INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

A Framework for Video Summarization using Visual Attention Technique

M Dhanushree[1]*, **R Priya**[2], **P Aruna**[2], **R Bhavani**[3]

1 Research scholar, Department of Computer Science and Engineering, Annamalai University, Chidambaram, Tamil Nadu, India
2 Professor, Department of Computer Science and Engineering, Annamalai University, Chidambaram, Tamil Nadu, India
3 Professor and Head, Department of Computer Science and Engineering, Annamalai University, Chidambaram, Tamil Nadu, India

OPEN ACCESS

Received: 17-02-2024

Accepted: 26-03-2024

Published: 12-04-2024

Citation: Dhanushree M, Priya R, Aruna P, Bhavani R (2024) A Framework for Video Summarization using Visual Attention Technique. Indian Journal of Science and Technology 17(15): 1586-1595. https://doi.org/10.17485/IJST/v17i15.456

* **Corresponding author**.

dhanushree269@gmail.com

Funding: Ministry of Social Justice and Empowerment of the Government of India

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To develop an efficient Video Summarization technique that aims to utilize the saliency map for mimicking the human way of selecting the important events in the given video. **Methods:** This paper proposes Histogram based Weighted Fusion (HWF) algorithm that uses spatial and temporal saliency maps to act as guidance in creating the summary of the video. The spatial saliency score and temporal saliency score obtained from the corresponding saliency maps are fused using the proposed HWF algorithm to obtain the frame level importance score. It tries to depict the visual attention of the human brain when watching a particular video. **Findings:** The experimental results show that the proposed HWF algorithm performs better than the state-of-the-art methods. **Novelty:** The use of Histogram intersection and the incorporation of the exponential function as the weight for the combined feature enhance the summarization ability of the proposed model.

Keywords: Video Summarization; Saliency Map; Histogram intersection; Contrast sensitivity function; Attention curves

1 Introduction

In the growing internet era, multimedia usage has become ubiquitous. Communication between people in a society is done through the sharing of images and videos. Nowadays, videos are the most shared medium on social media platforms like Facebook, YouTube, Instagram, etc. With technological development, millions of users are able to flood the data warehouses of social media platforms with different video clippings. This creates a problem where storing, indexing, and searching through such large amounts of video data becomes insanely difficult. Thus, video summarization has emerged as one of the most needed research areas in today's world. A video consists of several shots, and each shot contains semantically meaningful frames. The frames that are part of the final summary of the video are called keyframes. They depict the essence of the given input video without compromising the context. Thus, video summarization can also be viewed as a step towards video understanding. Video summarization has a broad spectrum of applications, depending on the type of video that is summarized. In a long surveillance video, it is difficult to search for some particular events like violence, abnormalities, unknown person identification, etc. In such cases, video summarization plays an impactful role in saving man's time. The summarized video automatically extracts the vital events according to the user's needs. In the case of medical video summarization, it is helpful for medical students and staff to quickly go through the surgical procedure and any diagnosis videos like endoscopy, colonoscopy, etc. Video summarization also helps to detect traffic rule violators, which is part of an intelligent traffic management system. With wider applications, video summarization also faces wider challenges, such as the fact that the extraction of important events from the video is very subjective. There are structured videos like sports and news videos, as well as unstructured videos like user generated videos. In structured videos, there will be a definite boundary between the shots, while in unstructured videos, there are no cuts in between the shots, and the camera movements are shaken. Developing a generalized framework for both types of videos is a challenging task.

In the work(1), the authors used both color and structure-based features to find the representative keyframes of social media videos. These features were eventually clustered using Kohonen's Self organizing map to obtain the final summary. In the work done by(2), the authors attempted to summarize commodity hardware videos. They used low level features to eliminate the unimportant frames. A tree-based model is used, which utilizes segment level features for training. The authors of the literature(3) proposed a novel framework called the pyramidal opponent color-shape model, whose aim is to detect the boundaries of the shots, including hard and soft transitions in a video. In the work proposed by(4), block based spatio-temporal features are used as video features, and the features undergo dimensionality reduction techniques. Then a self-motivated scoring mechanism is utilized for finding the frame importance score. A comprehensive review of the various techniques used is discussed in(5). The survey presents genre-wise datasets and technique wise methodologies in a broader way.

