Applied Acoustics 216 (2024) 109795

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Sports video summarization using acoustic symmetric ternary codes and SVM

Ameen Banjar[a,*], Hussain Dawood[b], Ali Javed[c], Bushra Zeb[c]

[a] *Department of Information Systems and Technology, University of Jeddah, Jeddah, Saudi Arabia*
[b] *Department of Information Engineering Technology, National Skills University Islamabad, Pakistan*
[c] *Department of Software Engineering, University of Engineering and Technology, Taxila, Pakistan*

ARTICLE INFO

Keywords:
Event detection
Excitement detection
SVM
Symmetric ternary codes
Video summarization

ABSTRACT

Broadcasters produce and transmit a vast number of sports videos in cyberspace due to immense viewership and potential commercial benefits. The analysis and processing of such a huge amount of video content are very challenging. This situation demands the development of effective and efficient summarization methods to manage the massive sports video repository while keeping the viewer's interest along with potential storage and transmission benefits. This paper presents an automated summarization framework based on excitement detection for sports videos i.e., cricket, soccer, etc. The audio stream of the sports video is analyzed to capture the significant events that are then used to produce the concise video. For effective representation of audio signals, we proposed an acoustic feature descriptor symmetric ternary codes and used them to train a binary Support Vector Machine classifier for excitement detection. Each audio frame is labeled as either an excited audio frame or a non-excited audio frame. The video frames corresponding to the excited audio frames represent the key-events in the sports videos and are marked as the key-frames. Each key-frame is appended with the neighboring frames to produce video skims for each key-event based on the user's required summary length. Finally, these video skims are sequentially arranged to produce the user-driven video summary. We evaluated our highlights generation method on our own diverse YouTube dataset of cricket and soccer videos, and a largescale SoccerNet corpus of soccer videos. The average accuracy of 97.7% and 91.23% on both datasets confirms the reliability of our method in terms of key-event detection for sports highlight generation.

1. Introduction

The exponential growth of multimedia content online makes it difficult to analyze and process such a massive content. Sports videos contribute to a large collection of available multimedia content in cyberspace. Sports broadcasters produce a massive number of videos daily. The analysis and processing of the available sports videos is a taxing activity for both humans and machines. Moreover, users don't have enough time to watch full-length sports videos rather they prefer to watch the highlights of the game. There is a strong potential and motivation to develop effective techniques for sports video content management due to commercial benefits and massive viewership. Video summarization techniques are frequently used to address the above-mentioned issues by providing a succinct representation of long-duration videos. Video summarization applications can be found in sports [1,2], surveillance [3,4], healthcare [5,6], media [7,8], and so on.

Event detection is a significant task in high-level semantic indexing and selective browsing of videos. Existing approaches for event detection and video summarization have used audio features, visual features, or a combination of both. Existing techniques [13–16,40] have used audio features for key-events detection in sports videos. Baijal et al. [13] used Mel-Frequency Cepstral Coefficients (MFCC) and delta-MFCC to detect the exciting segments in rugby videos. A multi-stage classification technique was employed for key acoustic event detection that was used to generate the game highlights. Tang et al. [14] used statistical rules in the first stage to detect key frames from different kinds of shots in tennis videos. In the second stage, important coefficients were assigned to these keyframes based on the audio energy feature. These keyframes were used to create the summarized video. Kolekar et al. [15] presented an automated highlights generation technique for broadcast sports video sequences extracted from the events and semantic concepts. Each sequence after extraction from the video was classified into a concept via

[*] Corresponding author.
E-mail address: abanjar@uj.edu.sa (A. Banjar).

https://doi.org/10.1016/j.apacoust.2023.109795
Received 26 June 2023; Received in revised form 15 November 2023; Accepted 29 November 2023
Available online 15 December 2023

sequential association mining. Kolekar [16] used the probabilistic Bayesian belief network technique for the automatic indexing of excited clips of sports videos. Islam et al. [40] employed the non-learning technique based on empirical mode decomposition for identifying the significant events in soccer videos.

Existing literature [1,17–20] has also used visual features to produce the highlights for sports videos. Javed et al. [1] proposed a replay and key-events detection method for sports video summarization. For this purpose, a thresholding-based gradient transition detection technique was employed for replay detection. Later, the Gaussian mixture model (GMM) was employed using the candidate key-events from replay frames for silhouette extraction and motion history image (MHI) generation against each key-event. Next, Confined Elliptical Local Ternary Patterns were used to extract the features from these MHIs and train the Extreme learning machine (ELM) classifier to identify the key-events. Wang et al. [17] presented a soccer video event annotation approach to synchronize video events with the text descriptions using high-level semantics and coarse-time constraint. A circle detection method was proposed to identify the soccer field zones. Mendi et al. [18] used motion features to develop a video summarization technique. The optical flow method was used to compute the motion metrics that were then used to detect the key-frames in the video. This approach is limited due to the dependency of key-frames on the perceived motion pattern of the video. Javed et al. [19] presented a replay detection-based summarization method for sports videos. In the first stage, gradual transitions were detected from the input videos using a dual threshold-based technique to identify the candidate replay segments. In the second stage, these candidate frames were analyzed to identify the live/replay video frame based on the presence/absence of score captions. Nguyen et al. [20] proposed a summarization method using intensive competitive scenes, players and audience reactions, and emotive instants to capture significant events in soccer videos. Javed et al. [41] presented a sports video analysis method to classify the shots via a lightweight Convolutional Neural Network (CNN) model. Next, local-octa patterns were proposed to train the ELM for replay detection. This method is more taxing due to the employment of two separate methods for shot classification and replay detection.

Existing literature [21–25,38,39] has also used audio-visual features to develop more effectual summarization approaches for sports videos. The increased accuracy is normally achieved at the expense of computational complexity. Javed et al. [21] developed an audio-visual features-oriented summarization framework for cricket videos. The audio stream was analyzed in the first stage where rule-based induction was applied to identify the excited audio segments. In the second stage, excited video frames analogous to the excited audio clips were fed to a decision tree to summarize the cricket videos. Merler et al. [22] proposed an automated highlights generation system for auto-curating sports highlights on golf and tennis videos by using both audio and visual features. The information on player's reactions, spectators, and commentators was used to find the exciting segments of the game. In addition, game statistics including the player names and hole numbers were also used for key-frames detection. Hasan et al. [23] used audio-visual features to present a sports video summarization method. Similarly, Raventós et al. [24] used audio-visual features to develop a video summarization technique for soccer. Low- and mid-level features were used to assign a relevance score to each shot. Finally, shots of significant scores were chosen to create the summarized video. Tomoki et al. [25] used audio-visual features to develop a CNN-oriented soccer video summarization approach. Javed et al. [38] presented an audio-visual features-oriented approach to summarize sports videos. Acoustic local binary patterns were extracted from the audio stream of sports video to train the SVM for the detection of excited audio segments. The excited audio frames were then used to choose the candidate key-video frames that were then fed to a decision tree-based classifier for key-events detection in the input cricket videos. Likewise, Khan et al. [39] developed an audio-visual feature-based deep neural network method to summarize sports videos.

