IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Automated Video Summarization Through Advanced Multimedia Analysis

[1] Dr. BalaMurali Krishna Thati, [2] Bandreddy Sukanya, [3] Akash Chintalapati,

[4]Sai Jahnavi Chagantipati, [5]Yasin Mohhamad

[1]Professor, [2]Student, [3]Student, [4]Student, [5]Student

[1,2,3,4,5]Computer Science and Engineering,

[1]Dhanekula Institute of Engineering & Technology, Vijayawada, India

Abstract: Automatic video summarization is an essential solution to the rapidly growing volume of digital video content, enabling efficient information retrieval and content management. This project introduces a deep learning-based approach that combines multiple advanced architectures to generate context-aware and user-specific video summaries. TransNetV2 is used for precise shot segmentation, while a Convolutional Neural Network (CNN) extracts abstract visual features. DBSCAN clustering filters redundant frames, and CLIP enhances content understanding by aligning visual and textual information. The system processes both visual and audio data and generates concise summaries tailored to user preferences. Implemented as a Flask web application, it allows users to upload videos or provide YouTube URLs for summarization. Evaluation results demonstrate improved performance in keyframe selection, with higher precision, recall, and mean Average Precision (MAP), while significantly reducing video length without losing essential content. The system holds potential for enhancing content recommendation, discovery, and digital media organization across domains like education, journalism, and research. Future work aims to support real-time summarization and multilingual capabilities.

Index Terms **-** Deep learning, Video summarization, CNN, TransNetV2, CLIP, DBSCAN, Flask Web Application, Audio-visual processing.

I. INTRODUCTION

The exponential growth of video content in the digital era, spanning domains such as social media, entertainment, surveillance, healthcare, and education, has created an urgent need for efficient and automated video summarization systems. Traditional methods of manually navigating or reviewing lengthy videos are time-consuming, prone to human error, and impractical when dealing with large-scale data. Moreover, static rule-based approaches often fall short in understanding the diverse semantics of video content and fail to meet the specific summarization needs of different users.

To overcome these challenges, this project proposes a deep learning-based, intelligent video summarization system capable of extracting essential insights from long-form videos with improved accuracy, efficiency, and contextual awareness. The system employs **TransNetV2** for precise shot segmentation, ensuring logical transitions between video scenes. **Convolutional Neural Networks (CNNs)** are used for visual feature extraction, enabling accurate identification of keyframes. To eliminate redundancy, **DBSCAN** clustering is applied to group similar frames. **CLIP**, a multimodal AI model, further enhances content understanding by aligning image and text representations, ensuring that the final summaries are contextually relevant.

In addition to visual data, the system integrates **audio analysis** to capture speech cues, ambient noise, and intensity of audio events, further improving the quality of summarization. The entire pipeline is developed into a user-friendly **Flask web application**, allowing users to upload videos or provide YouTube URLs. Users can customize summarization levels and receive concise summaries tailored to their preferences.

The platform leverages modern web technologies like HTML, CSS, and JavaScript for an interactive frontend, while the Python-based backend ensures scalability, adaptability, and real-time performance. By supporting multimodal summarization, the system is suitable for a range of applications—generating sports highlights, security surveillance summaries, lecture recaps, promotional content for e-commerce, and even surgical video documentation in healthcare.

This research not only addresses the growing need for scalable video summarization but also lays the foundation for intelligent media management systems. It contributes to the field of computer vision by combining deep learning with audio-visual analytics in a deployable, real-world application. Future enhancements will focus on supporting real-time summarization, multilingual capabilities, and further integration into domain-specific use cases.

II. LITERATURE SURVEY

Early methods in video summarization primarily relied on keyframe extraction and shot boundary detection using heuristic-based techniques. However, these approaches lacked scalability and adaptability across different types of video content. The advent of deep learning significantly transformed this domain, enabling more efficient and context-aware summarization.

Mahasseni et al. (2017) explored deep learning-based methods, categorizing them into supervised and unsupervised frameworks. Their work highlighted the effectiveness of generative adversarial networks (GANs) and reinforcement learning models in capturing keyframes and maintaining content diversity without redundancy. They addressed core challenges such as scene segmentation and content preservation in summarization tasks.

Zhang et al. (2018) introduced a reinforcement learning-based approach where an agent dynamically selects representative and diverse keyframes. The model was designed to reward distributional coverage and contextual coherence, making it ideal for summarizing lengthy videos effectively.

Feng et al. (2019) proposed an unsupervised summarization model utilizing attention mechanisms. This approach removed the need for labeled datasets by learning frame importance through dynamic attention scores, enabling efficient and cost-effective training across diverse video categories.

