**How to Cite**:
Harakannanavar, S. S., Kanabur, V. R., & Puranikmath, V. I. (2022). Machine learning based fusion algorithm to perform multimodal summarization. *International Journal of Health Sciences*, *6*(S3), 671–683. https://doi.org/10.53730/ijhs.v6nS3.5411

# Machine Learning Based Fusion Algorithm to Perform Multimodal Summarization

**Sunil S. Harakannanavar**
Department of Electronics and Communication Engineering, Nitte Meenakshi Institute of Technology, Yelahanka, Bangalore-560064, Karnataka, India
Email: sunilsh143@gmail.com

**Vidyashree R. Kanabur**
Department of Electronics and Communication Engineering, National Institute of Technology, Surathkal, Mangalore, Karnataka, India
Email: vidyashreerk1992@gmail.com

**Veena I. Puranikmath**
Department of Electronics and Communication Engineering, S. G. Balekundri Institute of Technology, Shivabasav Nagar, Belagavi-590010, Karnataka, India
Email: veenaip043@gmail.com

***Abstract*---**Video summarization is a rapidly growing research field which finds its application in various commercial and personal interests due to the massive surge in the amount of video data available in the modern world. The proposed approach uses ResNet-18 for feature extraction and with the help of temporal interest proposals generated for the video sequences, generates a video summary. The ResNet-18 is a convolutional neural network with eighteen layers. The existing methods don't address the problem of the summary being temporally consistent. The proposed work aims to create a temporally consistent summary. The classification and regression module are implemented to get fixed length inputs of the combined features. After this, the non-maximum suppression algorithm is applied to reduce the redundancy and remove the video segments having poor quality and low confidence-scores. Video summaries are generated using the kernel temporal segmentation (KTS) algorithm which converts a given video segment into video shots. The two standard datasets TVSum and SumMe are used to evaluate the proposed model. It is seen that the F-score obtained on TVSum and SumMe datasets are 56.13 and 45.06 respectively.

***Keywords*---**convolutional neural networks, kernel temporal segmentation, non maximum suppression, redundancy, ResNet-18.

International Journal of Health Sciences ISSN 2550-6978 E-ISSN 2550-696X © 2022.
***Corresponding author**: Harakannanavar, S. S.; Email: sunilsh143@gmail.com*
*Manuscript submitted: 18 Nov 2021, Manuscript revised: 09 Feb 2022, Accepted for publication: 27 March 2022*

## Introduction

In this modern world, with the rise in technology leading to tremendous growth in smart devices that are capable of recording videos with powerful sensors and the ability to share the contents on online social platforms like YouTube, Daily Motion, Instagram, Facebook, Twitter has caused a massive surge in the amount of data available which requires a technology that can facilitate the users to browse the constantly increasing collection of video data [1]. There are several applications where human intervention is required to analyze the large scale of a video data frame by frame which involves a great load of human effort. Video Summarization concept deals with the extraction of effective information from sets of video data from a large group of researchers. Video summarization aims to provide an automated way of generating short and informative versions of the original video by identifying the most important and relevant content [2]. The frames containing important information are selected as key frames from a given video sequence [3] to form the summary is shown in Figure 1.

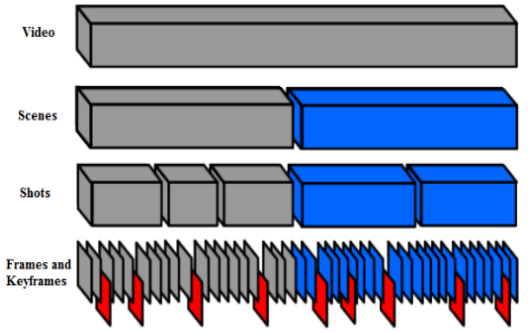


Figure 1. Scenes, shots, and Key Frames

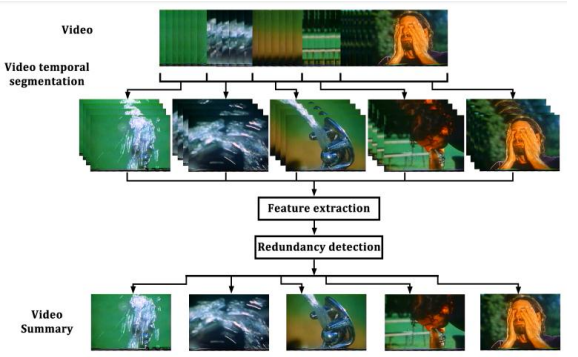


Figure 2. General approach for Video Summarization

The summary formed is a short version of the original video containing the frames which are having relevant important information [5]. The general approach for Video Summarization is shown in Figure 2. The paper is organized in various sections. Section 2 describes the detailed description of Video Summarization models. The proposed model is discussed in section 3. The results of proposed model on two popular datasets are analyzed in section 4. The proposed model is concluded in Section 5.

