IET Image Processing

Research Article

IET Journals
The Institution of Engineering and Technology

Multimodal framework based on audio-visual features for summarisation of cricket videos

ISSN 1751-9659
Received on 29th July 2017
Revised 29th May 2018
Accepted on 2nd January 2019
E-First on 6th March 2019
doi: 10.1049/iet-ipr.2018.5589
www.ietdl.org

Ali Javed[1] ✉, Aun Irtaza[2], Hafiz Malik[3], Muhammad Tariq Mahmood[4], Syed Adnan[2]

[1]Department of Software Engineering, University of Engineering and Technology, Taxila, Pakistan
[2]Department of Computer Science, University of Engineering and Technology, Taxila, Pakistan
[3]Department of Electrical and Computer Engineering, University of Michigan, Dearborn, USA
[4]School of Computer Science and Information Engineering, Korea University of Technology and Education, 1600 Chungjeolno, Byeogchunmyun, Cheonan, Republic of Korea
✉ *E-mail: ali.javed@uettaxila.edu.pk*

Abstract: Sports broadcasters generate an enormous amount of video content on the cyberspace due to massive viewership all over the world. Analysis and consumption of this huge repository urges the broadcasters to apply video summarisation to extract the exciting segments from the entire video to capture user's interest and reap the storage and transmission benefits. Therefore, in this study an automatic method for key-events detection and summarisation based on audio-visual features is presented for cricket videos. Acoustic local binary pattern features are used to capture excitement level in the audio stream, which is used to train a binary support vector machine (SVM) classifier. Trained SVM classifier is used to label audio frame as an excited or non-excited frame. Excited audio frames are used to select candidate key-video frames. A decision tree-based classifier is trained to detect key-events in the input cricket videos that are then used for video summarisation. Performance of the proposed framework has been evaluated on a diverse dataset of cricket videos belonging to different tournaments and broadcasters. Experimental results indicate that the proposed method achieves an average accuracy of 95.5%, which signifies its effectiveness.

1 Introduction

Exponential growth of sports videos production, sharing, and its availability in the cyberspace have sparked research activities to develop efficient video analysis and content management techniques. Analysis and consumption of available sports videos in the cyberspace is a challenging task. Video summarisation techniques [1, 2] are used to address the aforementioned challenges by providing a short synopsis video consisting of key-events. Motivation behind summarising sports videos are: the transmission requirements over low-bandwidth networks, storage cost, time constraints, and capturing viewer interest through exciting segments in the full-length video.

Video summarisation techniques have been proposed for various sports such as soccer [3], tennis [4], baseball [5], cricket [6, 7], basketball [8], rugby [9], and so on. The cricket videos are more difficult to address from video summarisation perspective than any other sports as the cricket matches are of the longest duration with high-frequency key-events. This might be the reason behind limited focus on cricket video analysis and summarisation. Moreover, cricket has unique field rules (i.e. no whistles etc.), therefore, we need to observe the external factors [e.g. excitement in commentary, score-captions (SCs), logos etc.] for video summarisation.

Existing state-of-the-art methods [2–5, 9–13] for sports video summarisation can be categorised into two broad categories, i.e. learning-based [2–5, 9, 10] and non-learning-based methods [11–13]. Learning-based methods [2–5, 9, 10] employ classification techniques to detect significant events for sports video summarisation. Learning-based techniques offer better results at the cost of increased computational complexity. Zawbaa *et al.* [3] proposed a learning-based video summarisation framework for soccer. Neural networks and SVM were trained to identify the logo frames for key-event detection. Javed *et al.* [10] applied the rule-based induction to identify the excited audio frames followed by using a decision tree to summarise cricket videos. Similarly, Midhu and Padmanabhan [6] detected various key-events to generate the summary of cricket videos.

Besides learning-based methods, non-learning-based methods [11–13] have also been proposed for video summarisation. Non-learning-based methods exploit the game observations, e.g. logo placement, game transitions, whistle detection, and so on to find the clear patterns for precise video summarisation [3, 14]. Meanwhile, non-learning-based methods are slightly static in nature, therefore, are unable to detect key-events if the observational information is improperly detected [15]. Chen and Chen [11] used statistical features to propose a non-learning technique for replay detection in basketball videos. Nguyen and Yoshitaka [12] proposed a non-learning technique using histogram difference and contrast to identify the logo frames for replay detection.

Existing sports video summarisation methods extensively use visual [16, 17] and audio features [18, 19] to detect the key-events. Visual features are usually computed through the estimation of colour [20], edges [21], texture [22], and salient-points [23]. Tavassolipour *et al.* [17] proposed a Bayesian network-based method for key-events detection and summarisation of soccer videos. Namuduri [24] applied a hidden Markov model (HMM) to generate the highlights for cricket videos. Histogram difference comparison was employed for shot boundary detection followed by performing the shot classification through MPEG-7 visual descriptors. At each frame, two histograms comparisons were performed that increases the computational cost of this method [24]. Wang *et al.* [25] proposed a framework for key-events annotation, event boundaries detection, soccer field classification, and whistle detection in soccer videos. Godi *et al.* [26] applied a deep convolutional neural network (CNN) to generate the highlights for ice-hockey. Similarly, Jiang *et al.* [27] applied a deep neural network based on CNN and recurrent neural network to detect four key-events in soccer videos.