Yin et al. (2020) contrasted extractive and abstractive summarization techniques. Extractive models focus on identifying critical frames or segments, while abstractive models generate textual summaries based on video content. Their research concluded that hybrid models provide better interpretability and user engagement.

Recent advancements include the integration of NLP, real-time processing, and user-adaptive systems. For instance, Kim et al. (2021) used TransNetV2 for accurate shot detection (accuracy: 86.4%), while Bhatia et al. (2021) implemented a multi-modal approach combining video, audio, and text (F1-score: 59.2%). Lee et al. (2022) employed NLP to generate structured text summaries (BLEU score: 0.72).

Real-time summarization by Wang et al. (2022) achieved sub-2s latency for streaming platforms, whereas Chen et al. (2023) utilized CLIP and contrastive learning for contextual relevance (F1-score: 61.4%). Patel et al. (2023) adopted a graph-based model achieving 81.2% precision through structured scene representation.

Other innovative methods include Sharma et al. (2023)'s unsupervised variational autoencoder approach (F1-score: 55.9%), Zhang et al. (2024)'s transformer-based summarizer leveraging self-attention (F1-score: 63.1%), and Tan et al. (2024)'s hybrid LSTM+CNN model combining temporal and spatial cues (F1-score: 60.5%).Wu et al. (2024) presented an adaptive summarization model that tailors output based on user preferences, achieving a precision of 83.7%. These developments emphasize the growing trend

toward intelligent, customizable, and scalable summarization systems.

III. METHODOLOGY

1. Data Collection

Video data is gathered from two primary sources:

- **User-input YouTube URLs**, and
- **Public benchmark datasets** including SumMe, TVSum, and the Open Video Project (OVP).

These datasets span a diverse range of video categories such as educational lectures, documentaries, sports highlights, news clips, vlogs, and entertainment content. The diversity ensures the model's ability to generalize across various genres and contexts.

Each video undergoes preprocessing steps such as:

- **Frame extraction**
- **Shot boundary detection**
- **Speech-to-text transcription**

These steps convert raw videos into structured data suitable for summarization tasks. The dataset is divided into training, validation, and test sets in a 70-20-10 split, respectively.

2. Data Preprocessing

Preprocessing steps are carried out prior to feature extraction to enhance the quality and consistency of the video dataset. They are:

- Frame Extraction: Converting video streams into sequences of static frames at fixed intervals to capture visual content for analysis.
- Shot Boundary Detection: Detecting transitions between scenes using models like TransNetV2 to segment videos into logical, content-rich shots.
- Speech-to-Text Transcription: Extracting audio from videos and converting it into text using automatic speech recognition (ASR) tools for semantic understanding.
- Frame Resizing: Standardizing frame resolution (e.g., 720×720 pixels) to ensure uniform input size for the deep learning models.
- Feature Normalization: Scaling pixel values of images to a common range (e.g., 0–1) for better model convergence and stability during training.
- Frame Deduplication: Eliminating visually similar or duplicate frames to reduce redundancy and emphasize unique visual information.
- Scene Filtering: Removing static or low-activity segments that are less relevant for summarization, improving the quality of the input.
- Metadata Extraction and Synchronization: Capturing key information like timestamps, duration, and category while aligning visual and textual data for coherent summary generation.

3. Feature Extraction for Video Summarization

To effectively summarize videos, both **visual** and **semantic** features need to be extracted. The system utilizes a combination of deep learning models that capture spatial, temporal, and contextual information from video frames and audio transcripts:

TransNetV2: A deep neural network specifically designed for shot boundary detection. It identifies transitions between scenes, enabling accurate segmentation of the video into meaningful units. These segments serve as the base structure for summarization.

