Automatic summarization of Sports highlights using Audio Processing

Ritwik Baranwa[1]; Ritesh Chaubey[2]; Dr. Neha Agrawal[3] and Meenu Garg[4]

[1]Information Technology, Maharaja Agrasen Institute of Technology, New Delhi, India.
[2]Information Technology, Maharaja Agrasen Institute of Technology, New Delhi, India.
[3]Assistent Professor, Maharaja Agrasen Institute of Technology, New Delhi, India,
[4]Assistent Professor, Maharaja Agrasen Institute of Technology, New Delhi, India.

Abstract: The problem of automatic excitement detection in Sports videos is considered and applied for highlight generation. This paper centers around distinguishing energizing occasions in video utilizing correlative data from the sound and video spaces. Initial, a technique for sound and video components separation is proposed. From there on, the "level-of-excitment" is estimated utilizing highlights like plentifulness, and phantom focus of gravity removed from the analaysts discourse's adequacy to choose the edge. Our analyses utilizing genuine cricket recordings show that these highlights are very much associated with human appraisal of volatility. At long last, sound/video data is melded by time-request scenes which has "sensitivity" to create features of cricket. The methods portrayed in this paper are conventional and pertinent to an assortment of point and video/acoustic spaces.

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

DOI of the Article: https://doi.org/10.46501/IJMTST0706035

Available online at: http://www.ijmtst.com/vol7issue06.html

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

To Cite this Article: Ritwik Baranwa; Ritesh Chaubey; Dr. Neha Agrawal and Meenu Garg. Automatic summarization of Sports highlights using Audio Processing. *International Journal for Modern Trends in Sceicen and Technology* 2021, *7*, 0706148, pp. 200-206. https://doi.org/10.46501/IJMTST0706035

Article Info. Received: 16 May 2021; Accepted: 11 June 2021; Published: 17 June 2021

INTRODUCTION

This Study centers around the issue of recognizing energizing occasions interactive media content. Our methodology breaks down discourse attributes that distinguish islands (or "problem areas") of forceful feeling. All in all, the capacity to naturally parse sight and sound substance and tag "fascinating occasions" is significant for some areas like games, security, films/TV shows, broadcast news, and so forth A few innovations like inquiry, rundown, and blend, can use "problem area" data to upgrade admittance to, just as route of substance. For instance, enthusiastic "problem areas" inside sports recordings are probably going to be "energizing" and this data can be utilized to direct the cycle of consequently creating features. This establishes the inspiration for this work, where programmed features of cricket recordings are produced utilizing passionate "problem area" recognition (or "energizing occasions" location).

Researchers have used sound and video streams to remove features that distinguish energizing plays in sports videos. Among video-based features, movement and density of cuts be useful for detection[1]. Then again, sound based features have been gotten from both speech (for the most part commentators) and foundation (general crowd), where crowd events like cheering/applause as well as the commentators' speech characteristics have demonstrated to be useful [2,3]. While video-based features will in general be more game-subordinate, sound based features distinguishing energizing plays. Research in sound based features have focused on feeling analysis of the commentator's speech .

METHOD

In our methodology the age is done on basic analysis of sound or sound processing, we realize that at whatever point a significant occasion occurs during a match the adrenaline rush can be seen into critique, so in simple terms it's an impulsive energy in short span of time. So, what we have done here is we have separatedeveryoneofthoseimpulsiveanalysis events and have seen the corresponding video at the same timestamp and created 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.

OBJECTIVE

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.

RELATED WORK

Video Summarization

There is a long history of research on video summarization [4], which aims at producing short videos or keyframes that summarize the main content of long full-length videos, by looking at eliminating redundancy either at signal level(feature dimensionality reduction [5]) or in semantic content [6]. Our work also aims at summarizing video content, but instead of optimizing for representativeness and diversity, as traditional video summarization methods do, our goal is to find highlights or exciting moments in the videos. A few recent methods address the problem of highlightandutilizethisdatawithheuristicstodistinguish energizingplays.

In this paper, we present a novel methodology for auto-curating sports highlights, showcasing its application for cricket match. Our approach combines data from the player, spectators, and the commentator to decide a game's most energizing moments. A normal attribute among most of the sports is that at whatever point an energizing second is occurring the commentators speak uproariously and the group cheers. We have used this as a signature for discovering those moments where significant or energizing things

are going on in the match. So we just expected to examine the sound of the match discover those moments where the group cheered or commentators are energized and removed those part.
detection in consumer videos [7]. Instead, our focus is on sports videos, which offer more structure and objective metrics than unconstrained consumer videos.

