ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/312336197

A hybrid approach for summarization of cricket videos

Conference Paper · October 2016

DOI: 10.1109/ICCE-Asia.2016.7804835

CITATIONS
20

READS
172

5 authors, including:

Ali Javed
University of Engineering and Technology, Taxila
109 PUBLICATIONS **953** CITATIONS
SEE PROFILE

Khalid Bashir Bajwa
University of Engineering and Technology, Taxila
13 PUBLICATIONS **309** CITATIONS
SEE PROFILE

Hafiz Malik
University of Michigan-Dearborn
107 PUBLICATIONS **1,124** CITATIONS
SEE PROFILE

Aun Irtaza
University of Engineering and Technology, Taxila
91 PUBLICATIONS **1,019** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Forensic Analysis, Machine Learning, and Information Retrieval (FAMLIR) Research Group View project

Advertiser's perception of Internet marketing for small and medium enterprises in Pakistan View project

All content following this page was uploaded by Ali Javed on 15 May 2019.

The user has requested enhancement of the downloaded file.

A Hybrid Approach for Summarization of Cricket Videos

Ali Javed[1], Khalid Bashir Bajwa[1], Hafiz Malik[2], Aun Irtaza[1], Muhammad Tariq Mahmood[3]

[1]*Faculty of Telecom and Information Engineering, University of Engineering and Technology-Taxila, Pakistan*

[2]*Department of Electrical and Computer Engineering, University of Michigan – Dearborn, USA*

[3]*School of Computer Science and Engineering, University of Technology and Education, Cheonan, Korea.*

E-mail addresses: ali.javed@uettaxila.edu.pk, khalid.bashir@uettaxila.edu.pk, hafiz@umich.edu, aun.irtaza@uettaxila.edu.pk, tariq@koreatech.ac.kr

Abstract

This study proposes an automatic method for key-events detection and summarization for cricket videos, particularly because of the longest match durations, broadcasting time concerns and largely available multimedia content. In the proposed work, first rule-based induction is applied to detect excited audio clips in cricket videos, and then a decision tree framework is designed for video summarization. The proposed method evaluated on a diverse dataset with average accuracy of 95% signifies the effectiveness in terms of video summarization. Hence, the cricket videos can reliably be broadcasted over the low-bandwidth networks and transmission with time constraints.

Keywords: Decision tree, event detection, score-caption, video summarization

1. Introduction

Sports broadcasters generate a massive collection of videos for online distribution over various networks. The processing, transmission, and storage of such a massive content is a challenging task. Video summarization [1-3] is commonly used to address the aforementioned challenges by providing a short synopsis of a full-length video. The reasons to summarize sports videos are: transmission time constraints, broadcasting requirements over low-bandwidth networks, storage cost, and viewership interest in only the exciting segments. Existing state-of-the-art for video summarization can be categorized into two broad categories i.e. learning-based[1, 2] and non-learning-based [3] methods. Learning-based methods employ various classifiers to detect significant events for video summarization. [2] presented a learning-based event detection and video summarization framework for soccer. Neural networks and SVM classifiers are trained to identify the logo frames that are then used for replay detection and highlights generation. [4] trained the HMM model to propose a replay event detection system based on slow motion segment identification and extraction. This method depends on the homogeneity of slow motion speed for entire replay. The accuracy of this approach degrades significantly in case of variation in replay speed. Beside learning-based methods, non-learning-based methods are also employed for video summarization [5]. [3] proposed a method based on matching the game statistics with the text content of overlays to detect key-events from the baseball videos. [6] used histogram difference and contrast features to identify logo frames for replay event detection. Performance of this method depends on the presence of logo frames in the input video.

Existing state-of-the-art for sports video summarization are mostly designed for soccer, baseball, tennis etc. Cricket is a challenging sport to address for video summarization due to long duration of the game that can range from few hours (i.e. T20 format) to around 40 hours (i.e. test format). There exist a need to develop effective techniques for video summarization that can successfully detect key-events and generate the summary of long duration cricket videos. Key-events consist of interesting segments in the game that excites the viewers. In cricket videos, some of the key-events are boundary, six, wicket, and replays. Generally, existing event detection methods have limitations of dependency on score-caption designs, camera variations, replay speed, logo design, size and placement.

In this work, we propose a hybrid method for summarizing the cricket videos. The hybrid structure reaps the benefits of learning and non-learning-based methods and finds the correlations between the audio and the visual features for an effective video summarization. The proposed method is able to detect replay events that are independent of camera variations, replay speed, logo design, score-caption designs and placement. The flowchart of the proposed system is shown in the Fig.1.

