Video Summarization for Sports

Shwethashree GC[1], Vishwas S Jois[2], Sudeep K[3] ,Tippesh Naik C[4] , Suraj S[5]

[1]*Assistant Professor, Dept of Computer Science and Engineering, JSS science & Technology University, Karnataka, India*

[2,3,4,5]*UG Student, Dept of Computer Science and Engineering, JSS science and Technology University, Karnataka , India*

---***---

Abstract - *This paper presents a transfer learning approach to summarize sports match videos by leveraging the single-stage detection principle of YOLO coupled with simple speech analysis. The main motive here is, to extract frames of importance in a sports video. The definition of importance of frame can vary according to the sport under consideration. Scope of these definitions are up to the individuals' needs, out of a sports match, Cricket and Football are the two sports under the consideration of this Paper. Crowd cheer is the event which can give an inkling to the important events that is taking place in the game. Thus, crowd cheer is considered as a common look-out event for both the games. Additionally, for football Red Card or Yellow Card shown by referee and a significant change in the score of the team within a short time, for cricket is considered as primary events in this Paper. Same are looked for, from video and corresponding sub-clips are concatenated to produce a summarized video, In layman's terms highlights of the match.*

Key Words: **Summarization, OCR ,YOLO ,Object Detection**

1. INTRODUCTION

Sports highlight generation is the task of creating a summary of a sport event that gives the viewer a summary of the game through its key moments. In general, they are created according to user description. There are many video retrieval techniques available most of which depend on user description, tags and thumb- nails which are not very descriptive. It also depends on the content creator who can add click-baits and irrelevant description which are not accurate. In this case, an automated tool that can generate the highlights can become handy.

Existing efforts in this regard consider only single features like background audio analysis or important frame detection which is being generic to all category of videos is used to generate highlights which can often miss out other important events due to obvious discrepancies in audio or definition of importance might change according to the sport. This would rather lead to decrease in efficiency of the outcome. Thus, here, through this paper we are aiming at gathering all the important events based on the pre-defined criteria by the addition of new efficient features to the model. The aim here is to come up with a simple web application where user can upload the video of the sport and get the video rendered on the same platform.

2. LITERATURE SURVEY

[1] Sanchit Agarwal, Nikhil Kumar Singh, Prashant Giridhar Shambharkar proposed a novel method for automatic annotation of events and highlights generation for cricket match videos. The videos are split into intervals denoting one ball clip using a Convolutional Neural Network and Optical Character Recognition. CNN detects the start frame of a ball. OCR is used to detect the end of the ball and to annotate it by recognizing the change in runs or wickets. This proposed framework is able to annotate events and generate sufficiently good highlights for four full length cricket matches. But this method requires more of processing techniques in order to annotate.

[2]Pushkar Shukla, Hemant Sadana,Apaar Bansal, Deepak Verma, Carlos Elmadjian, Balasubramanian Raman, Matthew Turk proposed a model that considers both event-based features and excitement-based features to recognize and clip important events in a cricket match. Replays, audio intensity, player celebration, and play field scenarios are examples of cues used to capture such events. Basically, all the events were grouped to form sub-highlights event-driven highlights and excitement-driven highlights. Events coming under event-driven highlights are boundaries, sixes, wickets, milestones and those under excitement-driven highlights are Audio cues and player celebrations.

[3] Ritwik Baranwal suggested 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 commentator speech’s amplitude to decide the threshold. 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

[4]Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkil and Naokazu Yokoya presented a video summarization technique for an internet video to provide a quick way to yoverview its content. They propose to use deep video features that can encode various levels of content semantics, including objects, actions, and scenes. A deep neural network that maps videos as well as descriptions to a common semantic space and jointly trained it with associated pairs of videos and descriptions was designed. To generate a video summary, the deep features were extracted from each segment of the original video and apply a clustering-based summarization technique to them.

[5]Ajeet Singh Ram Pathaka, Manjusha Pandeya, Siddharth Rautaraya worked on Application of Deep Learning for Object Detection. Deep learning frameworks and services available for object detection are also discussed in the paper. Benchmarked datasets for object localization and detection released inhas been discussed. State-of-the-art deep learning-based object detection techniques have been assessed and compared. Deep learning frameworks and services available for object detection are also enunciated

[6] P.K. Sandhya Balakrishnan, L. Pavithira in their paper aimed to optimize Convolution Neural Networks (CNN) using Simulated Annealing (SA), for optical character recognition. The results of a DBN are often highly based on settings in particular the combination of runtime parameter values for Deep Learning. Simulated Annealing is proposed to increase the results of Convolution Neural Network. The classification accuracy from the proposed method is lower than the original of CNN for variation of fonts. The proposed method could potentially be employed and tried for benchmark dataset.