A lot of research works [9, 14, 28, 29] have used audio features for sports video analysis. Baijal *et al.* [9] proposed a Gaussian mixture model (GMM) based method using audio cues to

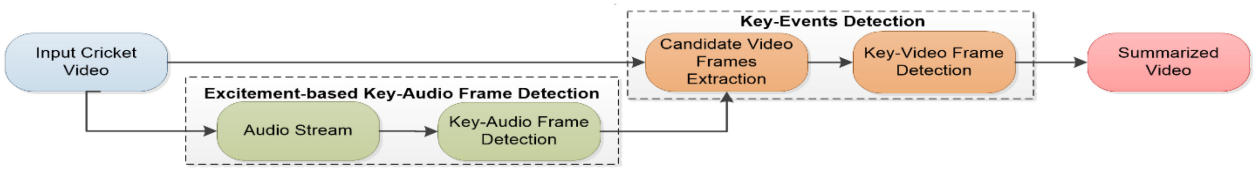

Fig. 1 *Block diagram of the proposed framework*

summarise the sports videos. Mel-frequency cepstral coefficients (MFCC) features were used to train a GMM classifier into two stages. In the first stage, input audio signal was classified into speech and non-speech components. Whereas, in the second stage, speech components were processed to classify between the excited and non-excited segments. Rui *et al.* [14] used energy features, MFCC, and audio pitch to develop an efficient summarisation method for baseball videos. However, the performance of this method [14] degrades significantly due to the deficiencies recorded in pitch detection method. Xu *et al.* [28] applied the rule-based heuristics using energy features to train the SVM classifier to detect different events in soccer videos. The performance of events detection in [28] is largely dependent on audio keywords recognition accuracy.

Existing techniques [30–32] have also combined the audio and visual features to summarise videos of various sports as fusion of the features result in improved accuracy. Kolekar and Sengupta [30, 31] have used audio-visual features to propose a hierarchical framework for sports video summarisation. Short-term energy features were used to analyse the audio stream of the input video. Whereas, colour and motion features were used to train a HMM to detect the replay segments. Raventos *et al.* [32] presented an automated framework for soccer video summarisation. For this, shot boundary detection was applied to segment the video into shots. Afterwards, low- and mid-level audio-visual descriptors were computed against each shot to acquire different relevance measures based on specific rules. Finally, these relevance measures were used in combination of user preferences to summarise the soccer videos.

The existing video summarisation methods show the computational inefficiency in external factor identification, i.e. logo detection, placement of SCs, illumination changes, replay speeds, camera variations, and so on. However, the proposed method is robust against the aforementioned limitations.

To reap the benefits of both learning- and non-learning-based methods, a hybrid approach is proposed in this paper to effectively deal with the summarisation of cricket videos. It has been observed that key-events result in a significant change in the excitement level of the audio stream. Likewise, key-events also result in change in the SCs for visual stream. A two-stage framework is proposed for key-event detection. The first stage uses audio stream to compute excitement score for each audio frame which is used to select the candidate key-event video frames. The second stage analyses the SCs of the candidate video frames to detect key-events in the cricket video. The proposed framework computes the length of each video skim against each key-frame to generate the summary of user-specified length. Moreover, SC and gradual transitions (GTs) are used to detect replay events that are also included in the summarised videos. A diverse dataset consisting of videos of different cricket tournaments and broadcasters are used for performance evaluation. Performance metrics such as precision, recall, accuracy, error rates, and F-1 score are used to evaluate the effectiveness of the proposed method. Effect of various recording parameters such as camera angle, SC design and placement, and replay speed variation on the performance of the proposed method has also been evaluated. Experimental results illustrate that the proposed method achieves an average accuracy of 95.5% which signifies its effectiveness for video summarisation.

2 Proposed framework

The proposed video summarisation method exploits audio-visual cues for key-event detection. Specifically, audio features are used to detect excitement through commentary voices and crowd cheers that is then used for candidate key-video frame selection. Selected candidate frames are analysed further to detect the key-events. The proposed system, therefore, can be divided into two stages: first stage uses audio stream corresponding to a video frame to compute excitement through SVM classifier by representing audio frames in the form of feature vectors. The key-audio frames are used to select associated video frames. The second stage uses selected video frames to detect the key-events. The architecture of the proposed video summarisation system is presented in Fig. 1 that is described in detail in further subsections.

2.1 Excitement-based key-audio frame detection

2.1.1 Problem formulation: Let $y[n]$ be the audio signal having N' samples associated with a cricket video containing K' video frames represented as $I'^{(i)}(x, y)$, and K video frames $I^{(i)}(x, y)$ are associated with N excited audio frames/windows where $N \ll N'$ and $K \ll K'$. For key-event detection, we analyse audio frames on the basis of audio recording sampling rate and representing them in the form of feature vectors. Once the excitement detection occurs through classification, we analyse the corresponding video frame for key-event detection.

2.1.2 Feature extraction: Effective feature extraction is an indispensable requirement to achieve higher classification accuracy. Acoustic features provide numerical representation of the information present in the form of sound waves. For implementation of this work, we extracted the acoustic-local binary pattern (acoustic-LBP) features from the audio stream of the input cricket videos. As described in [33], acoustic-LBP is a fast and computationally inexpensive mechanism for signal representation that distinctively marks certain signal features. Another reason to utilise the acoustic-LBP is that the acoustic-LBPs are never employed in excitement detection based video summarisation research [9, 10].

The signal features in the form of linear LBP codes can be adopted for signal segmentation and thumb-impression generation. The LBP examines the neighbourhood of data samples from a signal and assigns an LBP code to each centre sample after thresholding them against the neighbouring samples [33].

Let $y[j]$ be the central sample in the samples window with $P + 1$ elements in audio signal $y[n]$, where

$$j = \left[\frac{P}{2} : N' - \frac{P}{2}\right].$$

The acoustic-LBP can be defined as

$$\mathrm{LBP}_P(y[j]) = \sum_{m=0}^{(P/2)-1} \left\{ S\left[y\left[j + m - \frac{P}{2}\right] - y[j]\right] 2^m + \ldots S\left[y[j + m + 1] - y[j]\right] 2^{m+(P/2)} \right\} \quad (1)$$

where the sign function $S[\cdot]$ is given by

$$S[y] = \begin{cases} 1, & \text{for } y \geq 0 \\ 0, & \text{for } y < 0 \end{cases} \quad (2)$$

In acoustic-LBP, the sample $y[j]$ serves as a threshold for the neighbouring samples, and the sign function $S[\cdot]$ transforms the

