Proc. of Int. Conf. on Intelligence Computing & Information Technology , ICIT

Video Summarizer: Generating Abstract View of the Sports Videos

Vinay Rajpoot[1] and Sheetal Girase[2]
[1-2]Department of Information Technology, Maharashtra Institute of Technology, Pune, India
Email: vinayrajpootvinay@gmail.com, Email: sheetal.girase@mitpune.edu.in

***Abstract*—This work presents a methodology for summarization of sports videos (Cricket) using both audio and visual information. In recent years, there has been massive growth in recorded audio-visual content and on an average people are spending most of their time watching long sports videos. It will be useful if we can produce a glimpse of the video, by generating the summary of it. In this paper, we propose audio-video processing based approach for generating the summary of sports videos. The original video is divided into two processing tracks: video track and audio track. In the video track, video frames features are extracted using the convolutional neural network (CNN) and temporal sequences or scores are generated that represent how important the particular frame is. The threshold is applied and the sequences are learned using Long-short term memory (LSTM). In the audio track, audio samples are read in the interval of the sampling rate and intensities are calculated. The threshold is set and an array of seconds are aggregated which is aligned with the output of video track. After alignment, the summary is generated. The CNN trained with 74 videos of cricket sports actions. Our approach achieved much accuracy compare to previous methods.**

***Index Terms*— Convolutional neural network (CNN), Long-short term memory (LSTM) and Audio intensities.**

I. Introduction

Nowadays, there are a lot of video contents are generating and widely available in our daily life, due to advancement in digital video technology, storage space and high-resolution video recording cameras. Also, Easy access to the internet has led to the wide availability of digital video contents. Sports videos that attract a large population all over the world and have grown as a valuable video content that is watched over TV networks and Internet [1]. Millions of people are interested to watch sports videos like cricket, football, baseball, tennis, etc. Generally, the duration of sports videos ranges from few hours to many days and composed with the exciting events and boring events. On an average people are spending most of their time watching long videos to gain the important content from full sports videos. It will be useful if we can produce a glimpse of video, by generating video summary of the video. Here, we are aiming to generate the summary of cricket sports videos which is the sport of bat and ball. Also, played in most Asian countries. We are taking cricket as a case study but our model will work for all sports videos which consist of audio-video exciting events [2,3,4].

Our aim is to generate a summarized view of the input video and try to summarize the important occurrences

Grenze ID: 02.ICIT.2018.2.507

or exciting events by processing audio and video information. In our case, exciting events are six runs, four runs and wickets. These exciting events captured by processing the noise created by the audience and commentators. It has been noted that if any exciting or interesting event occurs, the crowd makes noise and this is the largest indication that shows important event in the sports videos [5].

In this work, we are presenting a model that uses audio-visual information-based approach for summarizing interesting events in cricket sports videos. The original video is divided into two tracks: Video and Audio. A video data is used for extracting visual features from the video and an audio data is used for detecting excitement levels in the video such as: crowd noise, commentators' excited speech, whistles, drums, clapping, etc. It has been observed that audio power level increases when an interesting event or action occurs. Video frames features are extracted using the convolutional neural network and audio intensities are calculated.

In our proposed model full video is taken as input. Video frames and audio samples are extracted from the original video. Video frames are processed in video track and audio samples in the audio track. In the video track, the video is divided into frames and fed into the trained convolutional neural network (CNN) for feature extraction. The output of CNN is fed into Long Short-Term Memory (LSTM) for sequence learning after applying a threshold. At the end of the video track, the output received is the number of frames having a score above a certain threshold, where a single score denotes how important the content in the corresponding frame is. In the audio track, audio samples are read from 0 to n seconds with an interval of the sampling rate (44.1 kHz for stereo or 22.050 kHz for mono channel) and wave intensities are calculated at each second. After calculating intensities, a threshold is set, so the intensities corresponding to seconds, that lies above a certain threshold are included in the audio output. At the audio track output, there is a vector of seconds which shows interesting events timing in the original video. Finally, the output of video track and audio track are aligned and final summary is generated. Results show that by combining both audio and visual analysis we get better accuracy as compared to the single type of analysis.

Our contribution and innovative content are as follows:

- To our knowledge, the first approach that summarizes the sports videos by using the combination of CNN, LSTM and audio analysis.
- The type of the LSTM that verifies the sequences that are generated by the convolutional neural network. LSTM shows, the sequences are good or not for video summarization. The first time this LSTM is being used for video summarization.
- An approach used for generating the summary of cricket sports videos and the results are much better than previous methods that were based on traditional methods.
- The Neural Network based summarization approach that utilizes audio information of the input video and the model shows more accuracy compared to the previous Neural Network based approach.
- The actual end-to-end tool that takes full input video and automatically generates the summary.

