Scientific Journal of Impact Factor (SJIF): 5.71

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 6, Issue 06, June -2019

Video Summarization - An improved Framework using CNN

Suvarna Patil[1], Dhanashree Phalke[2]

Department of Computer Engineering, SPPU, DYPCOE, Pune, Maharashtra, India

Abstract — *Summarization is an act of exhibiting the most significant thing which is used in every field like Video, Audio and Text summarization. Video Summarization is a concise and meaningful representation of a video which is mostly used in surveillance of data. Video summarization has been mainly preferred to enhance faster browsing of large videos collection and more effective contents accessing and indexing. The proposed system which is summarizes the video by using CNN technique. In this work, the input video is extracted number of frames, then applying Convolutional Neural Network the extracted frames are experienced with the trained dataset for recognizing important frames. From the sorted frames, the system generates the resultant summarized video. The proposed work generated more accurate video summary. Various cricket videos are collected as a dataset.*

Keywords*- Video Summarization, Key Frames, Convolutional Neural Network, Static Summary, Dynamic Summary.*

I. INTRODUCTION

The modern studies in compression techniques, accessibility of high speed connection and the diminishing cost of storage have promoted the creation, storage and distribution of videos. In recent years, video surveillance technology has become ubiquitous in every sphere of our life. Automated video surveillance system generates large quantities of data, which finally does rely upon manual inspection at some stage. Video reduction has been a field of active research for an extended time. However, the main attention was on either minimizing storage usage by compressing or removing redundant frames without loss of actual content.

The goal of Video Summarization is to generate a compact visual summary that wrapping the key components of the video. The Video Summarization methods are useful for various practical applications like video browsing, analyzing surveillance data, action recognition etc[12][13][14][15].

Now days, increasing the popularity of social media and expansion of video capturing devices, there are huge volumes of videos being captured and uploaded every second. So, in every seconds the huge amount of data has been generated. So need a well-organized way to handle these large video data.

Due to the increasing volume of video content on the Web, and the human attempt taken to process it, new technologies want to be researched in order to extend proficient indexing and search techniques to handle effectively and competently the huge amount of video data. One of the most rising research areas is Video summarization. As the name implies, video summarization is a mechanism to produce a short summary of a video to offer to the user a artificial and useful visual abstract of video sequence, it can either be a images (keyframes) or moving images (video skims).

Video summarization can be represented into two modes: A static video summary (story-board) and a dynamic video skimming.

A. Static video Summarization

The Static video summaries are created of a set of keyframes which is extracted from original video where the keyframe is a frame that represents the contents of a logical unit. This summary is generated on most relevant frames. This is also called a key frame based video summarization or still image abstract or storyboard [2][10].

B. Dynamic Video Summarization

The Dynamic video summaries are created of a set of shots where the shot represents a spatio-temporally coherent frame sequences which captures a continuous action from a single camera. This summary consists of most relevant small dynamic audio and video portion. The idea of dynamic summarization called as video skimming is a short video composed of informative scenes from original video presented to the user to receive in video format that is it condenses the original video into shorter form while preserving the important content of a video in short time. This technique summarize the video in which video and audio portion also consists significant audio or spoken words, instead of simply understanding the synchronized portion corresponding to the selected video frames[11].

The proposed system for video summarization that uses extracted Video Frames for preprocessing. These featured frames are thereafter analyzed via Convolutional Neural Networks (CNN). Using this analysis, features are extracted and calculated, which are used for generation of summarized videos.

II. LITERATURE REVIEW

There are various approaches for the video summarizations which can be found in the literature. These are discussed in this section.

Xuelong Li, Fellow, IEEE, Bin Zhao, and Xiaoqiang Lu[1] “A General Framework for Edited Video and Raw Video Summarization”, The proposed work is divided into three parts:- 1. To design the four models for capture the properties of video summaries like importance, representativeness, diversity, storyness. 2. To balance the influence of the four property models, a score function is developed with weighted combination of them. 3. To construct training set which can address the problem of lacking the training data and to reduce the structure mess in training set by the mixing coefficient. So Author built a general framework for both raw and edited video summarization.

