RESEARCH ARTICLE **OPEN**

Static Video Summarization Using Clustering

Dr. R. Priya*, Pavithra M**, Jothipriya J**, Janani M**, Jayasri R**
*Professor -Computer Science and Engineering,Annamalai University,Annamalai Nagar – 608 002,
**UG Students-Computer Science and Engineering,Annamalai University,Annamalai Nagar – 608 002,
India

ABSTRACT : The growing growth of digital multimedia content has made effective video summaries essential for quick analysis and retrieval of important information. This work demonstrates an intelligent video summarizing system that automatically extracts important keyframes and generates brief descriptions from input videos. The proposed system extracts spatial and temporal information from video frames using a variety of techniques, including Convolutional Neural Networks (CNN), a hybrid CNN–Long Short-Term Memory (LSTM) model, and ORB (Oriented FAST and Rotated BRIEF)+ BoVW (Bag of Visual Words) -based feature extraction. Frames are first obtained at a customisable sample rate and relevant attributes are computed in order to represent visual content. Keyframes are selected utilizing cluster centres and importance ratings following the clustering-based grouping of related frames. To enhance semantic comprehension, the system uses voice recognition algorithms for audio transcription. The entire system is built as an interactive dashboard that allows users to upload videos , analyse analytic findings and assess different summarizing strategies. Experimental results demonstrate that the CNN–LSTM strategy effectively captures temporal dependencies and produces more relevant summaries, whereas ORB+BoVW-based strategies offer faster processing. The recommended approach is efficient, scalable and suitable for real-world applications such as media management, content indexing and surveillance.

Keywords: *Keyframe extraction, CNN, LSTM, deep learning, computer vision, speech recognition, video summarization and multimedia analysis.*

I

. INTRODUCTION

Due to rapidly increasing use of digital video content such as social media platforms, surveillance systems, online streaming and education resources, managing and analysing the vast quantities of video content have become difficult. It is time consuming, inefficient, and often unproductive to watch an entire video for the purpose of retrieving useful and appropriate information , especially when the majority of what is shown is irrelevant to your objective. Therefore, the demand for intelligent video summarizing algorithms which can automatically identify and concisely present the most significant portion(s) of the video is growing.

Despite being computationally cheap to evaluate low-level features such as colour, texture and motion, traditional video summarization techniques usually do not have sufficient temporal and semantic context to create an appropriate video summation. As a result, the video summations created by these techniques are not always meaningful for the context. The recent advances in deep learning, particularly Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have enabled the automated analysis of video

to be more accurate and semantically meaningful by capturing both spatial and temporal information. However, most current video summarizing systems use only a single approach and therefore lack flexibility, interactivity and a user-friendly display.

The paper presents a video summarization (VS) system incorporating the following methods: i) CNN based deep feature learning for extracting semantic features.
ii) ORB + Bag of Visual Words (BoVW) based feature extraction.
iii) Hybrid CNN-LSTM based temporal analysis.

These techniques provide a solution to VS problems by automatically extracting frames, analysis of temporal and/or visual characteristics, and producing keyframe (KF) based summaries that accurately reflect the original video content. In addition, the addition of audio transcription through speech recognition significantly increases the overall semantic understanding of the video, therefore the summary generated from it is more informative. The described system is implemented as an interactive dashboard which enables users to test out different video summarization methods, upload their own videos, modify parameters, and observe results. The solution provides a balance between computational efficiency and summarization accuracy by combining classical methods and deep learning. Potential real world applications of the system include digital archiving, e-learning, media content management and video surveillance.

A. Video Summarization Using ORB+BoVW Features

Key point descriptors are used to represent each frame in a traditional yet efficient technique known as ORB (Oriented FAST and Rotated BRIEF)+Bag of Visual Words (BoVW)-based video summarization. This method encodes each frame as a histogram of visual word frequencies, builds a visual vocabulary using K-Means clustering, and extracts binary feature descriptors from identified key points in frames. This approach is computationally efficient and suitable for real-time applications with limited resources. It is more reliable than simple colour-based techniques because it catches local structural patterns and textural information beyond mere pixel-level statistics. In the proposed approach, ORB+BoVW-based summarization acts as a robust baseline technique, offering quick but significant frame representations.