II. Related Work

Audio is the most important content which describes the interesting events in the sports video. There has been a lot of reported work in the field of visual analysis as well as audio analysis in sports videos. Traditionally, video summarization was based on key-frames extraction from the video, the approaches for key-frame extraction is based on changes in video frame contents like motion activity [9] and colour histogram [12]. Another approach is clustering based on similarity of content using supervised and unsupervised learning [8]. Also, there are many techniques for video summarization such as multi-resolution key- frame extraction [10], object-based [6] by analysing trajectories of objects [11].

Recently, Deep learning is the area of interest for vision task. Most of the work has been reported for video summarization using Convolutional Neural Networks and Long-short term Memory. Kaiyang Zhou et al [7] proposed a deep summarization network for unsupervised video summarization, a probability is predicted for each video frame and selection of frames are decided based on the probability distribution. Also, feedback reward is calculated based on the effectiveness of summary. Jeff Donahue et al [13] presented a model for video description and the summary which is generated is in text format that describes the content in the video. The framework uses CNN for feature extraction and LSTM for sequence learning. Arnau Raventos et al [14] proposed a method for automatic soccer highlights generation based on audio and video descriptors. They have shown that audio gives the most important information about interestingness in the video. One more approach which uses the deep neural network for analysing audio and video is proposed in [16], the framework divides the video into audio and video tracks and features are extracted from both tracks.

III. Proposed Methodology

With the inclusion of audio analysis, the more accurate summary is generated compare to video analysis only. Audio data gives an indication of the most exciting events present in the input video. The methodology for summary generation is shown in figure 1. A method is divided into two processing tracks: Video track and audio track. An original video is given to the model as input where video frames are extracted for the video track and audio samples are obtained for the audio analysis track. In the video track, after obtaining video frames, each frame is fed into trained CNN model for recognizing action classes (with which action classes our input frames matches) and temporal sequences are generated at the output which is also known as scores. Scores are the values (0 to 1) which represent the importance of frame. At the output of CNN, scores are generated for each frame which gives information about the importance of frames. After aggregating scores, a threshold is applied for limiting the number of frames. Only the frames which have scores above a certain threshold are contained at the output of video track. LSTM is used for sequence learning. It takes sequences from CNN and gives an output that particular sequence for video summarization is good or not [15]. In the audio track, after obtaining audio samples from the video, an audio wave is read from 0 to n seconds in the interval of a sampling rate (44.1 kHz for stereo or 22.050 kHz for mono channel). Wave intensities are calculated for each second and threshold is set. So, the intensities that lie above a certain threshold are included in the audio track output. It is an array of seconds which shows interesting event's timing in the original video. Finally, the output of video track and audio track are aligned and summary is generated.

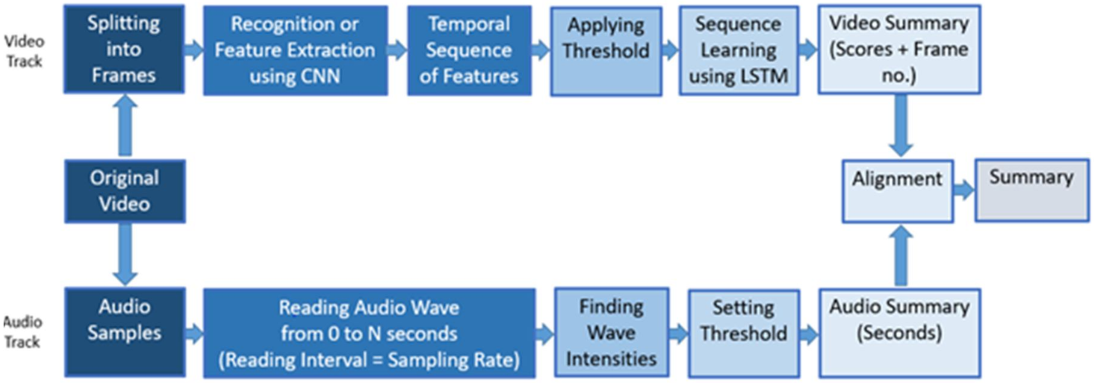
Video Track
Splitting into Frames
Recognition or Feature Extraction using CNN
Temporal Sequence of Features
Applying Threshold
Sequence Learning using LSTM
Video Summary (Scores + Frame no.)
Original Video
Alignment
Summary
Audio Track
Audio Samples
Reading Audio Wave from 0 to N seconds (Reading Interval = Sampling Rate)
Finding Wave Intensities
Setting Threshold
Audio Summary (Seconds)

