As per UGC guidelines an electronic bar code is provided to seure your paper

International Journal for Modern Trends in Science and Technology, 7(01): 48-53, 2021
Copyright © 2021 International Journal for Modern Trends in Science and Technology
ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST070111
Available online at: http://www.ijmtst.com/vol7issue01.html

Automatic Summarization of Cricket Highlights using Audio Processing

Ritwik Baranwal

Information Technology, Maharaja Agrasen Institute of Technology, New Delhi, India

To Cite this Article
Ritwik Baranwal, "Automatic Summarization of Cricket Highlights using Audio Processing", *International Journal for Modern Trends in Science and Technology*, Vol. 07, Issue 01, January 2021, pp.- 48-53.

Article Info
Received on 22-November-2020, Revised on 18-December-2020, Accepted on 22-December-2020, Published on 29-December-2020.

ABSTRACT

The problem of automatic excitement detection in cricket videos is considered and applied for highlight generation. This paper focuses on detecting exciting events in video using complementary information from the audio and video domains. First, a method of audio and video elements separation is proposed. Thereafter, the "level-of-excitement" is measured using features such as amplitude, and spectral center of gravity extracted from the commentators speech's amplitude to decide the threshold. Our experiments using actual cricket videos show that these features are well correlated with human assessment of excitability. Finally, audio/video information is fused according to time-order scenes which has "excitability" in order to generate highlights of cricket. The techniques described in this paper are generic and applicable to a variety of topic and video/acoustic domains.

KEYWORDS: *Video Segmentation, Audio Chunks, Short Time Energy.*

I. INTRODUCTION

This study focuses on the problem of identifying exciting-events in multimedia content. Our approach analyzes speech characteristics that identify islands (or "hot-spots") of strong emotion. In general, the ability to automatically parse multimedia content and tag "interesting events" is important for many domains such as sports, security, movies/TV shows, broadcast news, etc. A number of technologies such as search, summation, and mash-ups, can utilize "hot-spot" information to enhance access to, as well as navigation of content. For example, emotional "hot-spots" within sports videos are very likely to be "exciting" and this information can be used to guide the process of automatically generating highlights. This constitutes the motivation for this work, where automatic highlights of cricket videos are generated using emotional "hot-spot" detection (or "exciting events" detection).

Researchers have utilized audio and video streams to extract features that identify exciting plays in sports videos. Among video-based features, motion and density of cuts have been found to be useful for detection[1] . On the other hand, audio-based features have been derived from both speech (generally commentators) and background (generally audience), where audience-events like cheering/applause as well as the commentators speech characteristics have proven to be useful [2,3]. While video-based features tend to be more game-dependent, audio-based feature detecting exciting plays. Research in audio-based features have focused on emotion analysis of the commentator's speech and

employ this information with heuristics to identify exciting plays.

In this paper, we present a novel approach for auto-curating sports highlights, showcasing its application for cricket match. Our approach combines information from the player, spectators, and the commentator to determine a game's most exciting moments. A common trait among most of the sports is that whenever an exciting moment is happening the commentators speak loudly and the crowd cheers. We have used this as a signature for finding those moments where important or exciting things are happening in the match. So we just needed to analyze the audio of the match find those moments where the crowd cheered or commentators are excited and extracted those parts.

METHOD

In our approach the generation is done on basic analysis of sound or audio processing, we know that whenever an important event occurs during a match the adrenaline rush can be seen into commentary, so in simple terms it's an impulsive energy in short span of time. So, what we have done here is we have extracted all those impulsive commentary events and have seen the corresponding video at the same timestamp and generated the desired highlights.

STRUCTURE OF PAPER

The paper is organized as follows: In Section 1, the introduction of the paper is provided along with the structure, important terms, objectives and overall description. In Section 2 we discuss related work. In Section 3 we have the complete information about image processing tools. Section 4 shares information about the flexible YAML templating system created for it, its advantages and disadvantages. Section 5 tells us about the methodology and the process description. Section 6 tells us about the future scope and concludes the paper with acknowledgement and references.

OBJECTIVES

This project aims to address some of the problems in current systems by greatly minimizing the human intervention in the process and thus reducing costs and errors. The aim is to ease the task of both the technicians and audience.

II. RELATED WORK

Video Summarization.

