Query-Based Video Summariser

Shivam Gaikwad[1], Rabiya Farooq[2], Vartika Mishra[3], Jayesh Kulkarni[4], Prof. Virendra Bagade[5]

[1]Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India,

[2]Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India,

[3]Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India,

[4]Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India,

[3]Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India,

Abstract

In today's digital age, the exponential increase in video content across the web creates a challenge in efficiently retrieving relevant information. Traditional video summaries often fail to meet user-specific needs, resulting in time wasted searching through lengthy content. This paper explores a query-based video summarization system that generates dynamic, user-driven summaries using Natural Language Processing (NLP) and video analysis techniques. The system provides concise snippets based on user input, enabling fast and efficient access to specific content without viewing entire videos. We aim to reduce redundancy and streamline the retrieval of key information through advanced machine learning techniques.

Keywords: Video Summarization, NLP, Video Analysis, User Queries, Content Retrieval, Machine Learning.

1 Introduction

The proliferation of video content on platforms like YouTube, educational websites, news portals, and corporate archives has created vast repositories of digital media. As these repositories grow, users face an overwhelming amount of data to sift through when searching for specific information or insights. Traditional search engines provide metadata-based video retrieval, relying on titles, descriptions, and tags to match content with queries. However, these methods are insufficient when users need to locate specific moments or information within long videos, especially when the video content itself is not indexed or analyzed in depth. In this context, video summarization has emerged as a critical tool for reducing the cognitive load and time required to navigate video content. Video summarization refers to the process of generating a short, concise version of a video that captures its most important moments, highlights, or scenes. These summaries allow users to quickly glean the essential content without watching the entire video. The key challenge, however, is that most existing summarization systems are static and do not cater to the dynamic needs of individual users.

A. Motivation for Query-Based Summarization

With the rise of personalized content consumption, there is an increasing demand for systems that can provide user-driven, dynamic summaries based on specific queries. For example, in an educational video, a student might be interested in a particular topic, while another student might be searching for a specific example or demonstration. Similarly, in news or media analysis, journalists may require precise clips related to a particular event or speaker. A static summary is not sufficient to address such diverse needs, as it fails to adapt to the user's specific query. This project proposes a novel approach to video summarization, where the system generates dynamic, query-based summaries tailored to the user's specific request. By leveraging Natural Language Processing (NLP) techniques, the system is capable of understanding user queries in natural language and processing video content to extract relevant segments. This personalized approach enhances the user experience by focusing on the information that matters most to the user, without the need to sift through hours of irrelevant content.

B. Significance and Challenges

Query-based video summarization holds significant potential for various applications, including education, me-

dia, entertainment, and business. In educational settings, students often struggle to locate specific concepts within lengthy lecture videos. By allowing users to input queries like ”machine learning algorithms” or ”quantum mechanics principles,” the system can automatically retrieve and summarize the relevant sections of the video, reducing the time spent on manual search. Similarly, in the media industry, analysts and researchers can use query-based summarization to quickly extract specific events or statements from long interviews, debates, or news reports. Despite the potential benefits, developing an effective query- based video summarization system presents several challenges. First, there is the challenge of accurate query understanding. Users may phrase queries ambiguously or use synonyms, requiring the system to intelligently interpret the query's intent. This is where advanced NLP models, such as BERT (Bidirectional Encoder Representations from Transformers), come into play, as they can understand the context of user queries and expand them to include relevant terms. Another major challenge lies in the video processing component. Videos consist of both visual and auditory information, and relevant content might be distributed across various parts of the video. Traditional approaches that rely solely on frame-based analysis may overlook critical information conveyed in the audio or through visual transitions between scenes. To address this, our system integrates multiple layers of analysis, including audio, video frames, and transcripts, ensuring that the summarization process is comprehensive and accurate.