Figure 1. Block diagram of proposed methodology

IV. System Architecture

The overall system architecture is summarized in figure 2. A system takes a full video as audio samples and a sequence of frames. Video frames are fed into the CNN for feature extraction. The CNN contains five layers; the first three are convolutional layers and remaining two are fully connected layers [17]. The first convolutional layer takes original input frames, it has 32 filters with size 3x3 and stride 1. We have used ReLU activation for triggering different network functions. Feature maps are generated at the output of the convolutional layer. There are two types of pooling operation for down-sampling: Max pooling and average pooling. We have used max pooling for down-sampling operation. Filter size is 2x2 and the stride is 2 at the first pooling layer. The specification of the second and the third convolutional layers are same as previous except at the third convolutional layer has 64 filters. Also, the specification of the second and the third pooling layers are same as previous. There are two fully connected layers at the end of CNN with 128 output nodes. The output of fully connected layer is fed to a softmax that produces distribution over class labels. The LSTM is a memory cell and used for sequence learning. It learns sequences and gives an output that they are good or not for video summarization. Here, we are using a pre-trained model of LSTM and the type of LSTM which verifies the sequences, generated by the CNN. We have isolated LSTM model with main CNN model and build separately for easy interpretation and understanding. We have used 128 units of LSTM. At the audio track, wave intensity shows the amount of energy present in the audio sample. Wave intensities are found by calculating the energy of discrete samples by the following formula:

$$E(m) = \sum_{k=k1} |X_m(k)|^2$$

Where m is frame time index and k1 is a DFT sample.

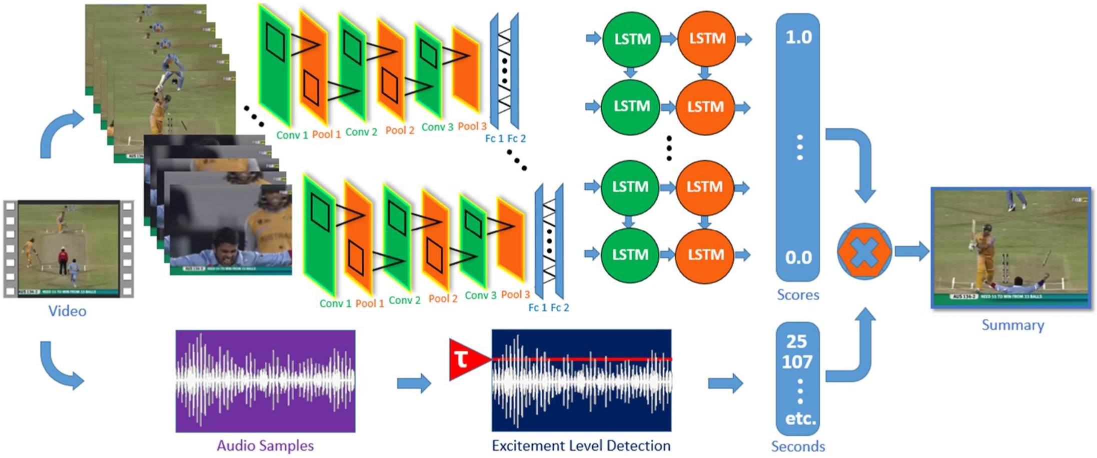

Figure 2. Overall system architecture

V. Dataset

We used YouTube-8M dataset which is the largest video classification dataset. It contains 8 million videos (500k hours of video) annotated with 4800 vocabularies. Also, it has multiple sports categories. We selected cricket sports category for getting the desired data. The dataset provides the link of youtube.com for downloading the videos and the duration of videos ranges from one minute to hours. In this work, the required videos must contain interesting events of cricket sport like sixes, fours and wickets with audio and video mode. So, to train the neural network we manually created action dataset of cricket sport. We downloaded the cricket videos from YouTube-8M dataset and manually snipped them to isolate them into distinctive actions. We created 74 videos for the training of CNN with duration ranges from five seconds to many minutes and each video contains actions of cricket. Some videos are in the MPEG-4 format (320x240) and remaining are in the 3GP format (320x180) in colour with sound. The frame rate is 25 fps. The neural network trained with these videos.