[7] Washington L.S. Ramos, Michel M. Silva, Mario F. M. Campos, Erickson R. Nascimento proposed a novel methodology to compose the new fast-forward video by selecting frames based on semantic information extracted from images. Proposed method analyzes the semantic score of each frame and segments the video into semantic and non-semantic parts. Based on the length of each segment and the desired speed-up for the final video, an optimization function is solved to select different speed-ups for each type of segment. The frame selection is performed by a shortest path algorithm.

3. PROPOSED METHOD

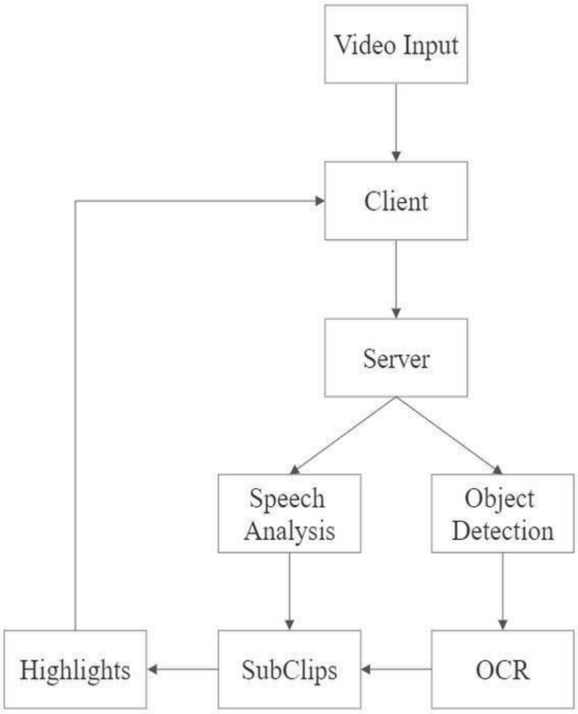

Fig-1 : System Architecture

The entire system is composed of four different modules. All of these modules are independent of each other. The Deep learning models for Object Detection and OCR are specific to each sport. Speech Analysis and User Interface module are common to both the sports. Dataset is custom built [13] containing images related to that particular sport. Images are annotated accordingly. The entire video is made to pass through the modules in the same manner listed down under.

A. Object Detection Module.

B. OCR module.

C. Speech Analysis module.

D. Video Processing.

A User Interface will be provided where the user is given choices of sports category, the video of which has to be summarized. User can then upload the video.

3.1 Object Detection Module

At first, the video enters Object Detection Module For football, all the annotated images are trained on yolo model using transfer learning method[9]. Video under consideration for summarization is split into frames [18] and each of these frames are passed through Object Detection model to check if card is present in the frame. If so, is the case, then a sub-clip with 15 seconds interval on either ends from that timestamp are extracted and marked as important event. Similarly for Cricket, all the annotated scoreboard images are trained on yolo model using transfer learning method[12]. The frames of the Video are then passed through Object Detection model in order to detect for scoreboard in that particular frame[14]. If so then the frame is considered for processing in OCR module.

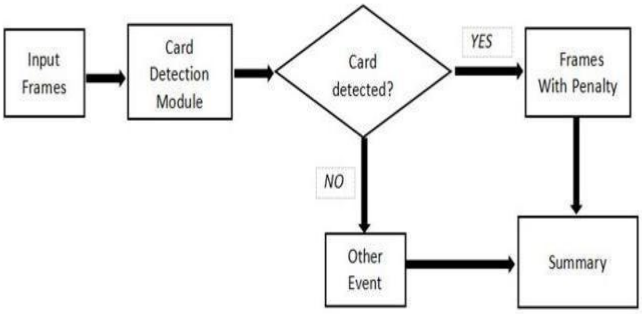

Fig-2 : Object Detection

3.2 OCR Module

In case of cricket, the images qualified by Object Detection model are processed to determine the score on scoreboard. The contents in the scoreboard are kept under scanner for every frame that comes in the interval of 10 seconds. After contents are determined, runs area and wickets area are bifurcated. Then, the runs and wickets are compared separately. If a significant change in runs is found i.e. if diff = 3,4,5,6 or if there is an increment in the wickets area, the sub clips of duration 30 seconds with 15 seconds on either side of the timestamp of that particular frame is extracted and considered as important event. In case of Football, if there is a change in goal counts area, the sub clips of duration 30 seconds with 15 seconds on either side of the timestamp of that particular frame is extracted and considered as important event.[10]

