Journal of Critical Reviews

ISSN- 2394-5125 Vol 7, Issue 8, 2020

Review Article

Survey and Evaluation on Video Summarization Techniques

G. PRIYANKA[1]*, M. PRASHA MEENA[2]

[1]Assistant Professor (Senior Grade), Department of CSE, Mepco Schlenk Engineering College, Sivakasi, Tamilnadu, India,
[2]PG Student, Department of CSE, Mepco Schlenk Engineering College, Sivakasi, Tamilnadu, India,
*Email: priyanka@mepcoeng.ac.in

Received: 11.03.2020 *Revised: 12.04.2020* *Accepted: 28.05.2020*

ABSTRACT

Processing Audio Visual (AV) content is complex and time consuming as they are synchronous and time continuous information. But most of the today's information and the data uploaded in the internetwork is AV information. AV occupies more space and are computationally intense. The main goal of the video summarization algorithm is to take video to summarize as input and to identify/recognize the important and significant parts of the input and to output a video that has only important parts of video that represents the input video. There are some disadvantages in video summarization process like lack of benchmark dataset, proper processing tools, efficient summarization algorithm, etc. This summarization technique has application in many real-time industries like sports, cinema, TV channels, etc., that has large video content to be handled in their day-to-day life. This paper discusses some of the video summarization techniques that can be employed in processing AV contents.

Keywords: Video summarization, RNN based summarization, CNN based summarization, Summarization by Deep Learning, Audio-Visual processing

DOI: http://dx.doi.org/10.22159/jcr.07.01.01

I. INTRODUCTION

Video summarization [1] is the process of identifying the significant segments of the video and produce output video whose content represents the entire input video. It has advantages like reducing the storage space used for the video. In scenarios like maintaining surveillance videos, only few frames have meaningful information while other frames are still and has same information. Surveillance video with less needed information occupies more space. There are many such videos that occupy more storage space.

Consider the sports video, TV Serials/Series, reality shows, dramas, movies, etc. that are hours length videos. Watching these hours length video is time consuming and boring for the viewers. To make the video content crisp and also in reducing the storage space, video summarization can be done. One additional advantage to create wide publicity for any massive event is to publish the event's promo in TV channels and social media one-week advance. The summary contains only the important content that represents/keeps the main content of the original video.

So far, these hours length videos are summarized using manual video editor who watches the entire video and summarizes. The summary created depends entirely on the mindset of the editor. They are prone to specialization (based on the his/her favorite content), error (non important content may be identified as important), etc. In order to generalize the summarization task, automation must be done. This paper discusses about various summarization techniques [8], [5] that can be applied on the AV contents from the real-time materialized world.

II. RELATED WORK

In today's computer world, the Deep Learning models and Techniques has penetrated and made considerable landmark in almost all specializations of IT sectors. It has made entry in the video summarization task also. The Deep Learning model trains to learn the representation of the important moment and identifies only those in the video given for the summarization. There are many methods in summarization namely,

1. Feature based video summarization: Video has features like color, motion, dynamic contents, gesture, audio-visuals, objects, speech transcripts, etc. This method of summarization is done if the end user wants to analyse the video contents based on the features. If the summarization is based on the feature1 (say, color), then the color feature of the video is learnt by the Deep learning model.

2. Video summarization using Clustering: In this mode summarization is done when the end user wants to summarize the video based in the similarity within frames or the characteristics. This method makes the browsing of video and retrieval of video based on the content easier. This is done by making use of clustering tools of data mining like K-Means, partitioned clustering and spectral clustering.

Michele Merler et al, [2] proposed a Deep Learning based method to learn the video representation. The audio content and the visual contents are analyzed using Deep Learning models like Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), etc. to learn the representation. Final summary is generated in the form of the video.

Haoran Li et al [3], proposed asynchronous text, image, audio and video based summarization for the video. Each of the asynchronous components are analyzed separately and some optimization techniques are applied on the summary and the final textual summary with increased accuracy is generated. Salience match is also done to make the summary to be more appropriate.