Automatic Trailer Generation

Another sub-region video summarization including multimodal video analysis that goes past content acknowledgment, and focusing instead on emotional responses evoked by the video, is film trailer age [8,9]. For instance, Evangelopoulos et al. [9] model and join sound, visual, and printed saliency to select the most important genes in a film. In this space, works focus on identifying content with the highest passionate effect based on the film type. For instance, with sickening apprehension movies, scenes summoning feelings of suspense or dread are significant [9].In our space of interest, then again, just positive emotions associated with the energy are applicable. Moreover, uniquely in contrast to this line of research, the focus of our work is on recognizing and measuring subjects reactions (players, group, and commentator)
straightforwardly in the video stream, instead of deducing reactions that are supposed to be evoked by inspected content which is considered as "impressive" [1].

Sports Highlights Generation

Several methods have been proposed to consequently extricate highlights from sports videos based on sound and visual cues. Model approaches incorporate the analysis of replays [1, 2, 4], swarm cheering [6, 3], movement features [5], and closed subtitling [4]. All the more as of late, Bettadapura et al. used relevant cues from the climate to understand the energy levels inside a basketball match-up. Tang and Boring [3] proposed to consequently create highlights by dissecting social media services such as Twitter. Decroos et al. [8] fostered a technique for forecasting sports highlights to accomplish more compelling inclusion of various games occurring at the same time. Not quite the same as existing methods, our proposed approach offers an interesting blend of fervor measures removed from live video streams to create highlights, including data from the spectators, the commentator, and the player response. As such, our system incorporates combines most of the data utilized by previous works (sound, visual, text).

It could also be easily reached out to coordinate different sources of consideration or fervor, such as social media feeds or creation cues (replays, closed captions, and so forth) Moreover, we empower personalized feature age or recovery based on a watcher's number one players.

Self-Supervised Learning

As of late, there has been significant interest in methods that learn profound neural organization classifiers without requiring a lot of physically explained preparing examples. Specifically, self-supervised learning approaches depend on helper tasks for include picking up, utilizing sources of supervision that are usually accessible "for nothing" and in huge quantities to regularize profound neural organization models. Examples of assistant tasks incorporate the expectation of inner self movement [4, 1], area and climate [4],
spatial setting or fix format [2, 4], picture colorization [4],and fleeting coherency [2].
Aytaretal. [5] investigated the characteristic synchronization among vision and sound to take in an acoustic representation from unlabeled video. We influence this work to assemble sound models for swarm cheering and commentator fervor using a couple of preparing examples and use those classifiers to constrain the preparation information assortment for player response acknowledgment. All the more interestingly, we misuse the identification of TV graphics as a free supervisory signal to learn include representations for player acknowledgment from unlabeled video.

AUDIO PROFILE GENERATION

1.MPEG Bitstream Processing TheFíschlár system captures television broadcasts and encodes the programs according to the MPEG- 1 digital video standard with the audio signal coded in line with the Layer-II profile [8]. Unlike many other audio compression algorithms, which make assumptions about the nature of the audio source, MPEG-1 Audio exploits the perceptual restrictions of the human

auditory system, via psychoacoustic weighting of the bit allocation for each frequency sub- band, to attain its compression[7]. The MPEG-

1 Layer-II compression algorithm encodes audio signals by dividing the frequency spectrum of the audio signal, bandlimited to 20kHz, into 32 sub-bands that approximate the ear's critical bands. The sub-bands are assigned individual bit- allocations according to the audibility of quantization noise within each sub-band. A psychoacoustic model of the ear analyses the audio signal and provides this information to the quantizer. Layer-II frames consist of 1152 samples; 3 groups of 12 samples from each of 32 sub-bands. A group of 12 samples gets a bit-allocation and, if this is non- zero, a scale factor. Scale factors are weights that scale groups of 12 samples such that they fully use the range of the quantizer (the encoder uses a different scale factor for each of the three groups of 12 samples within each sub-band only if necessary). The scale factor for such a group is determined by the next largest value(given in a look-up table) to the maximum of the absolute values of the 12 samples. Thus it indicates the maximum power exhibited by any one of the 12 samples within the group [9,1].

