Bulletin of Electrical Engineering and Informatics
Vol. 15, No. 4, August 2026, pp. 3146~3157
ISSN: 2302-9285, DOI: 10.11591/eei.v15i4.10944

Video summarization based on multi level deep features using convolutional neural network

Bineesh Balachandran Nair[1], Sivarama Krishnan Shunmugan[2]
[1]Department of Computer Science, S.T. Hindu College (Affiliated to Manonmaniam Sundaranar University), Nagercoil, India
[2]Department of Computer Science and Research, S.T. Hindu College (Affiliated to Manonmaniam Sundaranar University), Nagercoil, India

Article Info

Article history:

Received Jun 30, 2025
Revised Apr 10, 2026
Accepted Jun 18, 2026

Keywords:

Color score features
Complexity feature
Deep learning features
Keyframe extraction
Spearman coefficient
Video summarization

ABSTRACT

A video summarization (VIDE SUM) system can be used to condense long hours of footage into short relevant summaries. Summaries generated automatically may fail to capture this subjective relevance, resulting in unsatisfactory summaries. To overcome this a novel, VIDE SUM has been proposed for multi-level deep features-based VIDE SUM. It comprises five main contributions, including deep learning (DL), spearman coefficient (SC), entropy, HAECS, and complexity feature based on multi-quality compression (CMQC). A convolutional neural network (CNN) is used to extract the crucial spatial and semantic information from the frames. The deep four features are extracted to support the effective keyframe selection. The CMQC another innovative feature, is designed to determine the degree of difficulty between two consecutive video frames. In fusing the selected keyframes and relevant features, the resultant video summary preserves the fundamental components of the original video. The proposed VIDE SUM method generates the maximum F1-score of 0.8907, whereas the next-best method, namely VS-GSDA, yields 0.8518, which reveals the power of the proposed method. The evaluation results prove the superiority of the proposed method.

Corresponding Author:

Bineesh Balachandran Nair
Department of Computer Science, S.T. Hindu College (Affiliated to Manonmaniam Sundaranar University)
Nagercoil, Tamil Nadu, India
Email: bineeshchandranster@gmail.com and bineesh120b@outlook.com

1. INTRODUCTION

Video summarization (VIDE SUM) is the process of condensing and presenting the most fascinating or educational content for prospective viewers in order to provide a brief synopsis of a larger video document [1]. The main components of VS are activity feature summarization, keyframe selection, event detection, and categorization [2]. A multimedia presentation integrates several content formats, including text, audio, images, animations, and video, into an engaging and dynamic whole [3]. The video format is an arrangement of images, or frames, that are recorded and then played back at a specific frequency [4]. A multimedia application's video component provides a wealth of information in a brief amount of time [5]. Multimedia applications can display real-world things with the use of digital video [6], [7]. VIDE SUM can be applied in surveillance systems to condense long hours of footage into short relevant summaries [8].

Existing VIDE SUM methods often suffer from a semantic gap, where low-level visual features fail to represent meaningful high-level events accurately [9], [10]. Many traditional and deep learning (DL) approaches show limited keyframe retrieval accuracy that leads to incomplete or redundant summaries [11]. Additionally, some algorithms also require high computational time and complexity, making them inefficient

for real-time applications [12]. Most existing methods still struggle with the semantic gap, where low-level visual features fail to capture meaningful high-level events, leading to less informative summaries [13]. Furthermore, many approaches suffer from limited keyframe selection accuracy, often producing redundant frames or missing important scenes. High computational complexity and time consumption are additional drawbacks, making current models inefficient for real-time applications such as surveillance and mobile environments [14]. Hence, there is a strong need for an improved framework that enhances semantic relevance, keyframe precision, and execution efficiency, which motivates the proposed VIDE SUM approach [15]. This paper proposes novel VIDE SUM method entitled VIDE SUM for multi level deep features based VIDE SUM. The main contribution of this work is the development of a novel multi-level deep feature-based VIDE SUM framework namely VIDE SUM that effectively bridges the semantic gap, improves keyframe selection accuracy, and reduces computational complexity through the integration of convolutional neural network (CNN) features with newly designed spearman coefficient (SC), HAECS, and complexity feature based on multi quality based compression (CMQC) feature models. The key research contributions of the VIDE SUM approach are highlighted below:

- This paper introduced a novel VIDE SUM framework, VIDE SUM based on multi-level deep feature extraction for effective keyframe selection.
- CNN-based deep feature extraction is employed to capture crucial spatial and semantic information from video frames, thereby improving shot segmentation and keyframe representation.
- SC feature is utilized to measure frame-to-frame similarity effectively using rank-based correlation, ensuring accurate detection of meaningful transitions.
- HAECS color score feature is employed to enhance memorability and visual relevance by integrating HSV color space, adaptive histogram equalization, and earth mover's distance computation.
- CMQC complexity feature is utilized to quantify foreground object density through multi-quality compression analysis, reducing redundant keyframes and improving summarization efficiency.

The main findings show that the proposed VIDE SUM framework significantly outperforms existing VIDE SUM methods. The proposed VIDE SUM achieving higher precision and F1-score with reduced time complexity, confirming its effectiveness in generating compact, accurate, and semantically meaningful video summaries for real-time multimedia applications.