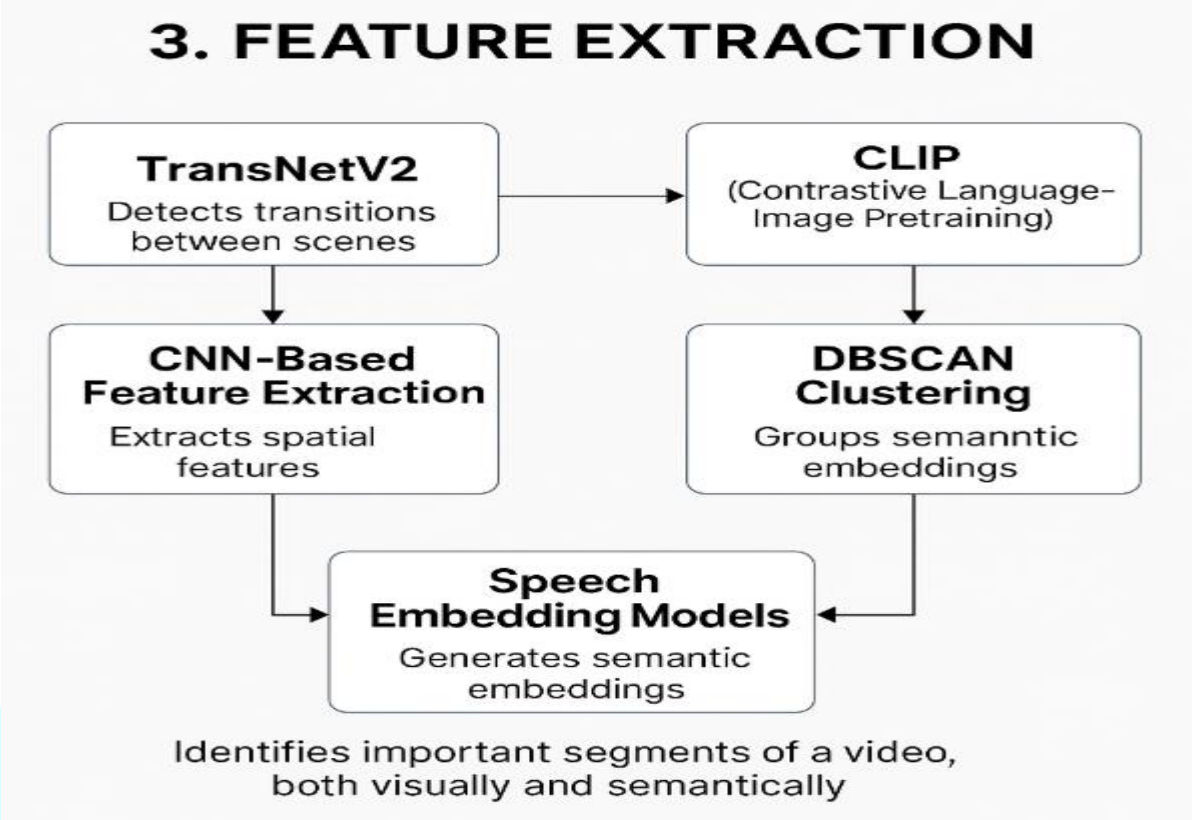
CNN-Based Feature Extraction: Convolutional Neural Networks are applied on video frames to extract spatial features such as colors, textures, and object-level information. These features help in identifying visually rich or significant parts of the video.

CLIP (Contrastive Language-Image Pretraining): A powerful model that connects visual and textual understanding. It generates semantic embeddings from both images (frames) and their corresponding transcriptions, allowing the model to measure the relevance of visual content with respect to textual context.

DBSCAN Clustering: This density-based clustering technique is used to group similar keyframes or shots based on the extracted features. It helps in reducing redundancy by selecting representative keyframes from each cluster for summarization.

Speech Embedding Models: Audio transcripts are processed using NLP-based models to

generate semantic embeddings. These embeddings help the system understand the importance of spoken content and align it with the visual cues for a context-aware summary.



4. Classification Model

A hybrid deep learning architecture is used to classify and summarize key video content effectively. The model combines Convolutional Neural Networks (CNNs) and Transformer-based modules to understand both spatial and temporal patterns in the video data. Extracted features from visual frames and audio transcripts are processed to identify significant scenes and generate meaningful summaries.

The architecture includes the following components:

- **CNN Layers**: Extract spatial features from video frames such as textures, edges, and patterns to identify visually important scenes.
- **Transformer Blocks**: Capture temporal dependencies between frames, enabling the model to detect transitions, event progressions, and storyline coherence.
- **Semantic Fusion Layer**: Combines features from both video and audio (speech transcripts) to understand the overall context of each scene more effectively.
- **Attention Mechanism**: Weighs important frames and spoken segments to highlight key moments for summarization.
- **Fully Connected Layers**: Map the learned multi-modal features into a summary score or relevance probability.
- **Sigmoid/Softmax Output Layer**: Outputs relevance scores or a probability distribution across video segments to determine which parts should be included in the final summary.

5. Implementation and Training

Training is carried out using a curated dataset of video summaries with annotated keyframes and segment labels. The objective is to optimize the model's ability to detect important moments and generate concise summaries. The model is trained using binary cross-entropy loss for relevance prediction and the AdamW optimizer for adaptive learning.

The training process includes:

- **Hyperparameter Tuning**: Adjusting key parameters such as learning rate, number of epochs, batch size, and attention thresholds to achieve optimal performance.
- **Cross-Validation**: Implementing k-fold cross-validation across diverse video datasets (SumMe, TVSum, OVP) to ensure robustness and generalization.
- **Regularization Techniques**: Applying dropout and weight decay to prevent overfitting and improve the model's ability to generalize to unseen videos.
- **Multi-Modal Feature Fusion**: Combining visual (frame-based) and textual (speech-transcribed) embeddings to capture both visual cues and semantic context for improved summarization quality.