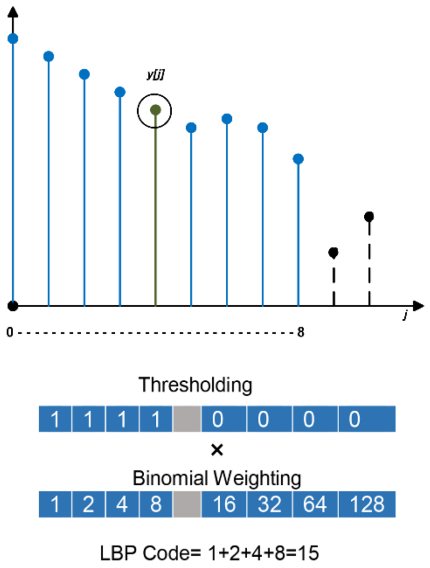

Fig. 2 *Acoustic-LBP calculation*

difference between $y[j]$ and the neighbourhood as a P-bit binary code where $P = 8$ in our implementation. The binomial weights are then multiplied to the acoustic-LBP code and summed to generate the acoustic-LBP value for the sample $y[j]$ (as shown in Fig. 2). The acoustic-LBP locally describes a sample using neighbourhood differences; for a constant signal these differences cluster near zero; whereas, at peaks and plateaus the difference is large. The acoustic-LBP codes are used to describe the local patterns as

$$H_m = \sum_{q=1}^{Q} \delta(\mathrm{LBP}_P(y[j]), m) \tag{3}$$

where $m = 1 \ldots z$, and z describes the histogram bins corresponding to each acoustic-LBP code and $\delta(i, j)$ is the Kronecker delta function. The parameter Q specifies the audio samples that correspond to the feature vector representing an audio frame that is defined as

$$Q = \omega \times \tau \tag{4}$$

where ω is the audio sampling rate (e.g. 48 kbps) and τ is the time interval that is set to 5 s in our implementation. Corresponding to the feature vector there are G video frames that are defined as follows:

$$G = \eta \times \tau \tag{5}$$

where η is the video frame rate (e.g. 25 fps).

2.1.3 Excited audio-frame detection: We classify excited and non-excited audio clips through SVM classifier [34]. The training data consisting of M excited and non-excited audio features is prepared as: $(\boldsymbol{x}^{(i)}, t^{(i)}), i = 1, \ldots, M$, where $t^{(i)} \in \{1, -1\}$ specifies the excited and non-excited audio classes. Hyperplanes linearly separating the two classes are given as

$$\begin{cases} \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^{(i)} + b \geq 1, & \text{if } t^{(i)} = 1 \\ \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^{(i)} + b < 1, & \text{if } t^{(i)} = -1 \end{cases} \tag{6}$$

where $\boldsymbol{w}$ is the weighting vector and b is the bias. The objective is to maximise the separation between two planes by minimising the norm $\| \boldsymbol{w} \|$ which can be formulated as a quadratic optimisation problem

$$\min_{\boldsymbol{w}} \| \boldsymbol{w} \| \quad \text{s.t.} \quad t\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^{(i)} + b\right) \geq 1 \tag{7}$$

The two events (excited and non-excited) can then be detected by using the discriminating function $f\left(\boldsymbol{x}^{(i)}\right) = \mathrm{sign}\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^{(i)} + b\right)$ such as

$$\begin{cases} \text{excited}, & \text{if } f\left(\boldsymbol{x}^{(i)}\right) = +1, \\ \text{non-excited}, & \text{if } f\left(\boldsymbol{x}^{(i)}\right) = -1 \end{cases} \tag{8}$$

The excited audio-frames are used to select the corresponding video frames which are candidates for key-events. The second stage analyses only the candidate video frames for key-event detection. The benefit of the proposed two-stage approach is that it significantly reduces the processing time required to detect key-events in the input video.

2.2 Excitement-driven key-event detection and video summarisation

At the second stage of video summarisation, video frames corresponding to the excited audio frames are processed further for key-event detection. Each step of the proposed framework is discussed in the subsequent sections.

2.2.1 Video-frame enhancement: The excited colour video frames are transformed into greyscale and every tenth frame is processed. Top-hat filtering [35] is applied for illumination adjustment that performs morphological opening with a structuring element SE of size α followed by the difference operation expressed as follows:

$$I_{o1}^{(i)}(x, y) = I^{(i)}(x, y) \circ \mathrm{SE} \tag{9}$$

$$I_{\mathrm{ad}\,j}^{(i)}(x, y) = I^{(i)}(x, y) - I_{\mathrm{open}}^{(i)}(x, y) \tag{10}$$

where $I_{o1}^{(i)}(x, y)$ and $I_{\mathrm{ad}\,j}^{(i)}(x, y)$ represent the morphed, and illumination adjusted frames, respectively. Whereas, $\circ$ is the opening operator. The size α of the structuring element SE is set to 3×3 in order to preserve the effectiveness of frame enhancement through illumination adjustment.

2.2.2 SC detection: As the SCs appear at fixed location in the frames of input cricket video, therefore, temporal image averaging is used to filter-out the SC region from the video frames. A sliding overlapped window of length L frames and step size W is used to compute temporal running average sequence expressed as follows:

$$I_{\mathrm{avg}}^{(i)}(x, y) = \frac{\sum \left\{ I_{\mathrm{adj}}^{(i)}(x, y) \right\}_{i-2}^{i+2}}{L} \tag{11}$$