In(6), the authors proposed a deep semantic and attentive network for video summarization. In this work, textual data is also incorporated, which helps in generating a semantically meaningful summary of the video. A self-attention mechanism is utilized to tackle the problem of long-term dependencies.(7) experimented in such a way by finding the shots in a video, followed by keyframe extraction. They used higher order color moments and the zero-normalized pixel correlation coefficient (ZNCC) for finding the keyframes as well as the shots. A detailed analysis of different techniques of video summarization is discussed in(8). A deep learning-based approach is undertaken by the authors of(9), in which the summaries are based on the object of interest. The video summary is produced by detecting the desired objects and then extracting those frames containing the objects.(10) summarized the campus surveillance videos using a deep CNN model. Activity recognition is performed initially using the attention based deep learning model. Using the attention scores that are generated, the keyframes are extracted.

The utilization of video summarization varies depending on the nature of the video being analyzed. In the contemporary setting, the prevalence of online educational videos has increased significantly, primarily due to their facilitation of remote learning. Consequently, the process of summarizing such instructional videos proves to be advantageous for learners during revision(11). Within the domain of intelligent traffic surveillance systems, the summarization of traffic videos is conducted based on various events like traffic violator detection and anomaly identification(12). Additionally, the summarization of extensive surveillance footage stands as another significant application of video summarization. In this context, the video background remains static while the methodologies employed revolve around the detection of individuals, objects, and events(13). One of emerging research subfield is the query focused video summarization(14) which is a type of personalized video summarization. The output summary is highly dependent on the user query(15). A recent advancement in this field involves the utilization of video summarization to assist in generating scene graphs(16), thereby enhancing video analysis capabilities. Furthermore, scene graphs find application in the domain of unsupervised machine translation processes(17). This type of translation is an essential component of comprehensive language models that are currently attracting significant attention. An essential aspect of video analysis and surveillance pertains to the detection of temporal activities, wherein activities within lengthy untrimmed videos are identified and localized simultaneously(18). Consequently, it is evident that video summarization encompasses a wide range of applications.

The major research gap in the above-mentioned literature is that the frame or images are considered as a whole and the features are extracted for further processing. Instead of utilizing the image in its entirety, a salient region of the image is taken in this research work. The best summarization results are achieved by finding the region of interest first and then doing further processing. The human brain tends to remember only things that are necessary, which can be applied to summarizing a video as well. People, while watching a video clip, remember or tend to capture those that are considered informative to them. Such a behavior can be modeled computationally using a visual attention technique. This paper presents a work that is based on visual attention both spatially and temporally using a proposed HWF method. By finding the saliency, the region of interest is obtained. The major contributions of this work are:

- A saliency-based feature extraction is utilized.
- A novel feature fusion algorithm based on histogram intersection is proposed.

- An efficient threshold based keyframe selection is used which plays a major role in efficient video summarization framework based on visual attention technique.

2 Methodology

Figure 1 shows the block diagram of the framework for video summarization using proposed HWF algorithm.

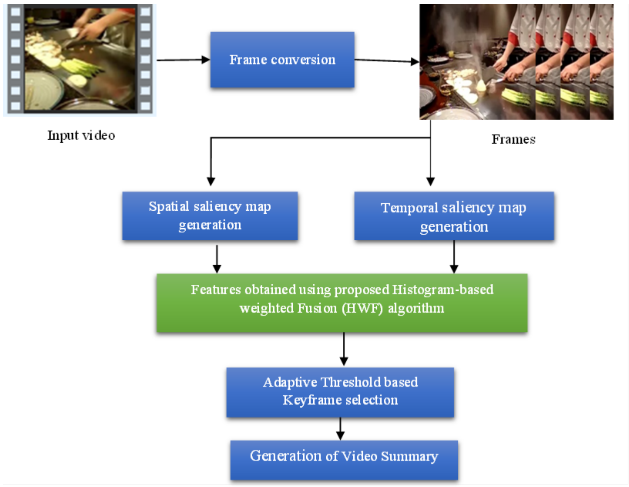

Input video
Frame conversion
Frames
Spatial saliency map generation
Temporal saliency map generation
Features obtained using proposed Histogram-based weighted Fusion (HWF) algorithm
Adaptive Threshold based Keyframe selection
Generation of Video Summary

Fig 1. Block diagram of the framework for video summarization using proposed HWF algorithm

2.1 Dataset description

For this work, two benchmark datasets are taken, named SumMe and TVSum. The SumMe dataset contains 25 videos, which include holidays, sports, and events. They are raw or minimally edited videos whose length varies from 1 to 6 minutes, and each video contains 25 frames per second. To ensure that video summaries are evaluated effectively, each of these videos has at least 15 user annotations. The title-based Video Summarization (TVSum) dataset consists of 50 videos from various genres, such as new, vlog, how-to, documentary, etc. Each video is annotated by 20 people, which is helpful in the automatic evaluation of various video summarization techniques. The average duration of the dataset is about 2 to 10 minutes.