Siyu Huang et al [4], proposed a ranking based mechanism to summarize video in multiple stages to learn the spatio-temporal representations. This method made their work, out perform the other state-of-art summarization techniques.

Ali Javed et al [7], proposed a method to summarize the audio-visual features of the cricket video. By identifying the key frame for the audio contents the key frames for the visual content is identified and the final summary is generated for the cricket match videos.

These summarization methods discussed earlier worked on the video input and produced the textual summary for the video. But reading the textual summary for a colorful entertaining event makes the summarization process less efficient. In order to make the summarization more interesting video summarization is mostly preferred. In this paper, the way of generating video summary is explained.

III. ARCHITECTURE OF SUMMARIZATION

The figure 1 shows the simple architecture of the video summarization process. The summarization algorithm or the method takes the video (entire event video) as input and outputs the summary as video that has only the important parts of the inputted video. The summary can be done by using different Deep Learning models in parallel and serial manner. It is discussed in the following sub-sections.

A. Recurrent Neural Network

The Recurrent Neural Network (RNN) based model is designed to learn the representation.

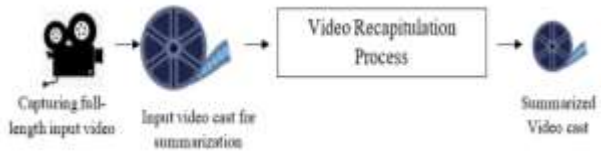

Fig. 1: Architecture of the Video summarization model

The content in the video is learnt in stages. First, the entire video is taken and preprocessed to be inputted. The frames of the video are inputted into VGG network to learn the representation. This learnt feature is feed as input to the Deep 1D FCN. Here the dimensionality of the input is reduced by applying pooling operations and the representation learnt is reduced to the dimension that can be feed to the LSTM architecture. The LSTM takes the output from the FCN network and outputs its prediction on the input. This is used in ranking the result.

From the rank obtained, the user-to-user consistency and summarization quality [4] are measured to check the correctness of the ranking done on the output. Based on the ranking, the output summary video is generated. The equation (1) and (2) shows the mathematical representation to compute the user-to-user consistency and summarization quality.

$$C(i,j) = \frac{1}{N}\sum_{n=1}^{N} F(S_i^n, S_j^n) \quad (1)$$

$$Q_i = \frac{1}{\|Z\|}\sum_{j=i,j\neq i}^{M} Q_j C(i,j) \quad (2)$$

Si is the summary of the video n created by user i = 1,2,3, …, M, where M is the number of users reviewing the video. The Equation 2 shows the summary quality of the user i. the consistency and the quality of a user is compared to decide the final quality of the user i. Siyu et al worked with SumMe [12] and TVSum [14] dataset and compared the results.

B. Using Convolutional Network

The contents of the video can be categorized as audio and visual information. The excitement score is computed for audio and the visual information to make the highlight. Now they are processed separately to reduce the computation complexity. The audio can be processed using Convolutional Neural Network to learn the cheering audio representation and the excitement tone of the commentator. The visual information can be processed to identify the action of the player.

For learning the audio representation, the Mel-Frequency Cepstral Coefficient (MFCC) is used. The frequency and the amplitude information for the cheer is learnt. From the learnt information, the CNN assigns the excitement score for the audio.

For the commentary processing, the CNN network is used to learn the voice tone of the commentator. Based on the voice tone the audio is identified to have excitement. The CNN assigns the excitement score to the audio based on the learnt voice tone representation.

For the visual information processing, the facial expression dataset like CK+ dataset [21] is used. The facial expression is learnt by using CNN network. The frames from the video are taken and preprocessed using Multi task Cascaded Convolution Neural Network (MTCNN) to confine the processing area to the region of frame having only the facial content. This preprocessed image is given to the CNN network. Based on the learn representation from the CK+ emotion facial dataset, the CNN network assigns the excitement score for the frames.

These excitement scores are combined using the weighted sum [2] of the excitement scores for each of the information from the video analyzed. The Equation 3 shows the mathematical representation of the weighted sum of the excitement score.

$$S(x) = \sum_{i=1}^{N} w_i E_n(x) \quad (3)$$

where wi is the weight assigned for audio and visual processing and En(x) is the excitement score for audio and the visual processing. This combined score is used in the highlight extracting process.

