iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** IV **Month of publication:** April 2019

DOI: https://doi.org/10.22214/ijraset.2019.4453

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com

Video Summarization using Convolutional Neural Network

Aishwarya Bhosale[1], Purva Badve[2], Radhika Gholap[3], Prajakta Joshi[4], Ms. S.P. Mone[5]
[1, 2, 3, 4, 5]*Marathwada Mitramandal College of Engineering.*

Abstract: ***Summarization techniques are used in every field like video, text, and audio. The video summarization techniques are mostly applied or used in surveillance data. The challenging task of surveillance video is to watch full video because it generates the data in huge amount. By apply summarization technique; it becomes more shorten then original and easy to watch.***
We propose a 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.
Keywords: Video Summarization, Video Skimming, Key frames.

I. INTRODUCTION

The development in video capturing devices and growing popularity of social media, there are huge volumes of videos being captured and uploaded every second. Video Shortening has been a field of active research for a long time. However, the main focus is on either minimizing storage usage by compressing or removing redundant frames without loss of actual content. Video Shortening has been a field of active research for a long time. However, the main focus was on either minimizing storage usage by compressing or removing redundant frames without loss of actual content.
Video summarization can be represented into two modes: A static video summary (story-board) and a dynamic video skimming. On one side, static video summary represents a video sequence in a static imagery form (one or more selected representative frames from the original video, or a synthesized image generated from the selected key frames).

A. Static Video Summarization

This is also called a key frame based video summarization techniques or still image abstract or storyboard. There are some criteria that come across for key frame based techniques, which are as follows:

1) *Redundancy:* frames with slight variations are selected as key frame
2) When there are different changes in content it is complicated to make clustering.

The following Figure shows selection of key frame.

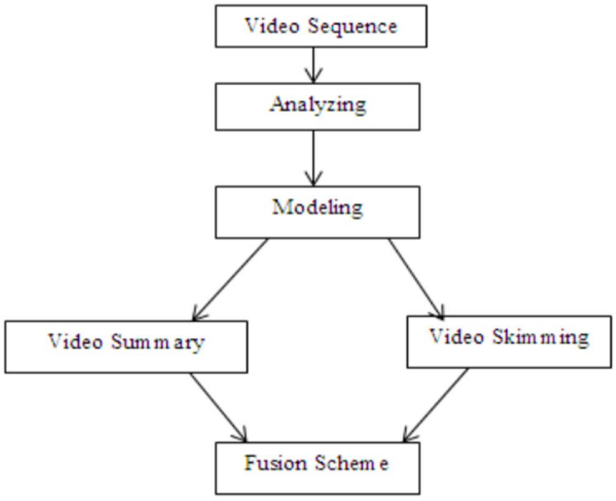


Fig. 1. Block Diagram for video Summarization.

The key frame based summarization can be classified in three different ways

a) Classification Based on Sampling: It selects key frame uniformly or randomly without considering the video content.

b) Classification Based on scene Segmentation: It extracts key frames using scene detection; it includes all semantic links in the video.

*c) Classification Based on shot Segmentation***:** It extracts first image and last image as a shot key frame.

B. Dynamic video Summarization

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. It also preserves the motion information.

Video Skimming as well as classification merge the Image and Language Understanding. 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.

The present work aims to address this ever increasing gap between the volumes of actual data generated and the volume that can be reasonably inspected manually. It is laborious and time consuming to scrutinize the salient events from the large video databases. We introduce smart surveillance by using video summarization for various applications. 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

For better understanding of the difference between the various approaches for the static video summarization and dynamic video summarization which can be found in the literature are discussed next.

Ana Garcia Del Molino et al [2] introduced the need for 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—storyboards, video skimming or fast-forwarding, as defined in Section. It consists of two steps: first is segmentation of the input data, and the second is a 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.

