%)	F1-Score (%)
Soc-1	98	98.20	98.22	1.78	98.10
Soc-2	97.10	96.98	97.72	2.28	97.06
Soc-3	98	98.22	98.25	1.75	98.11
Soc-4	97.48	96.58	97.32	2.68	97.08
Soc-5	98.90	97.30	96.20	2.8	98.10
Average (Soccer)	**97.89**	**97.46**	**97.54**	**2.46**	**97.68**
Cric-1	98.9	98.52	98.21	1.79	98.70
Cric-2	97.25	96.26	97.22	2.78	96.75
Cric-3	97.50	98.61	98.22	1.78	98.04
Cric-4	98.30	95.92	97.87	2.13	94.13
Cric-5	98.90	97.91	97.82	2.18	98.56
Average (Cricket)	**98.17**	**97.45**	**97.87**	**2.13**	**97.92**

5.3 Performance comparison with contemporary approaches

This comparative experiment was designed to assess the efficacy of our approach against the contemporary sports video summarization techniques (Javed et al. 2016; Merler et al. 2018; Raventos et al. 2015; Trinh et al. 2016; Javed et al. 2019; Islam et al. 2019; Shingrakhia et al. 2022; Banjer et al. 2024), and the results are mentioned in Table 2. The results of this comparative analysis revealed that the performance of method (Islam et al. 2019) was the lowest, by attaining an accuracy of 61.22%, (Merler et al. 2018) attained better than (Islam et al. 2019), but with 81.12% accuracy, came in second last spot. The rest of the comparative methods achieved the accuracies in the range of 92% to 97%, which shows better performance, but a few methods are limited to generating summaries of a single sport category. This shows that methods of such studies are tailored to accommodate a specific sport and fail to apply to multiple sports. On the contrary, our method has no dependency on any game and is capable of producing VS of multiple sports. Moreover, our VS method attained the best performance with an average accuracy of 97.71%, outperforming all the comparative approaches. The outcome of this comparative analysis signifies better generalization ability and performance of our approach over contemporary VS techniques.

Table 2. Comparative analysis with existing VS methods.

Video Summarization Methods	Precision (%)	Recall (%)	Accuracy (%)	Error (%)	F1-score (%)
(Javed et al. 2016)	91.87	89.85	95.01	4.99	90.84
(Merler et al. 2018)	80.65	81.21	81.12	18.88	80.93
Raventós et al. [13]	88	93	90.69	9.31	90.43
(Trinh et al. 2016)	91.38	92.38	91.67	8.33	91.83
(Javed et al. 2019)	98.80	97.60	97.70	2.36	98
(Islam et al. 2019)	60.34	61.10	61.22	38.80	60.72
(Shingrakhia et al. 2022)	96.82	95.41	96.32	3.68	95.67
(Banjer et al. 2024)	98.10	97.17	97.70	2.30	97.63
TMAD-LTP+SVM (Proposed)	**98.03**	**97.46**	**97.71**	**2.29**	**97.75**

5.4 Discussion

The proposed acoustics analysis-based sports VS approach effectively addresses the challenges of detecting key events in noisy outdoor environments using the novel TMAD-LTP features. The baseline LTP features are highly sensitive to background noise due to the use of a fixed threshold in computation. To better tackle this limitation, we introduced an adaptive thresholding strategy in our TMAD-LTP that dynamically adjusts to such signal variations locally. This ensures more stable and distinctive feature extraction from crowd reactions, commentary tone, and environmental sounds. The use of an SVM classifier complements the proposed features by providing

a robust and computationally efficient solution for binary excitement detection. Its kernel-based mapping improves class separation in non-linear feature spaces, resulting in enhanced accuracy and generalization, even with moderate training data.

The fact that our method only analyzes the audio stream for summary generation makes it computationally efficient for real-time sports video analysis compared to multimodal approaches. The summary generation mechanism of our method ensures maintaining the sequence of video skims, thereby improving the viewer experience. However, the system may sometimes misclassify non-exciting segments with high crowd noise or commentary cheering as key events, leading to false detections. Such misclassification scenarios can be better tackled by employing audio-visual features in the method. Despite this limitation, the proposed TMAD-LTP approach demonstrates strong performance for exciting events detection and summary generation, highlighting its potential for real-time applications. By offering a noise-resilient, adaptive, and efficient solution for sports VS, it shows that audio-driven methods can provide an effective alternative to complex multimodal systems.

6. Conclusion

This paper has presented an audio modality analysis-based approach for summarizing videos of multiple sports. We proposed a novel acoustics descriptor, trimmed mean absolute deviated-local ternary patterns, for audio representation and trained the SVM for classifying the exciting and non-exciting audio clips. The proposed method is tested on the audio dataset comprising soccer and cricket videos. Experimental results signify that the proposed framework can detect key-events with better accuracy due to the competency of our TMAD-LTP descriptor for capturing all salient traits from the exciting and non-exciting audios. Further, our approach is lightweight because of analyzing audio streams only for VS. Performance comparison against the contemporary VS methods indicates the efficacy and generalizability of our method for video summarization. Our technique sometimes produces false positives during any non-exciting part of the game, in case of crowd cheers loudly during camera focus on spectators, which can be better tackled by analyzing the visual stream as well.

