Significant Event Detection in Sports Video Using Audio Cues

Pradeep K
M. Tech Dept. of CSE
SITCOE,Yadrav
Ichalkaranji

***Abstract*— This work presents audio cue based approaches for event detection from sports video for construction of sports video summary. The approach has been proved effective applied to soccer and cricket videos. The approach extracts the event based on audio level, the sports matches always have mass spectators who show there response in the form of loud cheering, applause and referees signal also. These audio symbols can be detected by measure audio level and pattern which is normally higher than regular audio level. The co-relation between audio and video frames will help to extract frames and summarize these audio samples to combine event, all such events expected together can be used for summary. Proposed Methodology detects the peak values of the audio samples present in the sports video streams and looks for the video frame corresponding to the audio peak detected. The video event detection is generated as AVI streams using the video frames around the audio peaks detected. The experimental results found to match the manual examination and ground truth of the various sports videos.**

Index Terms— Broadcast video, semantic event detection, sports audio, Amplitude detection.

I. Introduction

The videos from web storages are presented to users from index, highlight, summary and thumbnails for user choice for watch importance of the video component. In recent years, the amount of digitized video content has been increasing rapidly and users need to access their content through various network solutions and digital equipments create summary form. Therefore, automatic video analysis, for extracting highlights in sports video to make it possible to deliver video clips into mobile devices or Web browsers becoming a need. Moreover, it would be very attractive if users can access and view the content based on their own preferences. To realize above needs, the source video has to be tagged with semantic labels. These labels must not only be broad to cover the general events in the video, e.g., goals scored, near-misses, fouls in soccer, it has to be also sufficiently deep to cover the semantics of the events, e.g., names of the players involved. This is a very challenging task and would require exploiting multimodal and multi context approaches.

According to different production and edition styles, videos can be classified into two major categories: scripted and non scripted [1] which are usually associated with different video mining tasks. Scripted videos, e.g., news and movies, are produced or edited according to a pre-defined script or plan, for which we can build a Table-of-Content (TOC) [2] to facilitate video search. Non-scripted videos such as meeting, sports, and surveillance videos do not have a pre-defined script, where however, all events happen spontaneously and usually in a relatively fixed setting. Therefore, detecting the highlights or events of interests is an issue for non-scripted videos. The ever increasing amount of video archives makes manual annotation extremely time consuming and expensive. Since humans tend to use high-level semantic concepts when querying and browsing video database, there is an increasing need for automatic video semantic annotation. Video annotation is also able to facilitate video semantic analysis such as video summarization and personalized video retrieval.

There are some works on general video highlight detection by replay shot detection [3], video activity analysis, audio analysis [4], structure analysis [5] and mid-level semantic extraction [6]. These works can only coarsely find highlights in sports video while most of the users look at the semantic events like “shoot on goal” in soccer, “goal” in basketball etc. For the reason that detecting events in general sports video is still an open problem at present, most of the researchers focus on a special kind of sports video, including American football [7], baseball [8], etc. in which soccer video is mostly concerned [9] for its high audience rating.

This works, presents a novel approach for *Significant Event Detection from Sports Video Using Audio Cues.* The audio peaks levels from the sports video are identified and corresponding events from video are extracted and video summary consisting of the events of importance is generated. The algorithm is tested on soccer and cricket videos. The proposed approach is generic and can be extended to other sports domains also the experimental results have proven the effectiveness of the algorithm.

The rest of the paper is organized into four sections as follows. First, we discuss current technologies and importances of event detection. In Section 2, we explain the related work use of event detection and in the next

Section we discuss the Proposed Algorithm is used to detect events in sports games. Finally, we present promising experimental results in Section 4 followed by a conclusion.

II. Related Work

The Significant event detection from sports video is essential for sports video summarization and retrieval. However, the existing sports video event detection approaches heavily rely on either video content itself, which face the difficulty of high-level semantic information extraction from video content using computer vision and image processing techniques, or manually generated video ontology, which is domain specific and difficult to be automatically aligned with the video content.