2.2 Frame conversion

Video summarization starts by converting the input video into frames using the uniform sampling method. The uniform sampling method is a process of selecting frames at a uniform distance d. Here, d=3 is taken i.e., every third frame is taken for further processing. As each video in the dataset used has a different dimension, the video frames are reduced to 256 x 256 for uniformity. This work is entirely based on the visual attention cues through which the video summary is generated. Two types of saliency maps, namely the spatial saliency map and the temporal saliency map, are generated for each input frame. From the saliency maps, the spatial attention score and temporal attention score are calculated, respectively. These attention score features are then fused together using the proposed Histogram-based Weighted Fusion (HWF) algorithm. Thus, each frame is assigned a particular attention score. The attention curves are then generated, and an adaptive threshold based keyframe selection is performed. These keyframes form the summary of the video.

2.3 Saliency map

This work aims at summarizing the video based on visual attention, i.e., video summarization is carried out from a user point of view, like the contrast of a frame or any motion present in between the frames. These contrasts and motions in the video model how human attention is drawn towards the important part or content in a video. Thus, a saliency map is used, which represents a salient or important part of the video. In this work, two types of saliency maps are used, namely the spatial saliency map and the temporal saliency map. The spatial saliency map gives an important region in the two-dimensional image space, while the temporal saliency map gives an important region across frames with time as the third dimension.

2.3.1 Spatial saliency map generation

The input frames are converted from RGB to L*a*b* color space. The choice of L*a*b* color space is due to the fact that it closely resembles the perception of colors by the human brain. People tend to give importance to those frames that attract them. When a color feature is considered, the contrast plays a major role in attracting visual attention. Thus, the saliency map is generated based on the contrast sensitivity function. The images that are converted to L*a*b* color space are then filtered using a contrast sensitivity function, which is given by the following equation:

$$S(r) = re^{-0.25r} \tag{1}$$

Where r is the pixel intensity value of image S.

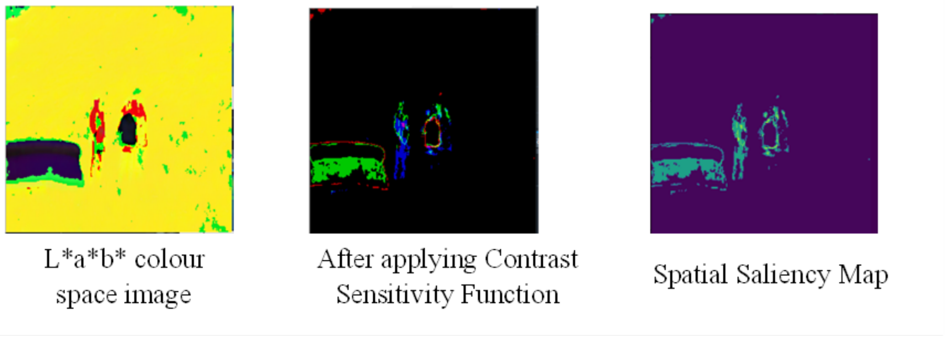
Finally, the spatial saliency map is generated by finding the spatial distance between the filtered image and the mean of the image which is given by the following equation.

$$SS(x,y) = \sqrt{(L^* - L^*{}_m)^2 + (a^* - a^*{}_m)^2 + (b^* - b^*{}_m)^2} \tag{2}$$

Where L*, a* and b* are the three channels of the L*a*b* color space.

$L^*{}_m$, $a^*{}_m$ and $b^*{}_m$ indicates the mean of the three-color channels respectively.

Figure 2 shows the output of the spatial saliency map in which the foreground is separated from the background.



L*a*b* colour space image

After applying Contrast Sensitivity Function

Spatial Saliency Map

Fig 2. Spatial saliency map generation

2.3.2 Temporal saliency map generation

Humans tend to get more visual attention when there is movement in the video. In order to model the motion of objects in a video, temporal saliency is calculated. A temporal saliency map is used to highlight those areas that are subjected to motion. The calculation of the temporal saliency map starts by finding the temporal gradient between the neighboring frames. Consider two frames, $Fr(t)$ and $Fr(t-1)$ at time t and $t-1$, respectively. The temporal gradient $TG(t)$ at time t is calculated using the formula given below.