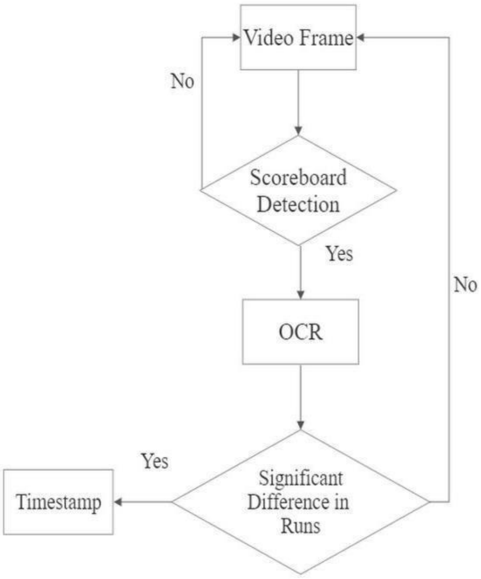

Fig-3: OCR

3.3 Speech Analysis Module

The entire sports video is fed into the librosa audio processor[8]. A threshold energy for chunk is defined. The duration of the video from which the amplitude of the crowd cheer ascends till it descends is noted down if the energy chunk crosses the defined threshold. Sub clips are extracted from the second the amplitude rises till the second amplitude falls down the threshold. [16]

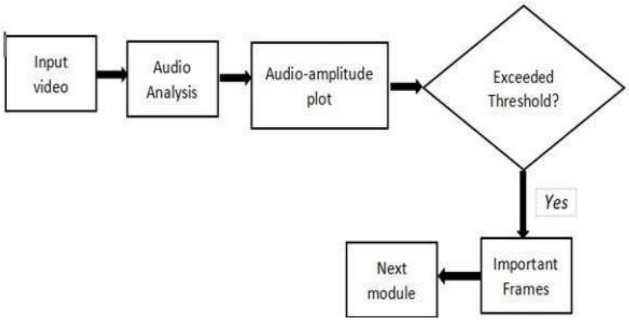

Fig-4 : OCR

3.4 Video Processing

All the timestamps from above modules are made to undergo union operation and sorted in ascending order. These timestamps are then used to extract sub clips from the Original Video. All the sub clips are concatenated to produce the highlights or summarized Video. The summarized video is then rendered in the same platform.

5. EXPERIMENT AND PERFORMANCE EVALUATION

5.1 Data Set

1) Football Cards Data set:600 annotated images containing cards shown by referee.

2) Score Cards Data set:250 annotated images consisting of scoreboards.

5.2 Transfer Learning

mAP	0.65
Threshold	0.5

Table -1: Card Detection Model

mAP	0.96
Threshold	0.5

Table -2: Scoreboard Detection Model

The Object Detection model for card detection was trained on pre-trained yolo-v3 model, with batch size 4 and num_iterations 10 on 500 images and that for scoreboard detection was trained for 180 images with aforementioned parameters used for training. Card Detection Model [20] achieved an mAP of 0.65 and Scoreboard Detection Model achieved and mAP 0.96 with IOU 0.5, nms threshold 0.5 and object threshold 0.3. On OCR end, images were initially resized to size 40*40 [19].Images were then resized to 4-D array for keras processing. A sequential model was chosen. The model has 1 CNN layer [11] with kernel size(3,3) with adam optimizer sparse categorical cross entropy as loss function. A hidden layer is present with 128 layers and ReLu activation unit. Final Output layer has 14 layers with SoftMax activation unit.

5.3 Audio Processing

Librosa sampling rate for processing the audio was set to 16000.The threshold energy chunk was at 12000.A plot was generated for energy chunks as well as distribution of amplitude over the duration of video [17]. Threshold value can be changed as required.

5.4 Editing

moviepy library was used to extract and concatenate sub clips. OpenCv was used to process each frame from the video. All the above functionalities were taking place on the top of flask server and highlighted video was being rendered on react application in the front end.

5.5 User Interface

Client User Interface was designed using React JS which communicates with back end flask server

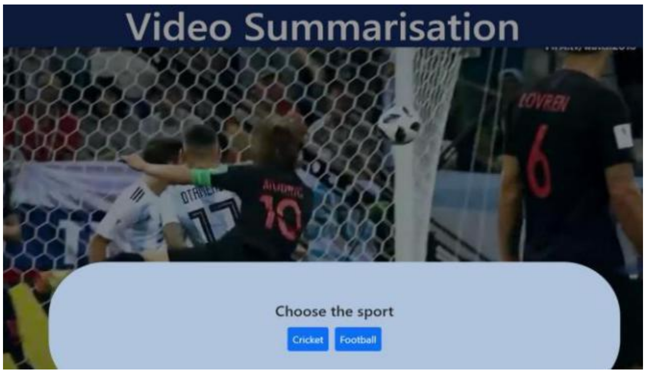

Fig-5 : User Interface

6. CONCLUSION

In this work, we presented a technique where a user can summarize sports video through a simple application. By examining each frame to get an idea if it could've been the part of important event, we were able to come up with pool of sub clips denoting important events in the match. The highlights detection was done by passing video through Object Detection, OCR and Audio modules. Subsequently, the detected highlights of the match were rendered to the user on the same platform. Although this model is specific to cricket and football, similar kind of models can be trained for different sports by having a data set accordingly.