There is a long history ofresearch on video summarization [4], whichaims at producing short videos or keyframes that sum-marize the main content of long full-length videos, bylooking at elminating redundancy either at signal level(feature dimensionality reduction [5]) or in semanticcontent [6]. Our work also aims at summarizing videocontent, but instead of optimizing for representative-ness and diversity, as traditional video summarizationmethods do, our goal is to find highlights or excitingmoments in the videos. A few recent methods addressthe problem of highlight detection in consumer videos[7]. Instead our focus is on sports videos,which offer more structure and objective metrics thanunconstrained consumer videos.

Automatic Trailer Generation.

Another sub-area ofvideo summarization involving multimodal video analysis that goes beyond content recognition, and focusinginstead on affective responses evoked by the video is movie trailer generation [8,9]. For example,Evangelopoulos et al. [9] model and combine audio visual and textual saliency to select the most relevantscenes in a movie. In this space, works focus on detectingcontent with the highest emotional impact based onmovie genre. For instance, in horror movies scenes evoking feelings of suspense or fear are important [9].In our domain of interest, on the other hand, only positiveemotions connected to excitement are relevant. Further-more, differently from this line of research, the focusof our work is on identifying and measuring subjectsreactions (players, crowd, and commentator)

directlyinthe video stream, rather than inferring reactions whichare supposed to be evoked by inspected content whichis deemed as "impressive" [1].

Sports Highlights Generation.

Several methods havebeen proposed to automatically extract highlights fromsports videos based on audio and visual cues. Exampleapproaches include the analysis of replays [1, 2, 4],crowd cheering [6, 3], motion features [5], and closedcaptioning [4]. More recently, Bettadapura et al. used contextual cues from the environment to understandthe excitement levels within a basketball game. Tangand Boring [3] proposed to automatically

produce high-lights by analyzing social media services such as twitter.Decroos et al. [8] developed a method for forecastingsports highlights to achieve more effective coverage ofmultiple games happening at the same time. Differentfrom existing methods, our proposed approach offers aunique combination of excitement measures extractedfrom live video streams to produce highlights, includinginformation from thespectators, thecommentator, andtheplayerreaction. As such, our system incorporatesand combines most of the information employed byprevious works (audio, visual, text).

It could also beeasily extended to integrate other sources of attentionor excitement, such as social media feeds or productioncues (replays, closed captions, etc.). In addition, we en-able personalized highlight generation or retrieval basedon a viewer's favorite players.

Self-Supervised Learning.

In recent years, there has been significant interest in methods that learn deep neural network classifiers without requiring a large amount of manually annotated training examples. In particular,self-supervised learning approaches rely on auxiliary tasks for feature learning, leveraging sources of super vision that are usually available "for free" and in large quantities to regularize deep neural network models. Examples of auxiliary tasks include the prediction of ego-motion [4, 1], location and weather [4], spatial context or patch layout [2, 4], image colorization [4], and temporal coherency [2]. Altar. [5] explored the natural synchronization between vision and sound to learn an acoustic representation from unlabeled video. We leverage this work to build audio models for crowd cheering and commentator excitement using few training examples, and use those classifiers to constrain the training data collection for player reaction recognition. More interestingly, we exploit the detection of TV graphics are free supervisory signal to learn feature representations for player recognition from unlabeled video.

III. AUDIO PROFILE GENERATION

a) MPEG BitstreamProcessingTheFischlár system captures television broadcastsand encodes the programmes according to theMPEG-1 digital video standard with the audio signalcoded in line with the Layer-II profile [8]. Unlikemany other audio compression algorithms, whichmake assumptions about the nature of the audiosource, MPEG-1 Audio exploits the perceptualrestrictions of the human auditory system, viapsychoacoustic weighting of the bit allocation foreach frequency subband, to attain its compression[7].The MPEG-1 Layer-II compression algorithmencodes audio signals by dividing the frequencyspectrum of the audio signal, bandlimited to 20kHzinto 32 subbands which approximate the ear's criticalbands. The subbands are assigned individual bit-allocations according to the audibility of quantisationnoise within each subband. A pyschoacousticmodelof the ear analyses the audio signal and provides thisinformation to the quantiser. Layer-II frames consistof 1152 samples; 3 groups of 12 samples from eachof 32 subbands. A group of 12 samples gets a bit-allocation and, if this is non-zero, a scalefactor.Scalefactors are weights that scale groups of 12samples such that they fully use the range of thequantiser (the encoder uses a different scalefactorforeach of the three groups of 12 samples within eachsubband only if necessary). The scalefactor for sucha group is determined by the next largest value(given in a look-up table) to the maximum of theabsolute values of the 12 samples. Thus it providesan indication of the maximum power exhibited byany one of the 12 samples within the group [9,1].