**State of the art methods**

In this section, the work related to existing methodologies for video summarization is discussed. Jiaxin Wuet al., [1] implemented video summarization based on understanding of the high-density peaks search approach. Pre-sampling of input video is done, where useless/redundant frames are removed to decrease complexity and intern saves the time and frames are selected. Then BoG model is considered to represent each frame. Next, they selected frames with fewer like generate video summary from the cluster. Shu Zhang et al., [2] developed the model to produce video summaries with most important video portions. Here the sparse coding model is used to gain an understanding of the feature dictionary. Using the adaptive background algorithms, the motion regions were detected for a set of videos. The developed model is evaluated on UCLA office dataset, VIRAT, SumMe and TVSum50 datasets. Ioannis et al., [3] proposed a multistage, multimodal summarization process for summarizing stereoscopic movies. In this case, each video frame is referred as feature vector based on luminance and disparity information to automatically extract the key frames per shot. These key-segments are combined to form the video skim. Marcos et al., [4] discussed a video summarization technique using VISCOM. Here, first frame sampling is done, then Co-occurring Matrix (CCM) descriptors are used to depict the video frames which represent the distribution of image intensity values. The entropy of all frames and if entropy is zero it is discarded. The model was evaluated on SBD, TRECVID dataset and Vidseg dataset.

Fairouz Hussein et al., [5] proposed the model based on latent structural SVM and graphical models. The model uses submodular functions for selecting key-frames. It is noted that, larger the generated summary and results in less the benefit of adding in a new element. Latent Structural SVM are used for the unsupervised and semi-supervised learning. The model is evaluated on ACE dataset and MSR DailyActivity3D dataset Xuelong Li [6] developed summarization framework for supervised video summarization of raw and edited videosThe optimal subset of video shot is found using the Accelerated Greedy algorithm. The model was evaluated their framework on TVsum50, Sum-Me and ADL dataset. Sinnu Thomas et al., [7] developed a new framework for perceptual video summarization. Firstly, the input video is down sampled to 5-10 frames/s and standard deviation is used to filter out the frames containing no information. The summarized video is obtained using the HVS system. The perceptual video summarization mainly focuses on combining Itti's top-down and bottom-up approach. Tongling et al., [8] generated summarization approach using global and local importance techniques based on multiple features. Initially, the video is split into multiple smaller video clips. Then, frames and centre parts of frames are extracted from each video clip. In the next step, for each frame and the centre part, the global and local importance scores are calculated. Finally, both the importance scores are combined to select optimal subset for generating a video summary.

Guilherme et al., [9] implemented video summarization using OPFSumm which is OPF with temporal and spatial enforcement to create video summaries, the step first is to sample frames from input video. Then they extract features using ACC,

CCV, GCH, and Border/interior pixel Classification (BIC). The GFD and Haar-HWD are used to test the spectral properties, the meaningless one colour frames are removed. Hana et al., [10] developed video summarization techniques using KEGC. It mainly focuses on local interest points information and graph clusters by approaching modularity values. The standard YUV dataset was used for experimentation. Yuan et al., [11] developed Crum model, where it follows the feature extraction, temporal modelling, and summary generated in the whole model. The CRNN is combined of RNN and CNN together for extraction of keyframe and LSTM is also used for feature extraction. Sinnu et al., [12] addressed optimization framework and then the frames which are identified to have activities in them are combined using an alpha matting algorithm to form a single frame. The idea behind this approach is to retain the context of the original video. Cheng et al., [13] explained the CapsNet technique for video summarization. It focuses on spatiotemporal features and generation of inter-frames motion curves. Self-attention mechanism is utilized to select key frames and the algorithm was tested on VSUMM, TvSumm, SumMe and RAI datasets. Zhong Ji et al., [14] developed a video summarization framework named attentive encoder–decoder networks for video summarization (AVS), the encoder uses a bidirectional long short-term memory (BiLSTM). Jiaxin Wu et al., [15] explained dynamic graph CNN that measures the importance and relevance of each video shot in the video itself as well as in the complete video collection. Each video shot is treated as a graph node and the pairwise relations are shown as graph edges. A dynamic graph convolutional network (mvsDGCN) is designed and used to generate the summaries of multiple videos. Bin Zhao et al., [16] used the recurrent neural network (RNN), in this approach they use a Dual learning framework for video summarization. Their first approach is to integrate the summary and the second is to Reconstruction of video and they give the name property-Constraint dual learning (PCDL).