C. State of the Art in Video Summarization

The field of video summarization has evolved rapidly over the past decade, driven by advancements in machine learning, computer vision, and natural language processing. Early approaches focused on simple extractive methods, where a predefined set of keyframes or clips were selected based on low-level visual features like color histograms or motion vectors. While these methods provided basic summaries, they often lacked contextual understanding and were not flexible enough to adapt to different user needs. In recent years, deep learning models, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have transformed video summarization by enabling more sophisticated feature extraction and temporal analysis. CNNs are particularly effective at identifying key objects, scenes, and actions within video frames, while RNNs help capture the temporal relationships between different parts of the video. These advancements have paved the way for more accurate and meaningful summaries. Further improvements have come from the integration of NLP techniques, allowing for text-based summarization that complements visual content analysis. Systems that incorporate speech recognition and automatic transcription can generate summaries that are sensitive to both the spoken dialogue and the visual elements of the video. This is particularly valuable in contexts where the spoken content provides crucial insights, such as educational or news-related videos. However, even with these advancements, most existing systems remain static and fail to address the need for real-time, dynamic summaries based on user queries. The use of transformers, which excel at handling sequential data, represents the latest frontier in this domain. Transformer-based models like BERT and GPT-3 can process both video transcripts and user queries to provide more contextually relevant and flexible summaries.

D. The Role of NLP and Machine Learning in Our System

Our query-based video summarization system leverages several key machine learning techniques to enhance its accuracy and usability. At the core of the system is an NLP model that interprets the user's query, transforming it into a machine readable format using methods like Bag of Words (BoW) and word embeddings. This ensures that the system can match user queries with relevant video segments, even if the query is phrased in an unconventional way. Furthermore, the system employs deep learning models, such as CNNs, to analyze video frames and identify key visual elements. These models allow the system to differentiate between relevant and irrelevant content, ensuring that the final summary is concise and targeted. Additionally, we incorporate audio processing tools to analyze the video's audio track, enabling the system to capture key moments based on both visual and auditory cues. By combining these technologies, our system offers a comprehensive solution to the problem of video summarization, addressing the needs of users in various domains, from education to entertainment and beyond.

E. Contribution and Scope

This paper contributes to the field of video summarization by introducing a system that generates personalized, query-based video summaries. The proposed system integrates NLP and machine learning techniques to provide accurate and dynamic content extraction. Unlike traditional video summarization methods that offer fixed, predefined summaries, our system allows users to interactively specify their information needs, making it highly adaptable and user-friendly. We also discuss the challenges and limitations of current approaches, proposing solutions to improve the scalability and performance of video summarization systems. The scope of this paper includes

an exploration of the theoretical underpinnings of video summarization, the practical implementation of the system, and an evaluation of its performance in real-world scenarios. We believe that our system has the potential to significantly reduce the time users spend searching for information in videos, making it a valuable tool for a wide range of applications.

F. Outline of the Paper

The remainder of this paper is organized as follows: Section II reviews related work in the field of video summarization and discusses recent advances in query-based systems. Section III provides an in-depth analysis of the system architecture, detailing the methods used for query encoding, video frame extraction, and relevance matching. Section IV outlines the experimental setup and performance evaluation, including user studies and benchmarks. Finally, Section V concludes the paper with a summary of findings and future research directions.

G. Problem Statement

With increasing video content, users often struggle to find specific sections within lengthy videos. Traditional search methods rely on metadata and cannot adequately identify relevant content within the video itself. This paper proposes a system that generates dynamic video summaries based on user queries, allowing users to access relevant parts of a video without viewing the entire content. Such a system improves efficiency, particularly in contexts where time-sensitive access to information is crucial.

H. Objectives

The main objectives of this project are:
• To provide concise and relevant video summaries based on user queries.
• To reduce the time spent searching through long videos for specific information.
• To simplify the retrieval of relevant content using NLP and video analysis techniques.
• To enhance the user experience by providing dynamic and personalized summaries.

2 Literature Survey

The field of video summarization has been evolving rapidly with the introduction of advanced machine learning models, NLP techniques, and video processing algorithms. In this section, we review key works that have contributed to the development of query-based video summarization.

A. Video Summarization Using NLP Techniques

In their 2019 paper, Sanjana et al. [1] explored the use of NLP for generating summaries from educational videos by analyzing transcripts. Published in the International Journal of Computer Applications (IJCA), their work primarily focused on extracting key phrases from video transcripts to summarize long instructional videos. This method allowed the extraction of relevant information based on specific educational content, making it highly effective for instructional videos. The primary learning from this paper was the effectiveness of transcript based summarization, particularly in educational environments where content can be highly structured.

