Superframe Segmentation for Content-based Video Summarization

Priyanka Ganesan[1], Senthil Kumar Jagatheesaperumal[2], Abirami R[3],
Lekhasri K[4], Silvia Gaftandzhieva[5], Rositsa Doneva[6]
Department of Computer Science and Engineering, Mepco Schlenk Engineering College, Sivakasi, India[1, 3, 4]
Department of Electronics and Communication Engineering, Mepco Schlenk Engineering College, Sivakasi, India[2]
Faculty of Mathematics and Informatics, University of Plovdiv "Paisii Hilendarski", Plovdiv, Bulgaria[5]
Faculty of Physics and Technology, University of Plovdiv "Paisii Hilendarski", Plovdiv, Bulgaria[6]

***Abstract*—Video summarization is a complex computer vision task that involves the compression of lengthy videos into shorter yet informative summaries that retain the crucial content of the original footage. This paper presents a content-based video summarization approach that utilizes superframe segmentation to identify and extract keyframes representing the most significant information in a video. Unlike other methods that rely solely on visual cues, our approach segments the video into meaningful and coherent visual content units while also preserving the original video's temporal coherence. This method helps keep the context and continuity of the video in the summary. It involves dividing the video into superframes, each of which is a cluster of adjacent frames with similar motion and visual characteristics. The superframes are then ranked based on their salient scores, which are calculated using visual and motion features. The proposed method selects the top-ranked super frames for the video summary. It has been evaluated on the SUMMe and TVSum datasets and achieved state-of-the-art results for F1-score and accuracy. Based on the experimental outcomes, it is evident that the suggested superframe segmentation method is effective for video summarization, which could be largely assistive for monitoring and controlling the student activities, particularly during their online exams.**

Keywords*—Video summarization; deep learning; super frame segmentation; keyframes; keyshot identification

I. INTRODUCTION

Video summary (VS), which creates a concise and precise representation of a video's visual information, has been a crucial tool for many video analytical activities. Two key characteristics define a qualitative video summary. It must be represented in the sense that it includes all the critical scenes from the original video, and it must also contain the bare minimum of redundancy. Various fields, such as electronic media, personal videos, medical videos, online databases, and surveillance applications, have witnessed the emergence of video summarization (VS) methods. These methods aims to facilitate the browsing of an increasing amount of video data in the field of surveillance and reduce the computational burden of video summarization. Despite efforts to improve video summarization accuracy using various techniques, such as novel edge inadmissibility measures for MST-based clustering and graph-based shot boundary detection, these methods have demonstrated limited success, as reported in previous research [3].

This paper proposes a new method for video summarization of long surveillance streams utilizing deep learning techniques. The summary consists of keyframes or video clips that have undergone some editing form to provide essential information from the original video in a condensed format, allowing users to assess the video's usefulness quickly. Caps-Net is used to avoid selecting transitional or similar frames in the same shot, improving the summary's quality. The proposed method addresses issues with summarizing multiple videos by relying solely on visual cues provided by video shots. By addressing the challenges of redundancy and transitional frame selection within shots, the proposed method offers a more efficient approach to summarizing multiple surveillance videos.

Event-based techniques can be employed to detect both regular and abnormal activities that occur in videos. For example, sudden changes in the environment, such as theft, robbery, or terrorist activities, detection can be achieved by using detection models to search for unusual or suspicious features. Once the frames with abnormal scenes are identified, they are combined using a video summarization algorithm to generate a video summary. P. Kalaivani and S.M. M. Roomi [6] described such helpful approaches for event recognition and creating summaries of the video. Kumar et al. [7] employed Bootstrap Aggregating to improve the accuracy of keyframe selection. Damjanovic et al. [8] proposed an event-based video summarization method that involves determining the energy of each frame by adding the absolute values of pixels in the current and reference frames, identifying frames during which events occurred, and producing a video summary for those frames. Thomas et al. [9] developed the Human Visual System (HVS) to create perceptual video summaries by identifying significant events in videos and eliminating redundancy. The paper is organized as follows: Section I proviedes an introduction to video summarization and outlines the research problems, and significance and contribution of the paper. Section II details the proposed method, including the utilization of capsule Networks and event-based techniques and Section III discusses the proposed method. Results and discussion is given in Section IV. Finally, Section V concludes the paper with a summary of findings and avenues for future research.

II. METHODOLOGY

In the past, unsupervised video summarization methods relied on shallow features and clustering techniques to group frames into clusters, with the cluster centres selected as keyframes. For instance, Ngo et al. [10] transformed each video into an undirected graph and clustered it. Cong et al. [11] used dictionary learning, while Zhou et al. [12] employed reinforcement learning and a reward function that considered representativeness and variety. Mahasseni et al. [13] introduced the first generative adversarial network (GAN) for video summarization, where an auto-encoder LSTM acted as the summarizer and a discriminator distinguished between the summarizer's reconstruction and the original video input. Rochan et al. [14] proposed an adversarial approach for learning summarization skills from unpaired data.

Supervised methods for keyframe selection in video summarization require human-labeled summaries [15]. One such method is the sequential Determinantal Point Process (seqDPP) developed by Gong et al. [16], which considers video summarization as a subset selection problem and uses a probabilistic model to choose representative and diverse subsets. Bulut et al. [17] proposed the key frame extraction method from a motion capture sequence. The important frames of a motion are selected to be the keyframes and the others are computed via the interpolation techniques nu using the keyframes. Zhao et al. [18] introduced the hierarchical recurrent neural network (H-RNN) method, which captures temporal dependencies from frame sequences and reduces information loss and computational complexity compared to other RNN models for video summarization.

Extracting keyframes from motion-based videos is a challenging task, particularly in the presence of cameras. The idea of using motion-based frames for keyframe extraction was first introduced by Wolf [19]. Li et al. [20] presented an approach that utilized relative motion for generating a video summary and analyzed spatial and emotional data to extract additional insights. Ajmal et al. [21] developed a technique that tracked human movements using the Kalman filter and analyzed the trajectory obtained. Almeida et al. [22] employed a colour histogram to create a distinct video summary by selecting the most representative frame. Zhang et al. [23] chose the first frame of each shot as the key frame and utilized colour histograms to identify other significant frames.

Object-based techniques have proven effective in identifying and summarizing specific objects in videos, including people, cars, and cats. Feng and Chong-Wah [24] used hierarchical hidden Markov models to produce a summary based on objects and events in rushed videos. Neeraj et al. [25] suggested the object-based video summarization method, which uses mathematical techniques to minimize redundancy, utilizing the loss function, summary variance, and score identification. However, such methods may not be as effective in summarizing fast-paced videos and could potentially miss significant items.

III. PROPOSED SYSTEM

In this proposed method, pre-processing is applied initially, frames are extracted, and the feature extraction is done using the superframe segmentation method. The superframe segmentation method is utilized to identify the boundary between temporal clusters, and it generates superframes by considering both motion similarity and the targeted number of clusters. A video summary can be generated by selecting representative frames or keyframes from each superframe. Keyframes can be selected based on various criteria, including visual saliency, diversity or importance to the overall video content. Fig. 1 shows the proposed method.

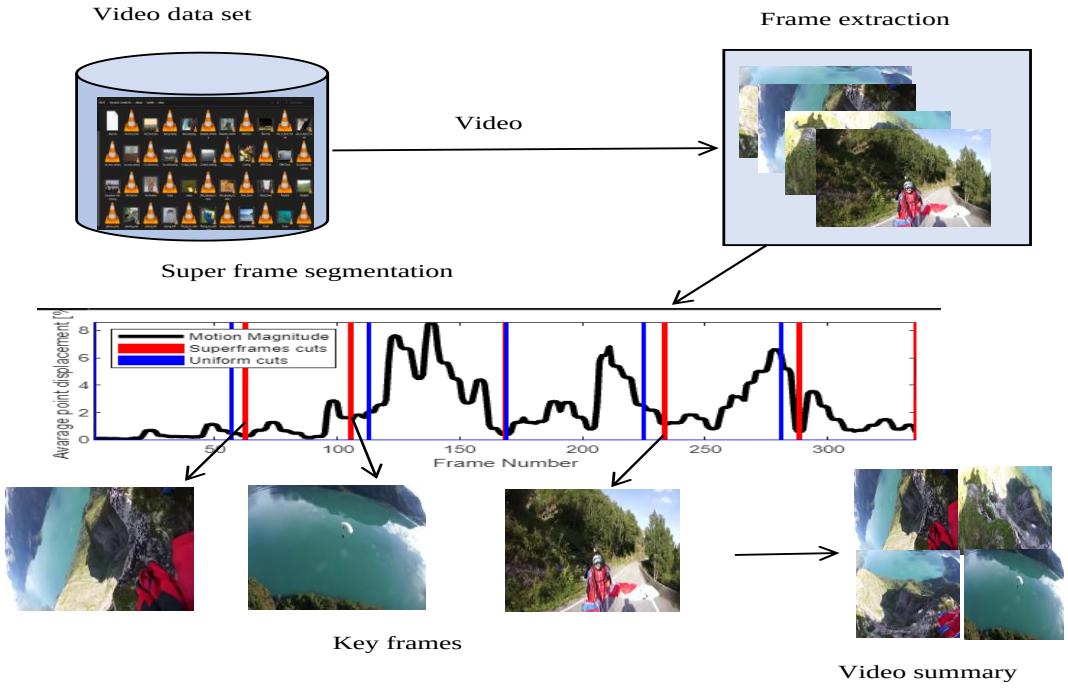


Fig. 1. Proposed system design.

A. Frame Extraction

Frame extraction involves capturing individual frames from a video, typically at a fixed rate or at specific points in time. This fact can be helpful for various applications, such as analyzing the contents of a video, detecting changes between frames, or creating a new video from selected frames. Algorithm 1 represents the frame extraction process from the video. The input for this module is a video data set which contains many user videos. Separate folder will be created for each video. Then, the frames are saved in the respective folder. Pre-processing refers to the techniques applied to data before the analysis to enhance its quality and suitability for downstream analysis. In this, we did rgbtohsi. Converting RGB to HSI can be a crucial step in image processing and computer vision tasks as it can help to separate the image information based on its colour properties. HSI colour model represents colours in terms of hue, saturation, and intensity, which are intuitive and perceptually meaningful to the human eye. The hue component can be used to segment objects based on their colour, the saturation component can be used to detect edges, and the intensity component can be used for brightness normalization. By converting an image from RGB to HSI, we can perform these operations more efficiently and accurately. It can provide a more meaningful and robust representation of colour information in an image.

ALGORITHM 1 Frame extraction(V)
//**Input:** Video V
//**Output:** Sequence of extracted frames
1.V←videoreader(videoname)
2. Initialize:
3. counter ← 0
4. while not at the end of V do:
5. read the frame
6. if not end of V:
7. break
8. else:
9. update the frames in the destined folder
10. counter← counter+1
11. end while

B. Super Frame Segmentation

For feature extraction, we use a superframe segmentation algorithm. Our superframe segmentation approach locates the boundary between temporal clusters in video frames. The superframe algorithm generates superframes based on the motion similarity and the required number of clusters. The following Algorithm 2 describes the superframe segmentation. At the beginning of the algorithm, cluster centres are initialized with a regular step size S and then adjusted to the position with the lowest gradient within a neighbourhood. This step aims to ensure that the clusters are initialized in a good position and to prevent them from getting stuck in local minima. The algorithm iteratively assigns frames to the nearest cluster centre using a distance measure as in Eq. (1). This could be any distance measure such as Euclidean distance, Manhattan distance or Cosine distance. After assigning frames to clusters, new cluster centres are computed using the L1 distance. The algorithm repeats this process until the error E falls below a threshold.

ALGORITHM 2 Video super frame clustering algorithm
//Inputs: Video Frames
1: **Initialize:**
a= 0.1 ∗ K.
Cluster centers $Cl_k = [x1 \ldots . xf]^T$ at regular step F
2: Perturb cluster centers in a neighborhood, to the lowest gradient position
3: repeat
4: $S_t \leftarrow S_t - 1$
5: for each Cl_k do
6: Assign best-matching frames from a 2S neighbourhood around Cl_k according to D_s
7: end for
8: Compute new cluster centres and error E (L1 distance)
9: until E ≤ threshold
10: Post-processing to remove very short clusters

Finally, post-processing is performed to remove very short clusters. This could be done by merging small clusters with neighbouring clusters or by removing them altogether. We employ D_s as a distance unit, which is denoted as follows: One way to measure the distance between cluster k and frame i is expressed by the following formula:

$$D_s = \sqrt{\Sigma}(X_k - X_i)^2 \quad (1)$$

where, X is the feature vector.

We might have a small number of clusters with very short lengths at the end of this operation. Fig. 2 shows the superframe cut during this segmentation process.

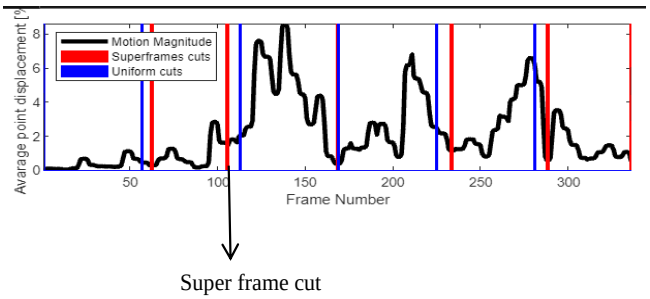


Fig. 2. Illustration of the superframe cut during the segmentation process.

C. Key Shot Identification

Based on characteristics, we will select n frames from the superframe segmentation. The key shots needed to create the summary were found in m frames. Conventional video summarization techniques primarily target edited videos, such as news reports, sports broadcasts, or movies that are composed of several short shots. Shot detection based on changes in the colour histogram is often adequate to segment such videos [25]. Such a technique cannot be applied in our case because we concentrate on user movies that are generally unedited and frequently only comprise one single shot. This issue was also addressed previously by [21], who offered to

partition egocentric films into shots by classifying the frames into static, in-transit, or head-movement categories.

This approach, however, is only appropriate for egocentric videos and produces shots that last for roughly 15 seconds, which is substantially longer than the average length of time for a video summary. Splitting a video into fixed-length segments is a commonly used technique, but it may not align with the meaningful units of the video. Furthermore, abrupt cuts caused by such randomly chosen shot boundaries are disliked by viewers due to the sudden changes in motion.

To achieve sub-shot segmentation, we propose a technique incorporating editing rules to identify moments of no motion or matching motion speed and direction between consecutive frames. These segments are then referred to as "superframes" and compared to superpixels. Additionally, we propose a method inspired by recent advances in image segmentation. The quality of super frames is measured using an energy function E(S_j) as,

$$E(S_j) = \frac{1}{1+\gamma C_{cut}(S_j)} \cdot P_l(|S_j|) \quad (2)$$

where, P_l is a length prior for the super frames and C_{cut} is the cut cost. The value of parameter γ determines how much weightage should be given to the cut cost versus the length prior in the energy function. By decreasing the value of the parameter, the superframes become more homogeneous. The cut cost is defined as,

$$C_{cut}(S_j) = m_{in}(S_j) + m_{out}(S_j) \quad (3)$$

The formula calculates the estimated motion magnitude of the first $m_{in}(S_j)$ and last frame $m_{out}(S_j)$ in a superframe as and, respectively. We obtain these estimates using the KLT technique to track points in the video and compute the mean magnitude of the translation. The cost incurred by a superframe is lower if its boundaries align with frames that have little or no motion. By applying a log-normal distribution to a histogram of segment lengths of the human-made summary selections, the length prior P_l is learned.

By using hill-climbing optimization, we locally maximize the energy of Eq. (2). First, using the segment length $|S_j|$ = argmax(P_l), the super frames are initialized and dispersed uniformly throughout the video/shot (P_l). Then, to improve Eq. (3), we iteratively update the borders between two superframes. This results in segments with boundaries that are aligned in places that are appropriate for cuts. The optimization process is performed in a step-by-step manner, starting with a coarse approach and gradually refining the results. The boundaries are adjusted by one frame at a time. If the adjustment improves the overall score of the energy function given by Eq. (1) for the two relevant superframes, the change is accepted. We begin at the initial value and update iteratively until the algorithm converges. The optimization is then carried out once again after it is reduced by one frame. Only a few iterations are required for this optimization to converge because it is local.

The super frame's interestingness rating S_i is just the total of the frames' degree of interest:

$$I(S_i) = \sum_{k=n}^{m} i_k \quad (4)$$

where the beginning and ending frames of the superframe are denoted as 'n' and 'm'. Although other scoring strategies were tested, including taking the maximum or considering the size of the cluster, it was found that the simple sum used in this method was the most effective. Fig. 3 illustrates the key shot identification process.

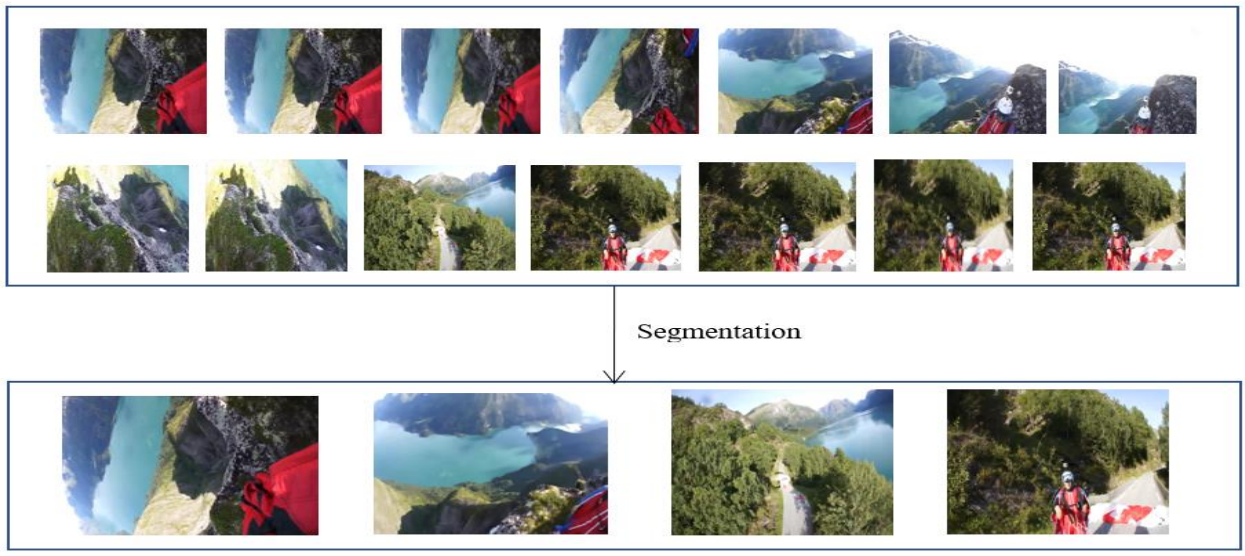


Fig. 3. Key shot identification of the video Base jumping in SUMMe dataset. The first row indicates the input video frame. The second row indicates the key shot identified from the input frames.

D. Key shot-based Summary Generation

Our objective is to find a subset of superframes, denoted as S, whose lengths are below a certain threshold (i.e., maximum), such that the sum of their interestingness scores is maximized.

$$\text{maximize} \sum_{i=1}^{n} x_i I(S_i) \quad (5)$$

X

$$\text{subject to} \sum_{i=1}^{n} x_i |S_i| \leq L_s \quad (6)$$

where, $x_i \in \{0, 1\}$ and $x_i = 1$ indicates that a super frame is selected. A summary is generated by combining the selected keyframes in a coherent and visually appealing way. The summary should provide an accurate representation of the original video's content while also being concise and easy to understand.

IV. Results and Discussion

A. Data Set

We evaluate our super frame segmentation on the most popular two datasets SUMMe and TVSum dataset. SUMMe (Summarization of Multiple Longer Videos) is a video summarization dataset that consists of 25 videos from YouTube. The videos are selected from different categories, such as sports, documentaries, and news. The TVSum dataset is a collection of 50 videos from various genres, such as news, documentaries, sports, and movies, suitable for video summarization research. This dataset, along with the SumMe dataset, includes multiple user annotations. To handle temporal redundancy and to comply with earlier efforts, we specifically downsampled all videos to 10 fps, initially shot at 30 fps, to minimize computation. The description of two video summarizing datasets is shown in Table I.

TABLE I. Description of Video Summarization Dataset Used

Dataset	Total videos	Content	Annotation	Duration(Min,Max,Average)
SumMe	25	User-generated Videos	Frame level score	38s, 324s, 146s
TVSum	50	Web videos	Frame level score	83s, 647s, 238s

B. Evaluation Metrics

For evaluation, we found F1-Score and Accuracy for the summarized video. The accuracy metric measures the percentage of correctly identified important frames or segments in the video summary compared to the ground truth summary. TP, TN, FP, and FN are employed in the context of binary picture segmentation. TP is the proportion of pixels in the expected binary picture and the ground truth that is properly classified as object pixels. The number of pixels in the anticipated binary picture as well as the ground truth that are properly classified as non-object pixels is known as TN. The number of pixels that are mistakenly classified as objects in the anticipated binary picture but are non-object pixels in the actual image is known as FP. FN is the number of pixels in the anticipated binary image that are mistakenly classified as non-object pixels but are object pixels in the actual picture. These metrics are used to determine how well a binary image segmentation algorithm performs, and they are frequently utilized to determine different evaluation criteria.

The formula for calculation accuracy, F1-score, precision and recall are discussed below.

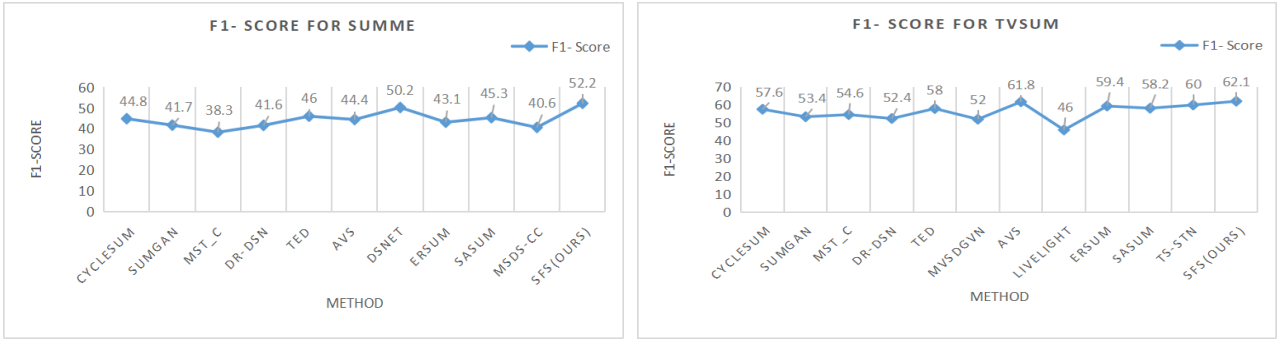
$$\text{Accuracy} = \frac{TP+TN}{FN+FP+TP+TN} \quad (7)$$

$$\text{Recall} = \frac{TP}{FN+TP} \quad (8)$$

$$\text{Precision} = \frac{TP}{FP+TP} \quad (9)$$

$$\text{F1- Score} = \frac{2*TP}{(2*TP+FP+FN)} \quad (10)$$

1) Our SFS approach was compared to other video summarization methods, including LiveLight [26], ERSUM [27], MSDS-CC [31], SUM-GAN[13], AVS [5], SASUM [28], DR-DSN [15], and TSSTN [29], on the SumMe and TVSum datasets. These methods were categorized into two groups: conventional and deep learning-based methods. The experimental results are presented in Table II under the canonical setting. The results show that the SFS approach outperforms all other methods, including state-of-the-art techniques, by at least 0.5% on both datasets. While methods such as MST_C, MSDS-CC, DR-DSN, and SUMGAN have the lowest F1-score, their performance lags behind that of the SFS approach by at least 4% on both SumMe and TVSum datasets. Fig. 4 shows the f1 score for each of the two datasets.



(a) (b)

Fig. 4. F1-Score (%) of both the datasets with other state-of-the-Art methods (a) TVSUM (b) SUMME.

TABLE II. PERFORMANCE MEASURE F1-SCORE(%) OF TVSUM AND SUMME DATASET WITH OTHER STATE-OF-THE-ART METHODS

Method	TVSUM	SUMME	Supervised/Unsupervised
CycleSum[13]	57.6	44.8	Unsupervised
SUMGAN[13]	53.4	41.7	Unsupervised
MST_C[4]	54.6	38.3	Unsupervised
DR-DSN[15]	52.4	41.6	Supervised
TED[2]	58	46	Supervised
mvsDGCN[3]	52.0	-	Supervised
AVS[5]	61.8	44.4	Supervised
LiveLight[27]	46.0	-	Unsupervised
ERSUM[28]	59.4	43.1	Supervised
SASUM[29]	58.2	45.3	Supervised
TS-STN[30]	60.0	-	Supervised
DSNet[31]	-	50.2	Supervised
MSDS-CC[32]	-	40.6	Unsupervised
SFS(ours)	62.1	52.2	Unsupervised

2) Comparison of Accuracy with other State-of-the-Art methods: We also use accuracy as the evaluation metric. The accuracy of a video summarization method is determined by calculating the percentage of significant frames or segments that are identified correctly in the generated summary compared to the ground truth summary. Table III and Fig. 5 present a comparison of the accuracy of our proposed method with other State-of-the-Art approaches. The experimental results demonstrate the superiority of our formulation.

TABLE III. COMPARISON OF ACCURACY METRIC WITH OTHER METHODS

Method	Methodology	Accuracy(%)	
		SUMME	TVSUM
CAVS [1]	The algorithm is developed to learn and update dictionaries of video features along with feature correlations	81.3	-
DR-DSN [15]	Dynamic graph node classification on videos is used to get the summary result.	89	90.6
SFS (ours)	Keyshot-based summary generation using superframe segmentation.	89.34	90.9852

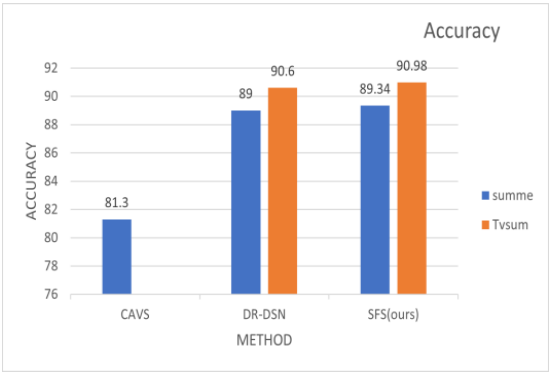


Fig. 5. Representative graph for the given dataset comparing with various methods.

3) We performed an investigation to assess the influence of long-range features, following a methodology comparable to the anchor-based approach. In this study, we analyzed the impact of using various feature extraction layers on performance metrics such as F-score, precision, and recall. The findings of this examination are illustrated in Table IV, indicating that our SFS technique surpasses the other temporal layers on both datasets, resulting in the most superior overall performance.

TABLE IV. COMPARISON OF F1-SCORE, PRECISION AND RECALL WITH ANOTHER STATE-OF-ART METHOD

Method	SUMME			TVSUM		
	F	P	R	F	P	R
LSTM	49.5	48.7	51.2	59.8	59.8	59.8
GCN[32]	50.5	50.0	51.3	59.8	59.8	59.8
Attention[33]	51.2	50.8	51.9	61.9	61.9	61.9
SFS(ours)	52.2	51.9	52.5	62.1	62.0	62.2

4) Exemplar key shot summaries: Exemplar key shot summaries are shown in Fig. 6. Video 47 discusses cleaning a dog's ears. We may observe that the key shot summaries of video 47 created by our SFS independently display the narrative details of washing a dog's ears. In the key shot summaries of video 47, our technique can skip a lot of unnecessary video shots. In comparison, DHAVS [34] and DR-DSN[15] key shots contain more frames that aren't significant. As a result, the suggested SFS can receive a higher F1 score.

Fig. 6. Generated summaries with F1-Score from video 47 in TVSum dataset.

C. Parameter Tuning

With varying γ values we found F1-Score for both the datasets. We set the initial data as γ=1 and then varied the values to get better performance evaluation for both datasets. The value sets of parameters γ are 0.4, 05, 0.7, 0.8. 1. 1.2 and 1.4 respectively. From Fig. 7, we observe that a rise in γ value leads to reduced performance.

Table V shows the varying F1-Score for SUMME and TVSUM datasets. The below table leads to the conclusion, for the value of 0.7 our model gives the highest F1-Score when compared with others for our datasets.

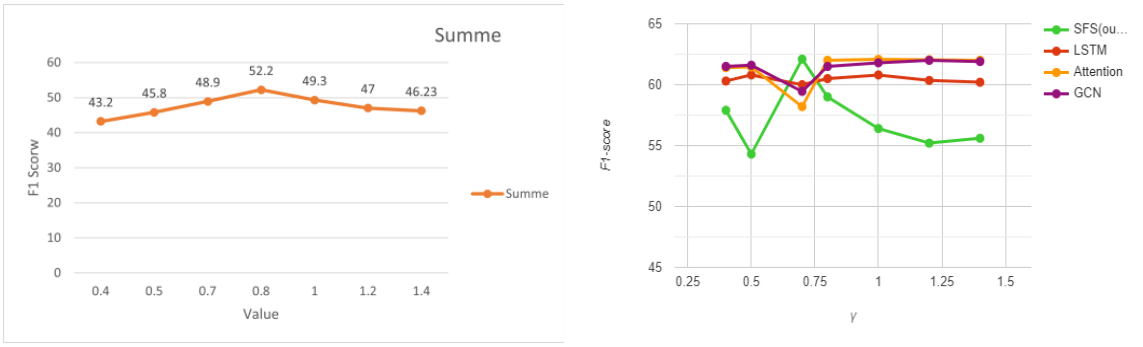
TABLE V. F1 SCORE FOR VARYING Γ VALUE (A) SUMME (B) TVSUM

Method	**Value**						
	0.4	**0.5**	**0.7**	**0.8**	**1**	**1.2**	**1.4**
LSTM	60.7	60.9	60.5	60	60.6	60.3	60.2
Attention [33]	61.2	61.4	61.8	61.9	62	61.9	61.9
GCN[32]	61.2	61.6	61.4	61.5	61.6	61.8	61.9
SFS(ours)	57.9	59.3	62.1	59	58.4	59.2	58.6

(a)

Value	**F1 Score**
0.4	43.2
0.5	45.8
0.7	48.9
0.8	52.2
1	49.3
1.2	47
1.4	46.23

(b)



(a) (b)

Fig. 7. Representative graph for both the datasets with varying γ values (a) SUMME (b) TVSUM.

V. CONCLUSION

The proposed approach in this study introduces a technique for segmenting user videos using temporal superframes and a method for generating informative video summaries. The main goal of this paper is to provide a better summary of surveillance video, which is helpful for law enforcement officials. So, we use a superframe segmentation process in which dividing the video into short, visually consistent segments called superframes and select representative frames from each superframe to construct the summary. In contrast to other methods, our approach can handle variations in camera movements and scene changes, which makes it different visual characteristics. We choose Accuracy, F1-score, Precision, and Recall as the assessment measures to compare our method fairly to existing video summarizing techniques. Our experimental findings demonstrate that the SFS method we proposed performs competitively on the SumMe and TVSum datasets. Furthermore, we conducted additional experiments to investigate the effect of the final summary length (L) on the performance metrics, including F-score, precision, and recall. Our results suggest that the best performance of our method can be achieved when the final summary length is set to 15% of the original video length on both datasets. This framework could be largely assistive for observing the activities of the students during their online exams. It could be a better alternative for the officials conducting the assessment and monitoring the suspicious activities of the students.

ACKNOWLEDGMENT

The paper is financed by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project № BG-RRP-2.004-0001-C01.

REFERENCES

[1] S. Zhang, Y. Zhu, A. Roy-Chowdhury, “Context-aware surveillance video summarization,” *IEEE Trans. Image process.*, vol. 25, no. 11, pp. 5469–5477, Nov. 2016, doi: 10.1109/TIP.2016.2601493.

[2] C. Huang, H. Wang. “A Novel key frames selection framework for comprehensive video summarization”, *IEEE Trans. Circuits And Sys. For Video Tech.*, vol. 30, no. 2, pp. 577-589, Feb. 2020, doi: 10.1109/TCSVT.2019.2890899.

[3] J. Wu, S. Zhong, Y. Liu. “Dynamic graph convolutional network for multi-video summarization”, *Elsevier Pattern Recog.*, vol. 107, no. 1, Art. no: 107382, Nov. 2020, doi: 10.1016/j.patcog.2020.107382.

[4] A. Sahu, A. Chowdhury, ”First person video summarization using different graph representations”, *Elsevier Pattern Recog. Lett.*, vol. 146, no. 1, pp. 185-192, Mar. 2021, doi: 10.1016/j.patrec.2021.03.013.

[5] Zh. Ji, K. Xiong, Y. Pang, X. Li. "Video summarization with attention-based encoder–decoder networks", *IEEE Trans. Circuits and Sys. for Video Tech.*, vol 30, no. 6, pp. 99, Aug. 2017, doi: 10.1109/TCSVT.2019.2904996.

[6] P.Kalaivani, S. Roomi. "Towards comprehensive understanding of event detection and video summarization approaches", *2017 Second Inter. Conf. on Recent Trends and Challe. in Comput. Models (ICRTCCM)*, pp. 61-66. IEEE, Feb. 2017, doi: 10.1109/ICRTCCM.2017.84.

[7] K. Kumar, D. D. Shrimankar, N. Singh. “Event BAGGING: A novel event summarization approach in multi-view surveillance video”, *Innova. in Electro., Signal Process. and Comm.(IESC), 2017 Inter. Conf.*, pp. 106-111, Apr 2017, doi: 10.1109/IESPC.2017.8071874.

[8] U. Damnjanovic, V. Fernandez, E. Izquierdo, J. Martinez. "Event detection and clustering for surveillance video summarization", *2008 Ninth Inter. Work. on Image Analy. for Multi. Inter. Serv.*, pp. 63-66. IEEE, May 2008, doi: 10.1109/WIAMIS.2008.53.

[9] S. Thomas, S. Gupta, V. Subramanian. "Perceptual video summarization—A new framework for video summarization", *IEEE Trans. Circuits and Sys. for Video Tech.*, vol. 27, no. 8, pp.1790-1802, Apr. 2016, doi: 10.1109/TCSVT.2016.2556558.

[10] Ch. Ngo, Y. Ma, H. Zhang. "Automatic video summarization by graph modeling", in *Procee. Ninth IEEE Inter. Conf. on Comp. Vision*, pp. 104-109. IEEE, Oct 2003, doi: 10.1109/ICCV.2003.1238320.

[11] Y. Cong, J. Yuan, J. Luo. "Towards scalable summarization of consumer videos via sparse dictionary selection", *IEEE Trans. Multi.*, vol. 14, no. 1 pp. 66-75, Sep 2011, doi: 10.1109/TMM.2011.2166951.

[12] K. Zhou, Y. Qiao, T. Xiang. "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward", in *Proc. of the AAAI Conf. on Arti. Intelli.*, vol. 32, no. 1, Apr. 2018, doi: 10.1609/aaai.v32i1.12255.

[13] B. Mahasseni, M. Lam, S. Todorovic, "Unsupervised video summarization with adversarial lstm networks", in *Proc. of the IEEE conf. Comp. Vision and Patt. Recog.*, pp. 202-211, Jul. 2017, doi: 10.1109/CVPR.2017.318.

[14] M. Rochan, Y. Wang. "Video summarization by learning from unpaired data", in *Proc. of the IEEE/CVF Conf. Comp. Vision and Patt. Recog.*, pp. 7902-7911, June. 2019, doi 10.1109/CVPR.2019.00809.

[15] K. Zhang, W. Chao, F. Sha, K. Grauman, "Video summarization with long short-term memory", in *Comp. Vision–ECCV 2016: 14th Euro. Conf., Amsterdam, The Netherlands, Oct. 11–14, 2016, Procee., Part VII 14*, pp. 766-782. Springer Inter. Publi., Oct. 2016, doi: 10.1007/978-3-319-46478-7_47.

[16] B. Gong, W. Chao, K. Grauman, F. Sha. "Diverse sequential subset selection for supervised video summarization", *Adv. in neural info. process. sys.*, pp. 2069-2077, Jan 2014.

[17] E. Bulut, T. Capin. “Key Frame Extraction from Motion Capture Data by Curve Saliency”, *Proceedings of 20th Annual Conference on Computer Animation and Social Agents.*, vol. 20, no. 5, Jun. 2007.

[18] B. Zhao, X. Li, X. Lu. "Hierarchical recurrent neural network for video summarization" In *Proceedings of the 25th ACM International Conference on Multimedia*, pp. 863-871. 2017.

[19] W. Wolf. "Key frame selection by motion analysis", in *IEEE Inter. conf. acoustics, speech, and sig. process. conf. proceedings*, vol. 2, pp. 1228-1231. IEEE, May. 1996, doi: 10.1109/ICASSP.1996.543588.

[20] C. Li, Y.T. Wu, S.S. Yu and T. Chen, “Motion-focusing key frame extraction and video summarization for lane surveillance system”, *16th IEEE International Conference on Image Processing (ICIP).*, pp. 7–10, Nov. 2009 doi: 10.1109/ICIP.2009.5413677.

[21] M. Ajmal, M. Naseer, F. Ahmad and A. Saleem. “Human Motion Trajectory Analysis Based Video Summarization”, *16th IEEE International Conference on Machine Learning and Applications (ICMLA).*, pp. 550-555, Dec. 2017 doi: 10.1109/ICMLA.2017.0-103.

[22] J. Almeida, R. D. S. Torres and N. J. Leite, “Rapid video summarization on compressed video", *IEEE International Symposium on Multimedia*., pp. 113- 120, Dec.2010, doi: 10.1109/ISM.2010.25.

[23] H.J. Zhang, J. Wu, D. Zhong and S.W. Smoliar, “An integrated system for content-based video retrieval and browsing”, *Pattern Recognition Elsevier.*, Vol. 30, pp.643–658, Apr. 1997, doi: 10.1016/S0031-3203(96)00109-4.

[24] F. Wang and C.W. Ngo, "Summarizing rushes videos by motion, object, and event understanding", IEEE Transactions on Multimedia., vol. 14, no. 1, pp.76-87, Aug. 2011, doi: 10.1109/TMM.2011.2165531.

[25] N. Baghel, S.C. Raikwar, C. Bhatnagar, "Image conditioned key frame-based video summarization using object detection," arXiv preprint arXiv: 2009.05269., Sep. 2020.

[26] B. Zhao and E.P. Xing, "Quasi real-time summarization for consumer videos", *Proceedings of the IEEE conference on computer vision and pattern recognition.*, pp. 2513-2520, Sep. 2014, doi: 10.1109/CVPR.2014.322.

[27] X. Li, B. Zhao, X. Lu. “A general framework for edited video and raw video summarization”, *IEEE Transactions on Image Processing.*, vol. 26, no. 8, pp. 3652–3664, Aug. 2017, doi:10.1109/TIP.2017.2695887.

[28] H. Wei, B. Ni, Y. Yan, H. Yu, X. Yang, C. Yao, “Video summarization via semantic attended networks”, *Proceedings of the AAAI conference on artificial intelligence.*, vol. 32, no. 1, 2018, pp. 216–223, doi: 10.1609/aaai.v32i1.11297.

[29] S. Huang, X. Li, Z. Zhang, F. Wu, and J. Han, “User-ranking video summarization with multi-stage spatio–temporal representation”, *IEEE Transactions on Image Processing.*, vol. 28, no. 6, pp. 2654–2664, Jun. 2019, doi: 10.1109/TIP.2018.2889265.

[30] W. Zhu, J. Lu, J. Li, J. Zhou, "Dsnet: A flexible detect-to-summarize network for video summarization", *IEEE Transactions on Image Processing.*, vol. 30, pp. 948-962, Dec. 2020, doi: 10.1109/TIP.2020.3039886.

[31] J. Meng, S. Wang, H. Wang, J. Yuan, Y.P. Tan, "Video summarization via multi-view representative selection", *Proceedings of the IEEE international conference on computer vision workshops.*, pp. 1189-1198, Jan. 2018, doi: 10.1109/ICCVW.2017.144.

[32] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” *arXiv preprint arXiv: 1609.02907*, Sep. 2016.

[33] A. Vaswant et al., "Attention is all you need", *Advances in neural information processing systems*, vol. 30. 2017, pp. 5998–6008.

[34] J. Lin, S.H. Zhong, and A. Fares, "Deep hierarchical LSTM networks with attention for video summarization", *Computers and Electrical Engineering Elsevier*, vol. 97, pp. 107618, Jan. 2022, doi: 10.1016/j.compeleceng.2021.107618.