Sandra Eliza Fontes de Avila, Ana Paula Brando Lopes, Antonio da Luz Jr. [2] proposed “VSUMM: A mechanism design to produce static video summaries and novel evaluation method ". In this paper, Author presents VSUMM, a methodology for the creation of static video summaries. It is simple and effective approach for automatic video summarization. The methods is based on the extraction of color features from video frames and unsupervised classification and also added new methodology for evaluation the video summarized called as comparison of user summarized i. e. CUS. In this method, the summaries are made by users and compared with other approached.

A. Rav-Acha, Y. Pritch, and S. Peleg [3] proposed “Making a long video short: Dynamic video synopsis”. Author introduce dynamic video synopsis where most of the activities in the video is combined by simultaneous viewing several actions even when that are originally occurred at different time. Author presented two approaches. First approach uses low level graph optimization where each pixel in the synopsis in the video is a node in the graph. In this approach directly obtain the synopsis but the complicity is high. In the second approach direct detect moving object and perform optimization on the detected object.

Y. Hadi, F. Essannouni, and R. O. H. Thami [4] proposed “Video summarization by k-medoid clustering”. Author introduce a video summarization algorithm by multiple extraction of key frames in each shot which is based on the K-medoid clustering algorithm for finding the best representative frame for each video shot. In this work the distance between frames is estimated using fast full search block matching algorithm based on the frequency domain. In addition, this approach generates different key frames even in the presence of large motion.

M. Gygli, H. Grabner, H. Riemenschneider, and L. V. Gool [5] proposed “Creating summaries from user videos”. Author introduces a novel approach and a new bench mark for video summarization. They proposed the novel temporal super frame segmentation for user video and methods to generate informative summary from them. Author propose a new method to calculate the interestingness of superframe and selecting summary from using 0/1 knapsack optimization. By evaluation proposed method it is generally show that to create good automatic summaries.

Ana Garcia del Molino [6] proposed “Summarization of Egocentric Videos: A Comprehensive Survey". In this paper the author introduced the need of video summarization techniques for the multiple egocentric contexts the characteristics of FPV, and how FPV summarization techniques differ from TPV. Then he also presented a general framework for FPV video summarization and review and organizes the literature according to it. The presented framework is data oriented, depending on the given input images or video and desired output story boards, video skimming or fast-forwarding, as defined in Section. It consists of two steps: First is segmentation of the input data, and second is selection of the relevant segments or key frames. The depth analysis in depth the datasets used for this task and the obtained results and evaluation approaches. They finalize by giving some insight on the promising research directions and challenges.

S. Zhang, Y. Zhu, and A. K. Roy-Chowdhury [7] proposed “Context-aware Surveillance video summarization”. The context-aware video summarization (CAVS) framework which is able to nd the most informative video portions, from video sequences is given. The sparse coding with generalized sparse group lasso is used to learn a dictionary of video features and a dictionary of spatiotemporal feature correlation graphs. Sparsity gives the most informative features from the video.

Sinnu Susan Thomas, Sumana Gupta [8] proposed “Perceptual video summarization - a new framework of video summarization”. In this paper, Author presents a new framework for video summarization i. e. perceptual video summarization. Author introduces for the first time features of Human visual system (HSV) and allow for the emphasis perceptually significant event while concurrently removing perceptual redundancy from the summaries. Author proposes to create an image like panorama registration based on some superior criterion for the choice of the reference frame from the video.

Yifang Yin, Roshan Thapliya et al [9] proposed method for automatic video summary generation with personal adaption. Author introduces a novel hierarchical dictionary name semantic tree (SeTree). SeTree is a hierarchy which captures the conceptual relationships between the visual scenes in the codebook. The author proposed the automatic content based feature encoding approach with semantic tree which is more effective for personalized adaption. In the proposed design of video summarization, it joins the personal interest and visual attention.