VI. Training of Neural Network

The actual time required to train the model is much high. It depends on the model and the amount of the training data. The training of the neural network did not take much time because we used only cricket sport videos. It trained in less than 15 hours. Since there was only one sports category, so the training data was less, in turn, less time required to train the model. The model trained with necessary data required to achieve the best performance and with a number of epochs needed to achieve stable performance. The training froze when performance not improved. The actual time required to train the model is much high. It depends on the model and the amount of the training data. The training of the neural network did not take much time because we used only cricket sport videos. It trained in less than 15 hours. Since there was only one sports category, so the training data was less, in turn, less time required to train the model. The model trained with necessary data required to achieve the best performance and with a number of epochs needed to achieve stable or best performance. The training froze when performance not improved. We used 74 cricket videos for training the

Table I. Training and Validation-Set

Total Frames	Training-set	Validation-set
25776	20621	5155

neural network and defined 74 action classes. All videos split into frames and these frames used for training the network. We divided 80 % data for the training set and 20 % data for the validation set. The division of training and validation frames are shown in table 1. We got a stable result in three iterations of training and observed that at each iteration training accuracy doubled than its previous iteration, in turn, improving validation accuracy. Validation loss decreased in each iteration.

VII. EXPERIMENTS & RESULTS

We performed an experiment by giving single video as input to the model and there are various results appear at each step. For example, input video "Australia vs India.mp4" (duration: 39:28 minutes, frames: 59182, width: 480, height: 360, size: 161 MB) is given to the model and the end summary result (duration: 01:12 minutes, frames: 1800, width: 480, height: 360, size: 64.1 MB) is shown in figure 3, with applying threshold (τ) = max – ((max * a)/100), where a = 25 and max is maximum value.

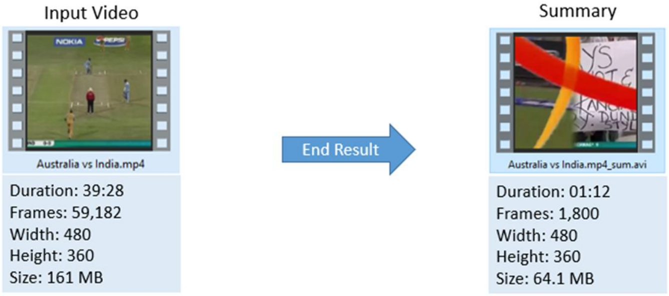

Figure 3. Full length video as input to the model and generated summary

There are following performance results:

A. CNN Category Distribution

After giving input video to the model, frames are extracted and given to the trained CNN for predicting video frames features to the classes or categories or to which feature categories the input video is matching. For the above input video, the CNN category distribution plot generated that is shown in figure 4, which shows how much percent of each class or category matches to the input video (outer portion of the plot shows the name of classes).

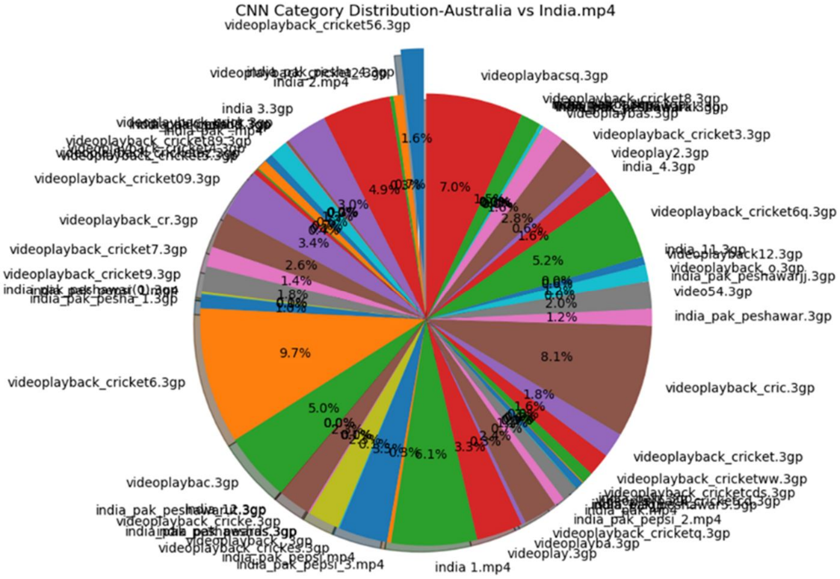

Figure 4. CNN category distribution

B. Timing vs Sound Intensities

The audio wave is extracted from the video and stored in the .wav file. The file is stored in lossless file format with storage space in the following format:

Storage space = Sample rate x Sample size x Time x Stereo sound.