Existing works also presented video summarization methods for a single sport. Tavassolipour et al. [27] presented a Bayesian network approach to identify important events for the summarization of soccer videos. In the first step, shot boundaries were identified for temporal video segmentation to create meaningful semantic units using the Markov model. In the next step, features were extracted from each semantic unit and trained on a Bayesian network to get the high-level semantic features for video highlights generation. Nguyen et al. [28] presented a replay-based video summarization for soccer. Histogram difference and contrast features were used to select the logo frames. The candidate frames between two successive logo frame transitions were identified as the replay segment. Bettadapura et al. [29] used leverage contextual cues from the game-playing environment to detect the exciting segments that were used to produce the summarized video for basketball. Seo et al. [30] proposed a method for basketball highlights that provided a live text scoreboard and related social media. Decroos et al. [31] developed a soccer video summarization approach by collecting leverages spatiotemporal event streams during the game. Trinh et al. [32] proposed an excitement detection method using the combination of Short Time Fourier Transform (STFT) bin strength, Low-Rank Matrix Recovery, and adaptive Gaussian Mixture Mode. Pushkar et al. [33] proposed a model that considered both events-based and excitement-based features to identify significant events in a cricket video. Shingrakhia et al. [42] have presented a hybrid method for the summarization of cricket videos. Firstly, an adapted threshold-based approach was used for the extraction of excited audio clips. Later, a stacked gated recurrent neural network was used for key event detection. Similarly, [34,43] also employed DL-based video summarization approaches for single sports only.

The diversity among multiple sports in terms of game rules, temporal structure, recurrent events, etc., introduces a massive challenge to develop a generic summarization framework for sports videos. Most of the existing video summarization techniques [9,10] have been designed for only one sports category. The proposed research work addresses this problem by developing a generic video summarization framework for multiple sports. In addition, video summarization approaches [11,12] have limitations of large overhead in terms of computational complexity. Therefore, there exists a need to develop efficient sports video summarization approaches. The proposed method addresses this problem by using computationally efficient audio features to detect the exciting segments in sports videos that are later used to create concise videos. In this paper, we presented an effective and efficient sports video summarization method by analyzing the audio streams. We analyzed the audio stream of the sports videos to develop an efficient video summarization framework. We proposed a novel feature descriptor to train the SVM for excitement detection. The excited audio frames are used to choose the key video frames. These key video frames are combined with their neighboring frames based on user-specified summary length to generate video skims for each key-event. Finally, we arranged these video skims in chronological order to create the highlights. The main contributions of the proposed work are:

1. We propose a novel acoustic symmetric ternary codes (STC) feature descriptor for effective representation of the audio.
2. We present an effective and efficient summarization method based on user-defined summary length for sports videos by detecting the excitement score in the audio stream.
3. Rigorous experimental validation on two different datasets reveals the efficacy of our method for sports video summarization.

2. Proposed method

The proposed sports video summarization approach based on novel time domain acoustic features is depicted in Fig. 1. The audio stream of the sports video is partitioned into an audio frame set consisting of 100

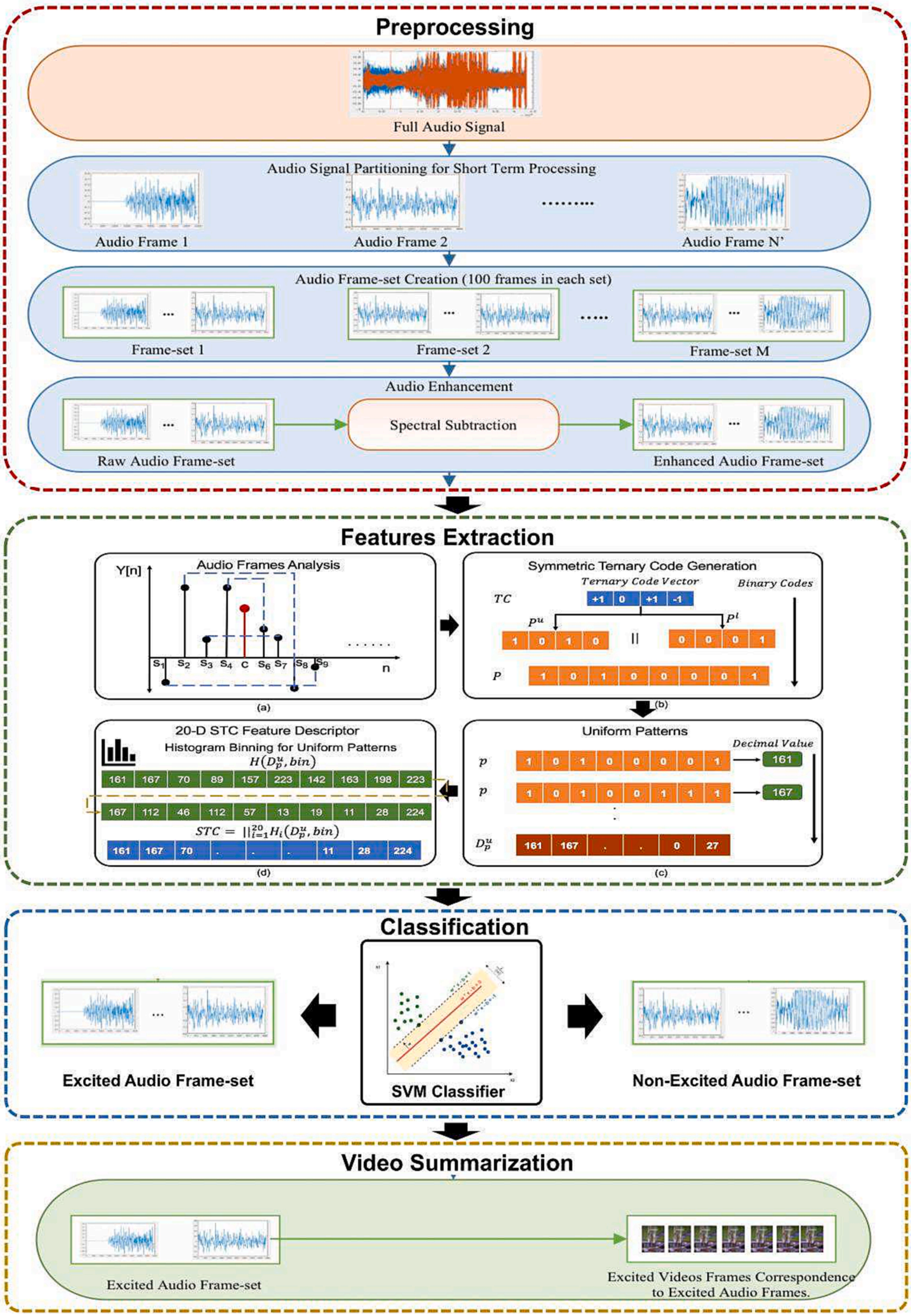

Fig. 1. Process Flow of the Proposed Excitement-based Key-Audio Frame Detection Framework.

audio frames. Next, the spectral subtraction technique is employed for enhancement of these audio frames-set. Later, we extract the STC features from the enhanced audio frames and train the SVM for excitement detection. The audio frames containing the excitement are marked as the key-audio frames. The video frames parallel to the excited audio frames are used to generate the summarized video for sports.