Many techniques have been proposed for video event detection and summarization based on supervised and unsupervised learning. Unsupervised framework mines the semantic audio-visual labels so as to detect "interesting" events. The proposed method [10] then use a Hidden Markov Model based approach to control the length of the summary. Two methods have been proposed for generating a summary of arbitrary length for large sports video archives. One is to create a concise video clip by temporally compressing the amount of the video data. The other is to provide a video poster by spatially presenting the image key frames which together represent the whole video content. These methods discussed in [11] deal with the metadata which has semantic descriptions of video content. Summaries are created according to the Significance of each video segment which is normalized in order to handle large sports video archives.

A novel algorithm for video scene segmentation has been explained in [12]. It models a scene as a semantically consistent chunk of audio-visual data. Central to the segmentation framework is the idea of a finite-memory model. Technique separately segments the audio and video data into scenes, using data in the memory. The audio segmentation algorithm determines the correlations amongst the envelopes of audio features. The video segmentation algorithm determines the correlations amongst shot key-frames. The scene boundaries in both cases are determined using local correlation minima. Then, it fuse the resulting segments using a nearest neighbor algorithm that is further refined using a time-alignment distribution derived from the ground truth.

Extensive research efforts have been devoted to sports video event detection in recent years. Exciting event detection in broadcast soccer video with mid-level description and SVM-based incremental learning is presented in [13]. In this method, video frames are firstly classified and grouped into views in terms of low-level playfield features. Mid-level description including view label, motion descriptor and shot descriptor are then extracted to present the characteristics of a view. By using the fixed temporal structure of views, SVM classification models are constructed to detected exciting events in a soccer match.

In [14] author present a novel approach for sports video semantic event detection based on analysis and alignment of webcast text and broadcast video. Webcast text is a text broadcast channel for sports game which is co-produced with the broadcast video and is easily obtained from the web. We first analyze webcast text to cluster and detect text events in an unsupervised way using probabilistic latent semantic analysis (pLSA). Based on the detected text event and video structure analysis, he employ a conditional random field model (CRFM) to align text event and video event by detecting event moment and event boundary in the video.

In [15] author proposes a new two-level framework to analyze high-level structure of video and to detect useful events automatically based on visual keywords. The first level extracts low-level features such as motion, color, texture etc to detect video segments boundaries and label segments as visual keywords. Next he apply an event detection grammar to the visual keywords sequence at the second level to detect video segments that match the predefined event model. The exact position of the segment that the event occurs can also be spotted. Author propose a novel approach for detecting highlights using easy-to-extract low-level visual features such as the color histogram(CH) or histogram of oriented gradients (HOG) in [16]. In particular, he focus on cricket which is an outdoor bat-and-ball team sport similar to baseball. It is played professionally in many countries, and has become the world’s second most popular sport after soccer. Even though our methodology does not use sport-specific features, he chosen cricket as a test case in part due to the availability of a large labeled dataset of video clips from a cricket tournament. There has been prior work on cricket highlights generation based on domain-specific modeling of semantic concepts and events, but it is highly customized for cricket. A novel framework for video summarization is proposed in [17] which obtains multiple index features from video frames and combines to describe the frame difference between consecutive frames. It is observed that certain frame difference features have more influence in generating a representative frame difference measure. Fuzzy Comprehensive Evaluation had been used to evaluate the efficiency of a particular frame difference measure based on the user’s feedback about summaries and thus generating the weights of each measure.

In this paper, we present a novel approach for ***Significant Event Detection from Sports Video Using Audio Cues.*** The audio peaks of the sports video are identified and corresponding events from video are extracted and video summary consisting of the events of importance is generated. We use soccer video and baseball video as our test-bed. The proposed approach is generic and can be extended to other sports domains. The experimental results found to match the manual examination of the various sports videos.

III. Proposed Algorithm

Video Highlights extraction from the sports videos can be achieved using various approaches such as analyzing shot changes, camera motion, Face detection, Text detection, phrase recognition, etc. An audio analysis based approach is proposed to extract key frames corresponding to audio level and finally build summary containing events of interest in the original video. The complete process of proposed methodology is depicted as shown figure.