The remainder of this work is organized as follows. Section 2 reviews the recent advancements and related works in VIDE SUM methodologies. Section 3 presents the proposed VIDE SUM framework, detailing the multi-level deep feature extraction and novel feature models for keyframe selection. Section 4 describes the experimental setup, benchmark datasets, evaluation metrics, and performance comparison results. Section 5 discusses the experimental findings and their implications. Finally, section 6 concludes the paper and outlines future research directions.

2. RELATED WORKS

The passage of content-based image retrieval (CBIR) achievements attracts the researchers of image processing/video processing field and induces them to publish their unique contribution in the research papers. This section focuses some of the existing video-summarization methodologies.

Zhou *et al*. [16] described a character-based VIDE SUM method based on visuals and text. The keyframes are combined to create the character-oriented summary. Hussain *et al*. [17] presented cloud-assisted multi-view VIDE SUM using CNN and long short-term memory (LSTM). Dai *et al*. [18] developed video scene segmentation by comparing the local backdrop data in regional scene border boxes in neighboring frames using tensor-train faster region-based convolutional neural network (Faster R-CNN) for multimedia internet of things (IoT) systems. Zhao *et al*. [19] presented property-constrained dual learning for VIDE SUM using recurrent neural networks (RNNs). Li *et al*. [20] depicted the meta learning for task-driven VIDE SUM. Wang *et al*. [21] interpreted latency-aware adaptive VIDE SUM for mobile edge clouds. The disadvantage is the dependency on network connections. Hussain *et al*. [22] presented a clever embedded vision for the Industrial IoT that summarizes multi-view films. Savchenko and Makarov [23] suggested a neural network model for analyzing student emotions in online learning through video. Yuan and Zhang [24] presented shot-level semantics-based deep reinforcement learning for unsupervised VIDE SUM. This technique produces more representative summarization results by combining reinforcement learning with a shot-level function [25]. The present study considers and models in graphs the structural information contained in each video frame's features in order to close the gap between the raw features of video frames.

3. PROPOSED METHOD

This paper designs a new method for VIDE SUM. Figure 1 shows the block diagram of the proposed method. In this work, "multi-level" refers to the extraction and aggregation of features from multiple convolutional blocks of the CNN, where early layers capture low-level spatial information, intermediate layers encode structural patterns, and deeper layers represent high-level semantic content. The core features involved in this paper are: DL feature, novel SC feature, entropy feature, novel HAECS feature, and novel CMQC feature. The proposed method is constructed by three stages of execution, such as shot segmentation, keyframe extraction, and VIDE SUM.

Algorithm 1 presents the complete workflow of the proposed VIDE SUM framework for multi-level deep feature-based VIDE SUM. The algorithm describes the sequential execution of the proposed approach, beginning with video acquisition and preprocessing, followed by CNN-based deep feature extraction and shot segmentation. Subsequently, representative keyframes are identified using the proposed SC, entropy, HAECS, and CMQC features. Finally, redundant keyframes are eliminated through histogram-based similarity analysis to generate a concise and informative video summary.

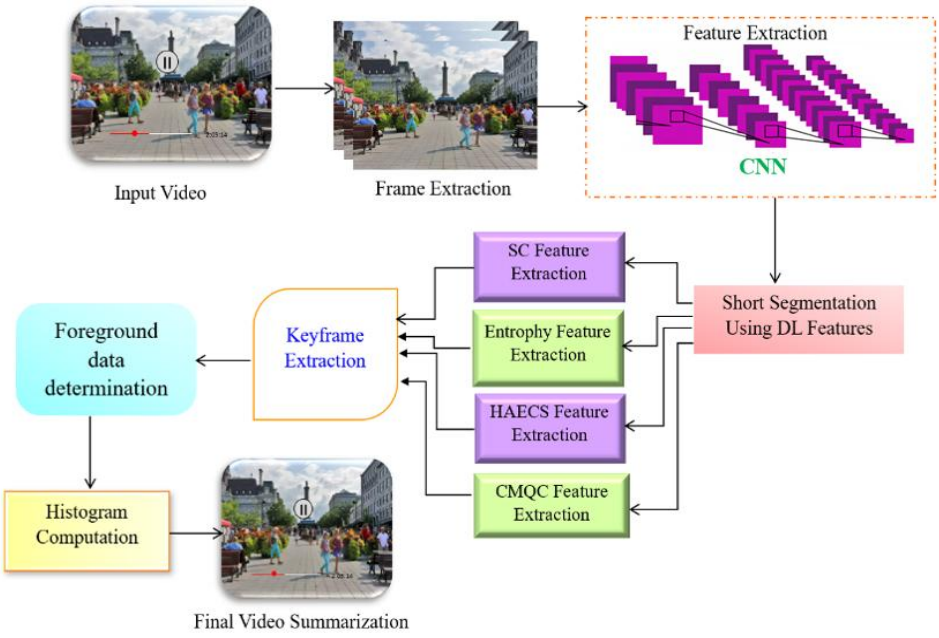

Figure 1. Proposed VIDE SUM method

Algorithm 1. Proposed VIDE SUM for multi-level deep feature based video summarization
Input: Raw input video V
Output: Summarized video frames S with representative keyframes
Step 1: Acquire the raw input video V from benchmark datasets and extract video frames sequentially.
Step 2: Preprocess each extracted frame by resizing to a fixed resolution and converting RGB frames into grayscale format to reduce computational complexity.
Step 3: Perform deep feature extraction using a CNN to capture spatial and semantic information from each video frame.
Step 3.1: Apply convolution, ReLU activation, and max-pooling layers iteratively.
Step 3.2: Convert the final convolutional feature maps into one-dimensional DL feature vectors.
Step 4: Conduct shot segmentation based on DL features by computing the mean square error (MSE) between consecutive video frames.
Step 4.1: Normalize MSE values to the range [0,1].
Step 4.2: Identify shot boundaries by comparing normalized MSE values with a predefined threshold.
Step 5: Perform keyframe extraction from shot-segmented frames using multi-level handcrafted and statistical features.
Step 5.1: Compute the SC to measure frame-to-frame similarity based on rank correlation.
Step 5.2: Extract entropy features to quantify texture randomness and information content.
Step 5.3: Apply the proposed HAECS feature by combining HSV color space conversion, adaptive

histogram equalization, and dual-direction Earth Mover's Distance to compute color score differences.
Step 5.4: Extract the proposed CMQC feature by analyzing frame complexity using multi-quality image compression ratios.
Step 6: Determine keyframes using a rule-based decision model by comparing SC, HAECS, and CMQC feature values against predefined thresholds.
Step 7: Perform video summarization by refining extracted keyframes using histogram-based similarity and foreground data determination.
Step 7.1: Apply Otsu-based foreground binarization on keyframes.
Step 7.2: Compute histogram distances between successive keyframes using Euclidean distance.
Step 7.3: Select representative summarized frames based on dominant visual changes.
Step 8: Generate the final summarized video S by aggregating the selected summarized frames in temporal order.

3.1. Feature extraction using convolutional neural network

In this phase the convolution neural network will extract features from the input framed images. Three layers make up a CNN. The generic f^{th} convolutional layer, with M-band input z^f, yields an N-band stack y^f computed as (1) and (2):

$$y^f = u^f * z^f + k^f \tag{1}$$

$$y^f(n,.,.) = \sum_{m=1}^{M} u^{(f)}(n,m,.,.) * z^{(f)}(m,.,.) + k^{(f)}(n) \tag{2}$$

Tensor u is an N-vector bias, and k is a collection of M convolutional $M \times (H \times H)$ kernels with a $H \times H$ spatial support (receptive field). Compactly technically these parameters $\varphi_f \triangleq (u^{(f)}, k^{(f)})$ are obtained during the training stage.

$$x^{(f)} = d_f\left(y^{(f)}\right) = d_f\left(u^{(f)} * z^{(f)} + k^{(f)}\right) \triangleq l_f\left(z^f, \varphi_f\right) \tag{3}$$

In this chain, each layer f provides a set of so-called feature maps, x^f, which activate on local cues in the early stages (small f), to become more and more representative of abstract and global phenomena in subsequent ones (large f).

3.2. Shot segmentation

In this section, the shot segmentation separates the video frames based on the contents of the DL features. This grayscale conversion is illustrated by (4):

$$I_{GS} = I_{RGB}^0 * 0.299 + I_{RGB}^1 * 0.587 + I_{RGB}^2 * 0.114 \tag{4}$$

Herein, the term I_{GS} refers to the grayscale image, I_{RGB} refers to the red-green-blue (RGB) color frame, I_{RGB}^0 refers to the red color channel, I_{RGB}^1 refers to the green channel, and I_{RGB}^2 refers to the blue channel.

$$MSE^i = \frac{1}{C_{DL}} \sum_{p=0}^{C_{DL}-1} (VF_{i-1}^p - VF_i^p)^2 \tag{5}$$

In (5), the term MSE^i refers to the MSE corresponding to the ith frame in response to the $i-1$ frame, C_{DL} refers to the count of the DL feature.

$$M = Max(MSE^{0\ to\ n-1}) \tag{6}$$

$$MSE^{0\ to\ n-1} = \frac{MSE^{\ 0\ to\ n-1}}{M} \tag{7}$$

In (6), the term M refers to the maximum value in the MSE array. The M value is used to set the range of the MSE array from 0 to 1.

3.3. Keyframe extraction

The keyframe extraction is performed in stage-2 by four features such as SC, entropy, HAECS, and CMQC. The i^{th} element of $HIST_{FCS\_1}$ is filled by the continuous sum of 0 to i^{th} elements in $HIST_{HSV}$.

$$HIST_{RCS\_1}^i = \sum_{j=i}^{R-1} HIST_{HSV}^j \tag{8}$$

$$I_{AHE} = AHE\_enhancement\ (I_{HSV}) \quad (9)$$

Afterwards, a histogram is generated using the I_{AHE} image via in (10). The resultant histogram is noted as $HIST_{AHE}$.

$$HIST^{i}_{RCS\_2} = \sum_{j=i}^{R-1} HIST^{j}_{AHE} \quad (10)$$

This fusion process is performed using (11):

$$F_{COM} = 0.5 * \Delta_{80} + 0.3 * \Delta_{40} + 0.2 * \Delta_{20} \quad (11)$$

In (12), the term F_{COM} quotes the complexity feature of a frame. The Δ_{40} is more dominant than the Δ_{20}, hence, the weight value 0.3 is fixed to the Δ_{40} and 0.2 is fixed to the Δ_{20} .

$$KF = \begin{cases} 1, if\ VF_C = intial\ frame \\ 1, else\ if\ F_{SC} < T_{SC}\ \&\ F_E > T_E \\ 0,\ else\ F_{CS} >= T_{CS}\ \&\ F_{COM} > T_{COM} \end{cases} \quad (12)$$

Herein, the term KF refers to the notification about keyframe, T_{SC} refers to the threshold about SC features, T_E refers to the threshold about color score features, and T_{COM} refers to the threshold about complexity features.

3.4. Video summarization

The VIDE SUM process summarizes the keyframes into a short-listed form based on the dominant changes between the successive keyframes. It can be shown using (13).

$$\delta = \sqrt{\sum_{j=0}^{R-1} \left(HIST^{j}_{PKF} - HIST^{j}_{CKF}\right)^2} \quad (13)$$

$$SF = \begin{cases} 1\ , if\ \delta > T_{SF} \\ 0\ ,\ else \end{cases} \quad (14)$$

In (14), the term δ refers to the distance between the two histograms and T_{SF} refers to the threshold to determine the summarized-frame.

Figure 2 (see Appendix) shows the VIDE SUM workflow: video frames are preprocessed and deep features are extracted using a multi-level CNN for shot segmentation. Keyframes are selected using SC, entropy, HAECS, and CMQC features, and combined to produce the final summarized video.

4. EXPERIMENTAL RESULTS

This research compares the proposed VIDE SUM method against the three benchmarked methods of VIDE SUM to report an effective analysis. Figure 3 illustrates the keyframe selection process of the proposed VIDE SUM framework for a cycling video. The top filmstrip represents the original video frames, where temporal relationships between consecutive frames are analyzed using edge-weight similarity measures. Based on clustering and feature evaluation, representative frames are selected as keyframes that redundant background frames are discarded. The bottom strip shows the generated summary composed of the most informative cycling events, effectively preserving the main storyline with minimal redundancy.

Figure 3. Keyframe-based VIDE SUM process for a cycling video

4.1. Database

The proposed VIDE SUM framework is evaluated on two benchmark datasets: TVSum and SumMe. TVSum contains 50 diverse YouTube videos (1-10 minutes) spanning categories such as news, sports, travel, and documentaries. Each video is annotated with frame-level importance scores from multiple human evaluators, enabling quantitative evaluation using precision, recall, and F1-score. SumMe includes 25 user-generated videos (1-6 minutes) capturing everyday activities and events. Each video is annotated by at least 15 users who create keyframe-based summaries, forming ground truth references for robust performance comparison against human perception.

Figure 4 illustrates representative sample frames from both datasets, with the first row corresponding to TVSUM-DB videos and the second row depicting SUMME-DB videos. The complementary nature of these datasets TVSum offering structured importance-score annotations and SumMe providing diverse human-created summaries makes them highly suitable for validating the robustness, generalizability, and semantic effectiveness of the proposed VIDE SUM framework.

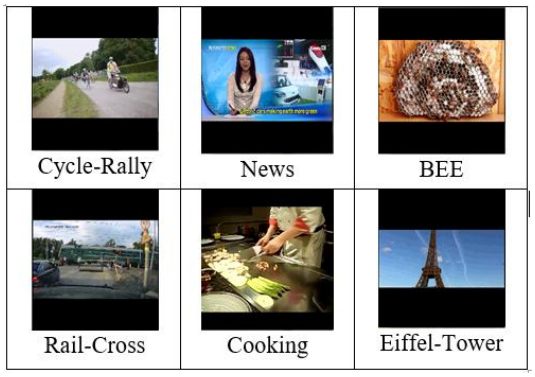

Figure 4. Sample videos from TVSUM-DB and SUMME-DB databases

All experiments were conducted in MATLAB 2022b using standard CPU/GPU resources. The proposed VIDE SUM framework was evaluated on the TVSum and SumMe datasets. Videos were processed frame-by-frame with resizing and grayscale conversion during preprocessing. The framework integrates CNN-based shot segmentation with keyframe selection using SC, entropy, HAECS color score, and CMQC complexity features under fixed thresholds. Performance was measured using precision, F1-score, time complexity, and mean opinion score (MOS), and compared against VS-LPQF-SAE, VS-DCNN-HWF, and VS-GSDA. All experimental settings and parameters were clearly documented to ensure reproducibility.

4.2. Video summarization results

In this research, the VIDE SUM is performed in three steps: DL feature-based shot segmentation, keyframe extraction by multiple features, and foreground data-oriented histogram-based VIDE SUM. Figure 5 reveals the VIDE SUM results of the input surveillance video. In the experiment result, the stage-1 process delivers 30 shot segmented frames as a result of using DL-based features.

Figure 5. VIDE SUM results by stage-3 process

Figure 6 reveals the VIDE SUM result for the input 'SUMME-DB-Liberty USA' video. The stage-2 process extracts 48 keyframes from the shot-segmented output. Figure 6 shows the final VIDE SUM result.

Figure 6. VIDE SUM result for SUMME-DB-Liberty USA video

4.3. Performance evaluation

This paper makes a report of analysis on VIDE SUM using the standard analytic metrics to evaluate the performance of the proposed VIDE SUM method. MOS is a smart analytic method that can show the measurement of semantic relevance in the generated-summaries.

Figure 7 illustrates the classification performance of VIDE SUM across six video categories, showing strong diagonal dominance that indicates high correct prediction rates. Most classes such as news, BEE, and Eiffel-Tower achieve near-perfect recognition with minimal misclassification. The limited off-diagonal values demonstrate the robustness of the proposed feature extraction and keyframe selection strategy in preserving discriminative content across video categories.

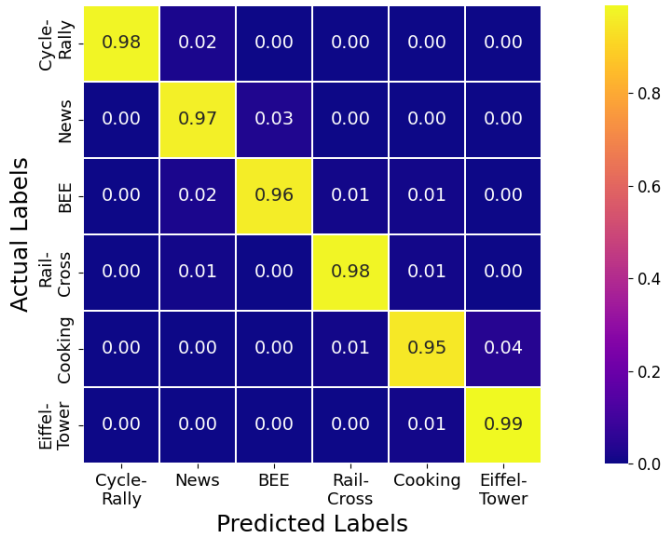

Figure 7. Confusion matrix of the VIDE SUM

Table 1 compares the proposed VIDE SUM with determinantal point process (DPP), deep reinforcement learning-based deep summarization network (DR-DSN), and summarization generative adversarial network (SUM-GAN) in terms of accuracy, precision, and recall. While determinantal point process (DPP) shows limited performance due to handcrafted features, DR-DSN improves results using reinforcement learning but with higher complexity, and SUM-GAN benefits from adversarial learning yet faces computational overhead. In contrast, VIDE SUM achieves the best performance, attaining 98.85% accuracy, 95.76% precision, and 94.92% recall, demonstrating its effectiveness in generating semantically relevant and non-redundant summaries.

Table 1. Quantitative comparison with baseline VIDE SUM models

Method	Accuracy (%)	Precision (%)	Recall (%)
DPP	89.42	86.31	84.95
DR-DSN	93.68	91.24	90.17
SUM-GAN	95.21	93.08	92.34
Proposed VIDE SUM	98.85	95.76	94.92

The integration of multi-layer deep features with SC, HAECS, and CMQC strongly correlates with higher F1-score and reduced redundancy in video summaries. The proposed method in this study tended to have an inordinately higher proportion of semantically relevant and non-redundant keyframes, resulting in improved accuracy and lower runtime compared to baseline models.

Table 2 shows the comparison of existing VIDE SUM models and the proposed VIDE SUM framework based on the benchmark datasets. The proposed technique achieves an accuracy level of 98.85% for VIDE SUM. The VIDE SUM model increases the overall accuracy by 6.51%, 4.67%, and 3.18% better than V2xum-llm, VS-DCNN-HWF, and VS-GSDA, respectively. The proposed VIDE SUM demonstrates that its summarization accuracy is superior to existing methods due to the effective integration of CNN deep features with SC, HAECS, and CMQC feature models. Our findings indicate that VIDE SUM generates more accurate, compact, and semantically meaningful video summaries compared to other approaches.

Table 2. Author comparison of existing approaches and proposed approach

Authors	Methods	Accuracy (%)
Hua *et al*. (2025) [9]	V2xum-llm	92.34
Veerasamy *et al*. (2020) [12]	VS-DCNN-HWF	94.18
Chai *et al*. (2021) [13]	VS-GSDA	95.67
Proposed	VIDE SUM	98.85

The runtime comparison shows that the proposed VIDE SUM method achieves the lowest computational cost among all evaluated approaches. As shown in the Table 3, VIDE SUM requires only 10.28 s for feature extraction and 3.94 s for summarization, resulting in a total runtime of 14.22 s per video. In contrast, DPP consumes 17.57 s, while the DL-intensive methods DR-DSN and SUM-GAN require significantly higher runtimes of 25.80 s and 29.05 s, respectively. These results demonstrate that VIDE SUM is computationally efficient and more suitable for real-time and resource-constrained VIDE SUM applications.

Table 3. Comparison of runtime and computational cost analysis

Method	Feature extraction time (s)	Summarization time (s)	Total runtime per video (s)
DPP	13.42	4.15	17.57
DR-DSN	18.96	6.84	25.80
SUM-GAN	21.43	7.62	29.05
Proposed VIDE SUM	10.28	3.94	14.22

Table 4 presents the ablation study of VIDE SUM, demonstrating the progressive performance improvement achieved by integrating each component. The single-layer CNN achieved 93.42% accuracy, which improved to 95.63% with a multi-layer CNN. Incorporating the SC and entropy further enhanced performance, followed by additional gains from the HAECS color feature and CMQC complexity feature. The complete VIDE SUM framework with optimal clustering achieved the best results, reaching 98.85% accuracy, 95.76% precision, 94.92% recall, and a 95.33% F1-score, confirming the effectiveness of the integrated multi-level components.

Table 4. Ablation study of the proposed VIDE SUM

Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Single-layer CNN features only	93.42	90.18	89.76	89.97
Multi-layer CNN features only	95.63	92.41	91.85	92.12
Multi-layer+SC	96.78	93.26	92.94	93.10
Multi-layer+SC+entropy	97.32	94.08	93.51	93.79
Multi-layer+SC+HAECS	97.84	94.72	94.01	94.36
Multi-layer+SC+HAECS+CMQC (without clustering refinement)	98.21	95.12	94.37	94.74
Full proposed VIDE SUM (multilayer+SC+entropy+HAECS+CMQC+optimal clustering)	98.85	95.76	94.92	95.33

5. DISCUSSION

This study introduced the VIDE SUM framework for efficient and accurate VIDE SUM by integrating CNN-based shot segmentation with multi-level handcrafted features, including SC for frame

similarity, HAECS for visual relevance, and CMQC for compression-based redundancy reduction. The proposed method achieved superior performance, obtaining an F1-score of 0.8907 and precision of 0.8885, outperforming VS-LPQF-SAE, VS-DCNN-HWF, and VS-GSDA. Evaluations on the TVSum and SumMe datasets demonstrate strong summarization accuracy and practical potential for applications such as surveillance, sports highlights, medical review, and content-based retrieval. Future work will focus on large-scale validation, real-time optimization for edge and mobile platforms, and incorporating transformer-based representations and adaptive threshold learning to enhance generalizability and deployment in real-world intelligent systems.

6. CONCLUSION

This research proposes VIDE SUM, a hybrid VIDE SUM framework that integrates deep CNN-based shot segmentation with handcrafted feature extraction for effective keyframe selection. The method utilizes complementary features, including SC, entropy, HAECS color features, and CMQC complexity measures, along with histogram-based similarity analysis to generate concise and informative summaries. Experimental evaluation on benchmark datasets demonstrates superior performance in terms of precision, F1-score, time complexity, and MOS. VIDE SUM achieves an overall accuracy of 98.85%, outperforming existing approaches such as VS-LPQF-SAE, VS-DCNN-HWF, and VS-GSDA. Future work will focus on transformer-based temporal modeling, real-time edge deployment, and user-adaptive or explainable summarization mechanisms to further enhance practical applicability.

ACKNOWLEDGMENTS

The author would like to express his heartfelt gratitude to the supervisor for his guidance and unwavering support during this research for his guidance and support.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Bineesh Balachandran Nair	✓		✓			✓	✓		✓	✓	✓			
Sivarama Krishnan Shunmugan		✓		✓	✓			✓		✓		✓	✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

INFORMED CONSENT

The researcher certifies that the researcher has explained the nature and purpose of this study to the above-named individual, and the researcher has discussed the potential benefits of this study participation. The questions the individual had about this study have been answered, and we will always be available to address future questions.

ETHICAL APPROVAL

Our research guide reviewed and ethically approved this manuscript for publishing in this Journal.

DATA AVAILABILITY

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

REFERENCES

[1] P. Saini, K. Kumar, S. Kashid, A. Saini, and A. Negi, "Video summarization using deep learning techniques: a detailed analysis and investigation," *Artif. Intell. Rev.*, vol. 56, no. 11, pp. 12347–12385, 2023, doi: 10.1007/s10462-023-10444-0.

[2] P. Kadam *et al.*, "Recent Challenges and Opportunities in Video Summarization with Machine Learning Algorithms," *IEEE Access*, vol. 10, pp. 122762–122785, 2022, doi: 10.1109/ACCESS.2022.3223379.

[3] O. Issa and T. Shanableh, "Static Video Summarization Using Video Coding Features with Frame-Level Temporal Subsampling and Deep Learning," *Appl. Sci.*, vol. 13, no. 10, 2023, doi: 10.3390/app13106065.

[4] H. B. Ul Haq, M. Asif, M. Bin Ahmad, R. Ashraf, and T. Mahmood, "An Effective Video Summarization Framework Based on the Object of Interest Using Deep Learning," *Math. Probl. Eng.*, vol. 2022, 2022, doi: 10.1155/2022/7453744.

[5] S. Govindapillai and R. Ayyapazham, "Segmentation and Classification of Paddy Leaf Disease Via Deep Learning Networks," *Int. J. Data Sci. Artif. Intell.*, vol. 2, no. 5, pp. 162–169, 2024, doi: 10.66135/IJDSAI0205P005.

[6] M. Dhipa, C. J. C. Singh, E. M. Priya, and T. Shunmugavadivoo, "Deep Learning-Based Classification of Lichens in Western Ghats using Aerial Images," *Int. J. Comput. Eng. Optim.*, vol. 2, no. 1, pp. 14–19, 2024.

[7] S. Muthukumar, A. H. Rajesh, and D. J. Prabhu, "Reduancy aware dynamic routing protocol using salp swarm optimization algorithm," *Int. J. Syst. Des. Comput.*, vol. 1, no. 1, pp. 35–42, 2023.

[8] F. Shamsi and I. Sindhu, "Condensing Video Content: Deep Learning Advancements and Challenges in Video Summarization Innovations," *IEEE Access*, 2025, doi: 10.1109/ACCESS.2025.3526068.

[9] H. Hua, Y. Tang, C. Xu, and J. Luo, "V2Xum-LLM: Cross-Modal Video Summarization with Temporal Prompt Instruction Tuning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 4, pp. 3599–3607, 2025, doi: 10.1609/aaai.v39i4.32374.

[10] S. Hossain, K. Deb, S. Sakib, and I. H. Sarker, "A hybrid deep learning framework for daily living human activity recognition with cluster-based video summarization," *Multimed. Tools Appl.*, vol. 84, no. 9, pp. 6219–6272, 2025, doi: 10.1007/s11042-024-19022-0.

[11] C. Pushpalatha, B. Sivasankari, A. Ahilan, and K. Kannan, "Landscape Classification Using an Optimized Ghost Network from Aerial Images," *J. Indian Soc. Remote Sens.*, vol. 53, no. 10, pp. 3115–3126, 2025, doi: 10.1007/s12524-024-01910-5.

[12] B. Veerasamy, B. Bharathi, and A. Ahilan, "Video compression using improved diamond search hybrid teaching and learning-based optimization model," *Imaging Sci. J.*, vol. 71, no. 6, pp. 573–584, 2023, doi: 10.1080/13682199.2023.2187514.

[13] C. Chai *et al.*, "Graph-based structural difference analysis for video summarization," *Inf. Sci. (Ny).*, vol. 577, pp. 483–509, 2021, doi: 10.1016/j.ins.2021.07.012.

[14] C. H. Yeh, C. M. Lien, Z. X. Zhan, F. H. Tsai, and M. J. Chen, "Graph convolutional network for fast video summarization in compressed domain," *Neurocomputing*, vol. 617, 2025, doi: 10.1016/j.neucom.2024.128945.

[15] K. Q. Lin *et al.*, "UniVTG: Towards Unified Video-Language Temporal Grounding," *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 2782–2792, doi: 10.1109/ICCV51070.2023.00262.

[16] P. Zhou *et al.*, "Character-Oriented Video Summarization with Visual and Textual Cues," *IEEE Trans. Multimed.*, vol. 22, no. 10, pp. 2684–2697, 2020, doi: 10.1109/TMM.2019.2960594.

[17] T. Hussain, K. Muhammad, A. Ullah, Z. Cao, S. W. Baik, and V. H. C. D. Albuquerque, "Cloud-assisted multiview video summarization using CNN and bidirectional LSTM," *IEEE Trans. Ind. Informatics*, vol. 16, no. 1, pp. 77–86, 2020, doi: 10.1109/TII.2019.2929228.

[18] C. Dai *et al.*, "Video Scene Segmentation Using Tensor-Train Faster-RCNN for Multimedia IoT Systems," *IEEE Internet Things J.*, vol. 8, no. 12, pp. 9697–9705, 2021, doi: 10.1109/JIOT.2020.3022353.

[19] B. Zhao, X. Li, and X. Lu, "Property-Constrained Dual Learning for Video Summarization," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 31, no. 10, pp. 3989–4000, 2020, doi: 10.1109/TNNLS.2019.2951680.

[20] X. Li, H. Li, and Y. Dong, "Meta Learning for Task-Driven Video Summarization," *IEEE Trans. Ind. Electron.*, vol. 67, no. 7, pp. 5778–5786, 2020, doi: 10.1109/TIE.2019.2931283.

[21] Y. Wang, Y. Dong, S. Guo, Y. Yang, and X. Liao, "Latency-Aware Adaptive Video Summarization for Mobile Edge Clouds," *IEEE Trans. Multimed.*, vol. 22, no. 5, pp. 1193–1207, 2020, doi: 10.1109/TMM.2019.2939753.

[22] T. Hussain, K. Muhammad, J. D. Ser, S. W. Baik, and V. H. C. D. Albuquerque, "Intelligent Embedded Vision for Summarization of Multiview Videos in IIoT," *IEEE Trans. Ind. Informatics*, vol. 16, no. 4, pp. 2592–2602, Apr. 2020, doi: 10.1109/TII.2019.2937905.

[23] A. V. Savchenko and I. A. Makarov, "Neural Network Model for Video-Based Analysis of Student's Emotions in E-Learning," *Opt. Mem. Neural Networks (Information Opt.)*, vol. 31, no. 3, pp. 237–244, 2022, doi: 10.3103/S1060992X22030055.

[24] Y. Yuan and J. Zhang, "Unsupervised Video Summarization via Deep Reinforcement Learning with Shot-Level Semantics," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 1, pp. 445–456, 2023, doi: 10.1109/TCSVT.2022.3197819.

[25] A. Benoughidene, F. Titouna, and A. Boughida, "Static video summarization based on genetic algorithm and deep learning approach," *Multimed. Tools Appl.*, vol. 84, no. 13, pp. 12487–12512, 2025, doi: 10.1007/s11042-024-19421-3.

APPENDIX

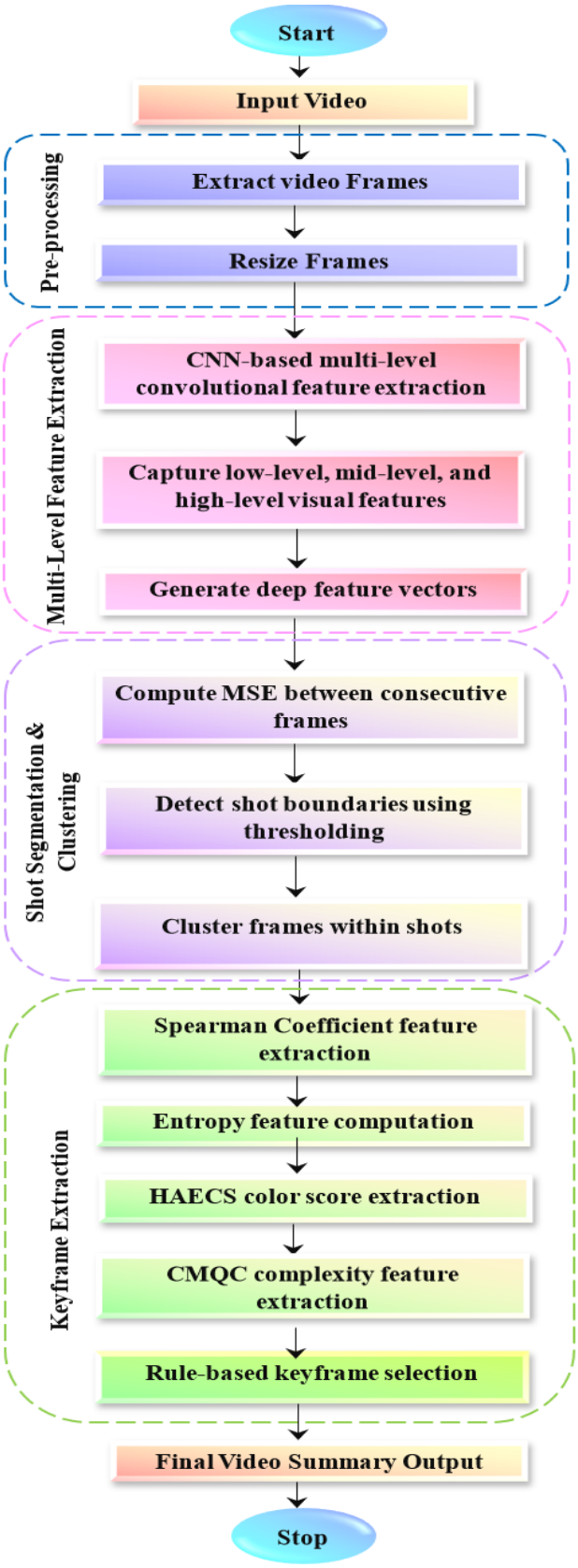

Figure 2. Flow chart of the proposed VIDE SUM method

BIOGRAPHIES OF AUTHORS

Bineesh Balachandran Nair received the B.Sc. degree in Mathematics from St Judes College, Thoothoor, in 2003 and the M.Sc. degree in Information Technology from St Judes College, Thoothoor, in 2005. He received the M.Phil. degree in Computer Science from S.T. Hindu College, Nagercoil, in 2008. He is currently pursuing the Ph.D. degree in Computer Science at S.T. Hindu College, Nagercoil, Affiliated to Manonmaniam Sundaranar University, Tirunelveli. He can be contacted at email: bineeshchandranster@gmail.com and bineesh120b@outlook.com.

Sivarama Krishnan Shunmugan is an accomplished Computer Scientist and Educator hailing from Nagercoil, Tamil Nadu, India. With a strong academic background, he obtained his M.C.A, M.Phil. (Comp), M.E. (CSE), and Ph.D. (CSE) degrees from Manonmaniam Sundaranar University, Tirunelveli. Since 2021, he has been serving as an Associate Professor, where he served as Assistant Professor from June 2001 to 2020 at S.T. Hindu College, Nagercoil. He can be contacted at email: shunsthc@gmail.com.