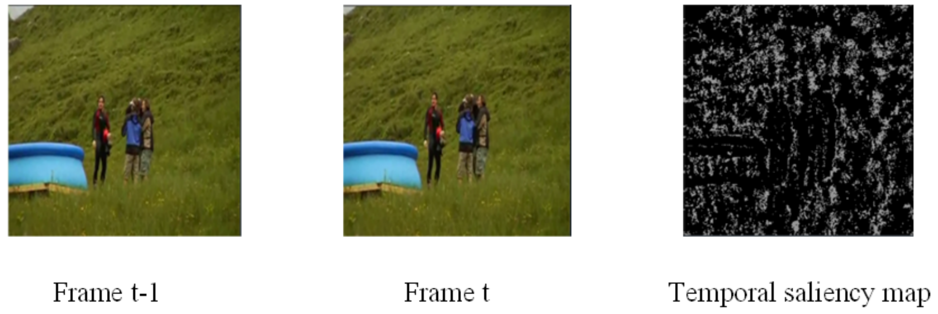
$$TG(t) = |Fr(t-1) - Fr(t)| \tag{3}$$

Thus, a temporal gradient is found for each frame between its neighboring frames. After finding the temporal gradient, the temporal saliency value at each pixel u of frame $Fr(t)$ is found by taking the sum of the absolute difference between the pixel u and its 7x7 neighboring pixels of the temporal gradient. The temporal saliency value at pixel u in frame $Fr(t)$ is found by the following equation:

$$TS_u = \sum_{v=1}^{N(u)} |TG_u(t) - TG_v(t)| \tag{4}$$

Where $N(u)$ is the number of neighbor pixel of the pixel u and v indicates its neighbor pixels.

Figure 3 shows Frame t-1, Frame t and the output of temporal saliency map. The highlighted pixels are due to the camera movement.



Frame t-1

Frame t

Temporal saliency map

Fig 3. Temporal saliency map generation

2.4 Features obtained using proposed Histogram based weighted Fusion (HWF) algorithm

The generation of spatial as well as temporal saliency maps helps to model human visual attention in creating the summary of the video. From these two saliency maps, attention scores are calculated as follows. The spatial saliency map for each frame is normalized by dividing each pixel value by the maximum value of that frame. The average of non-zero values in the normalized spatial saliency map is the spatial attention score (S_A). Similarly, temporal saliency maps are also normalized in the ranges 0 and 1. Temporal attention scores (T_A) are calculated by finding the average of the normalized temporal saliency map.

These attention scores are then fused using the proposed Histogram based weighted Fusion (HWF) algorithm. The first step is to find the histogram of each of the L*a*b* color space versions of the input frames. The histogram with 16 bins is generated for each of the color channels L, A, and B, respectively, and is concatenated to form a 48-bin histogram for each frame. The algorithm works by finding the histogram intersection D between the successive frames. Histogram intersection is one of the methods of finding the similarity between the histograms. This helps in calculating the amount of intersection between the histograms under study. By finding the histogram intersection between the frames, the frame similarity can be easily measured. The major reason for using histogram-based similarity is that it is computationally fast and efficient. Even a 2-minute video with 30 frames per second consists of 3600 frames. Extracting feature from all these frames exhibits computational overhead. Thus, an efficient histogram-based frame similarity is employed. Figure 4 depicts the histogram intersection for each frame of the “Jumps” video from the SumMe dataset.

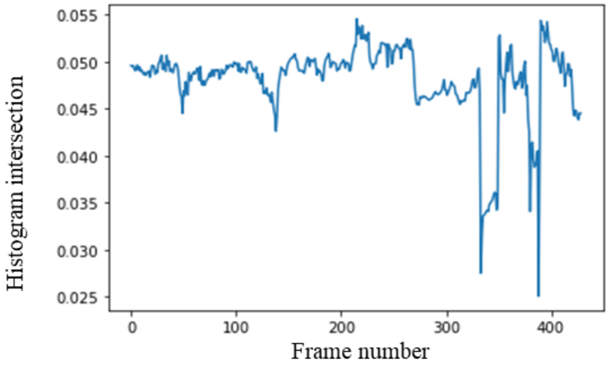


Fig 4. Histogram intersection of video ’Jumps’ from SumMe dataset

The value of D ranges from 0, indicating no overlap, to 1, indicating almost the same frames. Using the value of D helps give weight to a particular frame. The proposed algorithm, HWF, works effectively using the following feature fusion equation.

$$F_A(f_t) = e^{d(f_t)} * S_A(f_t) + e^{d(f_t)} * T_A(f_t) \quad (5)$$

Where,

$F_A(f_t)$ is the final attention score of frame f at time t,
$S_A(f_t)$ is the spatial attention score of frame f at time t,
$T_A(f_t)$ is the temporal attention score of frame f at time t and
$d(f_t)$ is the histogram intersection between the frames f and f' at time t and t-1 respectively.