References

Adnan, S. M., Irtaza, A., Aziz, S., Obaid Ullah, M., Javed, A., and Mahmood, M. T., “Fall Detection through Acoustic Local Ternary Patterns,” *Applied Acoustics*, Vol. 140, pp. 296–300, 2018.

Antonio, T., Nakashima, Y., Sato, T., Yokoya, N., Linna, M., and Rahtu, E., “Summarization of User-Generated Sports Video by Using Deep Action Recognition Features,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Apostolidis, E., Balaouras, G., Patras, I., and Mezaris, V., “Explainable Video Summarization for Advancing Media Content Production,” in *Encyclopedia of Information Science and Technology*, 6th ed., IGI Global, pp. 1–24, 2025.

Baijal, A., Cho, J., Lee, W., and Kim, B.-S., “Sports Highlight Generation Based on Acoustic Events Detection: A Rugby Case Study,” *IEEE International Conference on Consumer Electronics (ICCE)*, pp. 20–23, 2015.

Banjar, A., Dawood, H., Javed, A., and Zeb, B., “Sports Video Summarization Using Acoustic Symmetric Ternary Codes and SVM,” *Applied Acoustics*, Vol. 216, 109795, 2024.

Decroos, T., Davis, J., Van Haaren, J., and others, “Predicting Soccer Highlights from Spatio-Temporal Match Event Streams,” *AAAI Conference on Artificial Intelligence*, 2017.

Dongmahn, S., Kim, D., Park, H., and Kim, H., “User Generated Highlight System for Baseball Games with Social Media Activities,” *IEEE International Conference on Consumer Electronics (ICCE)*, 2014.

Evlampios, A., and Mezaris, V., “Fast Shot Segmentation Combining Global and Local Visual Descriptors,” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

Hasan, T., Bornemann, H., Sangwan, A., Hansen, J. H. L., “Multi-Modal Highlight Generation for Sports Videos Using an Information-Theoretic Excitability Measure,” *EURASIP Journal on Advances in Signal Processing*, 2013:173, 2013.

Hossain, S., Kaushik, D., Saadman, S., and Iqbal, H. S., “A Hybrid Deep Learning Framework for Daily Living Human Activity Recognition with Cluster-Based Video Summarization,” *Multimedia Tools and Applications*, Vol. 84, No. 9, pp. 6219–6272, 2025.

Islam, M. R., Paul, M., Antolovich, M., and Kabir, A., “Sports Highlights Generation Using Decomposed Audio Information,” *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 579–584, 2019.

Issa, O., and Tamer, S., “CNN and HEVC Video Coding Features for Static Video Summarization,” *IEEE Access*, Vol. 10, pp. 72080–72091, 2022.

Javed, A., Bajwa, K. B., Malik, H., and Irtaza, A., “An Efficient Framework for Automatic Highlights Generation from Sports Videos,” *IEEE Signal Processing Letters*, Vol. 23, No. 7, pp. 954–958, 2016.

Javed, A., Bajwa, K. B., Malik, H., Irtaza, A., and Mehmood, M. T., “A Hybrid Approach for Summarization of Cricket Videos,” *IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, pp. 1–4, 2016.

Javed, A., Irtaza, A., Khaliq, Y., Malik, H., and Mahmood, M. T., "Replay and Key-Events Detection for Sports Video Summarization Using Confined Elliptical Local Ternary Patterns and Extreme Learning Machine," *Applied Intelligence*, Vol. 49, No. 8, pp. 2899–2917, 2019.

Javed, A., Irtaza, A., Malik, H., Mahmood, M. T., and Adnan, S., "Multimodal Framework Based on Audio-Visual Features for Summarisation of Cricket Videos," *IET Image Processing*, Vol. 13, No. 4, pp. 615–622, 2019.

Khan, A., Shao, J., Waqar, A., and Saifullah, T., "Content-Aware Summarization of Broadcast Sports Videos: An Audio–Visual Feature Extraction Approach," *Neural Processing Letters*, 2020.

Ma, J., Wang, S., Wang, H., Yang, J., and Tan, Y.-P., "Video Summarization via Multiview Representative Selection," *IEEE Transactions on Image Processing*, Vol. 27, No. 5, pp. 2144–2156, 2018.

Merler, M., Joshi, D., Nguyen, Q.-B., Hammer, S., Kent, J., Smith, J. R., and Feris, R. S., "Automatic Curation of Golf Highlights Using Multimodal Excitement Features," *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

Merler, M., Mac, K.-N. C., Joshi, D., Nguyen, Q.-B., Hammer, S., and Kent, J., "Automatic Curation of Sports Highlights Using Multimodal Excitement Features," *IEEE Transactions on Multimedia*, 2018.

Min, S., Farhadi, A., Taskar, B., and Seitz, S., "Summarizing Unconstrained Videos Using Salient Montages," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, No. 11, pp. 2259–2270, 2017.

Nguyen, N., and Yoshitaka, A., "Shot Type and Replay Detection for Soccer Video Parsing," *IEEE International Symposium on Multimedia (ISM)*, 2013.