2.The Amplitude of the Speech Band Most of the energy in a speech signal lies between0.1kHz – 4kHz. As indicated by the MPEG-1 Layer-II sound standard, the most extreme passable recurrence segment in the sound signal is at 20kHz. At the encoder, the recurrence spectrum (0 – 20kHz) is isolated consistently into 32 sub- bands, each having a transfer speed of 0.625kHz [4]. Thus, sub-bands 2through 7 represent the recurrence range from0.625kHz – 4.375kHz. See Figure 1.Speech1 2 3 4 5 6 7 8

31 320.625kHz 4.375Hz

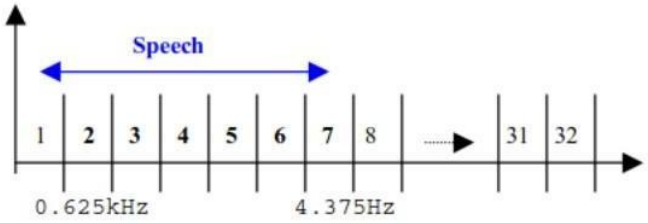


Figure 1: MPEG-1 Layer-II Frequency Subbands

Figure 1: MPEG-1 Layer-II Frequency Sub- bands Forsports program sound tracks, by strictly restricting the sound assessment to these sub-bands, which rough the scope of the speech band, we further concentrate the sound investigation commentator vocals. In this manner, the impact of the commentator on the age of the sound sufficiency profile is increased. This is desirable since it bolsters the assumption that the profile will be a precise pointer of the significance of the substance. It was normal that the assessment of sub- bands 2though 7 would accommodate a reasonable compromise between dismissal of low-recurrence foundation noise (commonly present in sports programs which would normally upset results) and the catch of the central recurrence for energizing speech.

Limit Detection One of the problems with the sound amplitudes procedure is caused by the inclusion of supplementary substance which regularly accompanies the headliner in a sports program. Features such as player profiles, highlights of late events, and so on will in general contain attributes such as commentator discourse and spectator noise, similar to that of the headliner. The issue is that these features by and large have sound amplitudes tantamount to that of the occasion of interest. To battle this issue, the system must have the option to distinguish the worldly boundaries of the principle highlight inside the general sports program. This is finished via searching through the sound track for expanded periods of sustained volume. Segments such as interviews, studio discussions, document video clips, and so on which make up the fringe content, are hailed by the discontinuous event of brief moments of silence. For instance, short silences exist in the middle of sentences spoken by an anchor person, when switching from anchor person to video clips, or between advertisements. In contrast, the headliner in a sports broadcast features moderately significant stretches of sustained volume because of the continuous presence of foundation noise. On this basis, it very well might be naturally distinguished from the supplementary substance. for example the fleeting boundaries of the headliner inside the general program might be distinguished. For the summary generation, the probing domain is restricted to lie within these boundaries.

CASE STUDY

a) Task

Coming up next is an illustration of the programmed age of a 10-minute summary of a terrestrial

transmission of a sports occasion through the discussed strategy. The trial subject is the UEFA Cup Final 2001 highlighting Liverpool FC Vs AlavesFC. This was an almost 3-hour soccer match broadcast, resulting in a 5-4 triumph for Liverpool FC. The program included the headliner plus studio discussions and analysis, player profiles, highlights of related events, and advertisement breaks.

b) Amplitude Profiles

A second-by-second sound plentifulness profile was established by a superposition of all the scale factors from sub-bands 2-7 over a window length of one second. See Figure 2.

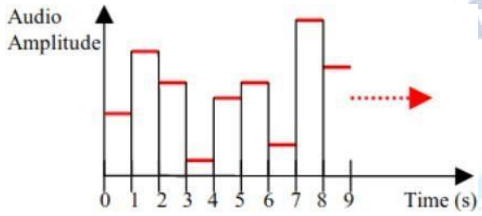


Figure 2: Per-Second Audio Amplitude Profile

c) Boundary Detection:

The general structure of the close to 3-hour subject, as caught by Físchlár, is described underneath. In terms of summary age, segments of interest are distinguished 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 exactly decided as