B. Feature Extraction Using CNN

Convolutional Neural Networks (CNNs) are widely employed in computer vision tasks to extract high level information from images and videos. In this study, CNN models are used to analyse video frames in order to extract semantic information such as objects, sceneries, and patterns. Keyframe selection is enhanced by CNN-based feature extraction, which provides a more meaningful representation of visual content than traditional methods. The retrieved features are subsequently analysed using clustering techniques to select sample frames for summarization.

C. Temporal Analysis Using CNN-LSTM

Video data inherently contains temporal information that shows the sequence and progression of events. The recommended method effectively captures this temporal dependency using a hybrid CNN–LSTM model. The LSTM network sequentially examines the spatial features that CNN collects from individual frames in order to understand the temporal connections between frames. This enables the algorithm to assign relevance values to frames and recognize significant scene changes. As a result, the CNN–LSTM approach produces video summaries that are more perceptive and context-aware.

D. Semantic Enhancement Based on Audio

It takes both visual and aural information to comprehend video material. The proposed system employs voice recognition techniques

for audio transcription to convert spoken input into text. This enhances the semantic comprehension of the video and provides more context for the summaries that are generated. Because audio and visual data are integrated, the technology is more dependable and suitable for applications requiring in-depth content analysis.

E. Interactive Visualization and Analysis Dashboard

To improve usability and accessibility, the system is built as an interactive dashboard. On the dashboard, users may upload videos , select summarization methods, alter parameters, and see results like timelines, keyframes and cluster distributions. It also let users compare different strategies and evaluate results based on summary quality and processing time. Because of its user-centric design, the system can be implemented in the real world.

II. LITERATURE SURVEY

Numerous studies are available regarding video summarization, which can compress a video while maintaining its most informative elements. Different methods have been utilized for this effect, ranging from traditional feature-based techniques to advanced deep learning-based methods.

Among these sets of research was the work of [1] M. Srinivas, M. M. Manohara Pai, and R. M. Pai in "An Improved Algorithm for Video Summarization: A Rank-Based Approach" present a rank-based method where frames are assigned importance scores to generate concise and informative summaries by reducing redundancy. [2] M. S. Afzal and M. A. Tahir in "Reinforcement Learning Based Video Summarization with Combination of ResNet and GRU" proposed a reinforcement learning approach that combines ResNet for feature extraction and GRU for sequence modeling to improve adaptive frame selection.
[3] S. S. Harakannanavar and co-authors in "Robust Video Summarization Algorithm Using Supervised Machine Learning" introduce a supervised learning method trained on labeled data to accurately identify and select meaningful frames for summarization.[4] Y. Zhang and others in "Unsupervised Object-Level Video Summarization with Online Motion Auto-Encoder" developed an unsupervised technique using motion auto-encoders to capture object-level and motion features without requiring labeled data. [5] O. Elharrouss et al. in "A Combined Multiple Action Recognition and Summarization for Surveillance Video Sequences" combine action recognition with summarization to focus on important activities, particularly in surveillance applications. [6] K. Zhang et al. in "Video Summarization with Long Short-Term Memory" utilize LSTM networks to capture temporal dependencies and maintain the logical sequence of events in video summaries.[7] B. Mahasseni, M. Lam, and S. Todorovic in "Unsupervised Video Summarization with Adversarial LSTM Networks" proposed an adversarial framework using LSTM networks that generates summaries without labeled data by learning video representations. [8] J. Song and co-authors in "Video Summarization via Actionness Ranking" introduce a ranking method that selects frames based on action importance to improve summary relevance. [9] T. A. Ngo et al. in "Video Summarization: A Survey" provide a comprehensive review of reinforcement learning-based summarization techniques, discussing models, challenges, and future directions. Finally, [10] H. Yuan, X. Zhang, and J. Li in "Deep Learning-Based Video Summarization: A Review" present an extensive overview of deep learning approaches, including CNN, RNN, and hybrid models, highlighting their effectiveness, limitations, and research trends.