2.1. Pre-processing

In this step full-length audio signal $Y = \{y(i)\}_{i=1}^{i=N}$ from sports videos are extracted having videos frames $\{K(i)\}_{i=1}^{i=N}$. The audio signal $Y = \{y(i)\}_{i=1}^{i=N}$ is divided into various non-overlapping frames F^i of 9 samples each. Further, 100 such audio frames are combined to make an audio frameset. The fact that audio signals are non-stationary and change swiftly due to the variation in commentary and audience cheer makes it necessary to partition the audio signal during processing.

Audio signals usually contain massive background noise making it difficult to process the audio. Sports videos also contain immense background noise that must be suppressed to effectually process the audio for the identification of excited segments. In the present work, the spectral subtraction method [24] is employed to clean the background noise by processing each audio frame in the corresponding frame set.

2.2. Features extraction

Effective feature extraction is vital to attain better classification performance. For the effective representation of an input audio signal $Y[n]$, we propose acoustic time domain symmetric ternary code features. STC features are calculated locally by encoding each frame of the audio signal $Y[n]$. Firstly, we divide the $Y[n]$ into n number of non-overlapping frames $F^{(i)}$ where each frame contains 9 samples including the central sample c. We tried different numbers of samples in a frame and selected the frame of 9 samples after extensive experimentations. Hence, each frame $F^{(i)}$ comprises a central sample c and four closest neighbors on each side (left and right) of c represented by s^k, where k represents the neighbor index around c as shown in Fig. 2. Next, we compare the values of symmetric neighboring samples surrounding the central sample as shown in Fig. 2. More explicitly, we compare the values of s^1 and s^9, s^2 and s^8, s^3 and s^7, and s^4 and s^6 around c (central or 5th sample in each frame), as illustrated in Fig. 2. This comparative analysis of the values of respective symmetric neighbors (s^k) as mentioned above is quantized to 1 if the value of the left symmetric neighbor is greater than twice the value of the right symmetric neighbor i.e., $s^k_{left} > 2 \times s^k_{right}$, -1 in case of the value of right symmetric neighbor is greater than twice the value of left symmetric neighbor i.e., $s^k_{left} < 2 \times s^k_{right}$, and zero (0) otherwise. Since the minor change in values of symmetric neighbors is possible even for samples in the non-excited frames, therefore, we used the criteria mentioned in Eq. (1) to generate the STCs as follows:

$$TC\left(s^k_{left}, s^k_{right}\right) = \left\{ \begin{array}{c} 1, if\, s^k_{left} > 2 \times s^k_{right} \\ -1, if\, s^k_{left} < 2 \times s^k_{right} \\ 0, otherwise. \end{array} \right\} \tag{1}$$

where $TC(s^k_{left}, s^k_{right})$ represents a three-valued function of the proposed STC features. This symmetric ternary code generation process involves four comparisons (s^1 with s^9, s^2 with s^8, s^3 with s^7, and s^4 with s^6), thus, we get a 4-bit ternary code against each audio frame. These ternary codes are further divided into upper binary p_u and lower binary p_l patterns. We generate the p_u patterns by retaining the value $+1$ in STC to 1 and replacing all the remaining values (-1 and 0) to 0 as follows:

$$p_u\left(s^k_{left}, s^k_{right}\right) = \left\{ \begin{array}{c} 1,\ if\ TC\left(s^k_{left}, s^k_{right}\right) = +1 \\ 0, otherwise. \end{array} \right\} \tag{2}$$

Similarly, we generate the p_l patterns by replacing the values of -1 to 1 and all other values to zero as follows.

$$p_l\left(s^k_{left}, s^k_{right}\right) = \left\{ \begin{array}{c} 1,\ if\ TC\left(s^k_{left}, s^k_{right}\right) = -1 \\ 0, otherwise. \end{array} \right\} \tag{3}$$

Next, we concatenate these upper and lower binary codes to generate an 8-bit representation (Fig. 2(b)) and computed the decimal values as:

$$p = p_u \| p_l \tag{4}$$

Where p represents the 8-bit concatenated vector of upper and lower binary patterns computed from the symmetric ternary patterns. Further, we used uniform patterns due to their capacity to capture the maximum characteristics of a signal over non-uniform patterns [26]. We extracted the uniform patterns from the p and computed the decimal values as follows:

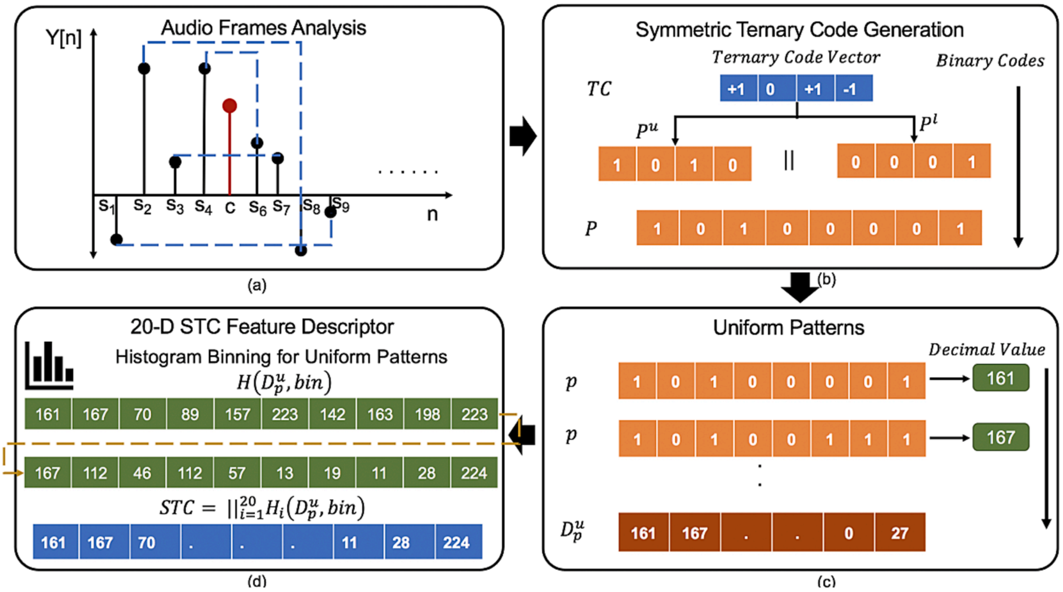

Fig. 2. STC Feature Generation Process.

$$D_p^u = \sum_{k=0}^{k=7} p \times 2^j \tag{5}$$

Here, D_p^u shows the decimal value of uniform symmetric ternary patterns. Finally, we assigned each uniform pattern to a histogram bin, and all non-uniform patterns were allotted to one histogram bin as follows:

$$H\left(D_p^u, bin\right) = \sum_{l=1}^{L} \delta\left(D_p^u, bin\right) \tag{6}$$