Sth = 0.033 * in general mean sound sufficiency

Using the per-outline sound abundancy profile and Sth, the whole sound track of the subject was analyzed for periods of continuous volume lasting, in any event 1-minute. It was tracked down that sustained volumes surpassing Sth happen during the accompanying video frames:

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 rejects all segments aside from three (distinguished by an asterisk), which correspond almost precisely to the segments of interest referenced previously (for example the worldly boundaries of the match play segments were precisely identified). Changing units to seconds these are:

- Segment-1:1050s – 4110s
- Segment-2 :4945s – 7958s
- Segment-3:8050s – 9763s

Just the substance which resides inside these boundaries is qualified for inclusion in the summary. Subsequently, further sound processing is restricted in like manner. The limit identification preface is not a critical part of the summarization method, for example in case of disappointment, the fundamental sound analysis system would still be relied upon to deliver a respectably successful summary. In any case, it is a useful apparatus that prevents the consideration of superfluous material and thus lightens the responsibility of subsequent procedures.

d)**Summary Generation** :The per-second sound sufficiency profiles of segments1-3 (above) were analyzed. A loudness threshold, Lth, was characterized and introduced to the worth corresponding to the largest pinnacle found. See Figure4.

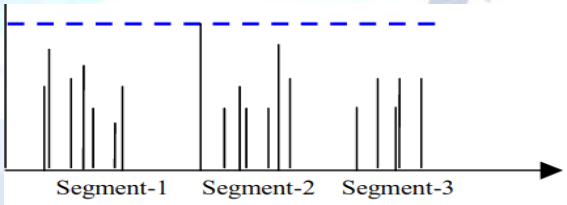


Figure 4: Examination of Segments 1-3

A sound amplitude top is characterized as uproarious on the off chance that it surpasses Lth. Overlooking segregated pinnacles, Lth was slowly diminished until it started to choose boisterous times of at any rate 3-seconds in length (sound floods). See Figure 5.

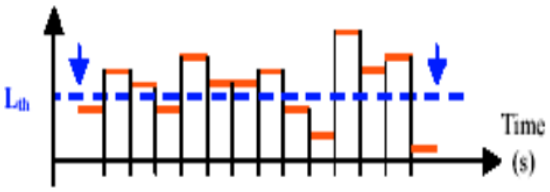


Figure 5: Decreasing $\mathbf{L_{th}}$ and detecting *audio surges*

Figure 5 shows three sections that extend beyond the current value of Lth. The second and third have periods

of 4 seconds and 3 seconds respectively. Thus both are recognized as audio surges. The first section is ignored since with a length of 2 seconds, it does not meet the minimum surge threshold of 3- seconds. Lth was further reduced until the number of detected surges was sufficient such that a 10- minute video summary could be produced. The summary was then generated by first matching up the video clips within the combined audio/video track which temporally aligns with the audio surges. Then, a pre-clip buffer of 1 shot and a post clip buffer of 2 shots was appended (to make viewing the amalgamation less visually disturbing). Finally, these clips were extracted from the audio/video stream and(chronologically) concatenated to generate the highlights summary. See figure 6.

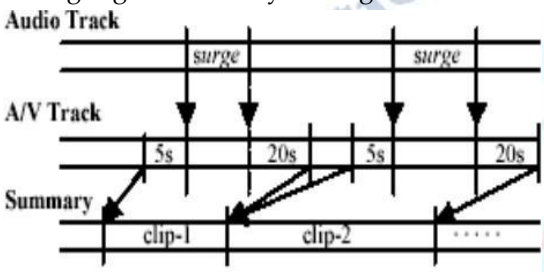


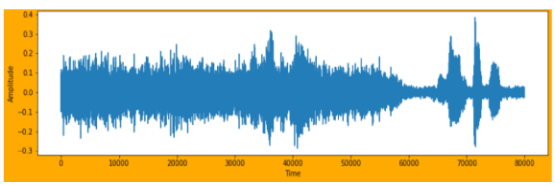
Figure 6: Summary generation

RESULTS

CRICKET HIGHLIGHT

The investigation returned 18 individual clips comparing to the accompanying portrayals, involving a rundown length of a little more than 5-minutes:

1. Wicket clip * 15. Catch Drop clip -
2. Boundary clip - 16. Boundary clip -
3. Boundary clip - 17. Wicket clip-
4. Wicket clip - 18. Boundary clip -
5. Wicket clip - 19. Boundary clip#
6. Boundary clip -
7. Boundary clip -
8. Boundary clip -
9. Boundary clip-
10. Boundary clip *
11. Boundary clip *
12. Boundary clip #
13. Wicket clip #
14 Boundary Clip#



SOUND INTENSITY THROUGH OUT VIDEO

For the reasons for assessment, the nineteen clasps returned were inspected and ordered into four classes as per importance. Twelve clasps appeared to portray extremely huge snapshots of the component and subsequently were depicted as positive features (-). The consideration of clear features in the rundown is constantly liked. Four of the clasps returned appeared to address snapshots of ostensibly lesser importance. These were depicted by the term semi features (#), and their incorporation in the outline is wanted once all clear features as of now have been. The framework returned three further clasps containing substance of impressively less importance, named lowlights (*). Incorporation of lowlights would commonly not go on without serious consequences aside from when the joined length of all positive and semi feature cuts neglects to fulfill the ideal length of the rundown.

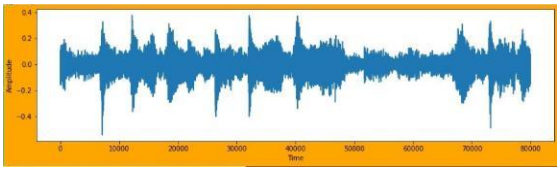
SNAPSHOT FROM THE HIGHLIGHT GENERATED

FOOTBALL HIGHLIGHT

The investigation returned 4 individual clips comparing to the accompanying portrayals, involving a rundown length of a 45 sec.

1. Goal Clip#
2. Big Chance Missed-
3. Goal Clip #
4. Goal Clip#

SOUND INTENSITY THROUGH OUT THE VIDEO



For the reasons for assessment, the 4 clips returned were inspected and ordered into four classes as per importance.

The covering of big chance missed is a very big point for us because as we know that fans or spectators are a little bit extra interested in knowing their team's mistakes. because usually, people saw highlights to show all the important chances created or missed because they can easily saw goals on social media. That's why is given(-) that sign. And the goal scored is given # because it was quite obivious to get covered.

SNAPSHOT FROM THE HIGHLIGHT GENERATED

FUTURE SCOPE AND CONCLUSION

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 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. This threshold is then used to extract the most exciting part of the audio to gather the highlights. Further, this methodology can be utilized with OCR to build the level of unmistakable features. Distinctive limit mixes can likewise be attempted to sift through better outcomes.

CRICKET HIGHLIGHTS VS FOOTBALL HIGHLIGHTS

Cricket Highlights	Football Highlights
Usually, clips generated from the cricket video are more as compared to football. Because there are more exciting moments produced in cricket than football.	Usually, clips generated from the video are less as compared to cricket. Because there arelessexciting moments produced in football than cricket.
the threshold for generating Cricket highlight is 300.	the threshold for generating Football highlight is 450.
The energy Impulse length for generating Cricket highlight is 10.	The energy Impulse length for generating Football highlight is 7.
It is more precise in case of cricket highlight generation.	Little bit less precise because it is not covering important tackles.

REFERENCES

1. C.Liu, Q. Huang, S. Jiang, L. Xing, Q. Ye, and W. Gao, "A framework for flexible summarization of racquet sports video using multiple modalities," Computer Vision and Image Understanding, vol.113, pp. 415–424, 2009.
2. E.Kijak, G. Gravier, P. Gros, L. Oissel, and F. Bimbot, "HMM-based structuring of tennis videos using visual and audio cues," income,2003.
3. R. Radhakrishnan, Z. Xiong, A. Divakaran, and Y. Ishikawa, "Generation of sports highlights using a combination of supervised and unsupervised learning in the audio domain," inPacific Rim Conference on Multimedia, 2003.
4. Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li, "A user attention model for video summarization," in ACM Multimedia,2002
5. J. Zhang, J. Yu, and D. Tao, "Local deep- feature alignment for unsupervised dimension reduction," IEEE transactions on Image Processing, vol. 27, no. 5, pp.2420–2432, 2018.
6. K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization 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," in ACM Multimedia, 2010
9. G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," IEEE transactions on Multimedia, vol. 15, no. 7, pp. 1553–1568, 2013
10. RitwikBaranwal, "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.