III.PROPOSED SYSTEM

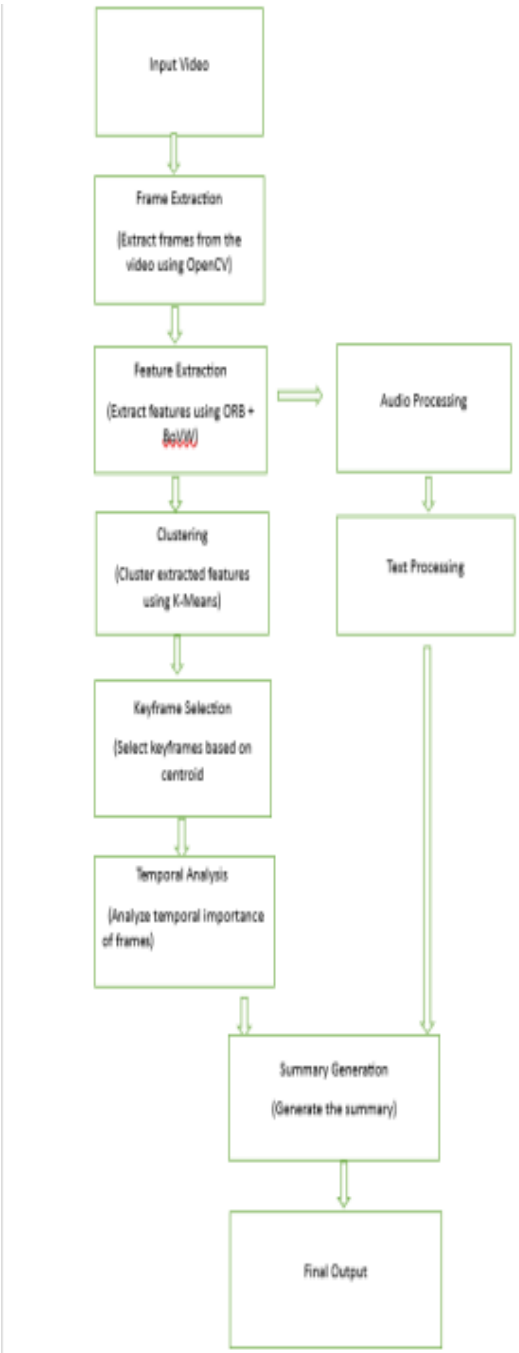

Fig.1. Block Diagram of the proposed system

Fig 1 shows the block diagram of the proposed system that it is developed to generate accurate, clear, and meaningful summaries of video content by combining traditional feature extraction techniques with advanced deep learning approaches. The process begins with the video input module, where users upload a video to be analyzed as part of the intelligent video summarization workflow. Once the video is provided, the system performs frame extraction at an adjustable sampling rate, ensuring that redundant frames are minimized while important visual information is preserved. These extracted frames are then processed in the feature extraction stage using two complementary methods. The first method, ORB+BoVW (Bag of Visual Words), captures low-level visual features by identifying keypoints and representing them using a visual vocabulary in the form of histograms. The second method uses Convolutional Neural Networks (CNN) to extract high-level semantic features such as objects, scenes, and contextual information present in the video. By combining both approaches, the system achieves a more comprehensive understanding of the video content.

The extracted features are then passed to a clustering module, where algorithms such as K-Means are used to group similar frames based on their visual characteristics. This clustering process helps in reducing redundancy and organizing the frames into meaningful groups. From each cluster, the keyframe selection module identifies the most representative frames that best describe the content of that group. These selected keyframes play a crucial role in visually summarizing the video. To further enhance the summary, a temporal modeling module based on a hybrid CNN-LSTM architecture is used to analyze the sequence of frames over time. This module detects important events, scene transitions, and temporal dependencies, ensuring that the generated summary maintains the logical flow of the original video.

At the same time, the system processes the audio component of the video in parallel. The audio extraction module separates the audio track, which is then enhanced to remove noise and improve clarity. The processed audio is passed through a speech-to-text module that converts spoken content into textual form, thereby adding semantic meaning to the visual summary. Finally, the outputs from both the visual and audio processing stages are integrated in the summary generation module to produce a comprehensive and concise result. The final output includes a set of keyframes, a condensed video clip, and text-

based information describing the video content. All results are presented through an interactive visualization dashboard, allowing users to easily view, analyze, and compare different summarization methods. This integrated approach ensures that the system produces efficient, accurate, and user-friendly video summaries suitable for various applications.

A. Frame Extraction

For further processing by the Frame Extraction module, the input video must be converted into a sequence of frames. The system gathers frames at a configurable frame rate to achieve a balance between computational efficiency and summary quality. By removing superfluous frames and focusing on representative samples, this module ensures efficient processing while preserving important visual information. The retrieved frames serve as the primary input for later feature extraction and analysis modules.