In Eq. (6), $\delta()$ is Kronecker delta function. By performing detailed experiments, we found that the first 20 uniform patterns are sufficient to hold the maximum characteristics in each audio frameset. Hence, we concatenate the histograms to create a 20-dim STC feature as follows:

$$STC = \|_{i=1}^{20} H_i\left(D_p^u, bin\right) \tag{7}$$

2.3. Classification of excited video frames

We employ the SVM in our method for the binary classification task of excited and non-excited audio segment classification. The SVM employs kernel tricks to project feature vectors to high dimensional space and finds an optimal separating hyperplane by minimalizing a cost function. In our method, we minimize the cost function for excited audio detection as follows:

$$J(\theta) = \min \frac{1}{m}\left[\sum_{i=1}^{m} \begin{matrix} t^{(i)}(-\log h_\theta(\boldsymbol{x}^{(i)})) + \ldots. \\ (1 - t^{(i)})(-\log(1 - h_\theta(\boldsymbol{x}^{(i)}) \end{matrix}))\right] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2 \tag{8}$$

where $h_\theta(\boldsymbol{x}) = \frac{1}{1+e^{-\theta^T x}}$ and θ represents the optimization parameter computed via gradient descent as:

$$\Delta\theta_j = \theta_j - \eta\frac{\partial J}{\partial\theta_j} = \theta_j - \eta\sum_i (t^{(i)} - h_\theta(\boldsymbol{x}^{(i)})\boldsymbol{x}_j^{(i)}) \tag{9}$$

$\frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$ is used to reduce overfitting. We modify the (8) as:

$$J(\theta) = \min C \sum_{i=1}^{m}[t^{(i)}\text{cost}_1(\theta^T\boldsymbol{x}^{(i)}) + (1 - t^{(i)})\text{cost}_0(\theta^T\boldsymbol{x}^{(i)})].. + \frac{1}{2}\sum_{j=1}^{n}\theta_j^2 \tag{10}$$

$$C = \frac{1}{\lambda}$$

By plotting a cost function shown in Fig. 3, a condition on the hypothesis $h_\theta(\boldsymbol{x})$ is enforced as:

$$R_1 = \left\{ \begin{matrix} Excited\ , & \theta^T\boldsymbol{x} \geqslant 1 \\ Non - Excited\ , & \theta^T\boldsymbol{x} \leqslant -1 \end{matrix} \right\} \tag{11}$$

By optimal parameter selection for $h_\theta(\boldsymbol{x}) \geqslant 1$ excited audios, and $h_\theta(\boldsymbol{x}) \leqslant 1$ for non-excited audios, the cost function returns a value ≈ 0 i.e., the minimum error, therefore, Eq. (10) will be reduced as:

$$J(\theta) = \min C(0) + \frac{1}{2}\sum_{j=1}^{n}\theta_j^2 \tag{12}$$

$$J(\theta) = \min\frac{1}{2}\sum_{j=1}^{n}\theta_j^2 = \min\frac{1}{2}\left(\sqrt{\theta_1^2 + \theta_2^2 \ldots + \theta_n^2}\right)^2 = \min\frac{1}{2}\|\theta\|^2 \tag{13}$$

Next, by projecting the feature vector $\boldsymbol{x}^{(i)}$ over the parameter vector $\theta^{(i)}$ as represented by $\boldsymbol{p}^{(i)}$, we obtain the separating hyperplane (Fig. 4) as:

$$\begin{matrix} \boldsymbol{p}^{(i)}\|\theta\| \geqslant 1, & if\ t^{(i)} = 1 \\ \boldsymbol{p}^{(i)}\|\theta\| \leqslant -1, & if\ t^{(i)} = 0 \end{matrix} \tag{14}$$

Next, we perform the predictions by transforming the feature space into higher dimensional via RBF kernel as:

$$C^{(i)} = \exp\left(\frac{\left\|\boldsymbol{x}^{(i)} - l^{(i)}\right\|^2}{2\delta^2}\right) \tag{15}$$

The feature vectors nearby the landmarks are labelled as one (1) while those farther away are labelled as zero (0). Thus, we get the trained classifier as:

$$C^*(\boldsymbol{x}) = C^{(i)} \tag{16}$$

We select the audio frameset of the given video using Eq. (1) to determine the excitement by transforming the audio into STC features. The excited audio frames are used to choose the corresponding video frames, which represent the excited video frames containing the key-events. The proposed method drastically decreases the computation time for the identification of key-events in the video.

2.4. Selection of key-frames for video summarization

The proposed framework uses the detected key-events to produce the summary of user-defined length. For each key-event, a video skim is produced using the summary length provided by the user. We computed the skim duration for each key-event as follows:

$$S_L = \frac{VL}{Tk} \tag{17}$$

where V_L and T_K represent the total length of the video summary chosen by the user and the total number of key-events detected in the video, respectively. S_L represents the length of each key-event. Lastly, these video skims are arranged in chronological order to create the final summary of the input sports video. The algorithm of the proposed method is provided below.

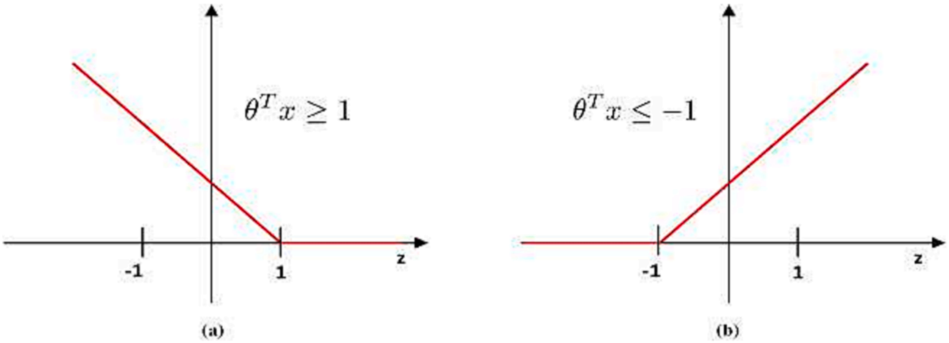

Fig. 3. Cost function: (a) Excited audio, (b) Non-Excited audio.

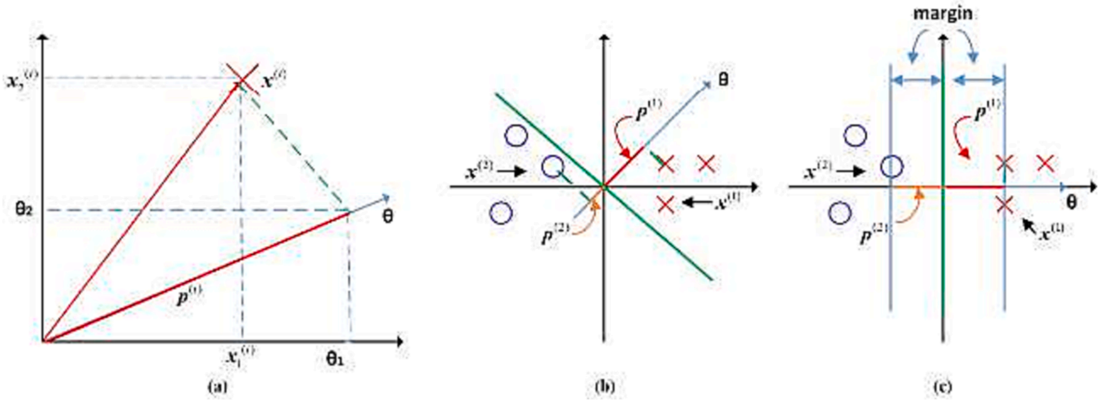

Fig. 4. (a) Projection of $\mathbf{x}^{(i)}$ on θ, (b, c) Hyperplanes scenarios.