Wencheng Zhu et al., [17] have developed a Detect-to-Summarize network (DSNet) framework which can be used for supervised video summarization. The anchor-based method produces temporal interest proposals, which solve the problem of the video segments being of variable lengths. Frame-level features are extracted, and a temporal modeling layer is used to capture long-range features. Madhushree et al., [18] extracted deep visual features by transfer learning method, a set of feature maps is used for k-means clustering using Euclidean distance and every frame is assigned a cluster label. The frame which is present at the class border is given the right to be a key frame and is selected as a part of the summary. Zhong Ji et al., [19] used SumVClip using supervised video summarization. The school given to frame is based on extended visual features. This feature is extracted by GoogleNet and BiLSTM encoder. Evlampios et al., [20] developed AC-SUM-GAN approach where the Actor-Critic model is embedded into GAN to find the optimal policy for selecting the key frames and forming the summary. Kasbgar et al., [21] presented spatio-temporal method which provides the linkages towards wavelet sub-bands for exaggerating the small motions of images and video frames. The wavelet transformed frames and Laplacian Pyramid are used to determine the pixels for images having motions. Otroshi Shahreza et al., [22] developed NR-VQA) technique that can predict only mean opinion score. In this model frame-level features are extracted and feed to recurrent neural network to predict opinion score in the last layer of the RNN.

Salih et al., [23] discussed about the scene change detection techniques for the videos for detection of useful parameters such as cut, dissolve, wipe, etc. The AFD, MAFD, MGV techniques are tested on video datasets having low and high object motion scenes to evaluate the performance of the model. Mehrgan et al., [24] recognized license plate from the video frames using a weighted interpolation method. A coarse registration is done using the features generated SURF features and then recognition of the license plate is performed using the phase correlation technique. Firouzian et al., [25] extracted the facial features from the videos using LBPs approachThe Lucas-Kanade algorithm is adopted to track the primary and secondary components of face such as eyes, nose, mouth, eyebrows etc. The matching scores of test and train samples is carried out using Chi-square similarity measure. Most of the people get addicted to videos through social networks. In such cases, video may not contain significant information and leads the user to download the video. Because of cost and bandwidth issues, the video requires a large bandwidth to download or view it. Thus, it is required to enable users to watch videos which helps to reduce costs. Video summarization approaches are the proposed solution to address the above issues.

**Proposed Methodology**

In this section, the steps in video Summarization and the architecture of the proposed model are shown in Figure 3 and Figure 4 respectively. The proposed model includes Resnet 18 feature extraction technique for features extraction. The Non-Maximum Suppression algorithm and Kernel Temporal Segmentation algorithm are used for Redundancy removal and generating the video summary. The model is tested on two popular databases such as SUMme and TVSum to evaluate the model.