B. Query-attentive Video Summarization

Ajinkya et al. [2], in their 2020 paper published in the ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), focused on video summarization methods that generate content based on user-specific queries. The authors introduced a query attentive mechanism that utilized attention models to selectively focus on important video segments. This work was particularly relevant to the field of personalized content retrieval, emphasizing the importance of dynamically generating summaries tailored to the user's needs. The key takeaway from this study was the use of attention models in improving the relevance of the summarized content.

C. Video Key Concept Extraction using CNN

Shraddha et al. [3], in a paper published in the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2021, explored the use of Convolutional Neural Networks (CNNs) to extract key video concepts, enabling the creation of summaries by identifying significant visual content within the video. Their work contributed to the

understanding of how deep learning models could automate the extraction of visual features, such as objects and scenes, to create meaningful summaries. The learning from this paper was the potential of CNNs in video frame analysis, which we incorporate into our system to improve accuracy.

D. Automatic Video Summarization by Machine Learning

Pallavi et al. [4], published in 2022 in the Journal of Machine Learning Research (JMLR), explored automatic video summarization using machine learning techniques. This paper proposed a novel approach that combined supervised learning with unsupervised clustering to summarize video content. The system they proposed was capable of learning from user feedback to refine its summaries over time. The main takeaway from this research was the incorporation of machine learning models that adapt to user preferences, which we plan to implement in future iterations of our system.

E. Speech Recognition-Based Summarization for Videos

In a 2021 paper by Sarah et al. [5], published in the IEEE Transactions on Multimedia, the authors reviewed speech recognition methods to create video summaries by extracting relevant transcripts. This technique leveraged automatic speech recognition (ASR) models to create text-based summaries for video content, particularly useful in contexts where audio plays a critical role in conveying information. The integration of ASR in video summarization provides an alternative approach to purely visual analysis, enhancing the system's versatility.

F. Abstractive Video Summarization Using Transformers

Shraddha et al. [6], in their 2022 paper presented at the Conference on Neural Information Processing Systems (NeurIPS), explored abstractive video summarization using transformers. They demonstrated how transformer models, originally designed for NLP tasks, could be adapted to generate abstract summaries of video content. The use of transformers in summarization tasks marked a significant leap forward, as these models can generate more concise and contextually rich summaries compared to traditional extractive methods.

3 Proposed Methods

In this section, we analyze the various techniques employed in our video summarization system and discuss their effectiveness in achieving accurate and meaningful summaries.

3.1 Audio Extraction Techniques

The choice of audio extraction technique significantly influences the quality of the subsequent transcription. FFmpeg is a robust tool that efficiently extracts audio, maintaining the integrity of the sound quality. It supports a wide range of audio and video formats, making it versatile for various applications. FFmpeg allows batch processing and can be easily integrated into automated pipelines, enhancing its usability for large datasets. Other alternatives, such as pydub, can also be considered for simpler tasks. Pydub provides a more user-friendly interface for audio manipulation but may not handle certain formats as efficiently as FFmpeg. While pydub is excellent for straightforward audio tasks, its performance can degrade with large files or complex audio structures. In scenarios where high fidelity is required, especially in noisy environments, FFmpeg's superior processing capabilities are favored.

3.2 Transcription Techniques

System transcription precision depends heavily upon audio quality combined with the reliability of its incorporated models. Our system uses semantic transformers and the L5 model to deliver contextual transcription services because of their advanced architecture design. Such models keep continuous sentence structure and ensure semantic appropriateness particularly when transcribing long or split up speech.

The existing audio processing tasks have benefited from using traditional architectures including LSTM (Long Short-Term Memory) and CNN (Convolutional Neural Networks) because LSTM shows excellence in sequence learning and CNN provides robust spatial feature extraction from spectrograms. These approaches currently fail to deliver satisfactory results when processing real-time audio with noise or uncertainty because they cannot properly retain long dependencies and conduct contextual reasoning.

Semantic transformers deployed in our system rectify these issues because they maintain long dependencies while preventing gradient collapse. The semantic models interpret full phrases to interpret meaning so they achieve

better accuracy in different linguistic environments and dialect variations. A post processing grammar correction system with these models produces outputs that remain accurate while also being grammatically correct in advance of summary generation and analysis operations.