Algorithm 1. Working of Proposed Method.

```
START
INPUT: VIDEO FRAMES, $Y = \{y(i)\}_{i=1}^{i=N}$
OUTPUT: SKIMMED VIDEO, $S_L$
// Preprocessing
$Y = \{y(i)\}_{i=1}^{i=N}$
$F^i \leftarrow Y$          // Frames creation of 9 samples each
$F^i \leftarrow$ spectral subtraction ($F^i$)          // noise reduction
// Feature Extraction
$s_{left}^k \leftarrow F^i$          // Four left neighboring samples w.r.t center sample
$s_{right}^k \leftarrow F^i$          // Four right neighboring samples w.r.t center sample
$TC\left(s_{left}^k, s_{right}^k\right) \leftarrow \left\{\begin{array}{c} 1, if s_{left}^k > 2 \times s_{right}^k \\ -1, if s_{left}^k < 2 \times s_{right}^k \\ 0, otherwise. \end{array}\right\}$          // Symmetric Ternary code (STC)
  generation
$p_u\left(s_{left}^k, s_{right}^k\right) \leftarrow \left\{\begin{array}{c} 1, if TC\left(s_{left}^k, s_{right}^k\right) = +1 \\ 0, otherwise. \end{array}\right\}$          // Left 4-bit STC generation
$p_l\left(s_{left}^k, s_{right}^k\right) \leftarrow \left\{\begin{array}{c} 1, if TC\left(s_{left}^k, s_{right}^k\right) = -1 \\ 0, otherwise. \end{array}\right\}$          // Right 4-bit STC generation
$p \leftarrow p_u \| p_l$          // 8-bit STC generation
$D_p^u \leftarrow \sum_{k=0}^{k=7} p \times 2^j$          // STC decimal value
$H\left(D_p^u, bin\right) \leftarrow \sum_{l=1}^{L} \delta\left(D_p^u, bin\right)$          // histogram Binning
$STC \leftarrow \|_{i=1}^{20} H_i\left(D_p^u, bin\right)$          // 20-D STC features
//Classification
C ← SVM(STC)          // Keyframe vs Non-key frame
  classification
// Video Skim Generation of user defined length
$v_L \leftarrow$ Length of video summary
$T_K \leftarrow$ Total number of key-events
$S_L \leftarrow \frac{v_L}{T_K}$
```

3. Results and discussion

This section presents the details of the dataset used for performance assessment. Moreover, it also provides the details of experiments and a discussion of the results obtained during the assessment of the proposed method. We have used the *precision, recall, F-1 score, accuracy,* and *error rate* to assess the performance of our method as the comparative methods also used the same metrics.

3.1. Dataset

A custom dataset consisting of cricket and soccer videos is created for performance assessment. The dataset videos are selected from different broadcasters including *PTV sports*, *star sports, ten sports*, *sky sports,* and *super sports*. All videos in the dataset have a frame rate of 30 fps and a spatial resolution of 640 × 480. For dataset creation, we selected the same policy as adopted by the comparative methods [19,22,24,32,38,40]. We have ensured the creation of a diverse dataset that is independent of video lengths, broadcasters, sports genres, tournaments, commentator's genres, etc. Cricket videos contain samples from the T20 series between *Bangladesh* and *Sri-lanka*, the Ashes series between *England* and *Australia*, 2014 ODI series between *Pakistan* and *New Zealand*, IPL 2014 between *Chennai Super Kings* and *Kolkata Knight Riders*, Pakistan Super League 2018 between *Islamabad United* and *Karachi Kings*, Women World cup 2017 between *India* and *South Africa*. Whereas, soccer videos contain samples from the 2014 friendly match between *Portugal* and *Argentina*, La Liga Cup 2016 between *Real Madrid* and *Barcelona*, Euro Cup 2012 between *Spain* and *Italy*, FIFA World Cup 2014 between *Argentina* and Netherlands, a friendly match between *France* and *England*, World cup 2014 between *Germany* and *Brazil*, Euro 2016 between *Wales* and *Belgium*, World cup 2010 between *Spain* and *Netherland*. Our dataset is publicly available at [36,37] for research purposes. Some images of our dataset are presented in Fig. 5.

Additionally, we have also used the SoccerNet dataset [35] for the performance evaluation of our method. The SoccerNet dataset consists of the videos of 500 games of 764 h duration acquired from different online sources in different encodings and frame rates of 25 to 50 fps. Moreover, all videos also contain audio segments which are used for performance assessment. The dataset videos are divided into training, validation, and testing collections of 300, 100, and 100 games, respectively. Moreover, each game contains two separate videos of each half, thus, containing 600 videos of training, 200 each for validation and testing collections.

3.2. Performance evaluation of excitement-based key audio frame detection

Evaluation on the proposed dataset: We designed an experiment to assess the performance of our video summarization framework on the audio dataset of our cricket and soccer videos. The fact that the proposed framework performs the acoustic analysis for key-events detection, therefore, we extracted the audios from the sports videos to create an audio dataset comprising excited and non-excited clips. We performed the audio clip annotation where the frames having high magnitude (i.e., loud audience cheers and commentary) were marked as excited audio frames. The rest of the frames were marked as non-excited ones. Our audio dataset is comprised of 3564 clips where half of the clips (1782) are soccer videos and the rest are cricket videos. For each sports genre (i.e., cricket, soccer) we used 80 % of the clips for training purposes where 50 % of clips belong to excited and the rest 50 % to the non-excited class. Whereas, the remaining 20 % of audio clips from each sports category are used for testing purposes where 50 % of clips belong to the excited and the rest 50 % to the non-excited class. We extracted the STC features of these excited and non-excited audio clips and used them with the SVM

Fig. 5. Snapshots of Dataset.

for classification. As SVM classifier has the ability to correctly classify the linearly separable classes, i.e., excited and non-excited in order to achieve higher objective evaluation. And, due to the involvement of the excitement factor in our case, both excited and non-excited classes are linearly distinguishable. Therefore, good excitement detection performance is expected from the SVM classifier.

We have reported the results obtained on the audio clips of 10 sports videos as shown in Table 1. The proposed method achieves **98.1 %, 97.17 %, 97.6 %, 97.7 %**, and **2.3 %** as average *precision, recall, F-1 score, accuracy,* and *error rate.* It can be clearly observed from Table 1 that the proposed video summarization method attains excellent results and can reliably detect the key events in sports videos. The proposed method performed marginally better on cricket videos than on soccer videos. This might be due to the fact that we often experience loud cheers from the audience in soccer videos even in the absence of key events.