Oliveira, J. D., Henrique, D. P., Helena, A. S., Ulbrich, D. P. S., Couto, J. C., Marcelo, A., Joaquim, S., Costa, M. M., Fabio, D. F., Tabalipa, O., and Nogueira, R. F., "Development and Evaluation of a Clinical Note Summarization System Using Large Language Models," *Communications Medicine*, Vol. 5, No. 1, 376, 2025.

Pan, Y., Hou, O., Ye, Q., Li, Z., Wang, W., Li, G., and Chen, Y., "Exploring Global Diversity and Local Context for Video Summarization," *IEEE Access*, Vol. 10, pp. 43611–43622, 2022.

Paul, J., Roy, A., Mitra, A., and Sil, J., "HyV-Summ: Social Media Video Summarization on Custom Dataset Using Hybrid Techniques," *Neurocomputing*, Vol. 614, 128852, 2025.

Pushkar, S., Hemant, S., Apaar, B., Verma, D., Elmadjian, L., Raman, B., and Turk, M., "Automatic Cricket Highlight Generation Using Event-Driven and Excitement-Based Features," *CVPR Workshops*, 2018.

Raventós, A., Quijada, R., Torres, L., and Tarrés, F., "Automatic Summarization of Soccer Highlights Using Audio-Visual Descriptors," 2015.

Shanmukhappa, A., and Vilas, N., "Entropy Based Fuzzy C Means Clustering and Key Frame Extraction for Sports Video Summarization," *IEEE International Conference on Signal and Image Processing (ICSIP)*, 2014.

Shingrakhia, H., and Patel, H., "SGRNN-AM and HRF-DBN: A Hybrid Machine Learning Model for Cricket Video Summarization," *The Visual Computer*, Vol. 38, No. 7, pp. 2285–2301, 2022.

Shiqi, T., and Zhang, M., "Summary Generation Method Based on Audio Feature," *IEEE International Conference on Software Engineering and Service Science (ICSESS)*, 2015.

Shoumi, M., Hamdhana, D., Harada, K., Oshita, H., Sakashita, S., and Inoue, S., "Extraction and Summarization from Visiting Nurse Transcriptions Using Improved Prompt Techniques," *International Journal of Activity and Behavior Computing*, Vol. 2025, No. 1, pp. 1–50, 2025.

Tavassolipour, M., Kasaei, S., and Mahmoodi, K., "Event Detection and Summarization in Soccer Videos Using Bayesian Network and Copula," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 24, No. 2, pp. 291–304, 2014.

Tomoki, H., Takahashi, S., Ogawa, T., and Haseyama, M., "Estimation of Important Scenes in Soccer Videos Based on Collaborative Use of Audio-Visual CNN Features," *IEEE Global Conference on Consumer Electronics (GCCE)*, 2018.

Trinh, T. D., Ma, X., Lee, H. K., Kim, J. Y., and Cho, S.-H., "Cheering Event Detection in Basketball Audio Stream Using Adaptive GMM Model and Low Rank Matrix Recovery," *Journal of Korean Institute of Information Technology*, Vol. 14, No. 10, 2016.

Veesam, S., and Aravapalli, R. M., "Design of an Integrated Model for Video Summarization Using Multimodal Fusion and YOLO for Crime Scene Analysis," *IEEE Access*, 2025.

Vinay, B., Pantofaru, C., and Essa, I., "Leveraging Contextual Cues for Generating Basketball Highlights," *Proceedings of ACM Multimedia*, 2016.

Biographies

Ahmad Alhabib earned a bachelor's degree in electrical engineering, with a concentration in power and energy, from Arizona State University. He received a master's degree in electrical engineering with a specialization in robotics from Wayne State University, where he is currently pursuing a Ph.D. His research interests include solid-state electronics, smart sensors, robotics, and artificial intelligence applications.

Muteb Aljasem is an assistant professor in school of engineering at Bowling Green State University, where he teaches courses in computer, electronic and robotics engineering. He earned a bachelor of science degree in electrical engineering from West Virginia University, a master of science degree in electrical engineering from University of Michigan, and a Ph.D. in electrical engineering from Wayne State University. His research interests centre on machine learning, computer vision, signal processing, and cybersecurity.

Gregory W. Auner earned his Ph.D. in Physics from Wayne State University in 1990. He is a Professor of Surgery at Wayne State University and also holds appointments in Electrical and Computer Engineering and Physics. Dr. Auner founded and directs the Smart Sensors and Integrated Microsystems (SSIM) program at Wayne State. His research includes microsystems, biomedical sensors, nanotechnology, and intelligent deep-learning systems for medical and environmental applications.

Zaid Aldoulah is an Assistant Professor at Alasala University in Dammam, Saudi Arabia. He received his Bachelor of Science, Master of Science, and Ph.D. degrees in Electrical Engineering from The University of Toledo, USA. His doctoral research focused on Artificial Intelligence, and his current work extends AI applications to electrical and biomedical engineering. Dr. Aldoulah is committed to advancing research and education that integrates intelligent systems with engineering innovation.