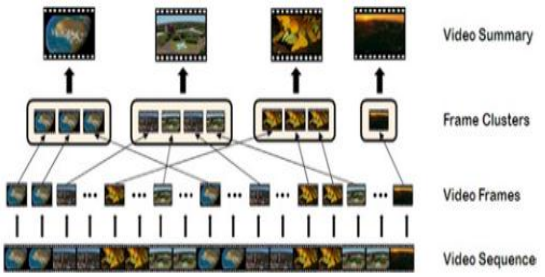


Figure 3. Steps in Video Summarization

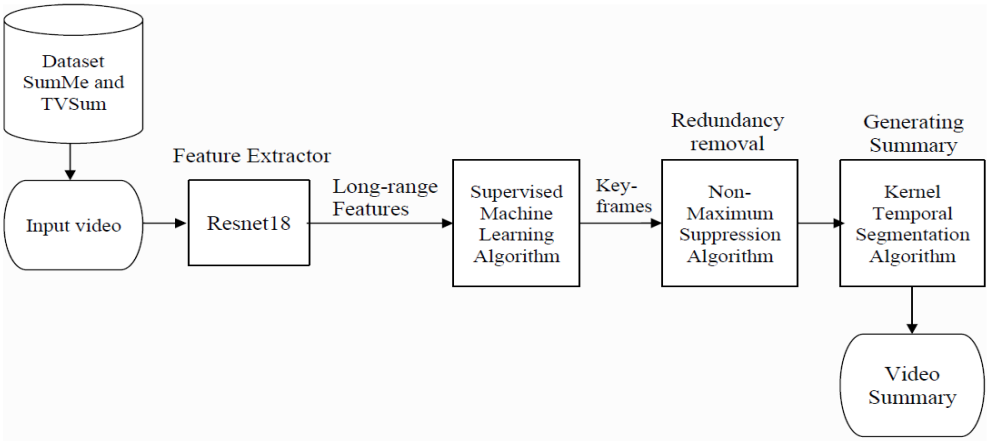


Figure 4. Proposed Methodology

**Dataset**

In this paper, SUMme and TVSum datasets, which are widely accepted benchmark datasets used in the evaluation of Video Summarization model. These datasets contain user generated videos and videos from popular video sharing websites like Youtube. The sample of SUMme and TVSum datasets are shown in Figure 5 and Figure 6.

Figure 5: Sample of TVSum dataset

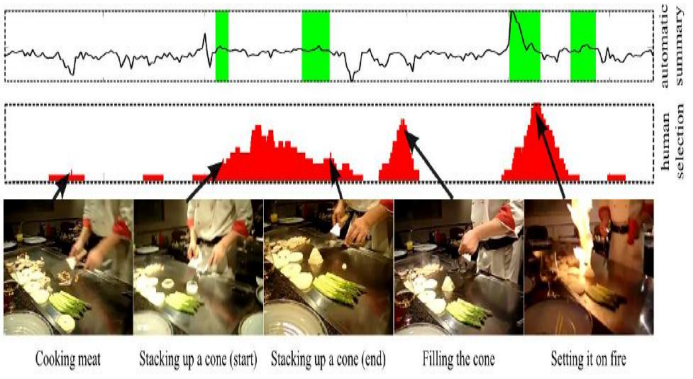


Figure 6: Sample of SumMe dataset

**Feature extraction**

Extracting features is an important part for understanding a video to identify the most representative video frames. In the proposed approach, ResNet-18 is used for feature extraction [26]. It is an 18-layer deep Residual Neural Network which is used to extract image features from each frame of the input video. ResNet-18 uses ReLU as its activation function which outputs the original input if it's positive otherwise it outputs zero. Table 1 shows the various layers of CNN present in the ResNet-18 architecture. There are 5 versions of the ResNet CNN, each having 18, 34, 50, 101 and 152 layers respectively. A million images can be trained on the ResNet convolutional neural network [27]. It is very efficient and useful in image classification. It can sort the images into thousand different object-categories by identifying them. Mathematically ReLU is defined as,

$$Y(z) = \max \{0, z\} \qquad (1)$$

Table 1
Various Layers of convolutional neural network

| Layer Name | Output Size | ResNet-18 |
|---|---|---|
| conv1 | 112 × 112 | 7 × 7, 64, stride 2 |
| conv2_x | 56 × 56 | 3 × 3 max pool, stride 2 $\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix} \times 2$ |
| conv3_x | 28 × 28 | $\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix} \times 2$ |
| conv4_x | 14 × 14 | $\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix} \times 2$ |
| conv5_x | 7 × 7 | $\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix} \times 2$ |
| average pool | 1 × 1 | 7 × 7 average pool |
| fully connected | 1000 | 512 × 1000 fully connections |

In the proposed model, ResNet-18 is pretrained on ImageNet dataset which has 1024 dimensional features of pool5 layer and normalization with a fully connected layer with dropout layer 0.5 and a ReLu fully connected activation function layer. The threshold of NMS value was made 0.5. The model was trained over 300 epochs using Adam optimizer with a weight decay of 10-5 and 5 x 105 of base learning rate. The model ResNet18 is implemented using PyTorch Library [28-29].