In (11), $I_{\mathrm{avg}}^{(i)}(x, y)$ represents average at ith frame, and L is the length of sliding window. Morphological opening is performed on the extracted SC region to further enhance the SC contents followed by computation of the first- and second-order statistics that generates the binary image as follows:

$$I_{o2}^{(i)}(x, y) = I_{\mathrm{avg}}^{(i)}(x, y) \circ \mathrm{SE} \tag{12}$$

$$\mu^{(i)} = \frac{1}{rc} \sum_{i=1}^{r} \sum_{j=1}^{c} I_{o2}^{(i)}(x, y) \tag{13}$$

$$\sigma^{(i)} = \frac{1}{rc} \sum_{i=1}^{r} \sum_{j=1}^{c} \left[I_{o2}^{(i)}(x, y) - \mu^{(i)} \right] \tag{14}$$

$$I_{\mathrm{bin}}^{(i)}(x, y) = \begin{cases} 0, & \text{if } \left(\mu^{(i)} - \beta * \sigma^{(i)}\right) \leq I_{o2}^{(i)}(x, y) \leq \left(\mu^{(i)} + \beta * \sigma^{(i)}\right) \\ 1, & \text{otherwise} \end{cases} \tag{15}$$

where $\mu^{(i)}$ and $\sigma^{(i)}$ represent the mean and standard deviation for morphed frame $I_{o2}^{(i)}(x, y)$, and $I_{\mathrm{bin}}^{(i)}(x, y)$ represents the binary image for ith frame, whereas β is a positive real constant that is set to 2.5

after detailed experimentation. The parameter β is used to select the range of intensity to classify each pixel into either foreground or background region. To remove the outliers, two passes of morphological thinning are applied on the binary image $I_{\text{bin}}^{(i)}(x, y)$ to generate thinned image $I_{\text{thin}}^{(i)}(x, y)$ as

$$I_{\text{thin}}^{(i)}(x, y) = I_{\text{bin}}^{(i)}(x, y) \otimes \text{SE} \tag{16}$$

To bridge the gaps and preserving fine details in characters, dilation is applied on $I_{\text{thin}}^{(i)}(x, y)$ as follows:

$$I_{\text{dil}}^{(i)}(x, y) = I_{\text{thin}}^{(i)}(x, y) \oplus \text{SE} \tag{17}$$

where $I_{\text{dil}}^{(i)}(x, y)$ is the dilated image for ith frame, and $\oplus$ is the dilation operator. The transformed image finally contains the SC contents that are more suitable to be processed by the optical character recognition (OCR). The processed SC region obtained after the dilation operation is passed to the OCR algorithm [15] to recognise the characters.

For OCR, connected component analysis is performed initially to store the outlines of the components together into blobs. These blobs are organised into text lines that are further partitioned into characters based on identifying the connected components. Finally, each character is passed to an adaptive classifier for recognition.

2.2.3 Decision-tree-based key-event detection: In cricket, SCs display score and wicket information by using one of two separators '/' or '-', i.e. score/wicket or score-wicket. Therefore, in this work, SCs are processed for key-event detection by designing a five-layer decision tree to detect various key-events, i.e. boundary, six, wicket, and replay. The layout of the proposed decision tree is presented in Fig. 3.

As the broadcasters usually omit SCs during replays due to the disagreement of game stats in live and replay frames, therefore, in the present work we used this observation to classify the frames as either active/live or replay frames. Another observation about the replay segments is that the replay frames are sandwiched between GT frames. Therefore, the detection of GT in the absence of SC is the representation of replay frame. The frames marked as replay frames are also used for video summarisation, i.e. to capture any interesting event (e.g. misfields, overthrows, player collisions etc.) other than the key-events in the form of boundary, six, or wicket. For GT detection, we used a dual-threshold-based method that is proposed in our earlier work [15]. Detected GT frames are used to extract the candidate replay segments. Shown in Fig. 4 are the GT frames of the input cricket video.

For replay event detection, a rule R_2 is defined to classify between the replay and closed frames (CFs) as follows:

$$R_2 = \{\text{if}(\text{SC} != \text{active} \wedge \text{GT} = \text{true}) \text{ then Replay} \tag{18}$$

Absence of the GT and SCs marks the frames as CF, i.e. the frames that cannot be used for the key-events detection. If the SCs are found in active frames, then score separator (SS) is used for separating score value (SV) and wicket value (WV), respectively. The W counter contains the difference between WVs from the current and the previous frames, whereas S counter stores the difference between the SVs from the current and the previous frames. For wicket event, a rule R_3 is defined as follows:

$$R_3 = \{\text{if}(\text{SC} = \text{active} \wedge \text{SS} = \text{WV} \wedge W > 0) \text{ then Wicket} \tag{19}$$

where W is expressed as

$$W = \text{WV}^{(i)} - \text{WV}^{(i-1)} \tag{20}$$

More specifically for wicket event, W counter must contain a value other than zero otherwise no key-event in the form of wicket is detected.

Similarly, for boundary event a rule R_4 is defined as follows:

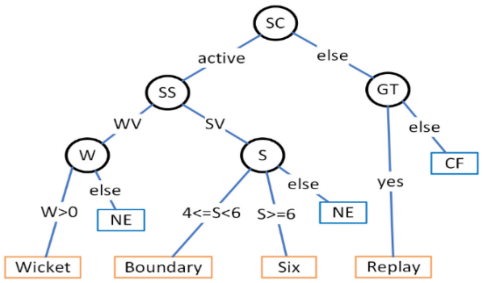

Fig. 3 *Decision tree for key-events classification*

Fig. 4 *GT frames of crick1 video*

$$R_4 = \{\text{if}(\text{SC} = \text{active} \wedge \text{SS} = \text{SV} \wedge 4 \leq S < 6) \text{ then Boundary} \tag{21}$$

where S counter represents the difference in the SV between the current and previous frame and expressed as follows:

$$S = \text{SV}^{(i)} - \text{SV}^{(i-1)} \tag{22}$$

More specifically, if S counter is greater than or equal to 4 and less than 6 then 'boundary' event is detected. The reason to use two thresholds, e.g. 4 and 6, is that: the boundary event can also occur on a no/wide ball and also in the form of misfield/overthrows and increments the score counter S by 5 instead of 4. Score counter is further analysed to detect the six-event by defining a rule R_5 as follows:

$$R_5 = \{\text{if}(\text{SC} = \text{active} \wedge \text{SS} = \text{SV} \wedge S \geq 6) \text{ then Six} \tag{23}$$

More specifically, if S is greater than or equal to 6 then a six-event is detected. The six-event increments the score counter by a factor of either 6 or 7 in a cricket video. In case of 7, six is scored on a no-ball in cricket. For each detected event, the corresponding video frame is marked as a key-frame. A video skim is created for each key-event followed by generating the summarised video according to the user-specified length as proposed in our earlier work [10].