III. PROPOSED SYSTEM

The main objective of the proposed system is to create the summary of original video. By using summarization techniques, it becomes shorten than the original video and easy to understand the important event in the original video.

- To give quick browsing and retrieval of large collection of video data without losing the significant aspects.
- Fetching highlights of video.
- To minimize storage usage by compressing or removing redundant frames without loss of actual content.

A. System Architecture

The proposed system is used to generate the summary of the input video using CNN i.e. Convolutional Neural Network. Fig.1 shows the overall architecture of the proposed system.

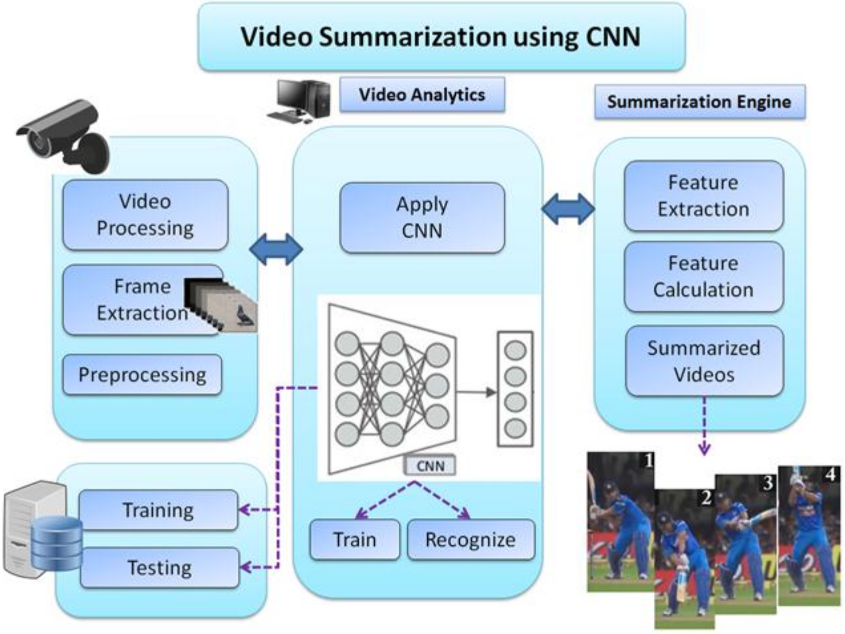

Figure.1. System Architecture

In the proposed system, first, the video is extracted into number of frames by using FFMPEG and each frame is saved along with the frame number for later summary generation. After that each frame is converted into grayscale image of 128*128 at preprocessing stage. Once the image is resized, it will be sent to CNN for matching. CNN will be used to train the video analytics engine for recognizing important frames in the video. Feature is extracted with the help of CNN (Convolutional Neural Network) and a network is trained with 3 layer CNN network. Frame number wise important frames will be stored, also extracted intensity of sound from the original video. The featured frame is analyzed by CNN as important frame that time audio intensity of the video is greater than 0.7, so this frame is selected and stored, thereafter before and after 5 second shot is captured for the summary. The proposed work generated more accurate video summary. Various cricket videos are collected as a dataset.

B. Proposed Algorithm

Input: Accept the frames with size W1×H1×D1

Output:- Recognized Featured Frame

Step 1. The input images are extracted from video and pass to the first convolutional layer.

Step 2. Convolution layer

This layer computes the output volume by computing dot product between all filters and image.