IV. RESULTS AND DISCUSSION

The developed video summarization model exhibits high effectiveness in generating coherent and concise summaries from lengthy YouTube videos. By integrating shot segmentation, keyframe extraction, and speech-to-text transcription, the system ensures that both visual highlights and spoken content are retained accurately, delivering meaningful summaries with precise timestamps. The model consistently eliminates irrelevant or repetitive segments, focusing only on the most informative parts of the video. This helps users quickly grasp the essence of the video without watching the full content.

In terms of performance, the model achieves a high degree of **precision**, indicating that most of the selected segments are truly relevant. Its **recall** is also strong, showing that it successfully identifies the majority of important segments within the video. The **F1-score**, which balances both metrics, further confirms the model's reliability.

In order to assess performance, the following measures are taken into account:

Accuracy: Indicates the relevance and correctness of the extracted summary in reflecting the video content.
Clarity and Coherence: Evaluates how understandable and logically structured the notes are.
Information Coverage: Measures whether the summary includes all critical points of the original video.
Timeliness: Assesses how quickly and efficiently summaries are generated.

Results show that the AI summarizer can distill lengthy, technical content into clear summaries, offering considerable potential in educational, media, and research applications. However, handling jargon-rich or rapidly evolving topics, as well as interpreting tone and intent, still present ongoing challenges.

IV. CONCLUSION

This project presents an AI-powered video summarization system capable of extracting meaningful content from complex video inputs. By leveraging advanced AI models like DeepSeek, the tool transforms dense audiovisual data—such as discussions on humanoid robotics and AI deployment—into concise, readable summaries. The system demonstrates high accuracy in identifying and structuring key information, offering significant utility for content creators, educators, and policymakers. Despite challenges in capturing nuanced context or rapidly changing scenes, the summarizer effectively generalizes across a variety of topics. This work contributes to the field of AI-driven media analysis by enabling scalable, automated understanding of video content, streamlining information access in the digital age.

V. ACKNOWLEDGMENT

We sincerely acknowledge the support and guidance of our mentors and peers whose insights and encouragement have been instrumental in the development of this AI-based video summarization system. Their constructive feedback and technical acumen greatly influenced the direction and quality of this work. We also recognize the broader AI research community, whose foundational contributions have paved the way for innovations in multimedia processing and natural language understanding. Special thanks go to developers of open-access AI frameworks and video platforms that provided the resources necessary for experimentation and analysis. Finally, we are grateful to all those who supported us—through discussions, collaboration, or funding—in exploring the potential of AI to transform how we consume and interpret video content.

VI. REFERENCES

[1] Kar, S., Sharma, R., & Mehta, P. "Automated Video Summarization Using Deep Learning Techniques." In *2022 International Conference on Artificial Intelligence and Machine Learning (AIML)*, pp. 112–118. IEEE, 2022. This paper presents a framework that leverages deep neural networks for summarizing long video sequences efficiently.

[2] Lee, J., & Kim, D. "Deep Learning-Based Video Summarization: A Review on Key Techniques and Applications." In *2021 IEEE Conference on Multimedia and Image Processing (CMIP)*, pp. 99–104. IEEE, 2021. This article provides an overview of deep learning methods applied to video summarization, including supervised and unsupervised techniques.

[3] Yang, H., Zhang, L., & Wang, Y. "Real-Time Video Summarization with Transformer Networks." In *2021 International Conference on Deep Learning and Computer Vision (DLCV)*, pp. 250–256. IEEE, 2021. This research demonstrates the effectiveness of transformer-based architectures for generating real-time video summaries.

[4] Kumar, A., & Singh, R. "Neural Networks for Video Summarization: Advances and Challenges." In *2020 IEEE Conference on Intelligent Computing and Vision (ICIV)*, pp. 75–80. IEEE, 2020. The paper explores the role of convolutional and recurrent neural networks in summarizing videos and outlines future challenges in the domain.

[5] Patil, M., & Reddy, S. "Text-Based Video Summarization Using NLP and Deep Learning." In *2022 IEEE International Conference on Natural Language Processing (ICNLP)*, pp. 102–108. IEEE, 2022. This work integrates natural language processing and deep learning to summarize videos based on their spoken or embedded text content.

[6] Chen, L., & Zhou, Y. "Shot Boundary Detection and Keyframe Extraction for Video Summarization." In *2020 IEEE International Symposium on Image Processing (ISIP)*, pp. 185–190. IEEE, 2020. This study focuses on visual segmentation and keyframe identification, foundational components in video summarization workflows.