3 Experiments and results

This section presents a detailed discussion of the results and various experiments that are designed to evaluate the performance of the proposed video summarisation system.

3.1 Dataset

Twenty real-world you-tube cricket videos of total duration of 10 h are used for performance evaluation. The dataset consists of videos from different sports broadcasters and tournaments following the similar approach adopted in sports video summarisation research [6, 7, 10, 30, 31]. Each video in the dataset has a frame resolution of 640 × 480 pixels and a frame rate of 25 fps. The cricket videos contain samples from 2006 One Day International (ODI) series between Australia and South Africa, 2014 test series between Australia and Pakistan, 2014 ODI series between South Africa and New Zealand, 2014 (T20) cricket world-cup tournament, and 2015 ODI cricket world-cup tournament. Some snapshots of our dataset are shown in Fig. 5.

As the proposed video summarisation method also performs the acoustic analysis of the audio stream associated with the input cricket videos, an audio dataset was created that consisted of excited and non-excited audio clips. For this, the annotation of the audio stream was performed, and the audio frames with loud commentator voices or crowd cheers were considered as the excited audio frames. Whereas, the other audio stream portions

were considered as non-excited audio frames. The dataset consists of 100 audio clips with 50 audio clips in each class. Thirty clips from both excited and non-excited classes were used for classifier training, whereas, the remaining 40 clips were used to evaluate the classification. The dataset we developed is publically available at [36].

3.2 Performance evaluation

For performance evaluation of the proposed method, precision, recall, F-1 score, accuracy, and error rates are computed. In order to justify the effectiveness of the proposed scheme, we also compared our method against state-of-the-art techniques for summarisation of the cricket videos.

3.2.1 Performance evaluation of excitement-based key-audio frame detection: Performance of the proposed video summarisation framework depends on accuracy of the underlying excitement-based key-audio frame detection. The goal of this experiment is to evaluate the performance of the proposed key-audio frame detection method. The results presented here are averaged over all the audio streams of the input cricket videos. The proposed excitement detection method can reliably detect the key-audio frames and achieved 97.76, 98.87, 97.60, and 98.7%, accuracy, precision, recall, and F-1 score, respectively. The higher evaluation scores are attributed to the robustness of acoustic-LBP features that enable SVM classifier to correctly classify excited and non-excited classes.

Fig. 5 *Snapshots of dataset*

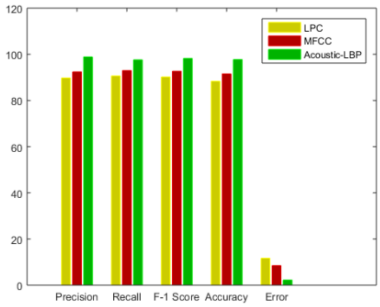

Fig. 6 *Performance comparison of excitement detection with LPC, MFCC, and acoustic-LBP features*

Table 1 Performance comparison of excitement detection methods against the proposed method

Excitement detection methods	Precision, %	Recall, %	Accuracy, %
Baijal *et al.* [9]	97.37	96.77	—
Merler *et al.* [18]	—	—	80
Raventos *et al.* [32]	88	93	—
proposed method	98.87	97.60	97.76

3.2.2 Performance comparison of acoustic-LBP features against spectro-temporal features: We designed an experiment to compare the performance of acoustic-LBP features against the spectro-temporal features [e.g. MFCC, linear predictive coding (LPC)] by training the SVM classifier. Impact of acoustic-LBP features selection on classification performance is highlighted in Fig. 6.

It can be observed from Fig. 6 that acoustic-LBP features significantly improve the classification performance of the excited audio frame selection method as compared to MFCC [9] and LPC [37] features.

3.2.3 Performance comparison of excitement detection against state-of-the-art methods: In this experiment, we evaluated the performance of the proposed excitement detection method in terms of key-events detection against state-of-the-art methods [9, 18, 32]. The brief description of the comparative methods can be found in Section 1. The reason for selecting these methods for comparison is that these methods also perform the excitement detection in audio streams for summarising the sports videos. From the results presented in Table 1, we can observe that the proposed method has the highest precision, recall, and accuracy rates. The performance difference against [18] is even significant, where, our method achieved ∼18% higher accuracy rates that clearly elaborate the effectiveness of the proposed scheme.

3.2.4 Key-event detection evaluation: In our second experiment, the proposed framework is evaluated in terms of key-events detection for cricket videos. The results are obtained by processing the candidate video frames selected by the excitement-based key-audio frame detection stage. The results presented here are averaged over all the 20 input cricket videos of our dataset. The objective evaluation statistics of the proposed system for each of the four key-events, i.e. boundary, six, wicket, and replay are shown in Table 2. Whereas, the graphical illustration of same experiment is presented in Fig. 7. The average precision, recall, accuracy, and error rate of 92.94, 91.04, 95.50, and 4.5% signifies the effectiveness of the proposed key-event detection method.

Confusion matrix (Table 3) is also provided to measure the classification performance of the proposed method for key-events detection. From confusion matrix it can be observed that the classification accuracy of the proposed system for key-events detection is remarkably well for all key-events. The slight degradation in the classification accuracy of boundary event is attributed to the fact that the appearance of advertisement bars in front of SCs. High accuracy of the replay detection, however, compensates this degradation.

3.2.5 Performance comparison against state-of-the-art methods: The goal of this experiment is to validate the performance of the proposed method in terms of video summarisation against state-of-the-art methods specifically designed for cricket videos. To achieve this goal, we compared the performance of the proposed method against [6, 7, 10, 11, 24–27, 31]. The details of the compared methods are already described in Section 1 (Introduction).