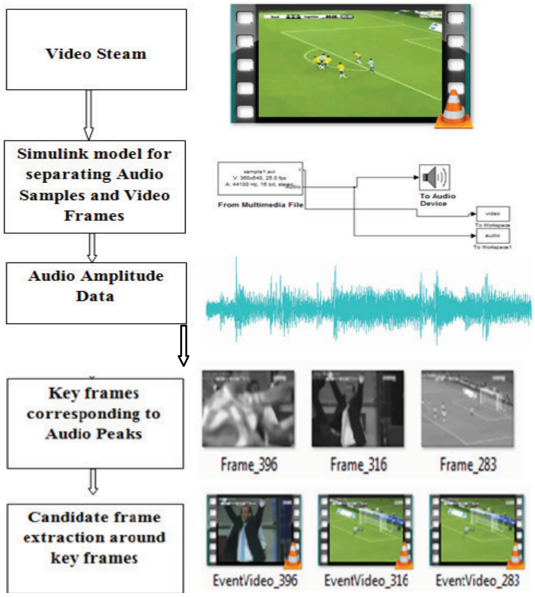


Figure 1: Proposed Methodology for Event Detection in Sports Video

The proposed methodology starts up by separating the input sports video stream into separate video and audio streams and then analyzing the audio stream for peak values present. The peak values corresponding to the video frame numbers is calculated and the frames corresponding to particular peak are extracted as peak frames. The peak frame numbers are further considered for event detection. As the event is a selective continuous part of the original video, an event is constructed by choosing some frames before peak frames and some frames after peak frames. The proposed algorithm is explained in following steps with figure.

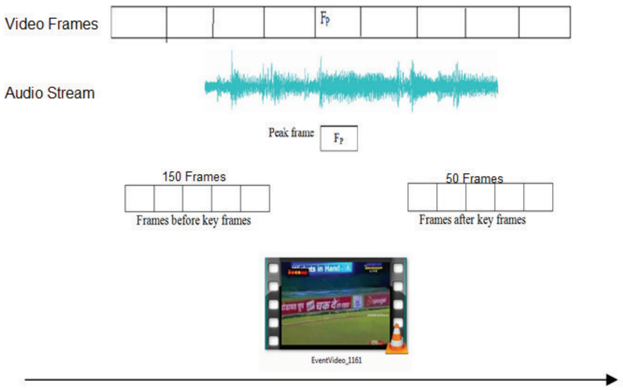


Figure 2: Mechanism of Event Detection in Sports Video

Step1: Read the input video stream and separate the audio samples and video frames. The separated audio samples and video frames are stored in the matlab module as arrays.

Step2: The number of audio samples corresponding to each video frame is calculated.

Step3: Determine the number of peaks in the input audio stream for event detection.

Step4: The numbers of higher audio levels are taken around which event can be extracted.

Step5: The video frame numbers corresponding to each of the amplitude peak are calculated and corresponding video frame is stored in as jpeg compressed image.

Step6: The video frame number corresponding to each peak is considered for video generation. The frames around 50-100 before and after the audio peak are considered for video. The output video in the form of AVI file is stored in the "Event Detection" folder.

The steps 1-6 are repeated for a fixed size of the input data for the complete sports video stream.

The total algorithm is shown in following flowchart:

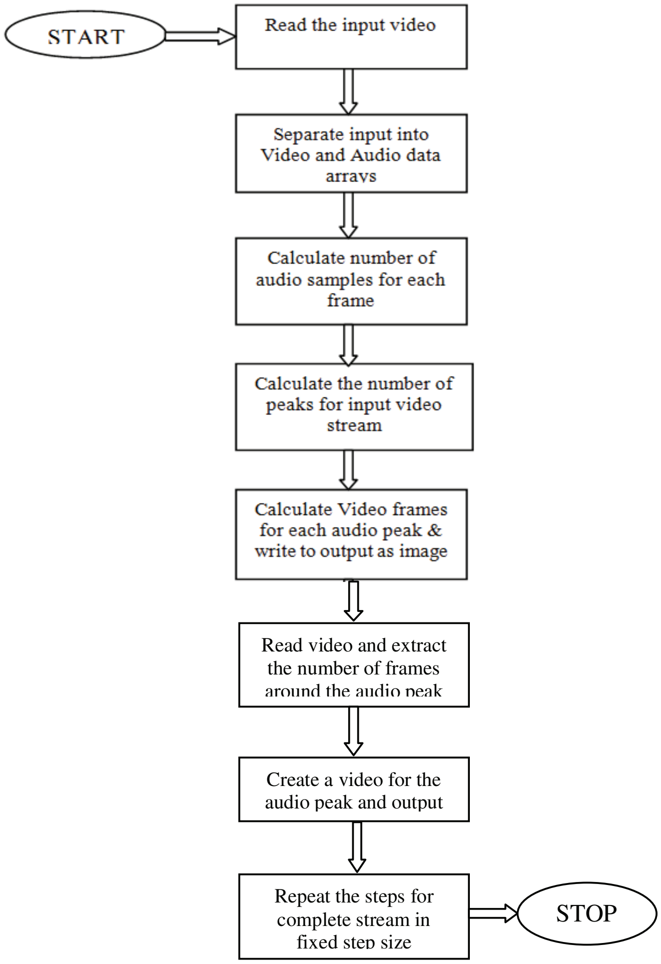

START
Read the input video
Separate input into Video and Audio data arrays
Calculate number of audio samples for each frame
Calculate the number of peaks for input video stream
Calculate Video frames for each audio peak & write to output as image
Read video and extract the number of frames around the audio peak
Create a video for the audio peak and output
Repeat the steps for complete stream in fixed step size
STOP

Figure 3: Flowchart of Proposed solution

The proposed algorithm is implemented in MATLAB 2010b version on *windows 7* operating system installed on a machine with 2 GB of primary memory and Intel I3 second generation processor. The videos used for the experimentation are of *.avi* and *.mpg* format. The experimental dataset used in this project are from the youtube action dataset, the Open Video Project.

4.1 Datasets

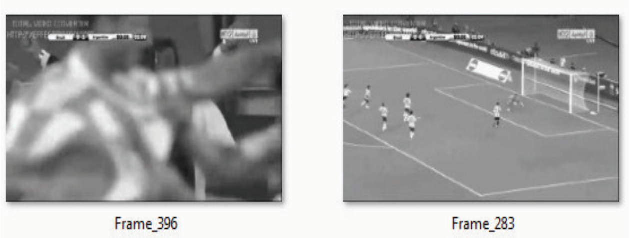
Various sports game video files used in our experiments were collected from a wide range of sources via the Internet. After excluding those video files that either have poor digital quality or do not contain any goal scene, there are 10 video files left, with different styles and produced by different broadcasters. The proposed model accepts the input video in the form of ".avi". The avi cutter is used to cut the large avi video in to samples of certain sizes ranging from 3MB to 90 MB. The corresponding sound tracks have been extracted from the

original video files. These video files and audio tracks serve as the test bed for the proposed framework. Rigorous testing has been carried out to test the above algorithms. To achieve cent percent accuracy is highly improbable because of the inherent inconsistencies in sports videos in itself.
Results for such a sample is explained below,

4.1.1 Football video (.avi file)

Figure 4.1 Input clip Sample 1
Number of audio Peaks: 2

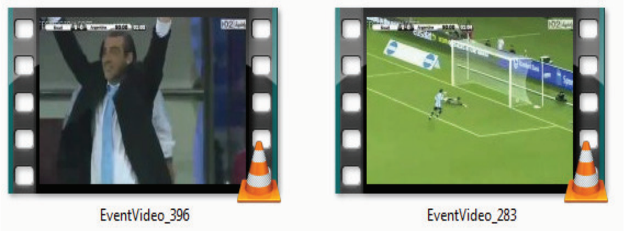
The sports video sample is a football clip where football player messi records a stunning goal and video shows players and coach celebrating the success. The sample has audio variations and manual examination reveals the maximum amplitude for audio during the goal and celebrations.
The problem solution is implemented by separating the input sports video stream into separate video and audio streams and then analyzing the audio stream for peak values present. The peak values corresponding to the video frame numbers is calculated and the frames corresponding to particular peak are extracted and stored as peak frames. The peak frame numbers are further considered for video generation. Key Frame detection based approach shows excellent detection accuracy and also results in saving of processing time. The algorithm gives following frames as the peak frames for this video.



Frame_396 Frame_283

Figure 4.2: Peak Frames Generated for given video

The event videos generated for the audio peaks for the above peak frame numbers and the summarized video generated from the event videos are shown below.