978-1-5090-2743-9/16/$31.00 ©2016 IEEE

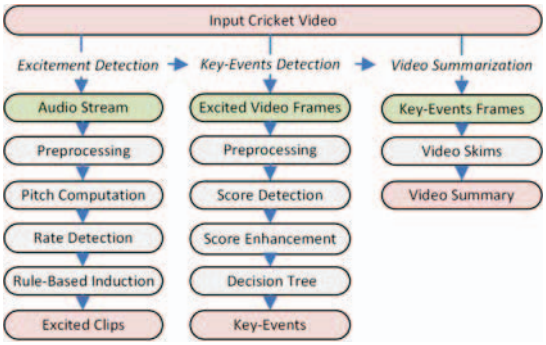

Fig. 1 *Block diagram of the proposed system*

2. Proposed Method

2.1. *Excitement Detection Framework*

The audio signals are divided into various non-overlapping short duration frames. Let $y(k)$ be the audio signal having K samples i.e., $k = 1, 2, \cdots, K$, associated with a cricket video containing a set of N' frames $\{I_i'\}_{i=1}^{i=N'}$. First, spectral subtraction method [7] is applied to reduce the background noise in the audio stream as a preprocessing step. Then, the audio stream is divided into M frame-sets such that each frame-set consists of the audio signal of $t = 10$ second duration and contains F numbers of frames. Then for each frame in the frame-set, the pitch p_i, $i = 1, 2, \cdots, F$ is computed through the autocorrelation based method [8]. The pitch measure is analyzed to identify the excitement in the commentary and audience cheers. Two quantitative measures high pitch rate $R_{hp} = \frac{1}{F}\left(C_{hp} \times 100\right)$ and low pause rate $R_{lp} = \frac{1}{F}\left(C_{lp} \times 100\right)$ are computed by using the pitch information. C_{hp} and C_{lp} are the counters for R_{hp} and R_{lp}, respectively that can be computed as,

$$C_{hp} = \begin{cases} C_{hp} + 1, & if \;\; p_i > T_1 \\ C_{hp}, & otherwise \end{cases} \tag{1}$$

and

$$C_{lp} = \begin{cases} C_{lp} + 1, & if \;\; p_i = 0 \\ C_{lp}, & otherwise \end{cases} \tag{2}$$

where $T_1 = \alpha \times \frac{1}{F}\sum_{i=1}^{F} p_i$ i.e. the threshold T_1 is computed by calculating the mean of the pitch values over all frames in each frame-set and multiplied by a constant $\alpha > 1$. If the maximum pitch of an audio frame exceeds the threshold T_1, then the C_{hp} is incremented by one. Similarly, if the pitch of an audio frame is equivalent to zero then C_{lp} is incremented by one.

Finally, rule-based induction is used to discriminate between excited and non-excited clips. For excitement detection, the following rule is defined,

$$\begin{cases} exited, & if \; R_{hp} \geq T_2 \;\; and \; R_{lp} \leq T_3 \\ non-excited, & otherwise \end{cases} \tag{3}$$

where $T_2 = \frac{1}{M}\sum_{i=1}^{M}\left(R_{hp}\right)_i$ and $T_3 = \frac{1}{M}\sum_{i=1}^{M}\left(R_{lp}\right)_i$ are two thresholds and are computed by averaging C_{hp} and C_{lp} over the all frame-sets. If the R_{hp} of an audio frame-set exceeds the threshold T_2 and R_{lp} lies below the threshold T_3, then the corresponding frame-set is marked as an excited clip otherwise the frame set is marked non-excited clip.

2.2. *Event Detection Framework*

At the second stage, the corresponding video frames of excited audio clips are fed to the proposed event detection framework to detect key-events for cricket. Let $\{I_i\}_{i=1}^{i=N}$ be the excited video frames, where $N << N'$. In preprocessing step, the input color video frames are transformed into grayscale images. The sequence of grayscale images are down-sampled by a factor of 2 and processed for illumination adjustment via top hat filtering. It is observed after watching enormous amount of cricket videos that the score-caption (*SC*) appear in every frame. Therefore, temporal image averaging is used to filter out the SC region by using a sliding window of length $L = 5$;

$$I_{avg}^{(i)} = \frac{I^{(i)} - I^{(i-1)} + I^{(i+1)}}{L} \tag{4}$$

Where $I_{avg}^{(i)}$ represents average at i^{th} frame, $I^{(i+1)}$ and $I^{(i-1)}$ represent the entering and exit frame in a sliding window. In order to enhance the *SC* region, morphological opening is used. The morphed image is subtracted from the extracted image of SC as follows:

$$I_{open}^{(i)} = \left(I_{avg}^{(i)} \circ se\right) \tag{5}$$

$$I_{sub}^{(i)} = \left(I_{avg}^{(i)} - I_{open}^{(i)}\right) \tag{6}$$

Where $I_{sub}^{(i)}$, $I_{avg}^{(i)}$, and $I_{open}^{(i)}$ represent the subtracted image, averaged grayscale image, and the morphed image respectively, $\circ$ is the opening operator, and *se* is the structuring element. In the next step, the

subtracted image $I_{sub}^{(i)}$ is transformed into the binary image $I_{bin}^{(i)}$. To remove the outliers, two passes of morphological thinning are also applied on the enhanced SC regions. The processed *SC* region $I_{thin}^{(i)}$ is passed to the optical character recognition (OCR) algorithm[9] to recognize the characters. In cricket, SCs score and wicket are displayed by using separator "/" or "-" i.e. score/wicket or score-wicket. In key event detection, a decision tree (set of rules) is designed to categorize the boundary, six, wicket, and replay events. The layout of the decision tree is shown in Fig. 2. A summary of rules for events Wicket, Boundary, Six, and Replay are given below.

$$\begin{cases} if\ (SC = active \wedge SS = WV \wedge W > 0)\ then\ \text{Wicket} \\ if\ (SC = active \wedge SS = SV \wedge 4 \le S < 6)\ then\ \text{Boundary} \\ if\ (SC = active \wedge SS = SV \wedge S \ge 6)\ then\ \text{Six} \\ if\ (SC! = active \wedge GT = true)\ then\ \text{Replay} \end{cases}$$

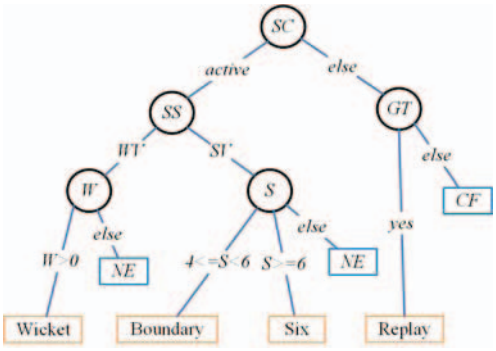

Fig. 2. Decision Tree for Key-Events Classification

At the root node *SC* is examined. 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 exploited 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 gradual transition (*GT*) frames. Therefore, the detection of *GT* in the absence of *SC* is the representation of replay frame. A dual threshold based method [10] is used to detect GT in the proposed work. Absence of the GT and SCs make the frames as closed frame (*CF*) i.e. the frames that cannot be used for the key-events detection. If the SCs are active frames then score separator (*SS*) is used for separating score value (*SV*) and wicket value (*WV*). The W counter contains the difference between wicket values from the current and the previous frames whereas, S counter stores the difference between the score values from the current and the previous frames. These counters are utilized to identify the wicket, six and boundary events. If W counter is greater than zero then there is a wicket event otherwise no key-event (*NE*) is detected. Similarly, if the *S* counter has value greater than or equal to 6 then there is "six" event, if the *S* value is between 4 and 6 then "boundary" event is detected otherwise frame is discarded. For each detected event, the corresponding video frame is marked as a key-frame. The thresholds for boundary, six, and wicket events are selected according to the expected increments in score and wicket values in a cricket video.

2.3. Video Summarization

A video skim against each detected key-event is generated and the video is summarized by obtaining the required length from the user. Video skim duration for each key-event is computed as,

$$L_{key} = \frac{L_{sum}}{N_{key}} \tag{7}$$

where L_{key} represents the length for each key-event, L_{sum} represents the length of video summary required by the user, and N_{key} represent the total number of key-events. The final length of the video skim for each key-event $\beta L_{key} + L_{key} + \lambda L_{key}$, $(0 < \beta, \lambda < 1)$ is obtained by including the frames at prior and after each key-frame. This procedure of video skim generation preserves the temporal information in the summarized video. All video skims are appended in the chronological order to generate the final summary for cricket videos.

3. Result

Performance of the proposed system is evaluated on a diverse dataset of various broadcasters consisting of 20 real-world cricket videos of a total duration of 362 minutes. Each video in the dataset has a frame resolution of 640 x 480 pixels and frame rate of 25 fps. Effectiveness of the proposed system is evaluated by detecting key-events to generate the summarized video using well-known qualitative measures including Precision, Recall, Accuracy and Error rate. The qualitative results are shown in Table. 1.

Table 1: Key-Event Detection results for Cricket Videos.

Key-Events	Precision	Recall	Accuracy	Error
Boundary	94.79%	89.65%	91.34%	8.66%
Six	90.41%	88%	95.53%	4.47%
Wicket	88.09%	90.24%	97.48%	2.52%
Replay	94.19%	91.53%	95.72%	4.28%
Average	**91.87%**	**89.85%**	**95.01%**	**4.99%**

From the results, it can be observed that the proposed system has detected key-events effectively and classified them with reasonable accuracy for video summarization. It is worth mentioning that the

event detection framework improves the overall performance of the system at the cost of relatively higher computational requirement.