Sandra Eliza Fontes de Avila et al [3] proposed VSUMM, a methodology for the creation of static video summaries. It is a simple and effective approach for automatic video summarization. The methods are based on the extraction of color features from video frames and unsupervised classification and also added new methodology for evaluating the video summarized called as a comparison of user summarized i. e. CUS. In this method, the summaries are made by users and compared with other approached.

S. Zhang, Y. Zhu et al.[4] proposed the context-aware video summarization (CAVS) framework which is able to find 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. Sparsely gives the most informative features from the video.

Yifang Yin, Roshan Thapliya et al. [6] proposed method for automatic video summary generation with personal adaption. The 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 a semantic tree which is more effective for personalized adaption. In the proposed design of video summarization, it joins the personal interest and visual attention.

Muhammad Ehsan Anjum et al. [8] introduced the mechanism to recognize highlights from videos are a basic and fundamental problem for indexing and retrieval applications. The different techniques to generate sports highlights from cricket video using techniques of optical character recognition. First, the score bar is fetched from the frames then the character recognition techniques are used to extract information for events like sixes, fours, and wickets. A short video summary is incorporating that includes the frames for the aforementioned significant events termed as Highlights. The process of sports highlights generation is automated resulting in a condensed summary for the viewer that reduces the time and space requirements.

Z. Lu and K. Grauman et al. [10] proposed a video summarization approach that discovers the story of an egocentric video. This work for the long video which selects a short chain of video sub-shots depicting the essential events. Author adapts a text analysis

technique that connects a new article to the visual domain. They also show that how to establish the influence of one visual event on another given their respective objects. The author introduces a novel temporal segmentation method design for egocentric video. They perform a large scale video on the proposed approach which has a better sense of story.

Y. J. Lee and K. Grauman et al. [12] proposed a video summarization approach for egocentric or wearable camera data. The proposed method produces a compact storyboard summary of the cameras wears day. The proposed method works as first, train a regression from labeled training videos that scores any region to an important person or objects. The author proposes two ways to adjust the compacted of the summary based on either target importance criteria of target summary length the main contribution of this approach is driven by predicted important people and objects.

III. SYSTEM ARCHITECTURE

There are very long duration videos which are of and viewers of that video dint have sufficient time to go through the whole video. The viewer just wants to know an overview of that video. It is a daily need of the day to save time and grasp just summary of the video which is in text format. This problem can be solved using Intelligent Summarization of Videos which will be useful to save and then anyone can have like notes of that video.

A. Developing a tool to monitor and analyze Video Summarization.
B. To increase gap between the volumes of actual data generated and the volume that can be reasonably inspected manually.
C. To abstract the video that reduces the time taken to send and watch the video.
D. To scrutinize the salient events from the large video databases.

IV. DETAILWORKING

A. Video Extraction

Video is a combination of images in sequence. FFMPEG is used to extract frames from any video format. Video frame extraction is done at 20 frames per second. Each frame is saved along with the frame number for later summary generation.

B. Preprocessing

Preprocessing involves converting the frame into gray scale image of 128x128. Once the image is resized, it will be sent to CNN for matching.