EventVideo_396 EventVideo_283

Figure 4.3: Event Videos Generated

Figure 4.2 and Figure 4.3 clearly demonstrate the extracted peak frames and corresponding video events. Frame 396 records the event peak in audio amplitude. There are totally 2 frames corresponds to the peaks, the frame numbers are 231,249,283,316,396,438.The event videos matches the expected event extraction as per manual examination. The event video is created by reading the original video stream and considering the number of frames before and after the peak frame, the target video is stored in the AVI Format.

4.1.2 Cricket video

The sports video sample shows the T20 world cup final sequence between India and Pakistan where a normal ball sequence where there is no important event is covered followed by next ball where batsman hits for a six. When the video is analyzed for 1 peak amplitude, the system accurately determined the portion of the video where batsman hits the ball for a six. The output for sample 2 video is as shown below,

Figure 4.4 Cricket T20 Final Sample 2
Number of audio Peaks: 1

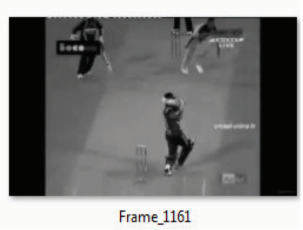


Figure 4.5: Peak Frame Generated for given video

For the input sample 2 generated 1 peak frame and 1 video corresponding to the peak frame. Frame 1161 records the highlight peak in audio amplitude. The highest peak frame number detected is 1161 and video is constructed using frames around frame number 1161.

Figure 4.6: Event Videos Generated

4.1.3 Baseball Video

The input sample is a clip taken from the World Series baseball championship last pitch sequence. The video shows the last pitch being bowled and crowd celebrates the win with loud of cheer. The system efficiently detected the peak as the crowd cheering for the win.

From figure 5.8 the peak frames are actually represents the last pitch ball and celebrations of the crowd and loud cheering. The numbers of peak frames are high as the highest peak values are of comparatively long duration corresponding to more adjacent frames. And analyzing the sample using audio peaks approach, the results usually captured the final celebrations and last moment of the ball is bowled.

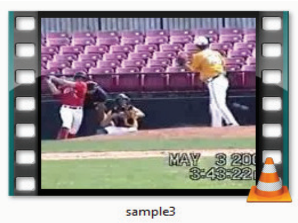


Figure 4.7 Sample 3 Baseball World Series
Number of audio Peaks: 2

Figure 4.8: Peak Frames Generated for given video

Figure 4.9: Event Videos Generated

For the evaluation purpose we have selected around 10 video clips downloaded from different datasets. The preprocessing step in the algorithm is omitted. *Table 4.1* shows the overview of the tests performed. It demonstrates the total frames, number peak frames detected. In the next columns it gives a number of peak frames extracted and finally the fidelity which describes the number of significant events with the original video.

Table 4.1 Experimental results

Index	**Name of files**	**No of frames**	**Size in (MBs)**	**Enter Peak value**	**No of peak frame detected**	**No of Event frames (Videos)**
1	**Sample1.avi**	**751**	**84.6**	**1**	**1**	**200(1)**
2	**Sample2.avi**	**850**	**9.16**	**1**	**1**	**200(1)**
3	**Sample3.avi**	**1199**	**6.42**	**2**	**2**	**400(2)**
4	**Sample4.avi**	**1196**	**40.5**	**1**	**1**	**200(1)**
5	**Sample5.avi**	**1201**	**37.2**	**1**	**1**	**200(1)**
6	**Sample6.avi**	**1001**	**20.2**	**1**	**1**	**200(1)**
7	**Sample7.avi**	**960**	**16.9**	**3**	**4**	**400(2)**
8	**Sample8.avi**	**1160**	**13.7**	**2**	**1**	**200(1)**
9	**Sample9.avi**	**1001**	**20.2**	**2**	**1**	**200(1)**
10	**Sample10.avi**	**1001**	**16.2**	**1**	**1**	**200(1)**

IV. CONCLUSION

The video abstraction and video summarization produce the comprehensive synopsis based on some general needs or user specific needs. The formal definition of video summarization is collection of frames which candidate represents video or part of video. The summarization can also be done by bringing together the significant events extracted from large sized video based on modality context or as per the query from user. In this direction an algorithm which extracts events based on audio cues is presented and experimented over various sports videos

The presented algorithm first extracted the Audio content from the video via a Simulink model, and then extracted the audio samples per frame, once got the audio samples, and then used the user input to extract the number of peaks required by user for Frame Extraction. The event video is considering the frames around the audio peak. The sports videos such as cricket, soccer, baseball contain meaningful audio markers such as applause, cheer, shouts near the event of interest. The result show that these events can be successfully detected and short meaningful summary can be constructed using proposed method.