The possible extension to this approach, which we have planned are as follows:

1.From football's perspective, additional important events can be defined like penalty, corner-kick, off-side, free-kick. In case of cricket, no-ball, byes, players' celebration can be thought of as important events.

2.Efficiency can be improved if models have relationship between them. An event marked as important by one module can be tested for the same in other modules to validate the outcome of that module. By doing this, false positives can be reduced.

3.It usually takes more time to process videos, we can come up with faster algorithms to pace up processing.

4.Broadcasters can think of implementing real-time highlights detection of a sports match. By doing which highlights can be generated instantaneously.

5.Instead of having to create separate models for every game, we can come up with a generic model altogether.

REFERENCES

[1]“Automatic Annotation of Events and Highlights Generation of Cricket Match Videos” by Sanchit Agarwal, Nikhil Kumar Singh, Prashant Giridhar Shambharkar,2019

[2]“Automatic Cricket Highlight generation using Event-Driven and Excitement-Based features” by Pushkar Shukla, Hemant Sadana, Apaar Bansal, Deepak Verma, Carlos Elmadjian, Balasubramanian Raman, Matthew Turk,2020.

[3]“Automatic Summarization of Cricket Highlights using Audio Processing” by Ritwik Baranwa,2020.

[4]“Video Summarization using Deep Semantic Features” by Mayu Otani, Yuta Nakashima, Esa Rahtu,Janne Heikkil¨a and Naokazu Yokoya,2016.

[5]“Application of Deep Learning for Object Detection” by Ajeet Ram Pathaka, *, Manjusha Pandeya,, Siddharth Rautaraya,2018.

[6]“Multi-font Optical Character Recognition Using Deep Learning” by P.K. Sandhya Balakrishnan, L. Pavithira,2019.

[7]“Fast Forward Video Based On Semantic Extraction” by Washington L.S. Ramos*, Michel M. Silva*, Mario F. M. Campos, Erickson R. Nascimento,2016.

[8]”Librosa: Audio and Music Signal Analysis in Python” by Brian McFee,Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenbergk, Oriol Nieto,Research Gate,2015

[9]"One-Shot Learning for Custom Identification Tasks; A Review" N. O’ Mahonya, Sean Campbella, Anderson Carvalhoa, L. Krpalkovaa , Gustavo Velasco Hernandeza, Suman Harapanahallia, D. Riordana and J. Walsha

[10]"Optical character recognition using deep learning techniques for printed and handwritten documents.", SSRN, 2020. Sanika Bagwe,Vruddhi Shah,Jugal Chauhan, Purvi Singh Harniya,Amanshu Shah Tiwari,Vartika Gupta,Durva Raikar,Vrushabh Gada, Urvi Bheda,Vishant Mehta ,Mahesh Warang ,Ninad Mehendale

[11]"Research Paper on Basics of Artificial Neural Network",IJRTC,2014,Ms. Sonali. B. Maind ,Ms. Priyanka Wankar

[12]"Object Detection and Recognition Using YOLO: Detect and Recognize URL(s) in an Image Scene" by John Temiloluwa Ajala

[13]"Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification" Anita Rácz, Dávid Bajusz and Károly Héberger,MDPI,2021

[14]“Yolov3: An incremental improvement,”J. Redmon , arXiv, 2018

[15]"ImageAI:Comparison Study on Different Custom Image Recognition Algorithms", Manuel Martins,David Batista Mota ,Francisco dey Morgado,Cristina Wanzeller,Research gate, 2021

[16]"Audio to Sign Language Conversion" Shivangi Saharia, Priyanka Kulkarni, Amardeep Bhagat, Himalaya Prakash, Amber Kesharwani , IJARCCE,2020

[17]"Data visualization: an exploratory study into the software tools usedby businesses", Michael Diamond,Angela Mattia,Journal Of Intstructional Pedagogies,2017

[18]"OpenCV for Computer Vision Applications",Naveen kumar Mahamkali,Vadivel Ayyasamy,Research Gate,2015

[19]"The NumPy Array: A Structure for Efficient Numerical Computation", Stéfan Johann van der rosie Walt,S. James Chris Colbert,Gael Varoquaux, ResearchGate,2011

[20]"A Tour of TensorFlow",Peter Goldsborough,Research Gate, 2016