3.3 Summarization Techniques

The summarization process employs a Semantic Transformer model that generates contextual summaries. The model creates fresh text from transcripts instead of picking individual phrases or sentences and it maintains both understandability and logical flow during text generation.

The generative model makes use of user intent matching and chronological segmentation together with language context while producing summary text that matches directly with user objectives. A unique feature of the L5 model allows users to manage language consistency through specific parameters while additional grammar correction tools operate during post generation tasks.

The powerful capabilities of abstractive summarization demand greater computing resources when performing the extraction operation. The optimization of transformer architecture and the use of GPU-backed inference provides our system a solution for reducing computational expense. The system enables users to switch to extractive logic during time sensitive situations yet the generated output might lack depth when compared to complete text content.

Users obtain tailored human like video summaries that are both syntactically correct and semantically relevant and flexible thanks to the integration of these elements.

4 Discussion

The integration of these techniques demonstrates the potential for creating a highly effective video summarization system. However, the choice of each technique must be carefully evaluated based on the specific use case, such as the type of videos being processed and the desired quality of the summaries. For instance, in environments where audio quality is poor, robust audio extraction and transcription techniques become essential to ensure accurate summarization. Furthermore, the choice between abstractive and extractive summarization methods will depend on the need for coherence versus computational efficiency. User feedback plays a critical role in iterating and refining the system to better meet the needs of its users. Continuous evaluation and adaptation of the chosen methods will be necessary to stay relevant in the fast-evolving landscape of video content.

4.1 Equations

To evaluate the performance of the video summarization system, we utilize several metrics that provide insights into transcription accuracy, summarization quality, and user satisfaction.

The BLEU score measures the overlap between the generated summary and reference summaries based on n-gram matching. It is computed as:

$$BLEU = BP \cdot \exp\Big(\sum_{n=1}^{N} w_n \log P_n\Big) \tag{1}$$

The Word Error Rate (WER) metric compares the number of substitutions (S), deletions (D), and insertions (I) in the transcription against the total words (N):

$$WER = \frac{S + D + I}{N} \tag{2}$$

The ROUGE-N score assesses extractive summarization quality by comparing n-gram counts:

$$ROUGE\text{-}N = \frac{\sum_{\text{ngram} \in \text{Reference}} \text{Count}_{\text{match}}(\text{ngram})}{\sum_{\text{ngram} \in \text{Reference}} \text{Count}(\text{ngram})} \tag{3}$$

The F1 score balances precision and recall:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Processing time measures end-to-end efficiency:

$$\text{Processing Time} = t_{audio} + t_{transcription} + t_{summarization} \tag{5}$$

- **BLEU Score**: Measures n-gram precision with brevity penalty.
- **Word Error Rate (WER)**: Measures transcription accuracy.
- **ROUGE Score**: Evaluates overlap of n-grams in extractive summaries.
- **F1 Score**: Balances precision and recall of summary segments.
- **Processing Time**: Sum of audio extraction, transcription, and summarization durations.

Table 1: Comparison of Different Techniques in Video Summarization

Technique	Description
FFmpeg	Fast processing with versatile format support; ideal for general audio extraction, though it has limited advanced features.
pydub	User-friendly interface for audio manipulation; slower than FFmpeg and less flexible for handling various formats.
LSTM	Suitable for sequential data and context handling; requires high computational resources and is complex, making it ideal for complex audio scenarios.
CNN	Offers high accuracy in feature extraction and is efficient in processing; however, it can be sensitive to noise and may miss context. Suitable for robust audio environments.
Transformer	Provides contextual understanding and state of the art performance; requires large datasets and is resource intensive, making it best for complex textual summarization.
Extractive Methods	Simple and quick to implement; may miss context and result in less coherent summaries, ideal for quick summaries or simpler tasks.

5 Conclusion

This paper presents a query-based video summarization system that enables users to retrieve relevant video snippets based on their specific queries. By integrating NLP, video frame extraction, and machine learning, the system reduces the time users spend searching through videos and enhances the user experience. Future work will focus on improving scalability, refining query understanding through advanced NLP techniques, and incorporating real-time processing capabilities.

References

[1] Sanjana, A., Gupta, R., amp; Sharma, P., “Video Summarization Using NLP Techniques,” International Journal of Computer Applications, vol. 182, no. 22, pp. 1-5, 2019.