For performance evaluation, we have used the similar experimental settings as adopted by the comparative techniques. In this regard, we used YouTube cricket videos as done by the comparative methods. For performance evaluation the only condition that exists in sports video summarisation research is that the dataset should be diverse in terms of content, i.e. it should

Table 2 Key-event detection results for cricket videos

Key-events	Precision rate, %	Recall rate, %	F-1 score, %	Accuracy rate, %	Error rate, %
boundary	95.33	90.64	92.92	92.17	7.83
six	91.78	89.33	90.53	96.08	3.92
wicket	90.47	92.68	91.56	98.04	1.96
replay	94.19	91.53	92.84	95.72	4.28
average	92.94	91.04	91.96	95.50	4.50

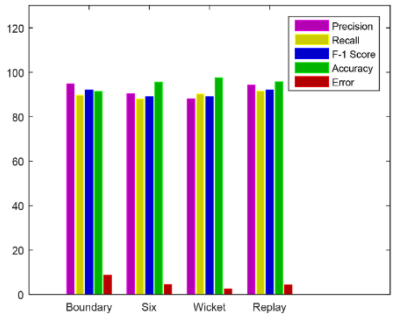

Fig. 7 *Objective evaluation of boundary, six, wicket, and replay events*

Table 3 Confusion matrix analysis

Actual class	Predicted class							
	Boundary		Six		Wicket		Replay	
Classes	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative
positive	182	21	66	09	37	04	292	27
negative	10	145	07	276	05	312	18	715

Table 4 Performance comparison with existing systems

Video summarisation methods	Dataset details				Precision rate, %	Recall rate, %	Accuracy rate, %	Error rate, %
	Frame rate, fps	Length, h	Resolution	Format				
Midhu and Padmanabhan [6]	30	03			88.01	87.91	—	—
Tang *et al.* [7]	25	07			—	—	86.54	13.46
Javed *et al.* [10]	25	06	640 × 480	AVI	91.87	89.85	95.01	4.99
Tavassolipour *et al.*[17]	25	09	640 × 368	MPEG-4	86.9	80.1	81.8	18.2
Namuduri [24]	30	04	352 × 240	—	—	—	87.42	12.58
Wang *et al.* [25]	25	17	—	—	91.3	91	—	—
Godi *et al.* [26]	30	02	100 × 100	—	69	84	78	22
Jiang *et al.* [27]	—	—	256 × 256	—	92.37	87.79	—	—
Kolekar and Sengupta [31]	10	—	—	—	—	—	85.44	14.66
proposed method	25	10	640 × 480	AVI	92.94	91.04	95.50	4.50

Fig. 8 *SC design and placement for cricket videos*

Table 5 Detection performance of key-events for SC design and placement

Videos	SC placement	Precision, %	Recall, %	Accuracy, %	Error, %
Crick4	top-left	93.76	92.31	96.36	3.64
Crick5	bottom	92.24	90.83	94.95	5.05
Crick6	bottom-left	91.15	90.87	92.50	7.50
average	—	92.38	91.34	94.60	5.40

cover different types of videos, in different playing scenarios (e.g. day/night timings, field conditions, playing kits etc.) that comprise of that sport. Therefore, the videos used for evaluation are taken in different time and illumination conditions, e.g. only day videos, only night videos, day and night videos, different pitch and ground conditions, different tournaments, different replay structure, different logos, camera variations, logo placement, SC types, replay speeds etc. Another reason to use YouTube dataset is that currently there is no any standard dataset for sports video summarisation [12, 15, 17].

The statistical comparison in terms of average objective evaluation criterion is shown in Table 4. From the results it can be observed that the proposed method outperforms the comparative methods in terms of precision, recall, accuracy, and error rate.

3.2.6 Robustness to SC detection: Cricket broadcasters use different designs and placement of SCs on the screen in various tournaments. Hence, a video summarisation system for cricket video must be robust against SC's design and placement. As shown in Fig. 8, various designs of the SCs and placement at different positions in the input cricket videos are used to evaluate the key-events detection performance of the proposed method. The average accuracy of 94% (Table 5) signifies the effectiveness of the proposed method for key-events detection irrespective of SC design and placement on the screen.

3.2.7 Camera angle variation: Broadcasters capture and display the live game and replays from different angles and cameras. To overcome the challenges associated with camera angle variations,

Table 6 Detection performance of key-events for camera angle variations

Videos	No. of frames	Camera Views		True positive	True negative	False positive	False negative	Precision rate, %	Recall rate, %	Accuracy rate, %	Error rate
Crick1	316	front	back	292	22	0	02	100	99.31	99.36	0.64
Crick4	341	front	back	290	28	0	23	100	92.65	93.25	6.75
average								100	95.98	96.30	3.70

Table 7 Detection performance of replay events for replay speed variations

Videos	No. of frames	GT start	GT end	True positive	True negative	False positive	False negative	Precision rate, %	Recall rate, %	Accuracy rate, %	Error rate, %	F-1 score
Crick1	316	4	312	292	22	0	02	100	99.31	99.36	0.64	0.99
Crick2	320	16	318	292	25	02	02	99.31	99.31	99.06	0.94	0.99
Crick3	731	71	658	420	294	0	17	100	96.11	97.67	2.33	0.98
average								98.79	95.67	96.78	3.22	0.97

Table 8 Detection performance of key-events for baseball and soccer

Videos	Precision, %	Recall, %	F-1 score, %	Accuracy, %	Error, %
baseball	91.48	91.83	91.65	92.9	7.1
soccer	94.30	92.50	93.39	95.8	4.2
average	92.89	92.05	92.52	94.35	5.65

there is a need to develop an effective key-events detection method independent of the camera variations. The proposed method effectively detects the key-events displayed from various cameras that depict different views of the same event. We designed an experiment to evaluate the robustness of the proposed method against camera variations for key-events detection. In our dataset, we obtained various views of the same event captured at different angles. The detection performance of the proposed method for key-events displayed from various camera angles is presented in Table 6.