The dataset used in this CNN based highlight identification is the matches from the Wimbledon 2019 series and US Open 2019 series. The preprocessing is done on the input video before processing. For training the audio modules the audio from the match is taken, chunked and labeled by analysing the frequency and the amplitude of the audio.

The visual model is trained using the CK+ facial emotion dataset and testing is done by preprocessing the frames from the video and identifying the frames with region of interest, that is, with facial images.

IV. EXPERIMENTAL RESULTS

The first method works on the SumME and TVSum dataset. The experimental results obtained are tabulated in the Table 1. The table shows the variation in the quality of the summarization. As the iteration count increases the Q's value converges to the needed value such that the summarized video is more accurate and interesting.

Table 1: Variation in User Summary Quality Q

Iteration.	Variation of Q	
	SumMe [12]	TVSum [14]
1	18%	17.9%
2	2.7%	2.2%
3	0.6%	0.4%

The Table 2 represents the evaluation results of the model with other different methods. Our multi-stage learning method out performs the other methods for TVSum and SumMe datasets. The CNN based summarization technique also results in a video summary for the given input video. The resulting video is given to the sports fan for the ranking and to check the correctness of the highlight generated.

Table 2. Comparison of performance with State-of-art methods

Method	SumMe [12]	TVSum [14]
LSTM [11]	37.6	54.2
SUM-GAN [9]	38.6	54.7
DR-DSN [8]	41.4	57.6
MultiStage + Ranking (our proposed system)	**48**	**62**

The evaluation done by the sports fan for the video clips. The clips are taken from the Wimbledon 2019 [15], [16], [17] and US Open 2019 [18], [19], [20]. The performance of the combined excitement score is evaluated by computing the Normalized Discounted Cumulative Gain (nDCG).

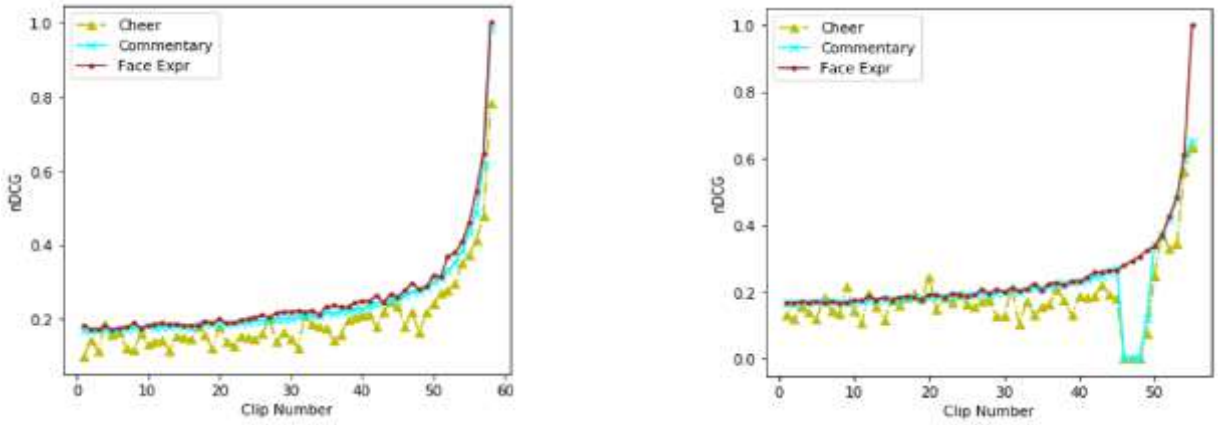

Fig. 2: Performance graph showing the plot of nDCG value against X-axis for each clip in the input sportscast in (a) Wimbledon 2019 and (b) US Open 2019

The equation 4 shows the mathematical representation of the nDCG computation.

$$nDCG(k) = \frac{1}{Z}\sum_{i=1}^{k}\frac{2^{rel_i}-1}{log_2(i+1)} \quad (4)$$

where rel_i is the score assigned by each CNN model for audio and visual analysis respectively and Z is the normalization parameter that brings the value of nDCG to 1. The following graph shows the nDCG value obtained for each of the analysis done on the input sportscast.