**Supervised machine learning**

To train the proposed model, a supervised Machine Learning algorithm is used. Here the training of the model is done using labeled data from the standard datasets (SumMe & TVSum). In this approach, temporal interest proposals are used to eliminate the problem of temporal inconsistency. Here a module consisting of classification and regression algorithms is employed followed by key shot detection techniques. In the proposed model, the ResNet-18 is employed to

extract the feature vectors of a particular video sequence. To extract the long-range representation, the self-attention mechanism [30-31] is used. In addition to this, LSTM and Bi-LSTM algorithms are adopted for understanding the effects of other sequences of long-range features. A neural network is used in the model to get fixed-length inputs required by the fully connected layers. The neural network consists of classification and regression modules. Tanh, dropout, and layer normalization along with two output branches form the module [32].

**Redundancy removal**

The summarized video may contain redundant frames which affect the final summary. The redundant frames increase the length of the generated video summary unnecessarily and hence make it a less efficient summary. To remove the redundant frames, the NMS algorithm is used, it is a technique which suppresses those candidate proposals which have a lower confidence score and selects only the frames with high-confidence scores. This reduces frame overlap and hence generates a more efficient summary.

**Generating video summary**

For generating the video summaries from the selected keyframes, segment them into shots and assign them their shot-level importance scores. Next, the kernel temporal segmentation (KTS) algorithm [33] is used, which is acting as shot detection method to segment the sequences into video shots and hence generate the final video summary.

**Training Process of proposed supervised Machine Learning technique**
The pseudocode outlines the training process of the proposed work. First, to extract frame-level features ResNet18 is used, which takes video sequence (v) as an input to the proposed supervised training method. Self-attention mechanism is used to effectively capture the long-range representations (wj) Summation is performed to combine the actual spatial feature vj along with the long-range representation wj.

**Input**: (videos $\{\boldsymbol{Vi}\}$, frame level annotations $\{\boldsymbol{u^*_i}\}$)
**Output:** Importance scores $\boldsymbol{p}$

**for** *epoch* ∈ {1,2, 3...., *N*} **do**
  **for** *video* $\boldsymbol{V} \in \{\boldsymbol{V_i}\}$ *and annotations* $\boldsymbol{u^*_i} \in \{\boldsymbol{u^*_i}\}$ **do**
    **//** *Apply KTS and knapsack*
    **for** *frame* $\boldsymbol{I_j} \in \boldsymbol{V}$ **do**
$\boldsymbol{v_j} \leftarrow$ ResNet18(**Ij**)
      **for** *k* ∈ {1, 2..., K} **do**
        $\mathbf{p_{jk}} \leftarrow$ AssignLabel($\mathbf{u^*}$)
      **end**
    **end**
    $\boldsymbol{v} \leftarrow$ Feature Vector
    $\boldsymbol{w} \leftarrow$ Temporal Feature
    $\boldsymbol{x} \leftarrow \boldsymbol{w} + \boldsymbol{v}$
    $\boldsymbol{x} \leftarrow$ Temporal Pooling

$\boldsymbol{p} \leftarrow$ Importance Scores
**end**
**end**

The importance scores pi for each frame of the video is generated using a Classification-Regression module which uses the sigmoid activation function. The above-mentioned method runs for N epochs for the videos in the input training data splits.

**Evaluation of video summaries**

The parameters such as TP, TN, FP, and FN are used for calculation of success rates and is given in equation 2, 3 and 4 respectively.

True Positives (TP): When the prediction is Yes, and the real output is also Yes.

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

True Negatives (TN): When the prediction is No, and the real output is also No.

$$TN = \frac{\text{TN}}{\text{TN+FP}} \quad (3)$$

False Positives (FP): When the prediction is Yes but it is No.

$$FP = \frac{\text{FP}}{\text{TN+FP}} \quad (4)$$

False Negatives (FN): When the prediction is No but it is Yes.

$$FN = \frac{\text{FN}}{\text{TN+FN}} \quad (5)$$

**Experimental results**

The proposed model is trained and tested on a machine which runs on AMD Ryzen 5600H Hexa-Core processor, with an 8 GB DDR4 RAM and NVIDIA GeForce GTX 1650. For evaluating the performance of the proposed model, two standard datasets such as SumMe and TVSum [35-36] are used. To remove redundant frames and reduce the video segments having low quality, 'non maximum suppression' (NMS) algorithm is used. To find out the best threshold for NMS an investigation is done by using various thresholds for NMS. It is observed that for a threshold of 0.4 the model gives the best results. So, the threshold for NMS has been set to 0.4.