The exponential function e^x is used as a weight for both the attention scores. The histogram intersection between the following frames determines these weights. The property of the exponential function is fully exploited in finding the key frames. If the frames are similar, the value of D is small, and hence the weight given is small, while in the case of dissimilar frames, the weight is large. This helps in finding the best summary of the video. The features obtained using the proposed HWF algorithm are then plotted as attention curves, as shown in Figure 5.

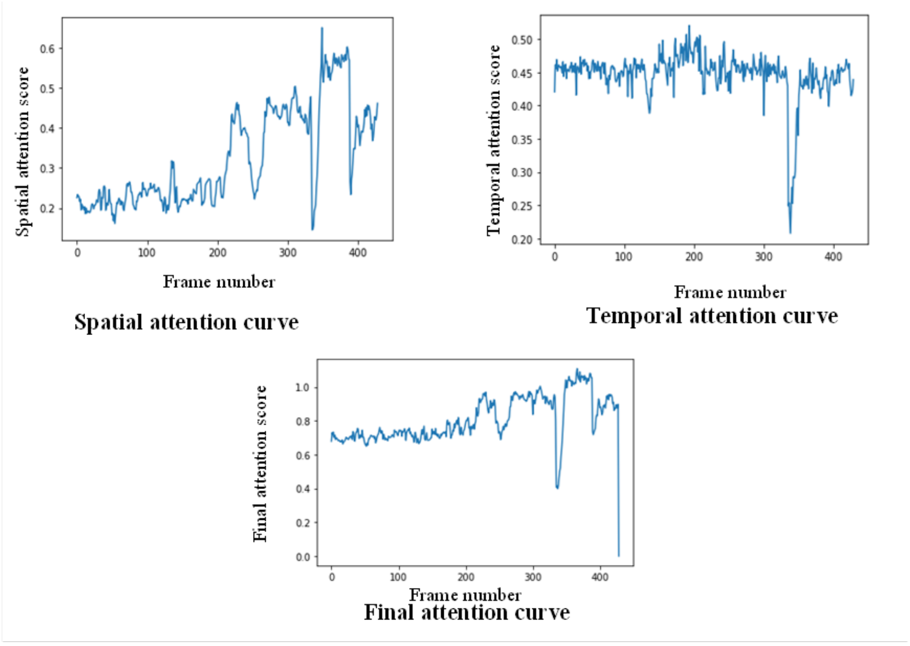


Fig 5. Attention curves for the video 'Jumps' from SumMe dataset

2.5 Adaptive threshold based keyframe selection

From the attention curve that is generated, the keyframes are selected based on a threshold value. The threshold value varies from 0.5 to 0.9 at an increment of 0.05, and the threshold that gives the maximum F-score value is selected. Frames whose final attention score is greater than the threshold is selected as keyframes. Figure 6 depicts the selection of key frames through an adaptive threshold-based technique.