b) Amplitude of the Speech BandMost of the energy in a speech signal lies between0.1kHz – 4kHz. According to the MPEG-1 Layer-IIaudio standard, the maximum allowable frequencycomponent in the audio signal is at 20kHz. At theencoder, the frequency spectrum (0 – 20kHz) isdivided uniformly into 32 subbands, each havingbandwidth of 0.625kHz [4]. Thus, subbands 2through 7 represent the frequency range from0.625kHz – 4.375kHz. See Figure 1.Speech1 2 3 4 5 6 7 8 31 320.625kHz 4.375Hz

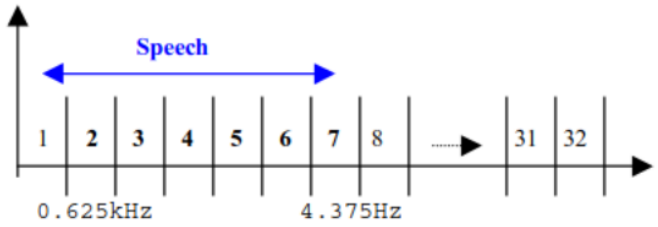

Figure 1: MPEG-1 Layer-II Frequency Subbands

Figure 1: MPEG-1 Layer-II Frequency SubbandsFor sports programme audio tracks, by strictlylimiting the audio examination to these subbands,which approximate the range of the speech band, wefurther concentrate the audio investigation oncommentator vocals. Therefore, the influence of thecommentator on the generation of the audioamplitude profile is increased. This is clearlydesirable since it bolsters the assumption that theprofile will be an accurate indicator of thesignificance of the content.It was expected that the examination of subbands 2though 7 would provide for a reasonable trade-offbetween rejection of low-frequency backgroundnoise (typically present in sports programmeswhichwould naturally upset results) and the capture of thefundamental frequency for excited speech.

c) Boundary DetectionOne of the problems with the audio amplitudestechnique is caused by the inclusion ofsupplementary content which typically accompaniesthe main event in a sports programme. Features suchas player profiles, highlights of recent events etc.tend to contain attributes such as commentatordialogue and spectator noise, similar to that of themain event. The problem is that these featuresgenerally have audio amplitudes comparable to thatof the event of interest.To combat this problem, the system must be able todetect the temporal boundaries of the main featurewithin the overall sports programme. This is done bysearching through the audio track for extendedperiods of sustained volume.Segments such as interviews, studio discussions,archive video clips, etc. which make up theperipheral content, are flagged by the intermittentoccurrence of brief moments of silence. For example short silences exist in between sentences spoken byan anchorperson, when switching from anchorperson to video clips, or between advertisements. In contrast, the main event in a sports broadcast features relatively long periods of sustained volume due tithe continuous presence of background noise. Ont his basis it may be automatically distinguished from the supplementary content. i.e. the temporal boundaries of the main event within the overall program may be detected. For the summary generation, the probing domain is restricted to lie within these boundaries.

IV. CASE STUDY

a) Task

The following is an illustration of the automaticgeneration of a 10-minute summary of a terrestrialTV broadcast of a sports event via the discussedtechnique. The experimental subject is the UEFACup Final 2001 featuring Liverpool FC Vs AlavesFC. This was a near 3-hour soccer match broadcast,resulting in a 5-4 victory for Liverpool FC. Theprogramme featured the main event plus studiodiscussions and analysis, player profiles, highlightsof related events and advertisement breaks.

b) Amplitude Profiles

A second-by-second audio amplitude profile wase stablished by a superposition of all the scalefactorsfromsubbands 2-7 over a window length of onesecond. See Figure 2.