Evaluation on the SoccerNet dataset: Additionally, we have also examined our method on SoccerNet dataset containing the soccer videos. We prepared the dataset for experiments by extracting the audio segments of the soccer videos in the same way as done for our own dataset. We used the excited and non-excited audio clips of the training set to train the SVM and testing set clips for evaluation purposes. We computed the STC features of these excited and non-excited audio segments and employed them to train the SVM for classification. Next, we evaluated the performance on the testing corpus and our method achieved a precision of 92.43 %, recall of 91.81 %, F1-score of 92.12 %, accuracy of 91.23 %, and error rate of 8.77 % for key-events detection. These results on the SoccerNet dataset illustrate the efficacy of our method to generate the summarization of sports videos.

Table 1
Detection results.

Video Summarization Methods	Precision Rate	Recall Rate	Accuracy Rate	Error Rate	F1-Score
Cricket 1	97.20 %	96.21 %	97.19 %	2.81 %	96.70 %
Cricket 2	99 %	98.42 %	97.13 %	2.87 %	98.70 %
Cricket 3	98.20 %	95.52 %	97.64 %	2.36 %	93.88 %
Cricket 4	97.10 %	98.51 %	98.13 %	1.87 %	97.79 %
Cricket 5	99.10 %	97.21 %	96.32 %	3.68 %	98.16 %
Soccer 1	97.40 %	96.52 %	97.26 %	2.74 %	96.95 %
Soccer 2	98 %	98.11 %	96.18 %	3.82 %	98.05 %
Soccer 3	98.20 %	95.93 %	97.64 %	2.36 %	97.05 %
Soccer 4	98 %	98.10 %	98.13 %	1.87 %	98.04 %
Soccer 5	99 %	97.21 %	96.20 %	2.8 %	98.09 %
Average	**98.1 %**	**97.17 %**	**97.7 %**	**2.3 %**	**97.6 %**

3.3. Performance comparison on different classifiers

To assess the significance of SVM for excited clip detection, we conducted a two-stage experiment where we used different backend classifiers with our STC features to examine the performance of the proposed method on our own custom YouTube dataset and SoccerNet dataset separately. More precisely, we compared the performance of SVM with KNN, Decision Trees, Naïve Bayes, and Ensemble model. We used our STC features to train the SVM, KNN, decision trees, Naïve Bayes, and Ensemble bagged trees separately for excited vs non-excited audio detection on our custom dataset and results are reported in Table 2. From the results, it can be seen that the Naïve Bayes performed the worst by obtaining the *precision, recall, F-1 score, accuracy,* and *error rate* of 73.45 %, 74.25 %, 73.85 %, 73.14 %, and 26.86 %, respectively. KNN and ensemble bagged trees performed comparatively similar but KNN obtained slightly better results by attaining the second-best *precision, recall, F-1 score, accuracy,* and *error rates* of 92.80 %, 92.70 %, 92.75 %, 92.47 %, and 7.53 %, respectively. More specifically, our method (STC-SVM) achieved the best results of 98.1 % *precision,* 97.17 % *recall,* 97.6 % *F-1 score,* 97.7 % *accuracy,* and a 2.3 % *error rate.*

Similarly, we assessed the performance of our STC features with different backend classifiers on the SoccerNet dataset, and the results are provided in Table 2. The Naïve Bayes when used with our STC features attained the lowest *precision, recall, F-1 score, accuracy,* and *error rate* of 64.13 %, 65.43 %, 64.78 %, 64.90 %, and 35.10 %, respectively. Ensemble bagged trees achieved the second-best results with *precision, recall, F-1 score, accuracy,* and *error rates* of 88.98 %, 89.51 %, 89.24 %, 89.32 %, and 10.68 %, respectively. Again, our STC features produced the best outcome with SVM by achieving a precision of 92.43 %, recall of 91.81 %, F1-score of 92.12 %, accuracy of 91.23 %, and error of 8.77 %. This comparative analysis on different classifiers shows that our STC features when used with the SVM can effectively be used to capture the excitement in the audio clips that ultimately lead to producing a reliable summarized video.

3.4. Performance comparison of proposed STC features with time domain and spectral features

To better examine the importance of our acoustic STC features for the effective representation of sports audio signals, we conducted an experiment to compare the performance of our STC features with both the time domain acoustic features i.e., acoustics-LBP [38], and spectral features i.e., MFCC, GTCC, and LPC. We selected the SVM as a backend classifier with all features. More precisely, we conducted this experiment to assess the performance of our proposed STC features with contemporary acoustic features for excitement detection in audios. We

Table 2
Detection results on different classifiers.

Dataset	Classifiers	Precision Rate	Recall Rate	Accuracy Rate	Error Rate	F1-Score
Proposed YouTube dataset	Naive Bayes	73.45 %	74.25 %	73.14 %	26.86 %	73.85 %
	KNN	92.80 %	92.70 %	92.47 %	7.53 %	92.75 %
	Decision Trees	84.10 %	83.21 %	83.94 %	16.06 %	83.66 %
	Ensemble Bagged Trees	92.10 %	92.11 %	92.17 %	7.83 %	92.11 %
	SVM	**98.1 %**	**97.17 %**	**97.7 %**	**2.3 %**	**97.6 %**
SoccerNet	Naive Bayes	64.13 %	65.43 %	64.90 %	35.10 %	64.78 %
	KNN	86.11 %	85.87 %	85.11 %	14.89 %	85.99 %
	Decision Trees	77.29 %	77.81 %	77.65 %	22.35 %	77.55 %
	Ensemble Bagged Trees	88.98 %	89.51 %	89.32 %	10.68 %	89.24 %
	SVM	**92.43 %**	**91.81 %**	**91.23 %**	**8.77 %**	**92.12 %**

used each of the selected features to train the SVM separately for excited audio detection and results are reported in Table 3. From the results, it can be seen that the LPC attained the lowest results by obtaining the *precision, recall, F-1 score, accuracy,* and *error rate* of 72.49 %, 69.27 %, 70.96 %, 71.84 %, and 28.16 %, respectively. Acoustics-LBP features ranked second by attaining the second-best *precision, recall, F-1 score, accuracy,* and *error rate* of 97.24 %, 96.84 %, 97.04 %, 96.97 %, and 3.03 %, respectively. Our STC features ranked the best among all comparative features by attaining 98.1 % *precision,* 97.17 % *recall,* 97.6 % *F-1 score,* 97.7 % *accuracy,* and a 2.3 % *error rate.* This experiment demonstrates the effectiveness of our STC features for better audio representation to reliably detect the excited audio clips that ultimately lead to producing a reliable summarized video.