V. CONCLUSIONS

This paper makes the survey of the different methods to summarize the video and shows the evaluation results of the sportscast summarization by using RNN and CNN networks to learn the representation. The performance of both the method is also combined together.

The work can be further extended by analyzing the textual transcript of the commentator to improve the prediction accuracy of the commentator analyzing network. Further the game analytics like the score that is displayed can also be analyzed to generate personalized (player specific) summary.

VI. REFERENCES

1. Ashenafi Workie, Rajesh Sharma, Yun Koo Chung, "Digital Summarization Techniques: A Survey", International Journal of Engineering Research and Technology, Vol. 9, Issue 01, Jan 2020.
2. Michele Merler, Khoi-Nguyen C. Mac, Dhiraj Joshi, Quoc-Bao Nguyen, Stephen Hammer, John Kent, Jinjun Xiong, Minh N. Do, John R. Smith, Rogerio Schmidt Feris, "Automatic Curation of Sports Highlights Using Multimodal Excitement Features", IEEE Transactions on Multimedia, Vol. 21, No. 5, May 2019.
3. Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, Chengqing Zong, "Read, Watch, Listen and Summarize: Multi-Modal Summarization for Asynchronous Text, Image, Audio and Video", IEEE Transaction on Knowledge and Data Engineering, Vol. 31, No. 5, May 2019.
4. Siyu Huang, Xi Li, Zhongfei Zhang, Fei Wu, Junwei Han, "User-Ranking Video Summarization with Multi-Stage Spatio-Temporal Representation", IEE Transactions on Image Processing, Vol. 28, No. 6, June 2019.
5. Tanuja Subba. Bijoyeta Roy, Ashis Pradhan, "AA Study on 'VIDEO SUMMARIZATION", International Journal of Advanded Research in Computer and Communication Engineering, Vol. 5, Issue 6, June 2016.
6. Jie Lei, Qiao Luan, Xinhui Song, Xiao Liu, Dapeng Tao, Mingli Song, "Action Parsing-Driven Video Summarization Based on Reinforcement Learning", IEEE Transaction on circuits and Systems for Video Technology, Vol. 29, No. 7, July 2019.

7. Ali javed, Aun Irtaza, Hafiz Malik, Muhammad Tariq Mahmood, Syed Adnan, "Multimodal framework based on audio-visual features for summarization of cricket videos", IET Image Processing, 2019.
8. K. Zhou and Y. Qiao, "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," in Proc. AAAI, 2018, pp. 7582–7589.
9. B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial LSTM networks," in Proc. IEEE Conf. CVPR, Jul. 2017, pp. 202–211.
10. Manasa Srinivas, M.M.Manohara Pai, Radhika M. Pai, "An Improved Algorithm for Video Summarization - A Rank Based Approach", 12th International Multi-Conference on Information Processing, 2016.
11. K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in Proc. ECCV, 2016, pp. 766–782.
12. Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "TVSum: Summarizing Web videos using titles," in Proc. IEEE Conf. CVPR, Jun. 2015, pp. 5179–5187.
13. M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in Proc. ECCV, 2014, pp. 505–520.
14. Tinumol Sebastian, Jiby J. Puthiyidam, "A Survey on Video Summarization Techniques", International Journal of Computer Applicaations, Vol. 132, No. 13, Dec 2015.
15. https://www.youtube.com/watch?v=T4S5YmO0KOU&list=PLwx9gNibGUz6SzJb1zHdstIDrBBy1ZB0b&index=5
16. https://www.youtube.com/watch?v=wZnCcqm_g-E
17. https://www.youtube.com/watch?v=zjX2sRXaGVk
18. https://www.youtube.com/watch?v=YsN1Wl290fs
19. https://www.youtube.com/watch?v=nH0a_LHQpws
20. https://www.youtube.com/watch?v=634UMLDrVzc
21. https://www.kaggle.com/shawon10/ck-facial-expression-detection