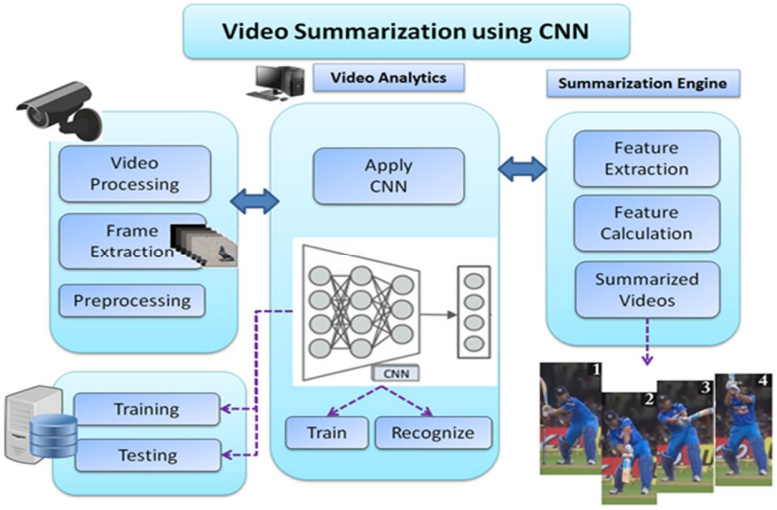


Figure: - System Architecture of Sensor Efficient Dynamic Street Light System.

C. Train Dataset

The proposed summarization framework is verified on 3 datasets, i.e., TVsum50, Sum Me, ADL, for edited videos, short raw videos and long raw videos, respectively. Based on these datasets, the proposed framework is compared with various popular approaches about edited video and raw video summarization.

Following, three types of datasets are used;

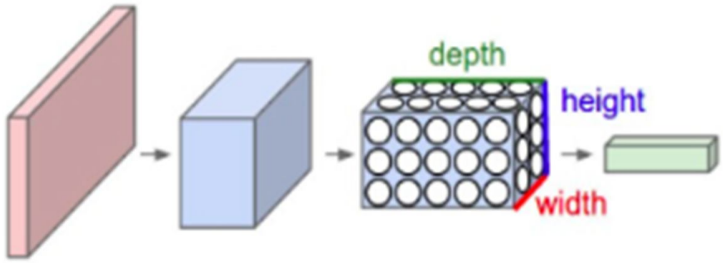
1) TVsum50 Dataset
2) Sum Me Dataset
3) ADL Dataset

D. CNN Training on Predefined Highlights Dataset

A CNN contain different layers first is input and second is output layer, and finally multiple hidden layers. The hidden layers of a CNN typically consist of different layers like; Convolutional layers, pooling layers, fully connected layers and normalization layers. CNN will be used to train the video analytics engine for recognizing important frames in the video.

E. Feature Extraction

Feature is extracted with the help of CNN and a network is trained with 3 layer CNN network.



F. Summarized Video

Frame number wise important frames will be stored. Each frame no will be mapped to that duration of video. An additional 5 second before and after duration will be added to catch the exact highlight. Generated Frame numbers are clubbed to form a 20fps video.

V. ALGORITHMS USED

A. Convolutional Neural Network (CNN)

Traditional feature learning methods rely on semantic labels of images as supervision. They usually assume that the tags are evenly exclusive and thus do not pointing out towards the complication of labels. The learned features endow explicit semantic relations with words.

We also develop a novel cross-modal feature that can both represent visual and textual contents. CNN itself is a technique of classifying images as a part of deep learning. In which we apply single neural network to the full image.

1) Accepts a volume of sizeW1×H1×D1
2) Requires four hyper parameters:
a) Number of filters K
b) Their spatial extent F
c) The stride S
d) The amount of zero padding P
3) Produces a volume of size W2×H2×D2 where:
a) W2=(W1−F+2P)/S+1
b) H2=(H1−F+2P)/S+1(i.e. width and height are computed equally by symmetry)
c) D2=K With parameter sharing, it introduces F*F*D1 weights per filter, for a total of (F*F*D1)*K weights and K biases.
4) In the output volume, the d-th depth slice (of size W2×H2) is the result of performing a valid convolution of the d-th filter over the input volume with a stride of S, and then offset by d-th bias.
5) A common setting of the hyper parameters is F=3, S=1, P=1 However, there are common conventions and rules of thumb that motivate these hyper parameters.

VI. DATA SET USED

The proposed summarization framework is verified on 3 data sets, i.e., TVsum50, Sum Me, ADL, for edited videos, short raw videos and long raw videos, respectively. Based on these datasets, the proposed framework is compared with various popular approaches about edited video and raw video summarization.