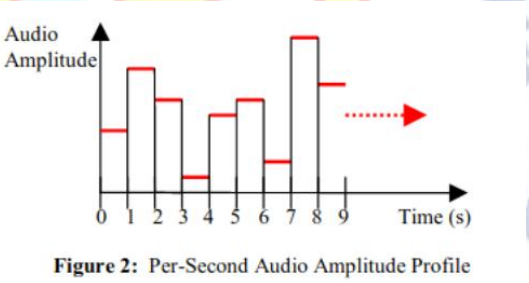

Figure 2: Per-Second Audio Amplitude Profile

A frame-by-frame audio amplitude profile wasestablished by a superposition of all the scalefactorsfromsubbands 2-7 over a window of lengthcorresponding to one video frame (≈1/25s). See Figure 3.

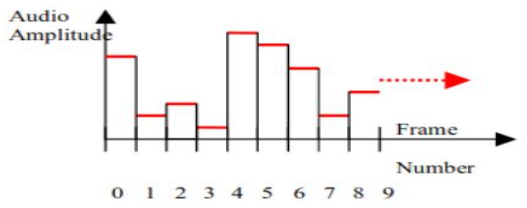

Figure 3: Per-Frame Audio Amplitude Profile

c) Boundary Detection

The overall structure of the near 3-hour subject, as captured by Físchlár, is described below. In terms of summary generation, segments of interest are identified by an asterisk.

*1st_half~51mins

Studio_analysis~14mins

*2nd_half~49mins

Studio analysis~4 mins

*Extra_time~26mins

Studio analysis~6 mins

A silence threshold was empirically determined as

Sth = 0.033 * overall mean audio amplitude

Using the per-frame audio amplitude profile and Sth,the entire audio track of the subject was examinedfor periods of continuous volume lasting, at least 1-minute. It was found that sustained volumesexceedingSth occur during the following videoframes:

309 – 3504~2mins

3867 – 6608~1min

6984 – 13577~4mins

15037 – 19467~3mins

19553 – 23405~2mins

*26248 – 102751~51mins

106225 – 109935~2mins

112696 – 115354~1min

*123629 – 198950~50mins
199586 – 201168~1min
*201252 – 244086~28mins
245527 – 247690~1min

Further thresholding at a length of 10-minutes rejectsall segments except for three (identified by anasterisk), which correspond almost precisely to thesegments of interest mentioned previously (i.e. the temporal boundaries of the match play segmentswere accurately detected). Changing units to secondsthese are:
•Segment-1:1050s – 4110s
•Segment-2 :4945s – 7958s
•Segment-3:8050s – 9763s

Only the content which resides within theseboundaries is eligible for inclusion in the summary.Hence, further audio processing is restricted accordingly. The boundary detection preamble is not a crucial component of the summarisation procedure, i.e. in the event of failure, the main audio analysisprocedure would still be expected to produce amoderately successful summary. However, it isbeneficial tool which prevents the consideration ofirrelevant material and thus lightens the workload ofsubsequent procedures.

d) Summary GenerationThe per-second audio amplitude profiles of segments1-3 (above) were examined. A loudness threshold,Lth, was defined and initalised to the valuecorresponding to the largest peak found. See Figure4.

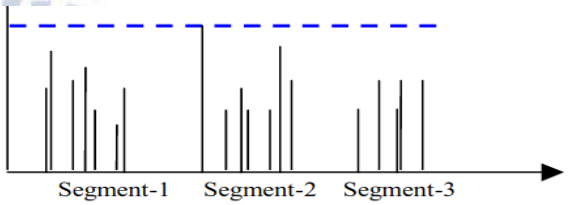

Figure 4: Examination of Segments 1-3

An audio amplitude peak is defined as loud if itexceeds Lth. Ignoring isolated peaks, Lth wasgradually reduced until it began to pick out loudperiods of at least 3-seconds in duration (audiosurges). See Figure 5.

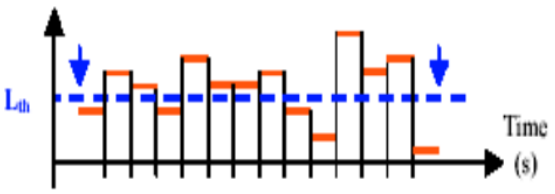

Figure 5: Decreasing L_{th} and detecting *audio surges*

Figure 5 shows three sections which extend beyondthe current value of Lth. The second and third havetime spans of 4 seconds and 3 seconds respectively.Thus both are recognised as audio surges. The firstsection is ignored since with a length of 2 seconds, itdoes not meet the minimum surge threshold of 3-seconds. Lth was further reduced until the amount ofdetected surges was sufficient such that a 10-minutevideo summary could be produced. The summarywas then generated by first matching up the videoclips within the combined audio/video track whichtemporally align with the audio surges. Then, a pre-clip buffer of 1 shot and a post clip buffer of 2 shotswas appended (to make viewing the amalgamationless visually disturbing). Finally these clips wereextracted from the audio/video stream and(chronologically) concatenated to generate thehighlights summary. See Figure 6.