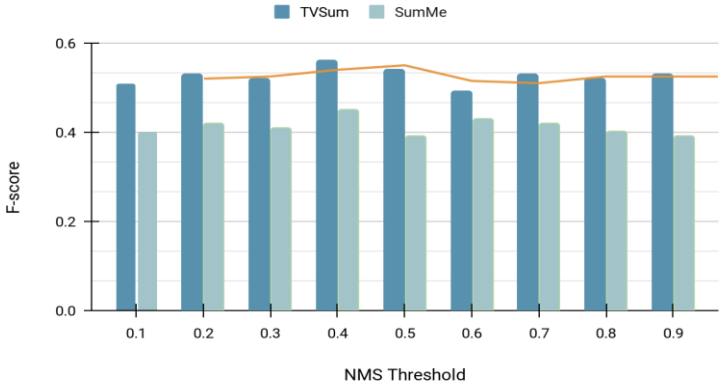


Figure 7. Effect of the NMS threshold on the F-score

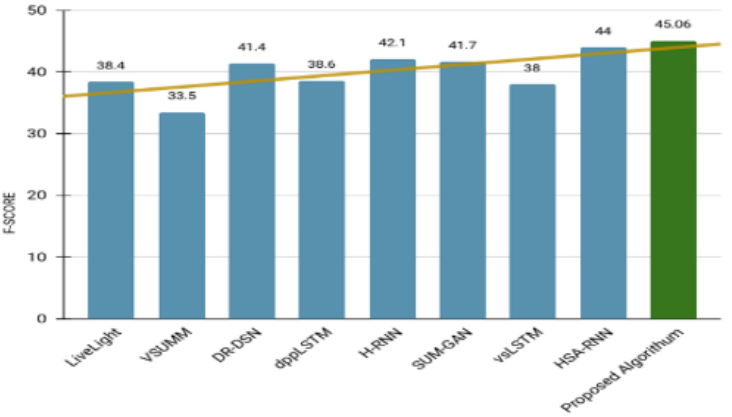


*Figure 8. Comparison of proposed model*

In Table 2, eight leading video summarization methods have been compared with the proposed method. Out of these eight methods DR-DSN and Live Light are based on unsupervised techniques whereas SUM-GAN, dppLSTM, VSUMM, vsLSTM, HSA-RNN and H-RNN along with the proposed method are based on supervised techniques. In this model, ResNet-18 is used, which is the latest state-of-the-art feature extraction architecture [37]. Compared to the feature extraction methods used in other approaches, ResNet-18 provides better results. The obtained F-score on the SumMe dataset is 45.06 and the F-score obtained on the TVSum dataset is 56.13. The effect of NMS threshold on F-score and Comparison of performance of several video summarization methods and Proposed model is shown in Figure 7 and Figure 8 respectively.

Table 2
Comparison of Proposed model with existing Video Summarization models

| Authors | Existing Models | SumMe (%) | TVSum (%) |
|---|---|---|---|
| S. Zhang et al.,[2] | vsLSTM | 38.0 | 54.0 |
| B. Zhao et al., [16] | VSUMM | 33.5 | 39.1 |

| | | | |
|---|---|---|---|
| Avila et al., [17] | LiveLight | 38.4 | 47.7 |
| K. Zhou et al., [18] | DR-DSN | 41.4 | 56.6 |
| K. Zhang et al., [19] | dppLSTM | 38.6 | 54.7 |
| B. Zhao et al., [20] | H-RNN | 42.1 | 54.9 |
| Mahasseni et al., [21] | SUM-GAN | 41.7 | 56.0 |
| B. Zhao et al., [22] | PCDLsup | 43.7 | 56.2 |
| A. Sahu et al., [23] | MST_C | 38.3 | 54.6 |
| Proposed | Model (ResNet18+NMS+KTS) | 45.06 | 56.13 |

**Conclusion**

In this paper, the proposed approach uses ResNet18 and temporal interest proposals. The feature extraction method used here i.e., ResNet-18, outperforms most of the similar feature extraction methods used in other works. The proposed method also maintains the temporal consistency of the video summary, which makes the resulting summaries being more representative of the original video and hence producing a better F-score. It is observed that the F-score obtained on the TVSum dataset is 56.13 and on the SumMe dataset is 45.06. In future, the plan is to find a way to combine all these separate steps into a single framework and test it on more video datasets.