A. TVsum50 Dataset

The experiments about edited video summarization are carried on the TVsum50 dataset. This video data set contains 50 edited videos, including news, documentary, etc. Their duration vary form 2-10 minutes. Each video is uniformly segmented into 2-second shots, and each shot is annotated with a user score, which is taken as the reference to generate the ground truth of video summary, i.e., high-score shots are selected as key-shots. Similar to existing approaches, the length of the summary is constrained to less than 15% of the video duration.

B. Sum Me Dataset

The Sum Me dataset is employed to verify the efficiency of our framework on raw video summarization. There are totally 25 videos, including holidays, events and sports. They are all raw videos, and the duration varies from 1 to 6 minutes. Each video is segmented into shots by the segmentation approach. Moreover, each video has about 15-18 human-made summaries as the ground truth. Their lengths vary from 5% to 15% of the video duration.

C. ADL Dataset

The ADL data set contains 20 videos; most of them are about 30min long. Each of them is generated with a chest-mount Go Pro camera, and records the wearers' daily life. During the shooting, the wearer is asked to do some daily activities, like combing hair, brushing teeth, etc. Generally, there are totally 32 kinds of actions in the dataset, and every video contains about 18 kinds of actions, which are temporally annotated in the video. Simply, we segment the video into shots for the length of every 3 seconds, since it is enough for people to recognize the activities.

VII. EXPERIMENTAL RESULTS

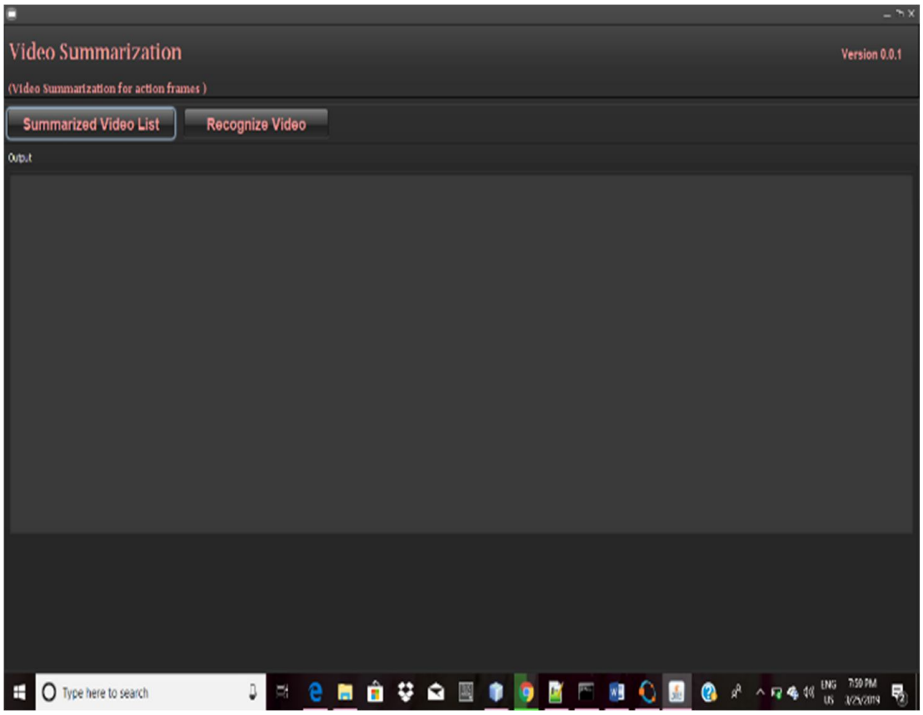


Figure: - GUI of Project

The project GUI contains two components that mainly include Summarized Videos and Recognized Videos respectively.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com