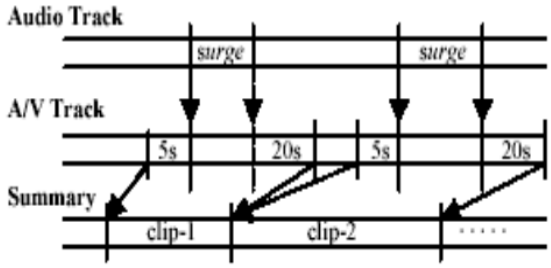

Figure 6: Summary generation

V. RESULTS

The analysis returned 18 individual clips corresponding to the following descriptions, comprising a summary length of just over 5-minutes:

1. Wicket *
2. Boundary -
3. Boundary -
4. Wicket -
5. Wicket -
6. Boundary -
7. Boundary -
8. Boundary -
9. Boundary -
10. Boundary *
11. Boundary *
12. Boundary #
13. Wicket #
14. Boundary #
15. Catch Drop -
16. Boundary -
17. Wicket -
18. Boundary -

For the purposes of evaluation, the nineteen clips returned were examined and classified into fourcategories according to significance. Twelve clips seemed to depict very significant moments of thefeature and hence were described as definite highlights (-). The inclusion of definite highlights inthe summary is always preferred.Four of the clips returned seemed to represent moments of arguably lesser significance. These were described by the term quasi-highlights (#), and theirinclusion in the summary is desired once all definitehighlights already have been. The system returned three further clips containing content of considerably less significance, labeled lowlights (*). Inclusion oflowlights would typically not be tolerated exceptwhen the combined length of all definite and quasi-highlight clips fails to satisfy the desired length of the summary.

VI. FUTURE SCOPE AND CONCLUSION

Further this approach can be used with OCR to increase the percentage of definite highlights. Different threshold combinations can also be tried to filter out better results.In this study, a novel methodology that uses estimates of excitability in sports video to create automatic highlights was presented. First, a method of audio and video elements seperation is proposed. Thereafter, the "level-of-excitement" is measured using features such as amplitude, and spectral center of gravity extracted from the commentators speech's amplitude.

REFERENCES

[1] C.Liu, Q. Huang, S. Jiang, L. Xing, Q. Ye, and W. Gao, "A frame-work for flexible summarization of racquet sports video using mul-tiplemodalities,"Computer Vision and Image Understanding, vol.113, pp. 415–424, 2009.

[2] E.Kijak, G. Gravier, P. Gros, L. Oisel, and F. Bimbot, "HMM basedstructuring of tennis videos using visual and audio cues," inICME,2003.

[3] R. Radhakrishnan, Z. Xiong, A. Divakaran, and Y. Ishikawa, "Gen-eration of sports highlights using a combination of supervised andunsupervised learning in audio domain," inPacific Rim Conferenceon Multimedia, 2003.

[4] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li, "A user attentionmodel for video summarization," inACM Multimedia,2002

[5] J. Zhang, J. Yu, and D. Tao, "Local deep-featurealignment for unsupervised dimension reduction,"IEEETransactions on Image Processing, vol. 27, no. 5, pp.2420–2432, 2018.

[6] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Videosummarization with long short-term memory," inECCV,2016.

[7] M. Sun, A. Farhadi, and S. Seitz, "Ranking domain-specific highlights by analyzing edited videos," inECCV,2014.

[8] G. Irie, T. Satou, A. Kojima, T. Yamasaki, andK. Aizawa, "Automatic trailer generation," inACMMul-timedia, 2010.

[9] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Mara-gos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Mul-timodal saliency and fusion for movie summarizationbased on aural, visual, and textual attention,"IEEETransactions on Multimedia, vol. 15, no. 7, pp. 1553–1568, 2013.