[2] Ajinkya, P., Patil, S., amp; Gawande, V., “Query-attentive Video Summa- rization,” ACM Transactions on Multimedia Computing, Communica- tions, and Applications, vol. 16, no. 2, pp. 1-24, 2020.

[3] Shraddha, R., Jain, S., amp; Kumar, A., “Video Key Concept Extraction using CNN,” IEEE Conference on Computer Vision and Pattern Recog- nition (CVPR), 2021.

[4] Pallavi, T., Nand, S., amp; Singh, M., “Automatic Video Summarization by Machine Learning,” Journal of Machine Learning Research, vol. 23, no. 1, pp. 1-15, 2022.

[5] Sarah, L., Chen, X., amp; Zha, H., “Speech Recognition-Based Summa- rization for Videos,” IEEE Transactions on Multimedia, vol. 23, pp. 1934-1945, 2021.

[6] Shraddha, R., Yadav, V., amp; Verma, H., “Abstractive Video Summariza- tion Using Transformers,” Conference on Neural Information Processing Systems (NeurIPS), 2022.

[7] Zhang, Y., Wu, S., amp; Li, H., “Video Summarization via Unsupervised Learning and Reinforcement Learning,” IEEE Transactions on Cybernetics, vol. 50, no. 3, pp. 1145-1156, 2020.

[8] Fan, Y., Yu, D., amp; Chen, J., "Deep Learning for Video Summarization: A Review," ACM Computing Surveys, vol. 54, no. 1, pp. 1-38, 2021.

[9] Sharif, M., Khan, M., amp; Ali, A., "A Comprehensive Review of Video Summarization Techniques," Journal of Visual Communication and Image Representation, vol. 72, pp. 103134, 2020.

[10] Jain, S., Gupta, A., amp; Saxena, A., "Video Summarization Using Deep Learning Techniques: A Survey," Computer Science Review, vol. 39, pp. 100-105, 2021.

[11] Lee, J., Park, S., amp; Kim, J., "Automated Video Summarization with Deep Learning: Challenges and Future Directions," Multimedia Tools and Applications, vol. 79, no. 1, pp. 243-268, 2020.

[12] Geng, J., Chen, Y., amp; Zhao, X., "Video Summarization via Attention Mechanisms: A Survey," IEEE Access, vol. 10, pp. 10833-10845, 2022.

[13] Singh, R., Singh, A., amp; Choudhury, R., "Context-Aware Video Summa- rization Using NLP Techniques," Journal of Ambient Intelligence

[14] Lin, Z., Liu, Q., amp; Wu, H., "Abstractive Video Summarization Using Reinforcement Learning and Neural Networks," IEEE Transactions on Multimedia, vol. 22, no. 12, pp. 3152-3166, 2020.

[15] Tran, M., Le, T., amp; Nguyen, T., "Video Summarization Using Temporal Attention Networks," International Journal of Computer Vision, vol.128, no. 4, pp. 973-985, 2020.

[16] Zhang, K., Liu, Y., amp; Wang, R., "A Review of Video Summarization Techniques: Challenges and Future Directions," Journal of Visual Com- munication and Image Representation, vol. 80, pp. 102992, 2022.

[17] Yang, H., Zhang, Y., amp; Wei, L., "Deep Learning-Based Video Summa- rization: A Comprehensive Review," IEEE Transactions on Cybernetics, vol. 50, no. 7, pp. 2995-3010, 2020.

[18] Raj, A., amp; Bansal, R., "A Review of Video Summarization Techniques and Approaches: Challenges and Future Directions," Multimedia Tools and Applications, vol. 80, no. 1, pp. 1-34, 2021.

[19] Chen, Y., Wang, R., amp; Wang, F., "Semantic Video Summarization Based on Deep Learning," Journal of Visual Communication and Image Representation, vol. 71, pp. 102812, 2020.

[20] Xu, C., Zhao, J., amp; Xu, H., "Video Summarization: A Survey of Algorithms and Applications," IEEE Transactions on Multimedia, vol. 23, pp. 777-795, 2021.

[21] Gupta, V., Kumar, R., amp; Jadhav, S., "Video Summarization Using Image Processing Techniques: A Review," International Journal of Computer Applications, vol. 975, no. 18, pp. 1-7, 2022.