The average accuracy of 96.4% signifies the effectiveness of the proposed method to successfully detect various key-events that are robust to the camera angle variations.

3.2.8 Replay speed variation: Cricket broadcasters display the replay event at various speeds. Replay detection methods thus must be able to detect the replays independent of the speed. We have designed an experiment to evaluate the performance of the proposed method to detect replay events of various speeds. For this purpose, we selected the cricket videos consisting of replays of different speeds from our dataset to measure the robustness of our method for replay speed variation. As shown in Table 7, experimental results indicate the effectiveness of the proposed method in terms of detecting replays of different speeds.

3.2.9 Performance evaluation on multiple sports: In this experiment, the proposed method is evaluated in terms of key-events detection for baseball and soccer videos. For baseball, we selected two key-events that are hit-&-run, and strikeout. Similarly, for soccer videos we selected goal and foul as key-events. The soccer videos are different from the cricket videos in a sense that the foul events are not observable from the SCs; therefore, our method relies on the acoustic analysis, where the foul events are detectable through a whisle.

The results obtained for key-event classes on baseball and soccer videos are shown in Table 8; where, the average precision, recall, F-1 score, accuracy, and error rate of the proposed method are 92.89, 92.05, 92.52 94.35%, and 5.65, respectively. The results clearly depict that the proposed method is independent of the sports categories and is not bounded only to the cricket videos.

3.3 Discussion

In the present work, two separate rules for boundary and six events detection are provided for the flexibility in the video summarisation. This scheme allows us to generate key-event specific summaries as well, e.g. all sixes, or all boundaries. The proposed replay detection framework exploits an occurrence of a significant event (e.g. six, boundary, or wicket) excluding *over throws*, *drop catches*, *missed run outs*; therefore, proposed framework has the ability to extract every important event in the generated video summary. The proposed replay detection method also serves as the stored energy hub to overcome the failure of rules for key-events detection. It is important to mention that the proposed replay detection method is unable to classify key-events. The key-event specific rules are therefore required for event-driven video summarisation.

It can be observed from Tables 5–7 that the proposed method is robust to video recording parameters including camera variations, replay speed, logo design (size, placement etc.), SCs design (size, placement etc.), broadcasters, game categories (i.e. ODI, test etc.), and lighting conditions (i.e. day time, night time). Due to the application benefits, it is reasonable to argue that the proposed hybrid scheme is an effective way to generate user-driven summaries for cricket videos.

4 Conclusion

This paper presents a computationally efficient and an effective sports video summarisation framework. The proposed multimodal framework exploits excitement-level variations in the audio stream to detect the excited key-audio frames that are then used to select candidate video key-frames. The selected video frames are analysed to detect and localise SCs and track its variations to detect the key-events such as boundaries, six, wicket, and replays. Each frame that falls in a key-event is labelled as a key-frame. A decision tree-based classifier is trained to detect key-events in the input cricket videos. The key-frames are used with their neighbouring frames to generate the video summary of user-specified length. The proposed framework effectively summarises long-duration cricket videos that facilitate the broadcasters to store and transmit these concise videos over low-bandwidth networks. Experimental results illustrate that the proposed method is capable of detecting key-events and generating exciting summaries. Cricket is selected as a test case because of the longest match durations and broadcasting time concerns that makes it more challenging compared against several other sports.

5 Acknowledgment

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education(2016R1D1A1B03933860)

6 References

[1] Meng, J., Wang, S., Wang, H., *et al.*: 'Video summarization via multiview representative selection', *IEEE Trans. Image Process.*, 2018, **27**, (5), pp. 2134–2145

[2] Cirne, M.V.M., Pedrini, H.: 'VISCOM: A robust video summarization approach using color co-occurrence matrices', *Multimedia Tools Appl.*, 2018, **77**, (1), pp. 857–875

[3] Zawbaa, H. M., El-Bendary, N., Hassanien, A. E., *et al.*: 'Machine learning-based soccer video summarization system'. Proc. Int. Conf. Multimedia, Computer Graphics and Broadcasting, Berlin, Heidelberg, 2011, pp. 19–28

[4] Boukadida, H., Berrani, S.A., Gros, P.: 'A novel modeling for video summarization using constraint satisfaction programming'. Proc. Int. Sym. Visual Computing, Las Vegas, USA, December 2014, pp. 208–219

[5] Yao, T., Mei, T., Rui, Y.: 'Highlight detection With pairwise deep ranking for first-person video summarization'. Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, Las Vegas, USA, June 2016, pp. 982–990

[6] Midhu, K., Anantha Padmanabhan, N.K.: 'Highlight Generation of Cricket Match Using Deep Learning'. In Hemanth, D., Smys, S. (eds): '*Computational Vision and Bio Inspired Computing*'. Lecture Notes in Computational Vision and Biomechanics, vol 28 (Springer, Cham, 2018), pp. 925–936

[7] Tang, H., Kwatra, V., Sargin, M.E., *et al.*: 'Detecting highlights in sports videos: cricket as a test case'. Proc. Int. Conf. IEEE Multimedia and Expo, Barcelona, Spain, July 2011, pp. 1–6

[8] Bertasius, G., Park, H.S., Stella, X.Y., *et al.*: 'Am I a baller? Basketball performance assessment from first-person videos'. Proc. IEEE Int. Conf. on Computer Vision, Venice, Italy, October 2017, pp. 2196–2204

[9] Baijal, A., Cho, J., Lee, W., *et al.*: 'Sports highlights generation based on acoustic events detection: a rugby case study'. Proc. IEEE Int. Conf. on Consumer Electronics, Las Vegas, USA, January 2015, pp. 20–23

[10] Javed, A., Bajwa, K.B., Malik, H., *et al.*: 'A hybrid approach for summarization of cricket videos'. Proc. IEEE Int. Conf. Consumer Electronics-Asia, Seoul, South Korea, October 2016, pp. 1–4