In our second experiment, we have computed the entropy to various rules in the decision tree for key-events detection. Table. 2 presents the detail of entropy values of various key-events. Low entropy values signify better predictability of the applied rules. Hence, the selected rules reliably detect the boundary, six, wicket, and replay events for cricket video summarization.

Table 2: Entropy for decision rule of key-events

Key-Event	Entropy
Boundary	0.46
Six	0.52
Wicket	0.44
Replay	0.4

In our last experiment, performance of the proposed system is compared with existing cricket video summarization systems. We compared the accuracy and error rates of the proposed system with existing cricket video summarization systems. Performance comparison of the selected [11-13] and proposed system is provided in Fig. 3. It can be observed from Fig. 3 that the proposed system has provided superior detection performance as compared to the existing state-of-the-art systems.

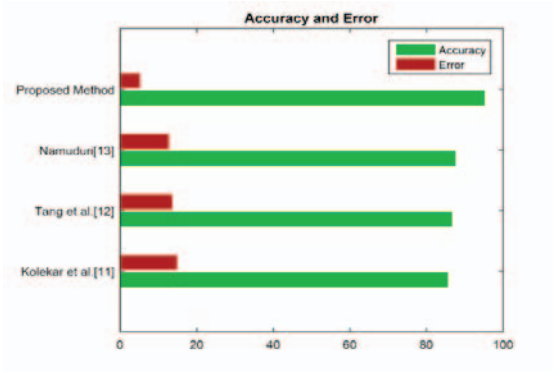

Fig. 3. Performance Comparison

4. Conclusion

In conclusion, in this work, we presented an effective video summarization framework for cricket videos by finding the correlations between audio and video frames. Rule-based induction is applied to detect the excited clips in the audio stream. Decision tree architecture is used to detect key-events in the input cricket videos that are used for video summarization. In the present work, Cricket is particularly chosen due to longest match durations and broadcasting time concerns that makes it even a more challenging problem as compared against several other sports.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science, ICT & Future Planning (MSIP) (2013R1A1A2008180).

References

[1] A. Ekin, A. M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *Image Processing, IEEE Transactions on,* vol. 12, pp. 796-807, 2003.

[2] H. M. Zawbaa, N. El-Bendary, A. E. Hassanien, and T.-h. Kim, "Machine learning-based soccer video summarization system," in *Multimedia, Computer Graphics and Broadcasting*, ed: Springer, 2011, pp. 19-28.

[3] N. Babaguchi, Y. Kawai, and T. Kitahashi, "Generation of personalized abstract of sports video," in *null*, 2001, p. 158.

[4] H. Pan, P. Van Beek, and M. I. Sezan, "Detection of slow-motion replay segments in sports video for highlights generation," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, 2001, pp. 1649-1652.

[5] H. H. Kim and Y. H. Kim, "Generic speech summarization of transcribed lecture videos: Using tags and their semantic relations," *Journal of the Association for Information Science and Technology,* vol. 67, pp. 366-379, 2016.

[6] N. Nguyen and A. Yoshitaka, "Shot type and replay detection for soccer video parsing," in *Multimedia (ISM), 2012 IEEE International Symposium on*, 2012, pp. 344-347.

[7] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *Acoustics, Speech and Signal Processing, IEEE Transactions on,* vol. 27, pp. 113-120, 1979.

[8] L. R. Rabiner, "On the use of autocorrelation analysis for pitch detection," *Acoustics, Speech and Signal Processing, IEEE Transactions on,* vol. 25, pp. 24-33, 1977.

[9] R. Smith, "An overview of the Tesseract OCR engine," in *icdar*, 2007, pp. 629-633.

[10] H. Zhang, A. Kankanhalli, and S. W. Smoliar, "Automatic partitioning of full-motion video," *Multimedia systems,* vol. 1, pp. 10-28, 1993.

[11] Kolekar, Maheshkumar H., and Somnath Sengupta, "A hierarchical framework for generic sports video classification," *In Asian Conference on Computer Vision,* pp. 633-642. Springer Berlin Heidelberg, 2006.

[12] Tang, H., et al. *Detecting highlights in sports videos: Cricket as a test case*. in *Multimedia and Expo (ICME), 2011 IEEE International Conference on*. 2011. IEEE.

[13] Namuduri, K. Automatic extraction of highloghts from a cricket video using MPEG-7 descriptors. in Communication systems and networks and workshops, 2009. *COMSNETS 2009. First International Conference on,* 2001. IEEE.

View publication stats