The input video is represented by this equation as a series of frames that are extracted at regular intervals. Every frame Fi acts as the basic building block for additional processing and analysis.

$$V = \{F_1, F_2, F_3, \ldots, F_n\} \quad (1)$$

Here, V represents the input video as a collection of frames, where F_1, F_2, …, F_n denote individual frames extracted sequentially from the video.

B. Feature Extraction

The Feature Extraction module is crucial for representing video material. The proposed system uses both CNN-based and ORB+BoVW-based feature extraction techniques. ORB+BoVW features identify and encode local key points into a visual lexicon, producing a reliable and condensed frame descriptor, whilst CNN models extract high-level semantic features like objects and scene context. This combination allows the technology to achieve a balance between speed and accuracy. The recovered features are then used for grouping and significance evaluation.

This uses the ORB detector to extract key point descriptors from each frame, then maps them to a visual vocabulary of k visual words using K-Means clustering. It provides a compact and distinctive frame representation by encoding local structural and texture features into a Bag of Visual Words histogram.

$$F_i = \{h_1, h_2, \ldots, h_k\} \quad (2)$$

Here, H_i represents the feature vector of the i frame, where h_1, h_2, …, h_k are the extracted feature values describing visual characteristics of that frame.

Distance:

Based on the properties of their ORB+BoVW feature vectors, the Euclidean distance calculates how similar two frames are. Higher similarity between frames is indicated by smaller distance values.

$$D(F_i, F_j) = \sqrt{\sum_{k=1}^{n} (H_{ik} - H_{jk})^2} \quad (3)$$

Here, $D(F_i, F_j)$ represents the Euclidean distance between frames F_i and F_j, where H_{ik} and H_{jk} denote the k-feature values of the respective frames, and n is the total number of features.

This formula illustrates the use of a convolutional neural network for deep feature extraction. The final feature vector, Xi extracts high-level semantic data from every frame.

$$X_i = f(F_i)$$

----(4)

$$
\in
$$

$$
\text{ -----}
$$

(5)

Here, X_i represents the deep feature vector of the i frame obtained using a CNN, where F_i is the input frame and $X_i \in R^d$ denotes a d-dimensional feature representation.

C. Clustering and Keyframe Selection

The Clustering module groups related frames based on their feature representations using techniques like K-Means. By clustering frames, the system discovers patterns in the video and removes duplication. The Keyframe Selection module then selects representative frames from each cluster, typically choosing frames that are closest to the cluster centroid. This ensures that the selected keyframes reduce repetition while providing a concise synopsis of the primary information of the video.

The distance between data points and the corresponding cluster centroids is reduced using this function. For effective summarization, it guarantees that related frames are clustered together.

$$
= \quad =
$$

1)^ $$
\in
$$

|| $$
-
$$

||^2 ---- (6)

Here, J represents the clustering objective function, where x_i denotes the feature vector of a frame in cluster C_k, μ_k is the centroid of cluster k, and K is the total number of clusters.

This chooses the representative keyframe to be the frame that is closest to the cluster centroid. It guarantees that the selected frame most accurately depicts the cluster's visual content.

$$
=
$$

$$
\in
$$

|| $$
-
$$

|| ----(7)

Here, F_key represents the selected keyframe, where X_i is the feature vector of frame F_i in cluster C_k, and μ_k is the cluster centroid, with the closest frame chosen as the representative.

D. Temporal Analysis Using CNN–LSTM

The Temporal Analysis module captures the sequential nature of video data using a hybrid CNN–LSTM model. While CNN gathers spatial information from individual frames, the LSTM network examines these properties over time to understand temporal correlations. This enables the system to recognize noteworthy occurrences, scene changes, and dynamic changes in the video. By assigning importance scores to the frames, this module enhances the quality and coherence of the final summary.

This formula uses LSTM to model temporal dependencies between successive frames. In video data, it aids in capturing significant transitions and sequential patterns.

$$
h_ \quad =
$$

$$
h_(
$$

1)) -----(8)

Here, h_t represents the hidden state at time t, where X_t is the input feature vector of the current frame and $h_{(t-1)}$ is the previous hidden state capturing temporal dependencies.

This uses learnt weights and hidden states to calculate a significance score for every frame. More important frames for the summary are indicated by higher ratings.

$$
=
$$

$$
h_ +
$$

-----(9)

Here, S_i represents the importance score of the i frame, where h_i is the hidden state, W and b are learned parameters, and higher values indicate more significant frames.