The wave file is read in the interval of sampling rate and wave intensity is calculated for each second. For above input video, the sound intensity is shown in figure 5.

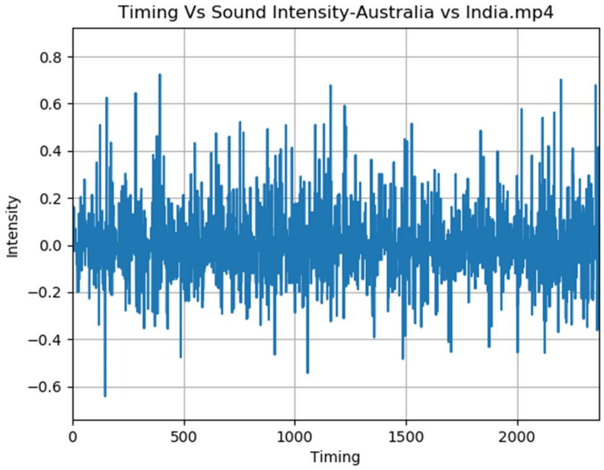

Figure 5. Timing vs sound intensities graph

C. Timing box plot

Timing box plot is shown in figure 6 for above input video, where the different intensities of an audio wave are shown in a single point of the x-axis. It shows the minimum and maximum peak of intensities present in the audio wave.

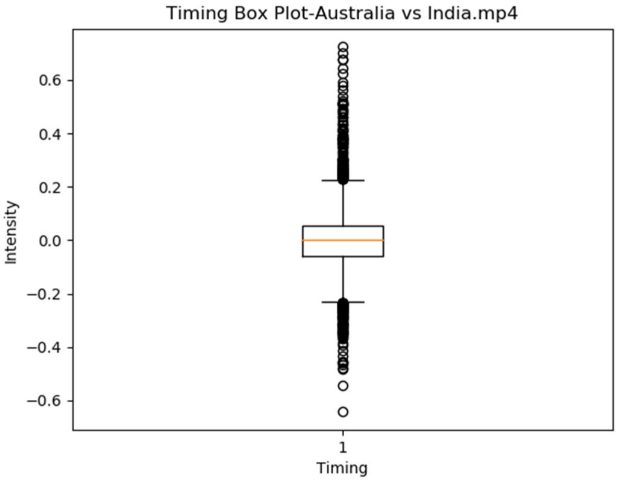

Figure 6. Timing box plot

D. Video Summary Timeline

Video summary timeline shows the segments of video which are included in the summary video. The video highlight timeline for above input video is shown in figure 7. Also, it shows the break duration of segments and gives information that at which time of input video the segment is taken for summary generation.