**References**


1. Jiaxin Wu, Sheng-hua Zhong, Jianmin Jiang and Yunyun Yang, “A novel clustering method for static video summarization”, Springer, Multimedia Tools and Applications, vol. 76, no. 7, pp. 9625-9641, 2016.
2. S. Zhang, Y. Zhu and A. K. Roy-Chowdhury, “Context-Aware Surveillance Video Summarization,” IEEE Transactions on Image Processing, vol. 25, no. 11, pp. 5469-5478, 2016.
3. I. Mademlis, A. Tefas, N. Nikolaidis and I. Pitas, “Multimodal Stereoscopic Movie Summarization Conforming to Narrative Characteristics,” IEEE Transactions on Image Processing, vol. 25, no. 12, pp. 5828-5840, 2016.
4. Mussel Cirne, M.V and Pedrini H, “VISCOM: A robust video summarization approach using color co-occurrence matrices”, Springer, Multimedia Tools and Applications, vol. 77, no. 1, pp. 857–875, 2017.
5. Fairouz Hussein and Massimo Piccardi, “V-JAUNE: A Framework for Joint Action Recognition and Video Summarization”, ACM Transactions in Multimedia Computer Communication Applications, vol. 13, no. 2, pp. 1-19, 2017.
6. X. Li, B. Zhao and X. Lu, “A General Framework for Edited Video and Raw Video Summarization,” IEEE Transactions on Image Processing, vol. 26, no. 8, pp. 3652-3664, 2017.
7. S. S. Thomas, S. Gupta and V. K. Subramanian, “Perceptual Video Summarization—A New Framework for Video Summarization,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 27, no. 8, pp. 1790-1802, 2017.

8. Hu T and Li, “Video summarization via exploring the global and local Importance”, Springer, Multimedia Tools and Applications, vol. 77, no. 1, pp. 22083–22098, 2018.
9. Martins Pereira and Almeida, “OPFSumm: on the video summarization using Optimum-Path Forest”, Springer, Multimedia Tools and Applications, vol. 79, no. 1, pp. 11195–11211, 2019.
10. Gharbi, Bahroun and Zagrouba, “E-Key frame extraction for video summarization using local description and repeatability graph clustering”, Signal, Image and Video Processing, vol. 13, pp. 507–515, 2019.
11. Y. Yuan, H. Li and Q. Wang, “Spatiotemporal Modeling for Video Summarization Using Convolutional Recurrent Neural Network,” IEEE Access, vol. 7, pp. 64676-64685, 2019.
12. S. S. Thomas, S. Gupta and V. K. Subramanian, “Context Driven Optimized Perceptual Video Summarization and Retrieval,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no. 10, pp. 3132-3145, 2019.
13. C. Huang and H. Wang, “A Novel Key-Frames Selection Framework for Comprehensive Video Summarization,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 2, pp. 577-589, 2020.
14. Z. Ji, K. Xiong, Y. Pang and X. Li, “Video Summarization With Attention-Based Encoder–Decoder Networks,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 6, pp. 1709-1717, June 2020.
15. Wu Jiaxin, Zhong, Sheng-hua and Liu, “Dynamic Graph Convolutional Network for Multi-video Summarization”, Elsevier Pattern Recognition” vol. 107, no. 1, pp. 1-13, 2020.
16. B. Zhao and E. P. Xing, “Quasi real-time summarization for consumer videos,” IEEE Conference in Computer Vision Pattern Recognition, pp. 2513–2520, 2020.
17. Avila, BrandãoLopes, A. Luz and Araújo, “VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method,” Elseveir Pattern Recognition Letters, vol. 32, no. 1, pp. 56–68, 2020.
18. K. Zhou, Y. Qiao, and T. Xiang, “Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward,” IEEE International Conference Artificial Intelligence, pp. 7582–7589, 2020.
19. K. Zhang, W.-L. Chao, F. Sha, and K. Grau man, “Video summarization with long short-term memory,” IEEE International Conference Computer Vision, pp. 766–782, 2020.
20. B. Zhao, X. Li, and X. Lu, “Hierarchical recurrent neural network for video summarization,” ACM Multimedia Conference, pp. 863–871, 2020.
21. B. Mahasseni, M. Lam, and S. Todorovic, “Unsupervised video summarization with adversarial LSTM networks,” IEEE Conference Computer Vision Pattern Recognition, pp. 2982–2991, 2020.
22. B. Zhao, X. Li, and X. Lu, “Property-constrained dual learning for video summarization,” IEEE Trans. on Neural Networks and Learning Systems, vol. 31, no. 10, pp. 3989–4000, 2020.
23. A. Sahu, A.S. Chowdhury, “First person video summarization using different graph representations”, Elsevier Pattern Recognition Letters, vol. 146, pp. 185–192, 2021.