E. Audio Processing and Transcription

The Audio Processing module extracts the audio track from the input video and converts it into text using voice recognition techniques. The audio signal is preprocessed to reduce noise and improve clarity before transcription. The generated text enhances the system's understanding of the video's content by providing it with more semantic context. This multi-modal approach makes the summarizing process more efficient overall.

This illustrates how audio signals are converted into characteristics in the time-frequency domain. It records speech's temporal and spectral features for semantic analysis.

$$M(f,t) = \sum_{n} x[n] \cdot w[n-t] \cdot e^{(-j2\pi f n)} \quad \text{-----(10)}$$

Here, M(f, t) represents the time-frequency representation of the audio signal, where x[n] is the input signal, w[n−t] is the window function, and f denotes frequency.

F. Visualization and Output Generation

This chooses frames whose importance score is higher than a certain cutoff. It guarantees that the final summary contains only the most pertinent frames.

$$\text{Summary} = \{F_i \mid S_i > \theta\} \quad \text{---(11)}$$

Here, Summary represents the set of selected frames F_i, where S_i is the importance score of each frame and θ is the threshold, such that only frames with scores greater than θ are included in the final summary.

The Visualization and Output module displays the final, condensed results to the user via an interactive dashboard. It creates a shortened video, displays particular keyframes, and uses audio transcription to provide textual insights. The module also enables users to evaluate performance based on processing time and quality by comparing different summary methods. Additionally, the system allows results to be exported as JSON files, images, and videos.

IV. RESULTS AND DISCUSSION

The proposed intelligent video summarizing system was evaluated using a range of videos, including surveillance footage, educational content, and general multimedia data. The system was tested using three different approaches: CNN-based feature extraction, ORB+BoVW-based feature extraction, and the hybrid CNN–LSTM model. The results demonstrate that the quality and applicability of the resulting summary are significantly improved by integrating geographical and temporal analysis. The ORB+BoVW-based method provided efficient performance and fast processing for simple videos with few scene changes. Although it outperformed basic colour-based approaches by utilizing local structural features, it still had some trouble capturing deep semantic meaning in complex scenarios. Conversely, the CNN-based approach demonstrated superior performance and more significant keyframe selection by extracting high-level information such as objects and scene context. The hybrid CNN–LSTM model greatly enhanced summarization and enabled precise identification of important events and transitions by effectively capturing temporal relationships between frames.

The K-Means clustering algorithm efficiently reduced redundancy and ensured that only representative frames were selected by merging related frames. The keyframe selection module reliably identified frames that preserved the overall context of the video. Additionally, the temporal analysis module improved coherence in the summaries generated by considering the sequential structure of the video data. The combination of speech-to-text transcription and audio processing greatly enhanced their usefulness by providing the visual summary with semantic context. The ability of the system to extract important textual information from the audio stream supplemented the visual analysis

and improved the results' interpretability. The interactive dashboard also allowed for the effective comparison and visualization of different summary techniques. Users could look at the number of keyframes, processing time, and summary quality across methods. The system demonstrated scalability and flexibility for a variety of video clip types, making it suitable for real-world applications.

Method	Video Duration (s)	Processing Time (s)	Accuracy (%)
ORB+ BoVW +CNN-Based Method+ CNN + LSTM (Proposed Method)	148	75	92%

Table 1. Performance Comparison of Different Methods

According to the results, traditional techniques like ORB+BoVW feature extraction are computationally efficient, but they don't provide the same level of high-level semantics as methods using deep learning. By detecting significant patterns found in the frames, deep learning methods (specifically CNN) have been shown to enhance the quality of the summary. While deep learning was previously used to build models that describe the order of events in a video using temporal modelling, LSTM models provide the most accurate and cohesive summaries..

In response, the proposed solution provides users with multiple methods to choose from based on their own needs, thereby establishing an overall balance between those two factors. The method is also enhanced by the incorporation of audio semantic analysis, which gives the summary generation process more robustness and reliability by combining textual and visual information in the summary.