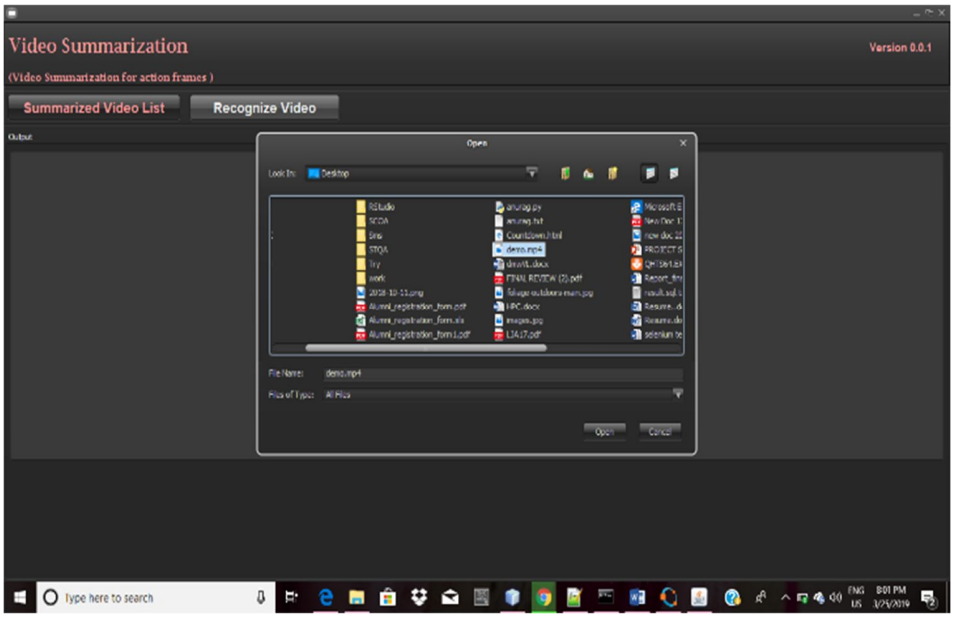


Figure: - Summarized video list

The above figure contains the list of videos that are in Summarized manner.