3.5. *Performance comparison with contemporary approaches*

We designed an experiment to assess the effectiveness of our approach as compared to the existing video summarization methods. For this, we conducted a comparative analysis of proposed and contemporary sports video summarization methods [19,22,24,32,38,40]. For performance evaluation, YouTube videos are used since no standard dataset exists for sports video summarization [10,11,17] and the same dataset selection strategy has been adopted by contemporary approaches. We tried to ensure that our audio dataset is distinct in terms of broadcasters, commentators, commentary language, etc. The results of this comparative analysis are presented in Table 4. The method [22] performed worst and achieved an accuracy of 81.12 % whereas, our proposed work performed equally best with our previous work [38] and achieved an accuracy of 97 %. However, our prior work [38] is limited in terms of sports genre and able to generate the summary of only Cricket videos, whereas, our proposed method can successfully generate the summaries of field sports such as Cricket and Soccer videos. This comparative analysis shows the superior performance of our method over existing summarization methods for field sports.

3.6. *Confusion matrix analysis*

In the last experiment, we portray the classification accuracy of our technique for key-events detection using the confusion matrix (CM) as presented in Table 5. The confusion matrix is employed to analyze the misclassification rate of the technique. More precisely, we employed the CM to examine the number of non-excited framesets detected as excited ones (false positives) and a number of excited framesets detected as non-excited ones (false negatives). From Table 5, we can see that the classification performance of our system is remarkable for all the key-events in cricket and soccer videos.

3.7. *Discussion*

In the proposed work, audio signal of the sports videos is analyzed to identify the exciting clips that correspond to key-events. The proposed method allows us to detect any kind of key-event in the game and hardly miss any significant event as it relies on excitement detection. Moreover, our method is applicable to make a video summary of any genre of sports video as the key-events detection mechanism is based on excitement identification. This attribute makes our method independent of the game rules and structure of any sport. From the results presented in Table 1 that are obtained on two diverse sports video datasets, it is proved that the proposed method is robust to sports genre, game structure, rules, broadcasters, commentators, audio recording parameters, etc. Due to the potential application benefits, we argue that our framework is very effective in creating useful summaries of sports videos. Additionally, the proposed scheme is computationally efficient as it analyses only the audio signal of the input sports video and does not require extracting the computationally complex visual features. However, the performance of our method can be expected to drop for indoor games with small audiences having lower cheers and less excitement in the commentator's voice. It is to be noted that the proposed approach can be extended to classify different key-events. For the classification of key events of these types of sports, we require game-specific rules for event-based video summarization.

4. Conclusion

This paper has presented an effective and efficient video summarization framework for multiple sports videos by detecting the excited clips. For excitement detection, we proposed novel symmetric ternary code features to represent the audio stream and used them to train the SVM classifier. The video frames corresponding to the excited audio frames are used to produce the summarized video. Experimental validation on two diverse datasets including the comparative analysis on cricket and soccer videos revealed the reliability of our method for sports video summarization. The proposed work can be extended to summarize all types of indoor and outdoor games.

Table 3
Detection results on proposed STC and comparative features.

Features	Precision Rate	Recall Rate	Accuracy Rate	Error Rate	F1-Score
MFCC	83.35 %	84.25 %	83.15 %	16.85 %	83.80 %
GTCC	85.11 %	84.91 %	85.22 %	14.78 %	85.01 %
LPC	72.49 %	69.27 %	71.84 %	28.16 %	70.96 %
Acoustics-LBP	97.24 %	96.84 %	96.97 %	3.03 %	97.04 %
STC (Proposed)	**98.1 %**	**97.17 %**	**97.7 %**	**2.3 %**	**97.6 %**

Table 4
Performance comparison of proposed framework with existing system.

Video Summarization Methods	Dataset Statistics			Precision (%)	Recall (%)	Accuracy (%)	Error (%)	F1-score (%)
	Frame Rate	Resolution	Length (hours)					
Javed et al. [19]	25fps	640 × 480	10	91.87	89.85	95.01	4.99	90.84
Merler et al. [22]	25fps	–	124	–	–	81.12	18.88	–
Raventós et al. [24]	–	–	–	88	93	–	–	90.43
Trinh et al. [32]	–	–	–	91.38	92.38	–	–	91.83
Javed et al. [38]	25fps	640 × 480	10	98.80	97.60	97.70	2.36	98
Islam et al. [40]	–	–	–	–	–	61.22	38.80	–
Shingrakhia et al. [42]	25fps	640 × 480	–	96.82	95.41	96.32	3.68	95.67
Proposed Method	25fps	640 × 480	10	98.10	97.17	97.70	2.30	97.63

Table 5
Confusion matrix analysis.

		Predicted Class			
	Class	Positive	Negative	Positive	Negative
		Cricket		Soccer	
Actual Class	Positive	**20399**	241	**15249**	551
	Negative	718	**30002**	403	**17197**

CRediT authorship contribution statement

Ameen Banjar: Conceptualization, Methodology, Software, Formal analysis, Investigation, Resources, Writing – original draft, Writing – review & editing, Funding acquisition. **Hussain Dawood:** Methodology, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Supervision. **Ali Javed:** Conceptualization, Methodology, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Supervision, Project administration. **Bushra Zeb:** Methodology, Data curation, Writing – original draft, Writing – review & editing, Visualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Link of datasets are provided in paper

Acknowledgment

This work was funded by the University of Jeddah, Jeddah, Saudi Arabia, under grant No. (UJ-22-DR-9). The authors, therefore, acknowledge with thanks the University of Jeddah for its technical and financial support.