[11] Chen, C.M., Chen, L.H.: 'A novel method for slow motion replay detection in broadcast basketball video', *Multimedia Tools Appl.*, 2015, **74**, (21), pp. 9573–9593

[12] Nguyen, N., Yoshitaka, A.: 'Shot type and replay detection for soccer video parsing'. Proc. Int. Symp. Multimedia, Irvine, USA, December 2012, pp. 344–347

[13] Chen, C.M, Chen, L.H.: 'Novel framework for sports video analysis: A basketball case study'. Proc. Int. Conf. in Image Processing, Paris, France, October 2014, pp. 961–965

[14] Rui, Y., Gupta, A., Acero, A.: 'Automatically extracting highlights for TV baseball programs'. Proc. ACM Int. Conf. Multimedia, California, USA, October 2000, pp. 105–115

[15] Javed, A., Bajwa, K.B., Malik, H., *et al.*: 'An efficient framework for automatic highlights generation from sports videos', *IEEE Signal Process. Lett.*, 2016, **23**, (7), pp. 954–958

[16] Hu, H.N., Lin, Y.C., Liu, M.Y., *et al.*: 'Deep 360 pilot: learning a deep agent for piloting through 360 sports video'. Proc. Int. Conf. in Computer Vision and Pattern Recognition, Honolulu, Hawaii, July 2017, pp. 3–15

[17] 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), pp. 291–304

[18] Merler, M., Joshi, D., Nguyen, Q.B., *et al.*: 'Automatic curation of golf highlights using multimodal excitement features'. Proc. Int. Conf. in Computer Vision and Pattern Recognition Workshops, Honolulu, Hawaii, July 2017, pp. 57–65

[19] Ma, Y.F., Hua, X.S., Lu, L., *et al.*: 'A generic framework of user attention model and its application in video summarization', *IEEE Trans. Multimedia*, 2005, **7**, (5), pp. 907–919

[20] Kapela, R., McGuinness, K., O'Connor, N.E.: 'Real-time field sports scene classification using colour and frequency space decompositions', *J. Real-Time Image Process.*, 2017, **13**, (4), pp. 725–737

[21] Kolekar, M.H., Sengupta, S.: 'Bayesian network-based customized highlight generation for broadcast soccer videos', *IEEE Trans. Broadcast.*, 2015, **61**, (2), pp. 195–209

[22] Homayounfar, N., Fidler, S., Urtasun, R.: 'Sports field localization via deep structured models'. Proc. Int. Conf. in Computer Vision and Pattern Recognition, Honolulu, Hawaii, July 2017, pp. 5212–5220

[23] Hannane, R., Elboushaki, A., Afdel, K., *et al.*: 'An efficient method for video shot boundary detection and keyframe extraction using SIFT-point distribution histogram', *Int. J. Multimedia Inf. Retr.*, 2016, **5**, (2), pp. 89–104

[24] Namuduri, K.: 'Automatic extraction of highlights from a cricket video using MPEG-7 descriptors'. Proc. Int. Workshop on Communication Systems and Networks, Bangalore, India, January 2009, pp. 1–3

[25] Wang, Z., Yu, J., He, Y.: 'Soccer video event annotation by synchronization of attack–defense clips and match reports with coarse-grained time information', *IEEE Trans. Circuits Syst. Video Technol.*, 2017, **27**, (5), pp. 1104–1117

[26] Godi, M., Rota, P., Setti, F.: 'Indirect match highlights detection with deep convolutional neural networks'. Proc. Int. Conf. on Image Analysis and Processing, Catania, Italy, September 2017, pp. 87–96

[27] Jiang, H., Lu, Y., Xue, J.: 'Automatic soccer video event detection based on a deep neural network combined CNN and RNN'. Proc. Int. Conf. in Tools with Artificial Intelligence, San Jose, CA, USA, November 2016, pp. 490–494

[28] Xu, M., Maddage, N.C., Xu, C., *et al.*: 'Creating audio keywords for event detection in soccer video'. Proc. Int. Conf. IEEE Multimedia and Expo, Baltimore, USA, July 2003, pp. 281–284

[29] Tang, S., Zhi, M.: 'Summary generation method based on audio feature'. Proc. IEEE Int. Conf. in Software Engineering and Service Science, Beijing, China, September 2015, pp. 619–623

[30] Kolekar, M.H., Sengupta, S.: 'Semantic concept mining in cricket videos for automated highlight generation', *Multimedia Tools Appl.*, 2010, **47**, (3), pp. 545–579

[31] Kolekar, M.H., Sengupta, S.: 'A hierarchical framework for generic sports video classification'. Proc. Asian Conf. on Computer Vision, Berlin, Germany, January 2006, pp. 633–642

[32] Raventos, A., Quijada, R., Torres, L., *et al.*: 'Automatic summarization of soccer highlights using audio-visual descriptors', *Springer Plus*, 2015, **4**, (1), pp. 1–19

[33] Chatlani, N., Soraghan, J.J.: 'Local binary patterns for 1-D signal processing'. Proc. European Conf. in Signal Processing, Aalborg, Denmark, August 2010, pp. 95–99

[34] Potapov, D., Douze, M., Harchaoui, Z., *et al.*: 'Category-specific video summarization'. Proc. Int. Conf. in European Conf. on Computer Vision, Zurich, Switzerland, September 2014, pp. 540–555

[35] Kushol, R., Kabir, M.H., Salekin, M.S., *et al.*: 'Contrast enhancement by top-hat and bottom-hat transform with optimal structuring element: application to retinal vessel segmentation'. Proc. Int. Conf. Image Analysis and Recognition, Montreal,Canada, July 2017, pp. 533–540

[36] 'Dataset Link', Available at: http://www-personal.engin.umd.umich.edu/~hafiz/projs/avs.htm, accessed May 2018

[37] Zeng, C., Dou, W.: 'Audio keywords detection in basketball video'. Proc. Int. Conf. in Audio Language and Image Processing, Shangai, China, November 2010, pp. 1765–1770