V.CONCLUSION & FUTURE WORK

To summarize, the innovative intelligent video summarization method offers an integrated approach to producing effective short forms of video material's content, through combining classical approaches with deep learning based approaches for producing informative yet concise summaries. The video summarization system relies on ORB+BoVW-based feature extraction, spatial analysis of videos via CNN, and temporal analysis of video data using CNN-LSTM techniques that provide excellent precision in identifying important events and frames from a given video; this is achieved through a clustering process and selection of keyframes to reduce redundancy within the summary, whilst preserving the critical content within the summary, resulting in informative and concise video summaries.

The reality that the video summarization method enhances the overall understanding of video material through the use of audio processing and transcription of text within any video to create semantic meaning gives users a greater awareness of video content. The visualization and analysis of various forms of video summarization techniques will occur easily via the interactive dashboard provided by the video summarization system making this system very user-friendly and versatile. The experimental findings show that the hybridization of CNN-LSTM approach delivers improved performance compared to previous methods for generating video summarization with respect to quality and coherence of generated summaries. Consequently, the intelligent video summarization method is a practical and dependable large scale video summarization solution across a range of applications, including but not limited to, surveillance, media storage and retrieval, education, and content indexing.

The new video summarization technology will be updated based on efficiency, scalable, and on-time capabilities in creation of summaries of the video. Transformers' attention mechanism allows for the
ability of the transformer networks to learn the temporal features in the video data, thereby providing a tool for the identification of long-

term occurrences in the video. To realize a robust and generalizable video summarization system, a more significant and more diverse real-world scenarios data set will be needed. After continuing to improve the ability to perform real-time video summarization for live application areas such as video streaming, video conferencing, and security monitoring by performing video summarization for real-time video processing technologies.

Further enhanced video summarization systems could be developed by using multimodal fusion methods that integrate vision, audio, and text information. Improved processing speed through scalability, in addition to implementing cloud and edge technologies, will provide for additional performance gains with video summarization systems. Additional user-learning through user preferences and personalized summaries for each will help to provide users with video summaries relevant to their interests. The establishment of additional advanced evaluation metrics and automated quality assessment will create increased reliability with the video summarization process. These enhancements will ultimately help to establish a viable and successful adaptive, efficient, and effective large-scale production tool in many real world systems.

VI. REFERENCES

[1] M. Srinivas, M. M. Manohara Pai, and R. M. Pai, "An Improved Algorithm for Video Summarization: A Rank-Based Approach," Procedia Computer Science, vol. 89, pp. 812–819, 2016.

[2] M. S. Afzal and M. A. Tahir, "Reinforcement Learning Based Video Summarization with Combination of ResNet and Gated Recurrent Unit," in Proceedings of the International Conference on Computer Vision and Image Processing, Karachi, Pakistan, 2019.

[3] S. S. Harakannanavar, S. R. Sameer, V. Kumar, S. K. Behera, A. V. Amberkar, and V. I. Puranikmath, "Robust Video Summarization Algorithm Using Supervised Machine Learning," Journal of Visual Communication and Image Representation, vol. 73, pp. 102933, 2020.

[4] Y. Zhang, X. Liang, D. Zhang, M. Tan, and E. P. Xing, "Unsupervised Object-Level Video Summarization with Online Motion Auto-Encoder," Pattern Recognition, vol. 104, pp. 107–117, 2020.

[5] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, A. Bouridane, and A. Beghdadi, "A Combined Multiple Action Recognition and Summarization for Surveillance Video Sequences," Journal of Ambient Intelligence and Humanized Computing, vol. 12, no. 3, pp. 1–15, 2020.

[6] K. Zhang, W. Chao, F. Sha, and K. Grauman, "Video Summarization with Long Short-Term Memory," in *European Conference on Computer Vision (ECCV)*, Amsterdam, Netherlands, 2016, pp. 766–782.

[7] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised Video Summarization with Adversarial LSTM Networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 202–211.

[8] J. Song, Y. Yang, Z. Huang, H. T. Shen, and R. Hong, "Video Summarization via Actionness Ranking," *IEEE Transactions on Image Processing*, vol. 27, no. 12, pp. 5863–5874, 2018.

[9] T. A. Ngo, Z. Ma, and H. Zhang, "Deep Reinforcement Learning for Video Summarization: A Survey," *IEEE Access*, vol. 9, pp. 123–135, 2021.

[10] H. Yuan, X. Zhang, and J. Li, "Deep Learning-Based Video Summarization: A Review," *IEEE Access*, vol. 10, pp. 12345–12360, 2022.