References

[1] Javed A, Irtaza A, Khaliq Y, Malik H, Mahmood MT. Replay and key-events detection for sports video summarization using confined elliptical local ternary patterns and extreme learning machine. Appl Intell 2019;49(8):2899–917.
[2] Tejero-de-Pablos A, Nakashima Y, Sato T, Yokoya N, Linna M, Rahtu E. Summarization of User-Generated Sports Video by Using Deep Action Recognition Features. IEEE Trans Multimedia 2018;20(8):2000–11.
[3] Rameswar P, Amit K, Chowdhury R. Multi-View Surveillance Video Summarization via Joint Embedding and Sparse Optimization. IEEE Trans Multimedia 2017;19(9):2010–21.
[4] Xuelong L, Wang Z. Xiaoqiang Lu. Surveillance Video Synopsis via Scaling Down Objects. IEEE Trans Image Process 2016;25(2):740–55.
[5] Khan M, Ahmad J, Sajjad M, Baik SW. Visual saliency models for summarization of diagnostic hysteroscopy videos in healthcare systems. Springerplus 2016;5:1–13.
[6] Mehmood I, Sajjad M, Bai SW. Video summarization based tele-endoscopy: a service to efficiently manage visual data generated during wireless capsule endoscopy procedure. J Med Syst 2014;38:1–9.
[7] Luming Z, Xia Y, Mao K, Ma H, Shan Z. An Effective Video Summarization Framework Toward Handheld Devices. IEEE Trans Ind Electron 2015;62(2):1309–16.
[8] Yang C, Liu J, Sun G, You Q, Li Y, Luo J. Adaptive Greedy Dictionary Selection for Web Media Summarization. IEEE Trans Image Process 2017;26(1):185–95.
[9] Min S, Farhadi A, Taskar B, Seitz S. Summarizing Unconstrained Videos Using Salient Montages. IEEE Trans Pattern Anal Mach Intell 2017;39(11):2256–69.
[10] Jingjing M, Wang S, Wang H, Yuan J, Tan YP. Video Summarization Via Multiview Representative Selection. IEEE Trans Image Process 2018;27(5):1189–98.
[11] Shanmukhappa A, Naik V. Entropy Based Fuzzy C Means Clustering and Key Frame Extraction for Sports Video Summarization. In: ICSIP 2014-2014 Fifth International Conference on Signal and Image Processing (ICSIP), p. 271-279.
[12] Evlampios A, Mezaris V. Fast shot segmentation combining global and local visual descriptors. In: ICASSP 2014-2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), p. 6583-6587.
[13] Anant B, Cho J, Lee W, Ko BS. Sports highlight generation based on acoustic events detection: A rugby case study. In: ICCE 2015-2015 IEEE International Conference on Consumer Electronics (ICCE), p.20-23.
[14] Shiqi T, Zhi M. Summary generation method based on audio feature. In: ICSESS 2015-2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS), p. 619-623.
[15] Kolekar MH, Sengupta S. Semantic concept mining in cricket videos for automated highlight generation. Multimed Tools Appl 2010;47(3):545–79.
[16] Kolekar MH. Bayesian belief network based broadcast sports video indexing. Multimed Tools Appl 2011;54(1):27–54.
[17] Zengkai W, Yu J. Soccer Video Event Annotation by Synchronization of Attack-Defense Clips and Match Reports with Coarse-Grained Time Information. IEEE Trans Circuits Syst Video Technol 2016;27(5):1105–17.
[18] Mendi E, Clemente HB, Bayrak C. Sports video summarization based on motion analysis. Comput Electr Eng 2013;39(3):790–6.
[19] Javed A, Bajwa KB, Malik H, Irtaza A, Mehmood MT. A hybrid approach for summarization of cricket videos. In: ICCE-Asia 2016-2016 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia), p. 1-4.
[20] Nguyen, Ngac KY, and Yoshitaka A. Soccer video summarization based on cinematography and motion analysis. In: MMSP 2014-2014 IEEE 16th International Workshop on Multimedia Signal Processing (MMSP), p. 1-6.
[21] Javed A, Bajwa KB, Malik H, Irtaza A. an Efficient Framework for Automatic Highlights Generation from Sports Videos. IEEE Signal Process Lett 2016;23(7):954–8.
[22] Merler M, Mac KNC, Joshi D, Nguyen QB, Hammer S, Kent J, et al. Automatic Curation of Golf Highlights using Multimodal Excitement Features. IEEE Trans Multimedia 2018;21(5):1147–60.
[23] Hasan T, Bořil H, Sangwan A, John H, Hansen L. Multi-modal highlight generation for sports videos using an information-theoretic Excitability measure. EURASIP Journal on Advances in Signal Processing 2013;2013(1):1–17.
[24] Raventós A, Quijada R, Torres L, Tarrés F. Automatic Summarization of Soccer Highlights Using Audio-visual Descriptors. Springerplus 2015;4:1–19.
[25] Tomoki H, Takahashi S, Ogawa T, Haseyama M. Estimation of Important Scenes in Soccer Videos Based on Collaborative Use of Audio-Visual CNN Features. In GCCE 2018-2018 IEEE 7th Global Conference on Consumer Electronics (GCCE), p. 710-711.
[26] Javed A, Malik KM, Irtaza A, Malik H. Towards protecting cyber-physical and IoT systems from single-and multi-order voice spoofing attacks. Appl Acoust 2021;183:108283.
[27] Tavassolipour M, Karimian M, Kasaei S. Event Detection and Summarization in Soccer Videos Using Bayesian Network and Copula. IEEE Trans Circuits Syst Video Technol 2014;24(2):291–304.
[28] Ngoc N, Yoshitaka A. Shot Type and Replay Detection for Soccer Video Parsing. In ISM 2013-2013 IEEE International Symposium on Multimedia (ISM), p. 344-347.
[29] Bettadapura V, Pantofaru C. Essa I. Leveraging Contextual Cues for Generating Basketball Highlights. In ACM MM 2016-2016 Proceedings of ACM Multimedia (MM), p. 908-917.
[30] Dongmahn S, Suhyun K, Park H, Ko H. User generated highlight system for baseball games with social media activities. In: ICCE 2014-2014 IEEE International Conference on Consumer Electronics (ICCE), p. 349-350.
[31] Decroos T, Dzyuba V, Haaren JV, Davis J. Predicting Soccer Highlights from Spatio-Temporal Match Event Streams. In AAAI 2017-2017 Proceedings of the AAAI Conference on Artificial Intelligence, p. 1302-1308.

[32] Trinh TD, Ma X, Lee HJ, Kim JY, Choi SH. Cheering Event Detection in Basketball Audio Stream Using Adaptive GMM Model and Low-Rank Matrix Recovery. J Korean Inst Information Technol 2016;14(10):87–96.

[33] Pushkar S, Sadana H, Bansal A, Verma D, Carlos E, Elmadjian L, Raman B, Turk M. Automatic Cricket Highlight Generation Using Event-Driven and Excitement-Based Features. In: CVPR 2018-2018 IEEE conference on computer vision and pattern recognition workshops, p. 1800-1808.

[34] Sanabria M, Precioso F, and Menguy T. Hierarchical multimodal attention for deep video summarization. In: ICPR 2020-2020 25th International Conference on Pattern Recognition (ICPR), p. 7977-7984.

[35] Giancola S, Amine M, Dghaily T, Ghanem B. Soccernet: A scalable dataset for action spotting in soccer videos. In: IEEE CVPR 2018-2018 IEEE conference on computer vision and pattern recognition workshops (CVPR), p. 1711-1721.

[36] Javed A, Zeb B, Irtaza A. "https://datadryad.org/review?doi=doi:10.5061/dryad.9c87t47".

[37] Javed A, Zeb B, Irtaza A. " https://datadryad.org/review?doi=doi:10.5061/dryad.n9d4100".

[38] Javed A, Irtaza A, Malik H, Mahmood MT, Adnan S. Multimodal framework based on audio-visual features for summarisation of cricket videos. IET Image Proc 2019;13(4):615–22.

[39] Khan AA, Shao J, Ali W, Tumrani S. Content-Aware Summarization of Broadcast Sports Videos: An Audio-Visual Feature Extraction Approach. Neural Process Lett 2020;52:1945–68.

[40] Islam MR, Paul M, Antolovich M, Kabir A. Sports Highlights Generation using Decomposed Audio Information. In: ICMEW 2019-2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), p. 579-584.

[41] Javed A, Khan AA. Shot classification and replay detection for sports video summarization. Frontiers of Information Technology & Electronic Engineering 2022;23(5):790–800.

[42] Shingrakhia H, Patel H. SGRNN-AM and HRF-DBN: a hybrid machine learning model for cricket video summarization. Vis Comput 2022;38(7):2285–301.

[43] Agyeman R, Rafiq M, Choi GS. Soccer video summarization using deep learning. In: MIPR 2019-2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), p. 270-273.