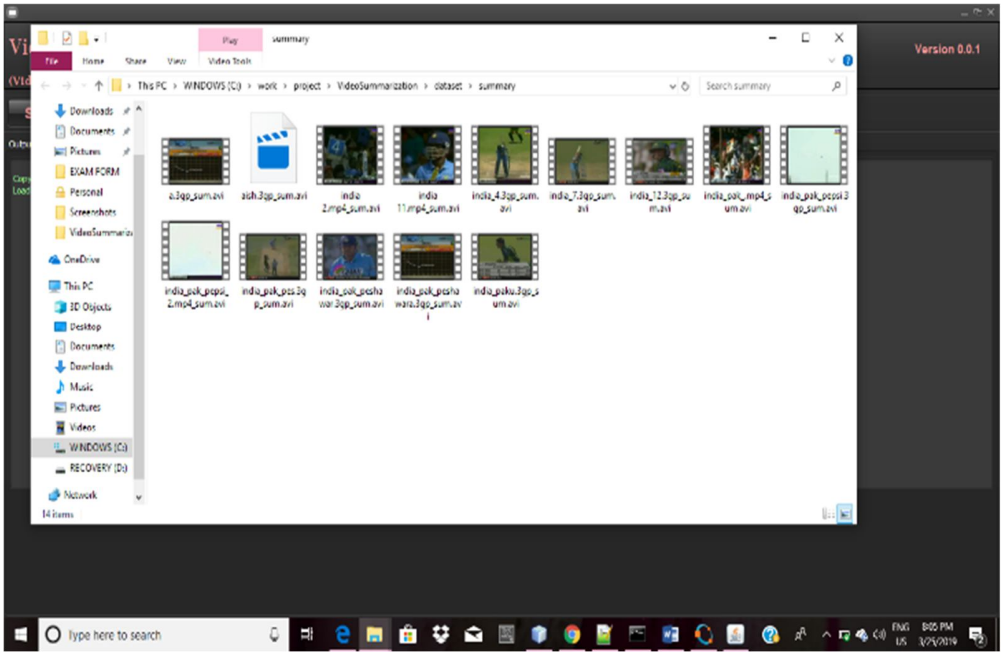


Figure: - Collection of Videos

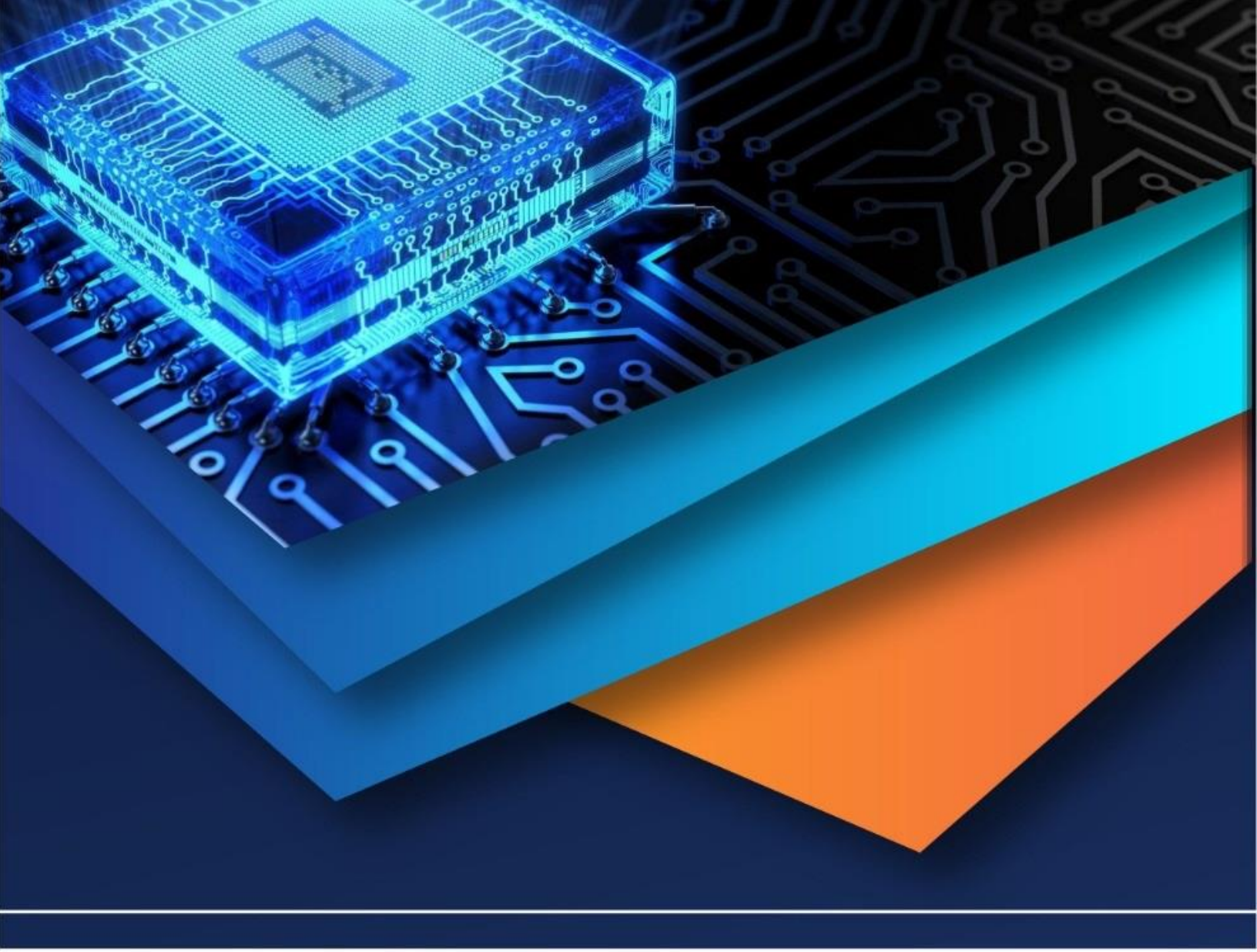
VIII. CONCLUSION

Summarization is a system that produces a condensed representation of its inputs for user consumption. With the explosion of abundant data present on social media, it has become essential to examine the text for searching information and use it as an advantage of various application and people. The present work aims to address this ever-increasing gap between the volumes of actual data generated and the volume that can be reasonably inspected manually.