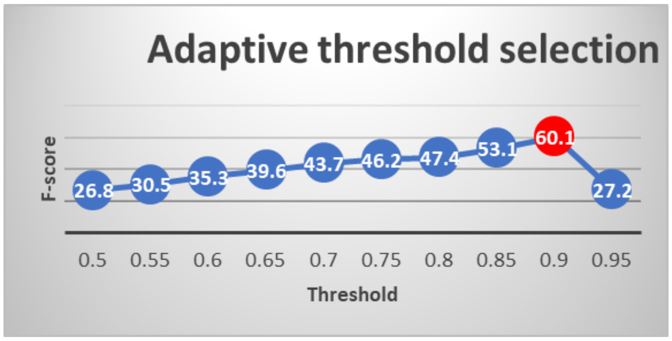


Fig 6. Adaptive Threshold selection for the video 'Jumps' from SumMe dataset

Algorithm 1: Video summarization using proposed HWF algorithm

```
Input: Video
Output: Summarized video
Step 1: Covert the video V into Frames Fr.
Step 2: Select every third frame for further processing.
Step 3: Resize the frames to 256 x 256.
Step 4: Convert RGB color space to L*a*b* color space.
Step 5: Generate Spatial attention map using the following steps:
Step 5.1: Apply Contrast Sensitivity filter using the following equation
$S(r) = re^{-0.25r}$.
Step 5.2: Find the mean of the image $L^*_m$, $a^*_m$and $b^*_m$ for all three channels of L*a*b* color space respectively.
Step 5.3: Find the Euclidian distance between the filtered image and the mean of the filtered image as follows to obtain the spatial saliency map (SS)
$SS(p) = \sqrt{(L(p)-L_m)^2+(A(p)-A_m)^2+(B(p)-B_m)^2}$.
Step 6: Generate Temporal attention map using the following steps:
Step 6.1: Find the absolute difference between the frames in L*a*b* colour space.
Step 6.2: These frame differences of each colour component are added together to obtain a single channel.
Step 6.3: The temporal saliency TS at pixel ‘p’ is found by summing the absolute difference between the center pixel ‘p’ and its 7 x 7 neighborhood pixel
$TS(p) = \sum_{i=1}^{N_b} |p-i|$.
Step 7: Calculate the Spatial attention score $S_A$ for each frame which is the average of the Spatial saliency map SS for each frame.
Step 8: Calculate the Temporal attention score $T_A$ for each frame which is the average of the Spatial saliency map TS for each frame.
Step 9: Find the histogram for each frame in L*a*b* color space with 16 bins for each channel respectively.
Step 10: Calculate the histogram intersection D between the consecutive frames using the following equation
$D(h_1,h_2) = \sum_{i=1}^{N} min(h_1(i), h_2(i))$.
Step 11: Final attention score $F_A$ for each frame $f_t$ is calculated using the following proposed equation
$F_A(f_t) = e^{d(f_t)} * S_A(f_t) + e^{d(f_t)} * T_A(f_t)$.
Step 12: Find threshold $th$ using threshold-based adaptive keyframe selection.
Step 13: Select those keyframes having $F_A$ greater than $th$ which gives the final summary of the input video.
```

3 Results and discussion

3.1 Performance Measures

The proposed video summarization framework is evaluated based on the quantitative measure of Precision, Recall and F-score. Precision and recall indicate the quality of the summarization.

$$Precision\ (P) = \frac{|S_U \cap S_A|}{|S_U|} \tag{6}$$

$$Recall\ (R) = \frac{|S_U \cap S_A|}{|S_A|} \tag{7}$$

Where,

S_U (User Summary) indicates the summary generated by user.

S_A (Automated Summary) indicates the summary generated by the proposed work.

F-score is measured as pairwise f-score proposed by[1] which considers the consistency of the generated summary with that of each summary produced by the human. For the automated summary, the average pairwise f-score is calculated using the following equation:

$$\bar{F} = \frac{1}{N-1}\sum_{i=1}^{N} 2\frac{P_i R_i}{P_i + R_i} \tag{8}$$

Where,

N is the total number of user summary for a particular video.

P_i is the precision of automated summary using user summary *i*.

R_i is the recall of automated summary using user summary *i*.

Table 1. Experimental results of the Proposed HWF algorithm for SumMe and TVSum dataset

Dataset name	Video name	Precision	Recall	F-score	Length of original video (seconds)	Length of Summarized video (seconds)
SumMe	Air_Force_One	52.01	55.7	53.79	179	36
	Base jumping	43.1	55.2	48.41	158	63
	Bearpark_climbing	51.21	55.2	53.13	133	46
	Bike Polo	44.8	45.9	45.34	103	21
	Bus_in_Rock_Tunnel	52.48	65.4	58.23	171	33
TVSum	Video_1	66.2	84.6	74.27	148	105
	Video_2	66.9	71.1	68.93	141	48
	Video_3	58.21	82.9	68.39	194	110
	Video_4	65.21	96.8	77.92	167	117
	Video_5	65.2	66.9	66.03	191	76

Table 1 shows the quantitative measures of Precision, Recall and F-score for the SumMe and TVSum dataset. It also shows the total video length and the summarized length for the videos taken.

3.2 Comparative analysis

Tables 2 and 3 shows the comparative analysis for SumMe and TVSum dataset respectively. For comparison, average F-score of the entire dataset is taken.

Table 2. Comparative analysis of the Proposed HWF algorithm for SumMe dataset

Dataset	Approach	Avg F-score
SumMe	RBVS [19]	45.06
	MST_C [20]	38.30
	SUMGAN [21]	41.70
	MBVS [22]	30.58
	SB2S3 [23]	42.80
	Proposed HWF	**47.82**

Table 3. Comparative analysis of the Proposed HWF algorithm for TVSum dataset

Dataset	Approach	Avg F-score
TVSum	RBVS [19]	56.13
	MST_C [20]	54.60
	SUMGAN [21]	56.00
	MBVS [22]	33.41
	SB2S3 [23]	57.80
	Proposed HWF	**58.50**

From the above tables, it is inferred that the proposed video summarization technique performs efficiently when compared to other methods. The main reason for the efficient performance is that the proposed video summarization framework depicts human behavior in choosing the salient frames from the video and fusing the salient features using histogram intersection. Here, histogram intersection acts as a frame similarity measure, which is eventually used in finding the key shots. The use of saliency maps as regions of interest and exponential functions as weights plays a vital role in the efficient performance of the proposed HWF algorithm, because the weight of each frame gets altered according to the similarity of the frames. Thus, the proposed video summarization technique performs better than the existing works.

4 Conclusion

Video summarization helps find the most important and informative content in a video. It abstracts the entire content of the video into short duration, covering the salient frames of the video. A video summary is useful for people who wish to watch only a glimpse of the video. In this work, a visual attention-based framework is used that tries to mimic how a human brain captures the vital content from the video. A histogram based weighted fusion algorithm and adaptive threshold based keyframe selection are proposed, which effectively summarize the video content. Moreover, the use of a saliency map removes unnecessary backgrounds while processing the video, which is considered a major strength of the proposed HWF algorithm. This contributes to the significant improvement in the summarization results when compared to existing works. Experimental results show that the proposed video summarization architecture outperforms the other state-of-the-arts method with an average F-score of 47.82% for the SumMe dataset and 58.5% for the TVSum dataset. The drawback of the proposed algorithm is that it shows a slightly lower performance for the unedited videos from the SumMe dataset when compared to structured videos like sports and news videos. In future work, the summarization results can be further improved using shot boundary detection as a preprocessing step, particularly in unstructured videos. As the SumMe dataset contains the most unedited and raw videos, it can be inferred that there is still room for improvement. The major concern with unstructured videos is that they are shaky and don't have solid boundaries between the shots. Thus, the identification of unclear boundaries helps in obtaining better results for unstructured videos.

Acknowledgement

MD gratefully acknowledges the funding agency, Ministry of Social Justice and Empowerment of the Government of India, for providing financial support for doing this research work.

References

1) Rani S, Kumar M. Social media video summarization using multi-Visual features and Kohnen's Self Organizing Map. *Information Processing & Management*. 2020;57(3):102190. Available from: https://dx.doi.org/10.1016/j.ipm.2019.102190.
2) Taylor W, Qureshi FZ. Real-time Video Summarization on Commodity Hardware. In: Proceedings of the 12th International Conference on Distributed Smart Cameras. ACM. 2018;p. 1–8. Available from: https://doi.org/10.1145/3243394.3243689.
3) Sasithradevi A, Roomi SMM. A new pyramidal opponent color-shape model based video shot boundary detection. *Journal of Visual Communication and Image Representation*. 2020;67:102754. Available from: https://dx.doi.org/10.1016/j.jvcir.2020.102754.
4) lin Li W, Zhang T, Liu X. A static video summarization approach via block-based self-motivated visual attention scoring mechanism. *International Journal of Machine Learning and Cybernetics*. 2023;14(9):2991–3002. Available from: https://dx.doi.org/10.1007/s13042-023-01814-9.
5) Gupta D, Sharma A. A comprehensive study of automatic video summarization techniques. *Artificial Intelligence Review*. 2023;56(10):11473–11633. Available from: https://dx.doi.org/10.1007/s10462-023-10429-z.
6) Zhong SH, Lin J, Lu J, Fares A, Ren T. Deep Semantic and Attentive Network for Unsupervised Video Summarization. *ACM Transactions on Multimedia Computing, Communications, and Applications*. 2022;18(2):1–21. Available from: https://dx.doi.org/10.1145/3477538.
7) Mounika BR, Prakash O, Khare A. Fusion of Zero-Normalized Pixel Correlation Coefficient and Higher-Order Color Moments for Keyframe Extraction. In: Recent Trends in Communication, Computing, and Electronics;vol. 524 of Lecture Notes in Electrical Engineering. Singapore. Springer. 2018;p. 357–364. Available from: https://doi.org/10.1007/978-981-13-2685-1_34.
8) Saini P, Kumar K, Kashid S, Saini A, Negi A. Video summarization using deep learning techniques: a detailed analysis and investigation. *Artificial Intelligence Review*. 2023;56(11):12347–12385. Available from: https://dx.doi.org/10.1007/s10462-023-10444-0.
9) Haq HBU, Asif M, Ahmad MB, Ashraf R, Mahmood T. An Effective Video Summarization Framework Based on the Object of Interest Using Deep Learning. *Mathematical Problems in Engineering*. 2022;2022:1–25. Available from: https://doi.org/10.1155/2022/7453744.
10) Muhammad W, Ahmed I, Ahmad J, Nawaz M, Alabdulkreem E, Ghadi Y. A video summarization framework based on activity attention modeling using deep features for smart campus surveillance system. *PeerJ Computer Science*. 2022;8:1–21. Available from: https://doi.org/10.7717/peerj-cs.911.
11) Cagliero L, Canale L, Farinetti L. Data-Driven Analysis of Student Engagement in Time-Limited Computer Laboratories. *Algorithms*. 2023;16(10):1–26. Available from: https://dx.doi.org/10.3390/a16100464.
12) Pramanik A, Pal SK, Maiti J, Mitra P. Traffic Anomaly Detection and Video Summarization Using Spatio-Temporal Rough Fuzzy Granulation With Z-Numbers. *IEEE Transactions on Intelligent Transportation Systems*. 2022;23(12):24116–24125. Available from: https://dx.doi.org/10.1109/tits.2022.3198595.
13) Shambharkar PG, Goel R. From video summarization to real time video summarization in smart cities and beyond: A survey. *Frontiers in Big Data*. 2022;5:1–14. Available from: https://doi.org/10.3389/fdata.2022.1106776.
14) Hu W, Zhang Y, Li Y, Zhao J, Hu X, Cui Y, et al. Query-based video summarization with multi-label classification network. *Multimedia Tools and Applications*. 2023;82(24):37529–37549. Available from: https://dx.doi.org/10.1007/s11042-023-15126-1.
15) Puthige I, Hussain T, Gupta S, Agarwal M. Attention Over Attention: An Enhanced Supervised Video Summarization Approach. *Procedia Computer Science*. 2023;218:2359–2368. Available from: https://dx.doi.org/10.1016/j.procs.2023.01.211.

16) Chang X, Ren P, Xu P, Li Z, Chen X, Hauptmann A. A Comprehensive Survey of Scene Graphs: Generation and Application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2023;45(1):1–26. Available from: https://dx.doi.org/10.1109/tpami.2021.3137605.

17) Li M, Huang PY, Chang X, Hu J, Yang Y, Hauptmann A. Video Pivoting Unsupervised Multi-Modal Machine Translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2022;45(3):3918 –3932. Available from: https://dx.doi.org/10.1109/tpami.2022.3181116.

18) Zhang L, Chang X, Liu J, Luo M, Li Z, Yao L, et al. TN-ZSTAD: Transferable Network for Zero-Shot Temporal Activity Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2022;45(3):3848 –3861. Available from: https://dx.doi.org/10.1109/tpami.2022.3183586.

19) Harakannanavar SS, Sameer SR, Kumar V, Behera SK, Amberkar AV, Puranikmath VI. Robust video summarization algorithm using supervised machine learning. *Global Transitions Proceedings*. 2022;3(1):131–135. Available from: https://dx.doi.org/10.1016/j.gltp.2022.04.009.

20) Sahu A, Chowdhury AS. First person video summarization using different graph representations. *Pattern Recognition Letters*. 2021;146:185–192. Available from: https://dx.doi.org/10.1016/j.patrec.2021.03.013.

21) Mahasseni B, Lam M, Todorovic S. Unsupervised Video Summarization with Adversarial LSTM Networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. 2017;p. 202–211. Available from: https://doi.org/10.1109/CVPR.2017.318.

22) Alam I, Jalan D, Shaw P, Mohanta PP. Motion Based Video Skimming. In: 2020 IEEE Calcutta Conference (CALCON). IEEE. 2020. Available from: https://doi.org/10.1109/CALCON49167.2020.9106488.

23) Ma M, Mei S, Wan S, Wang Z, Feng DD, Bennamoun M. Similarity Based Block Sparse Subset Selection for Video Summarization. *IEEE Transactions on Circuits and Systems for Video Technology*. 2021;31(10):3967–3980. Available from: https://dx.doi.org/10.1109/tcsvt.2020.3044600.