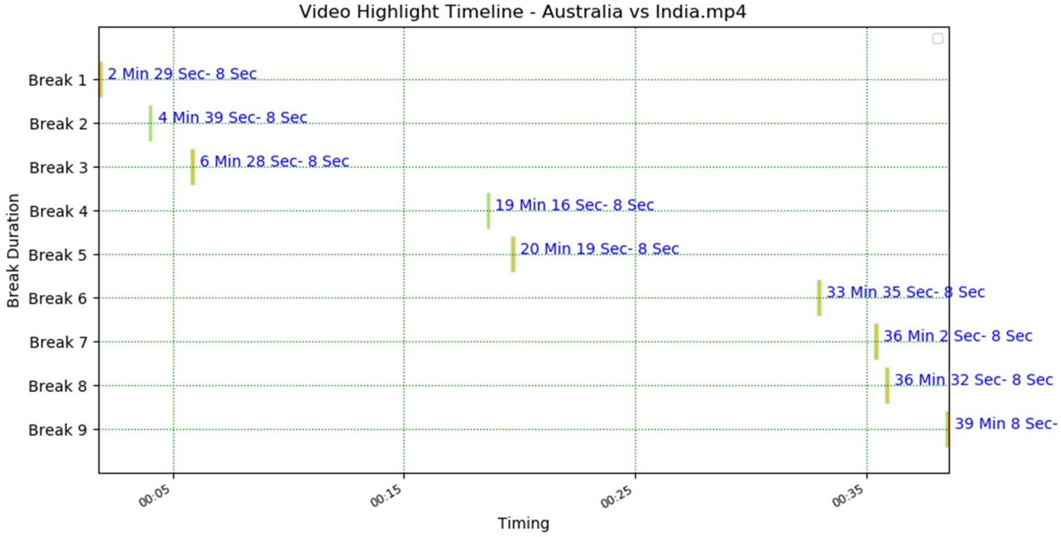

Figure 7. Video summary timeline

VIII. Conclusions

In this paper, a novel approach is present for video summarization using neural networks. Visual information, as well as audio information, is utilized for summary generation. The input video is processed in two processing tracks. Video frames are passed through CNN and LSTM for feature extraction and sequence learning respectively. Using CNN is an efficient way to handle features of video frames. Audio samples are processed in the audio track, where wave intensities are calculated and exciting events are detected in the full-length input video.

It has been observed, at each iteration training accuracy has been doubled than its previous iteration, in turn improving Validation accuracy. The summary contains most interesting events that are present in original input video. The processing speed can be increased by utilizing a powerful computer with GPU processors.

References

[1] H. Tang, V. Kwatra, M. E. Sargin and U. Gargi, "Detecting highlights in sports videos: Cricket as a test case," *2011 IEEE International Conference on Multimedia and Expo*, Barcelona, pp. 1-6, 2011.

[2] D. Yow, B. L. Yeo, M. Yeung, and B. Liu, "Analysis and presentation of soccer highlights from digital video,*" In ACCV*, Singapore, pp. 1–6, December 1995.

[3] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for tv baseball programs," *In ACM Multimedia*, Los Angeles, CA, pp. 105–115, October 2000.

[4] S. Nepal, U. Srinivasan, and G. Reynolds, "Automatic detection of 'goal' segments in basketball videos,*" In ACM Multimedia*, Ottawa, Canada, pp. 261–269, October 2001.

[5] A. Hanjalic, "Adaptive extraction of highlights from a sport video based on excitement modeling,*" IEEE Transactions on Multimedia*, vol. 7, no. 6, pp. 1114–1122, 2005.

[6] C. Kim and J. N. Hwang, "An integrated scheme for object based video abstraction," *In Proceedings of the eighth ACM international conference on Multimedia*, ACM, pp. 303–311. 2000.

[7] K. Zhou, Y. Qiao, T. Xiang, "Deep Reinforcement Learning for Unsupervised Video Summarization with Diversity-Representativeness Reward", *In CVPR*, Feb. 2018.

[8] Y. Zhuang, Y. Rui, T. S. Huang and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," *Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No.98CB36269)*, Chicago, IL, vol. 1, pp. 866-870, 1998.

[9] W. Wolf, "Key frame selection by motion analysis," *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, Atlanta, GA, vol. 2, pp. 1228-1231, 1996.

[10] P. Campisi, A. Longari, and A. Neri, “Automatic key frame selection using a wavelet-based approach,” *In SPIE’s International Symposium on Optical Science, Engineering, and Instrumentation*, International Society for Optics and Photonics, pp. 861-872, 1999.

[11] A. Stefanidis, P. Partsinevelos, P. Agouris and P. Doucette, "Summarizing video datasets in the spatiotemporal domain," *In Proceedings 11th International Workshop on Database and Expert Systems Applications*, London, pp. 906-912, 2000.

[12] H. J. Zhang, J. Wu, D. Zhong, and S. W. Smoliar, “An integrated system for content-based video retrieval and browsing,” *Pattern recognition*, vol. 30, issue 4, pp. 643–658, 1997.

[13] J. Donahue et al., "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description*," In IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 677-691, April 1 2017.

[14] A. Raventós, R. Quijada, L. Torres, F. Tarrés, E. Carasusán and D. Giribet, "The importance of audio descriptors in automatic soccer highlights generation*," 2014 IEEE 11th International Multi-Conference on Systems, Signals & Devices (SSD14)*, Barcelona, pp. 1-6, 2014.

[15] J. Brownlee, Book title: “Long Short-term Memory Networks with Python: Develop Sequence Prediction Models with Deep Learning,” *Jason Brownlee*, 2017.

[16] D. S. Sachan, U. Tekwani, and A. Sethi, “Sports Video Classification from Multimodal Information Using Deep Neural Networks*,” In AAAI Fall Symposium Series*, 2013.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks*,” In Proceedings of the 25th International Conference on Neural Information Processing Systems* - Volume 1 (NIPS'12), F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.), Vol. 1. Curran Associates Inc., USA, ACM, pp. 1097-1105, 2012.