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

REFERENCES

[1] J. Tang, K. Wang, and L. Shao, “Supervised matrix factorization hashing for cross-modal retrieval,” IEEE Trans. Image Processing, vol. 25, no. 7, pp. 3157–3166, 2016.
[2] Ana Garcia del Molino “Summarization of Egocentric Videos: A Comprehensive Survey", IEEE transactions on Video Technology 2016.
[3] 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.
[4] S. Zhang, Y. Zhu, and A. K. Roy-Chowdhury “Context-aware surveillance video summa- rization.", IEEE transactions on Video Technology 2015
[5] J. Han, K. Li, L. Shao, X. Hu, S. He, L. Guo, J. Han, and T. Liu, “Video abstraction based on fmri-driven visual attention model,” Inf. Sci., vol. 281, pp. 781–796, 2014.
[6] Yifang Yin, Roshan Thapliya “Encoded Semantic Tree for Automatic User Proling Ap-plied To personalized Video Summarization", IEEE transactions on Video Technology 2016
[7] J. Qin, L. Liu, Z. Zhang, Y. Wang, and L. Shao, “Compressive sequential learning for action similarity labeling,” IEEE Trans. Image Processing, vol. 25, no. 2, pp. 756–769, 2016.
[8] Muhammad Ehsan Anjum, Syed Farooq Ali, Malik Tahir Hassan, Muhammad Adnan “Video Summarization Sports Highlights Generation", IEEE conference 2014.3
[9] J. Shen, D. Tao, and X. Li, “Modality mixture projections for semantic video event detection,” IEEE Trans. Circuits Syst. Video Techn., vol. 18, no. 11, pp. 1587–1596, 2008.
[10] Z. Lu and K. Grauman “Story-driven summarization for egocentric video", IEEE Conference CVPR 2013
[11] X. Zhen, L. Shao, D. Tao, and X. Li, “Embedding motion and structure features for action recognition,” IEEE Trans. Circuits Syst. Video Techn., vol. 23, no. 7, pp. 1182–1190, 2013.
[12] Y. J. Lee and K. Grauman “Predicting important objects for egocentric video summarization", International Journal of Computer Vision, 2015.
[13] A. Rav-Acha, Y. Pritch, and S. Peleg “Making a long video short: Dynamic video synopsis",IEEE Computer Society Conference CVPR 2006.32
[14] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, “Category-specific video summarization,” in European Conference on Computer Vision, 2014, pp. 540–555.
[15] Y. Hadi, F. Essannouni, and R. O. H. Thami “Video summarization by k-medoid cluster-ing",Research Gate 2006.
[16] M. Gygli, H. Grabner, H. Riemenschneider, and L. V. Gool “Creating summaries from user videos", Springer 2014.
[17] Padmavathi Mundur,Yong Rao, Yelena Yesha “Keyframe-based video summarization using elaunay clustering", Springer 2006.
[18] C. Ngo, Y. Ma, and H. Zhang, “Automatic video summarization by graph modeling,” in IEEE International Conference on Computer Vision, 2003, pp. 104–109.
[19] D. Tao, L. Jin, Y. Wang, and X. Li, “Rank preserving discriminant analysis for human behavior recognition on wireless sensor networks,” IEEE Trans. Industrial Informatics, vol. 10, no. 1, pp. 813–823, 2014.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY