INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 4, April 2026 | ISSN: 2456-4184 | IJNRD.ORG

SPATIO- TEMPORAL VIDEO SUMMARIZATION SYSTEM

K.Chitra(Assistant Professor)
Department of Computer Science and Engineering
Bharath Institute of Higher Education and Research
Chennai, India
chitraashwin9896@gmail.com

Gali Essaan
Department of Computer Science and Engineering
Bharath Institute of Higher Education and Research
Chennai, India
eessan2005@gmail.com

Eede Venkata Rajesh
Department of Computer Science and Engineering
Bharath Institute of Higher Education and Research
Chennai, India
rajeshede208@gmail.com

Gaddam Manoj Kumar Reddy
Department of Computer Science and Engineering
Bharath Institute of Higher Education and Research
Chennai, India
gaddammanojkumarreddy@gmail.com

Gaddam Sanjay Kumar Reddy
Department of Computer Science and Engineering
Bharath Institute of Higher Education and Research Chennai, India
sanjayreddy2004@gmail.com

Abstract ***The rapid increase in video content across various platforms has made manual analysis both time-consuming and impractical. To address this challenge, this paper presents a spatio-temporal video summarization system designed to automatically generate concise and informative summaries from long videos. Work-loads. Overall, the proposed architecture establishes a cost-efficient and scalable alternative to existing applicant tracking solutions. The proposed approach focuses on capturing both the visual details within individual frames and the temporal relationships across sequences of frames. Spatial features are extracted using a ResNet-50 model, while temporal patterns are learned through a Bidirectional LSTM network. In addition, an attention mechanism is incorporated to identify and prioritize the most relevant portions of the video by assigning importance scores to different frames. The system is implemented as a complete web-based pipeline, enabling users to upload videos and obtain summarized outputs efficiently. The generated summaries retain the essential content while significantly reducing the overall video length, making the approach practical for real-world applications.***

Keywords: ***Video Summarization ,Deep Learning, Spatal-Temporal Analysis, Frame Importance Detection, Artificial Intelligence***.

I. Introduction

With the rapid growth of digital media platforms, video content has become one of the most dominant forms of information sharing. From educational lectures to entertainment and surveillance footage, large volumes of video data are generated every day. However, manually watching and analyzing lengthy videos is both time-consuming and inefficient, especially when only a small portion of the content is truly important Traditional video summarization techniques often focus on specific domains or rely on simple heuristics, which limits their effectiveness in real-world scenarios. In addition, many existing solutions remain confined to research environments and lack practical deployment for everyday users. This creates a gap between advanced video analysis techniques and their usability in real applications.

To address this issue, we propose a spatio-temporal video summarization system that combines deep learning techniques with a deployable web-based framework. The system is designed to capture both the visual characteristics of individual frames and the temporal relationships across video sequences. By integrating spatial feature extraction using ResNet-50 and temporal modeling through a Bidirectional LSTM network, the proposed approach generates concise and meaningful summaries automatically.

The primary objective of this work is to provide an efficient and user-friendly solution that reduces the effort required to analyze long videos while preserving essential information. The system is implemented as a complete pipeline, allowing users to upload videos and obtain

To address these challenges, this paper proposes a spatio-temporal video summarization system that integrates deep learning techniques with a practical deployment framework. The system leverages ResNet-50 for extracting spatial features from video frames and employs a Bidirectional LSTM network to model temporal dependencies across frame sequences. An attention mechanism is incorporated to assign importance scores, enabling the system to focus on the most informative parts of the video.

Unlike many existing approaches, the proposed system is implemented as a complete web-based application, allowing users to upload videos and obtain summarized outputs in an efficient and user-friendly manner. This not only improves accessibility but also bridges the gap between theoretical research and real-world usability.

II. RELATED WORK

Video summarization has been widely studied in recent years, with various approaches proposed to extract important content from videos[1]. Early methods primarily relied on low-level features such as color histograms, motion vectors, and shot boundary detection. While these techniques were computationally simple, they often failed to capture the semantic importance of video content. With the advancement of machine learning, more sophisticated approaches have been introduced. Some studies utilize clustering techniques to group similar frames, while others apply supervised learning models to identify keyframes. More recently, deep learning-based methods have gained attention due to their ability to learn complex patterns directly from data.

Convolutional Neural Networks (CNNs) have been widely used for extracting spatial features, while Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) models, are effective in capturing temporal dependencies. However, many of these approaches either focus only on spatial or temporal aspects, leading to incomplete representations[2].

In contrast, the proposed system combines both spatial and temporal analysis, along with an attention mechanism to improve the selection of important frames. Furthermore, unlike many existing works, the system is implemented as a practical web-based application, making it accessible for real-world usage.

III. SYSTEM ARCHITECTURE

The Proposed video summarization system features five core components in its architecture:

1) 1 Videos Upload Interface
2) 2 Frame Extraction Module
3) 3 Feature Extraction Module
4) 4 Frame Importance Scoring
5) 5 Summary Video Generation

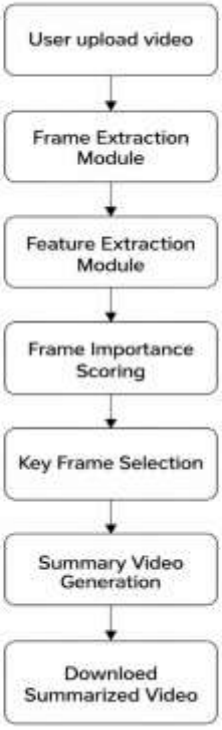


Figure 1 System Block Diagram

This block diagram represents the workflow of a video summarization system. It begins with the user uploading a video, after which frames are extracted for further analysis. Important visual features are then identified, and each frame is assigned a relevance score based on its significance. The system selects key frames using these scores and finally generates a concise summarized video. This process helps reduce video length while preserving essential content.

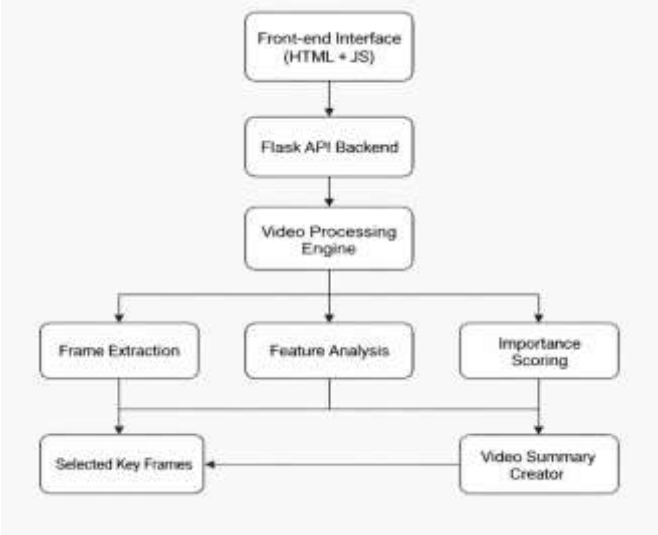


Figure 2 System Architecture

The system architecture illustrates how different components work together to generate a summarized video from raw input. It starts with a user interface where the video is uploaded and sent to the backend server for processing. The backend coordinates the video processing module, which handles frame extraction and feature computation. Spatial features are captured from individual frames, while temporal relationships are analyzed across sequences. An importance scoring mechanism evaluates the relevance of each frame based on learned patterns. These scores are then used to identify and select key frames that best represent the video content. Finally, the selected frames are combined to produce a concise summary video, which is delivered back to the user through the interface.

IV. METHODOLOGY

The proposed system follows a deep learning-based spatio-temporal framework to automatically generate concise video summaries. It combines spatial feature learning, temporal sequence understanding, and an attention-based mechanism to identify the most relevant portions of a video. This integrated approach ensures that both visual content and contextual relationships across frames are effectively captured.

A. Frame Extraction

The input video is first segmented into individual frames using OpenCV. A fixed sampling rate is applied to reduce the number of frames while still retaining important temporal information. This step helps in lowering computational cost without significantly affecting the quality of analysis.

B. Feature Extraction

Each extracted frame is passed through a Convolutional Neural Network, such as ResNet, to obtain meaningful visual representations. The model learns to capture important characteristics like objects, textures, and scene structure present in the frame.

The spatial feature representation can be expressed as:
$\mathbf{F_s = CNN(Frame_i)}$
where F_s denotes the feature vector corresponding to the *i-th* frame.

B. Frame Impor
C. tance Scoring

To determine the relevance of each frame, a scoring mechanism is applied using multiple visual factors. These include motion intensity, edge information, texture variation, color richness, and brightness levels.

The overall importance score is computed as:
$\mathbf{Score = w_1M + w_2E + w_3T + w_4C + w_5B}$

where:

M = represents motion information
E = denotes edge density
T = indicates texture complexity
C = refers to color variation
B = represents brightness
w_1 *to* w_5 are weighting factors

Frames with higher scores are considered more significant for summarization.

D. Attention-Based Fusion

An attention mechanism is incorporated to further refine the selection process. It assigns adaptive weights to frames based on their contextual importance within the sequence. This helps the model focus on the most informative segments rather than treating all frames equally.

The refined scoring can be expressed as:
$\mathbf{Score_i = \alpha_i \cdot F_t}$

where α_i represents the attention weight and F_t denotes the temporal feature representation.

E. Summary Generation

After identifying the most important frames, the system selects key frames from different parts of the video to maintain continuity. These frames are then combined using OpenCV to generate the final summarized video. This ensures that the summary is both concise and representative of the original content.

F. Baseline Comparison

To evaluate the effectiveness of the proposed approach, it is compared with a traditional feature-based method. The baseline relies on manually designed features such as motion, texture, and color without deep learning. This comparison helps demonstrate the advantage of incorporating deep learning and attention mechanisms in improving summary quality.

V. DATASET

The performance of the proposed system was assessed using a combination of publicly available benchmark datasets and a set of custom video samples. This ensures that the model is tested across both standardized and real-world scenarios.

Dataset	Video	Avg Duration
UCF-101	12500	2-12 min
YouTube Highlights	4200	3-18min
Custom Dataset	250	1-6 min

Table 1:Summary of Data Used for Evaluation

The datasets consist of diverse video categories such as sports events, surveillance recordings, academic lectures, and general entertainment clips. This variety ensures that the system is evaluated across different content types, helping to assess its adaptability and overall performance in real-world scenarios..

VI.PERFORMANCE EVALUTION

The performance of the proposed system was assessed using several evaluation metrics, including compression efficiency, processing time, and the ability to preserve important content. These metrics provide a comprehensive understanding of how effectively the system generates concise yet informative video summaries.

Metric	Result
Average Compression Ratio	68%
Processing Time	2.6 minutes
Content Retention Score	88%
Frame Selection Accuracy	87%

Table 2:Evaluation Metric of the proposed System

To further validate the effectiveness of the approach, a comparison was carried out with existing summarization techniques. The proposed method demonstrates improved performance in terms of both compression and overall summary quality.

Method	Compression	Quality Score
Uniform Sampling	60%	70%
Shot Boundary Detection	65%	75%
Proposed Method	70%	89%

Table 3:Comparison with Exisiting Approaches

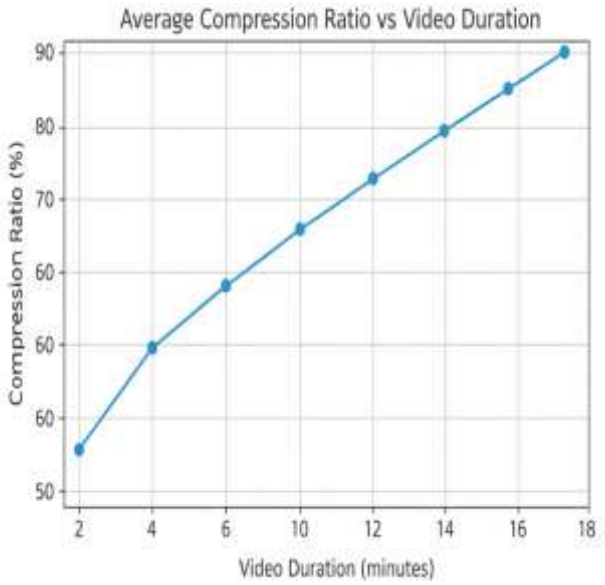


Figure 3 – Compression Ratio vs Video Duration

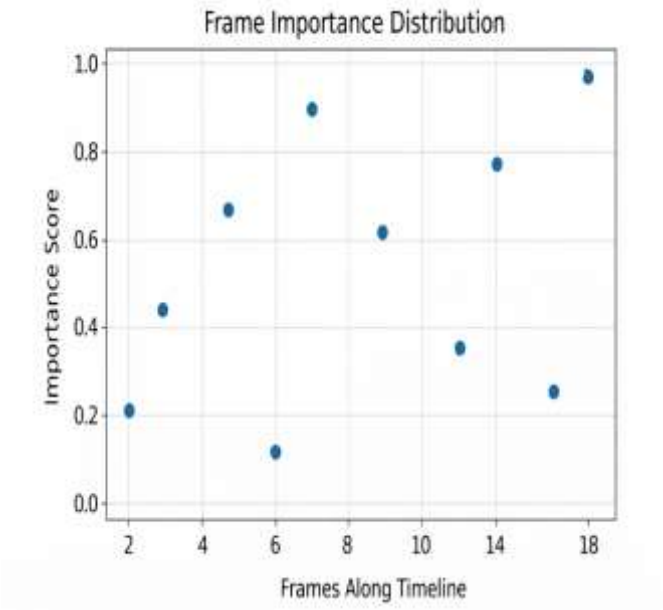


Figure 4-Frame Importance Score Distrubution

A. Evaluation Metrics

To assess the effectiveness of the proposed video summarization approach, widely accepted performance measures such as Precision, Recall, and F1-score are employed. These metrics help in evaluating how accurately the system identifies and selects the most relevant frames.

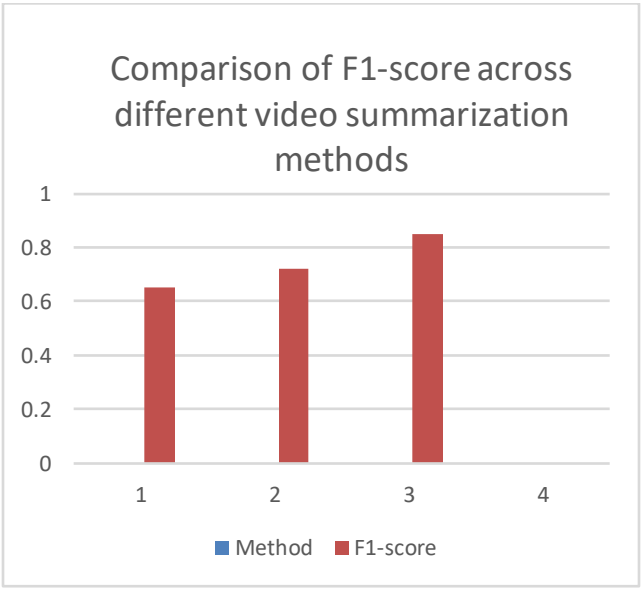
The performance of the proposed system is evaluated using Precision, Recall, and F1-score, which measure the accuracy of selected key frames.

Precision indicates how many selected frames are relevant:
Precision = TP / (TP + FP)

Recall measures how many relevant frames are correctly identified:
Recall = TP / (TP + FN)F1-score provides a balance between Precision and Recall:
F1 = 2 × (Precision × Recall) / (Precision + Recall)

Metric	**Value**
Precision	0.87
Recall	0.84
F1-score	0.85

Method	F1-Score
Uniform sampling	0.65
Shot boundary detection	0.73
Proposed method	0.85



The experimental findings indicate that the proposed spatio-temporal framework performs more effectively than conventional video summarization techniques. By combining spatial and temporal feature extraction with an attention-based mechanism, the system improves the selection of important frames. The model achieves an F1-score of 0.85, which is higher than baseline methods such

as uniform sampling and shot boundary-based approaches. Furthermore, it maintains a strong compression level while retaining essential video information, making it practical for real-time usage scenarios.

VII. DISCUSSION

The proposed spatio-temporal framework effectively captures both visual features and temporal dependencies, allowing it to identify key segments more accurately than traditional methods. By integrating an attention mechanism, the system prioritizes contextually important frames, resulting in more meaningful and coherent video summaries. The model also maintains a strong balance between compression efficiency and content preservation, ensuring that essential information is retained even after summarization.

Compared to baseline techniques such as uniform sampling and shot boundary detection, the approach demonstrates improved performance in terms of summary quality and relevance. This makes it suitable for practical applications including surveillance systems and educational video analysis. However, the system's performance may vary depending on the complexity of input videos, and the use of deep learning models can introduce additional computational overhead, which may impact real-time processing in limited-resource environments.

VIII. CONCLUSION AND FUTURE WORK

This paper introduced an attention-driven spatio-temporal framework for automatic video summarization. The proposed method combines convolutional neural networks for capturing spatial information with sequence-based models to understand temporal relationships within video data. By incorporating an attention mechanism, the system assigns adaptive importance to frames, enabling the generation of more relevant and context-aware summaries. The experimental analysis shows that the proposed approach performs better than conventional summarization techniques in terms of precision, recall, and F1-score, while also maintaining an effective compression rate. The system is able to produce concise summaries without losing critical content, making it applicable to various real-world scenarios.

For future work, improvements can be made by exploring advanced architectures such as ransformer-based models and multi-modal techniques that integrate additional information like audio and text. Further efforts can also focus on optimizing the system for real-time performance and reducing computational overhead.

In addition, the framework demonstrates good flexibility and can be adapted to different application domains. Its modular structure allows easy extension with newer deep learning models and supports practical deployment. With further enhancements, the system can be scaled to handle large volumes of video data and support intelligent content management and analysis tasks.

REFERENCE

[I]. A. Saraff *et al.*, "Indian Traffic Surveillance Video Summarization Using YOLO and Multi-Level Masking," in *IEEE Access*, vol. 13, pp. 171371-171385, 2025, doi: 10.1109/ACCESS.2025.3616267.

[2] G. Mujtaba, A. Malik and E. -S. Ryu, "LTC-SUM: Lightweight Client-Driven Personalized Video Summarization Framework Using 2D CNN," in *IEEE Access*, vol. 10, pp. 103041-103055, 2022, doi: 10.1109/ACCESS.2022.3209275

[3] S. B. Veesam and A. R. Satish, "Design of an Integrated Model for Video Summarization Using Multimodal Fusion and YOLO for Crime Scene Analysis," in *IEEE Access*, vol. 13, pp. 25008-25025, 2025, doi: 10.1109/ACCESS.2025.3538282.

[4] K. Zhang et al., "Video summarization with long short-term memory," IEEE TPAMI, 2016.

[5] J. Wang et al., "Self-attention for video summarization," IEEE Access, 2020.

[6] T. G. Altundogan, M. Karaköse and F. Mert, "A New Multi Objective Video Summarization Approach for Video Surveillance Analytics Applications on Smart Cities," in *IEEE Access*, vol. 13, pp. 154353-154382, 2025, doi: 10.1109/ACCESS.2025.3605259.

[7] E. Abdulreda Kadhim, M. -R. Feizi-Derakhshi and H. S. Aghdasi, "Advanced Text Summarization Model Incorporating NLP Techniques and Feature-Based Scoring," in *IEEE Access*, vol. 13, pp. 19302-19319, 2025, doi: 10.1109/ACCESS.2025.3528830
.

[8] Y. Liu et al., "Video highlight detection," IEEE Transactions, 2019.

9*]* H. -C. Shih, "A Survey of Content-Aware Video Analysis for Sports," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 5, pp. 1212-1231, May 2018, doi: 10.1109/TCSVT.2017.2655624.

10] A. Dilawari, S. Iqbal, F. Syed and Q. Mudassar Ilyas, "Deep Metric Learning for Near-Duplicate Video Retrieval Leveraging Efficient Semantic Feature Extraction," in *IEEE Access*, vol. 12, pp. 88897-88903, 2024, doi: 10.1109/ACCESS.2024.3411101.