24. W. Zhu, J. Lu, J. Li and J. Zhou, “DSNet: A Flexible Detect-to-Summarize Network for Video Summarization”, IEEE Transactions on Image Processing, vol. 30, pp. 948-962, 2021.
25. C. Huang and H. Wang, “A Novel Key-Frames Selection Framework for Comprehensive Video Summarization”, IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 2, pp. 577-589, 2020.
26. Z. Ji, K. Xiong, Y. Pang and X. Li, "Video Summarization with Attention-Based Encoder–Decoder Networks”, IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 6, pp. 1709-1717, 2020.
27. Wu, Jiaxin Zhong, Sheng-hua Liu, Yan, “Dynamic Graph Convolutional Network for Multi-video Summarization”, Elsevier Pattern Recognition, vol. 107, pp. 382, 2020.
28. B. Zhao, X. Li and X. Lu, “Property-Constrained Dual Learning for Video Summarization”, IEEE Transactions on Neural Networks and Learning Systems, vol. 31, no. 10, pp. 3989-4000, 2020.
29. W. Zhu, J. Lu, J. Li and J. Zhou, “DSNet: A Flexible Detect-to-Summarize Network for Video Summarization,” IEEE Transactions on Image Processing, vol. 30, pp. 948-962, 2021.
30. Basavarajaiah, Madhushree Sharma and Priyanka, “GVSUM: generic video summarization using deep visual features”, Springer Multimedia Tools and Applications, vol. 80, pp. 108-120, 2021.
31. Z. Ji, X. Yu, Y. Yu, Y. Pang and Z. Zhang, “Semantic-Guided Class-Imbalance Learning Model for Zero-Shot Image Classification, “IEEE Transactions on Cybernetics, pp. 1-12, 2021.
32. Evlampios, Eleni, Alexandros, Mezaris and Ioannis Patras, “Video Summarization Using Deep Neural Networks: A Survey”, IEEE International Conference on Circuits and Systems for Video Technology, pp. 1-6, 2021.
33. Kasbgar, Mokhtari and Shojaedini, “A New Wavelet Based Spatio-temporal Method for Magnification of Subtle Motions in Video”, International Journal of Engineering, vol. 29, no. 3, pp. 313-320, 2016.
34. Otroshi Shahreza, Amini, and Behroozi, “Predicting the Empirical Distribution of Video Quality Scores Using Recurrent Neural Networks”, International Journal of Engineering, vol. 33, no. 5, pp. 984-991, 2020.
35. Salih and George, “Dynamic Scene Change Detection in Video Coding”, International Journal of Engineering, vol. 33, no. 5, pp. 966-974, 2020.
36. Mehrgan, Ahmadyfard, and Khosravi, “Super-resolution of License-plates Using Weighted Interpolation of Neighboring Pixels from Video Frames”, International Journal of Engineering, vol. 33, no. 5, pp. 992-999, 2020.
37. Firouzian, Firouzian, Hashemi and Kozegar, “Pain Facial Expression Recognition from Video Sequences Using Spatio-temporal Local Binary Patterns and Tracking Fiducial Points”, International Journal of Engineering, vol. 33, no. 5, pp. 1038-1047, 2020.