REFERENCES

[1] Z. Xiong, X. Zhou, Q. Tian, Y. Rui, and T. Huang. Semantic retrieval of video - review of research on video retrieval in meetings, movies and broadcast news, and sports. IEEE Signal Processing Magazine, 23(2):18–27, March 2006.

[2] Y. Rui, T. Huang, and S. Mehrotra. Constructing table-of-content for video. In Proc. ACM Multimedia conference, 1998.

[3] H. Pan, P. Van Beek and M.I. Sezan. "Detection of Slow motion replay segments in sports video for highlights generation," In Proc. IEEE ICASSP 2001.

[4] A. Hanjalic, "Generic approach to highlights extraction from a sport video," ICIP 2003.

[5] E.A. Murat Tekalp and A. M. Tekalp, "Generic play-break event detection for summarization and hierarchical sports video analysis," to appear in Proc. IEEE ICME 2003.

[6] L.Y. Duan, M. Xu, T. Chua, Q. Tian, C.S Xu, "A mid-level representation framework for semantic sports video analysis", ACM Multimedia 2003.

[7] N. Babaguchi, Y.Kawai, T. Kitahashi, "Event Based Indexing of Broadcasted Sports Video by Intermodal Collaboration," IEEE Trans. On Multimedia 2002.

[8] S. Nepal, U. Srinivasan, G. Reynolds, "Automatic detection of 'Goal' segments in basketball videos," International conference on ACM Multimedia 2001.

[9] Y. Gong, T.S. Lim, and H.C. Chua, "Automatic Parsing Of TV Soccer Programs", IEEE International Conference on Multimedia Computing and Systems, May, 1995

[10] Regunathan Fbdhakrishant,Ziyou Xiong, Ajay Divakaran, Yasushi Ishikawa Mitsubishi Electric Research Labs, Cambridge, MA, USA"Generation of Sports Highlights Using a Combination of Supervised & Unsupervised Learning in Audio Domain",2003

[11] Yoshimasa Takahashi, Naoko Nitta, and Noboru Babaguchi, Graduate School of Engineering, Osaka University,2-1 Yamadaoka, Suita, Osaka 565-0871 Japan,"Video Summarization For Large Sports Video Archives", 2005

[12] Hari Sundaram,Shih-Fu Chang,Dept. Of Electrical Engineering, Columbia University,New York, New York 10027,"Video Scene Segmentation Using Video And Audio Features", 2011

[13] Qixiang Ye , Wen Gao，Shuqiang Jiang "Exciting Event Detection in Broadcast Soccer Video With Mid-level Description and Incrementa Learning" MM'05, November 6–11, 2005, Singapore.

[14] Changsheng Xu, Yi-Fan Zhang, Guangyu Zhu, Senior Member, IEEE, "Using Webcast Text for Semantic Event Detection in Broadcast Sports Video" IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 10, NO. 7, NOVEMBER 2008

[15] Yu-Lin Kang, Joo-Hwee Lim, "Soccer video event detection with visual keywords" Institute for Infocomm Research 21 Heng Mui Keng Terrace Singapore 119613 Dec 2003

[16] Hao Tang, Vivek Kwatra2, Mehmet Emre Sargin, Ullas Gargi, "Detecting Highlights In Sports Videos Cricket As A Test Case", HP Labs, Palo Alto, CA USA.

[17] Tommy Chheng, Department of Computer Science, University of California, Irvine," Video Summarization Using Clustering", 2005

[18] Shi Lu,Department of Computer Science and Engineering,The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR,"Video summarization by video structure analysis and graph optimization", 26 Nov 2003

[19] Naveed Ejaz and Sung Wook Baik , College of Electronics and Information Engineering, Sejong University, Seoul, Republic of Korea. "Weighting low level frame difference features for key frame extraction using Fuzzy comprehensive evaluation and indirect feedback relevance mechanism".

[20] L. Bai,Centre for Sensor Web Technologies, Dublin City University,Glasnevin, Dublin 9, Ireland, "Automatic Summarization of Rushes Video using Bipartite Graphs", 2008.