I. Accepts a volume of size W1×H1×D1
(Where W is width, H is Height and D is Depth)

II. Need four hyper parameters:
- Number of filters K
- Their spatial extent F
- The stride S
- The amount of zero padding P

III. Generate a volume of size W2×H2×D2 where:

$$W2 = (W1-F+2P)/S+1$$
$$H2 = (H1-F+2P)/S+1$$
$$D2=K$$

Step 3. ReLu Layer

In this layer, activation function apply to the output of convolution layer. Activation function such as max (0, x) thresholding at zero.

```
If   x<0 then
    f(x)=0
Else
    f(x)=x
```

Step 4. Polling Layer

This layer is used to shrink the image stack into the smaller size.

i. Pick a window size
ii. Pick the stride
iii. Move this window to across the filtered image
iv. From Each window to take the max value

Step 5. Fully Connected Layer

This layer takes the input from previous layer and computes the class score. This output represented in 1 D array which size is equal to the number of classes.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

1.1 Dataset

For analysis and testing purpose video data is collected from youtube. In which dataset, The Data files contains the cricket videos with MP4 file format. They contain different classes like Batting shot, Catch, LBW, Sixer, Boundary hit, Normal. For training and testing purpose uses 74 video.

1.2 Results

The proposed system classified each frame into different classes like Batting shot, Catch, LBW, Sixer, Boundary hit, Normal. Table shows that the result values of TP, TN, FP and FN. By observing the Table 1, the True-Positive value is greater so to obtain better accuracy. TP denotes True-Positive, TN denotes True-Negative, FP denotes False-Positive and FN denotes False-Negative.

Table 1. Classwise Classification

Classes	**TP**	**TN**	**FP**	**FN**
Classes N	2971	7501	264	704
Classes L	2940	7632	380	489
Classes R	2891	7231	227	718
Classes A	2786	7486	301	666
Classes P	2958	7514	309	727
Classes V	2226	7876	496	594

The following metrics used to evaluate the classification performance: accuracy, precision and recall. Accuracy indicates as the sum of correct classifications over the total number of input instances. The precision is the percentage of documents that are correctly classified as positive out of all the documents that are classified as positive, and the recall is the percentage of documents that are correctly classified as positive out of all the documents that are actually positive.

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (1)$$

$$\text{Recall} = \frac{TP}{TP+FN} \qquad (2)$$

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \qquad (3)$$

$$\text{F Measure} = \frac{2X\,\text{Precision}\ x\ \text{Recall}}{\text{Precision}+\text{Recall}} \qquad (4)$$

Table 2 show the results value of accuracy and precision. Accuracy indicates as the sum of correct classifications over the total number of input instances. The precision is the percentage of documents that are correctly classified as positive out of all the documents that are classified as positive.

Table 2. Classwise Accuracy

Classes	**Accuracy**
Classes N	0.915385
Classes L	0.924045
Classes R	0.914611
Classes A	0.91396
Classes P	0.909976
Classes V	0.902609

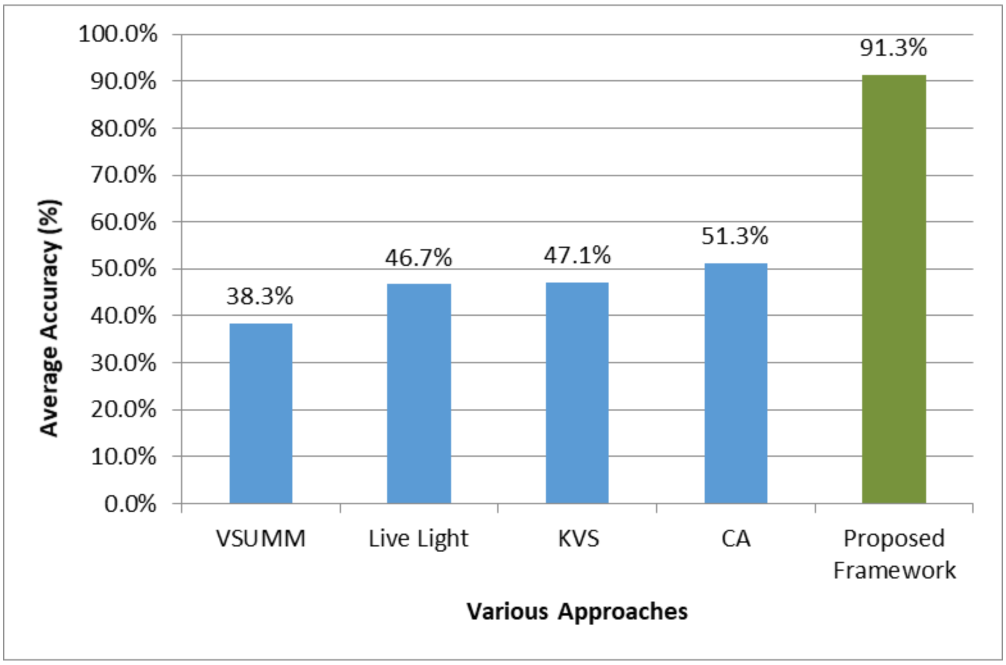

Figure 2. Average accuracy of Proposed Framework vs. Existing Framework

Fig 2 shows the accuracy of proposed system with the existing approaches like VSUMM, LiveLight, KVS and CA. It can be observed that the proposed framework achieves better performance than the other approaches.

V. CONCLUSION

The central focus was on either minimizing storage usage by compressing or removing redundant frames without loss of actual content. The goal of this work is to rescue the time of observer and moreover provide an effective summary of the original video. The proposed system which is summarizes the video by using CNN technique. In this work, the input video is extracted number of frames, then applying Convolutional Neural Network the extracted frames are experienced with the trained dataset for recognizing important frames. From the sorted frames, the system generates the resultant summarized video. The proposed work generated more accurate video summary. In future, it is also possible to add more activities like different sports, daily activities etc., also possible to use this system for smart surveillance system.

REFERENCES

[1] Xuelong Li, Fellow, IEEE, Bin Zhao, and Xiaoqiang Lu, Senior Member,IEEE , “A General Framework for Edited Video and Raw Video Summarization”, IEEE

[2] Sandra Eliza Fontes de Avila, Ana Paula Brando Lopes, Antonio da Luz Jr. “VSUMM : A mechanism design to produce static video summaries and novel evaluation method ”, ELSEVIER 2010.

[3] A. Rav-Acha, Y. Pritch, and S. Peleg “Making a long video short:Dynamic video synopsis”,IEEE Computer Society Conference CVPR2006.32

[4] Y. Hadi, F. Essannouni, and R. O. H. Thami “Video summarization by k-medoid cluster-ing ”, Research Gate 2006.

[5] M. Gygli, H. Grabner, H. Riemenschneider, and L. V. Gool “Creating summaries from user videos ”, Springer 2014.

[6] Ana Garcia del Molino “Summarization of Egocentric Videos: A Comprehensive Survey ”, IEEE transactions on Video Technology 2016.

[7] S. Zhang, Y. Zhu, and A. K. Roy-Chowdhury “Context-aware surveillance video summa- rization.”, IEEE transactions on Video Technology 2015

[8] Sinnu Susan Thomas, Sumana Gupta \Perceptual video summarization - a new framework of video summrization", IEEE transactions on image processing 2017

[9] Yifang Yin, Roshan Thapliya “Encoded Semantic Tree for AutomaticUser Proling Ap-plied To personalized Video Summarization ”, IEEEtransactions on Video Technology 2016

[10] Padmavathi Mundur,Yong Rao, Yelena Yesha “Keyframe-based video summarization using elaunay clustering ”, Springer 2006.

[11] Y. Zhang, G. Wang, B. Seo, and R. Zimmermann “Multi-video summary and skim generation of sensor-rich videos in geo-space”, ACM-2012

[12] Muhammad Ehsan Anjum, Syed Farooq Ali, Malik Tahir Hassan, Muhammad Adnan “Video Summarization Sports Highlights Generation”, IEEE conference 2014.3

[13] Z. Lu and K. Grauman “Story-driven summarization for egocentric video” IEEE Conference CVPR 2013

[14] Y. J. Lee and K. Grauman “Predicting important objects for egocentric video summarization”, International Journal of Computer Vision, 2015.

[15] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, “Category-specific video summarization ”, in European Conference on Computer Vision, 2014, pp. 540555.