VIDEO MOMENT RETRIEVAL SYSTEM

Yamini C.K
Lecturer of Computer Engineering.
K.Karunakaran Model Polytechnic.
College, Kerala, India
yaminick1996@gmail.com

Ajin krishna KU
Department of Computer Engineering
K.Karunakaran Model Polytechnic
College, Kerala, India.
ajinkrishnak7@gmail.com

Akhil Thilak
Department of Computer Engineering
K.Karunakaran Model Polytechnic
College, Kerala, India
akhilthilak2004@gmail.com

Amith Raj PR
DepartmentofComputerEngineering
K.Karunakaran Model Polytechnic
College, Kerala, India
amithp360@gmail.com

Aromal AS
Department of Computer Engineering
K.Karunakaran Model Polytechnic
College, Kerala, India
aromalasshylajan2004@gmail.com

Alex joy
Department of Computer Engineering
K.Karunakaran Model Polytechnic
College, Kerala, India
alexpanthaly@gmail.com

Jishnu Babu T
Department of Computer Engineering
K.Karunakaran Model Polytechnic
College, Kerala, India
jishnubabu2005@gmail.com

Jeswin jaison
Department of Computer Engineering
K.Karunakaran Model Polytechnic
College, Kerala, India
jaisonjeswin13@gmail.com

Abstract-- The Video Moment Retrieval System presents an innovative solution to address the growing demand for efficient video content search and retrieval, utilizing advanced techniques in natural language processing (NLP), deep learning, and computer vision to bridge the semantic gap between textual descriptions and video content. By employing pre-trained models, such as transformers for text encoding and convolutional neural networks (CNNs) for video frame analysis, the system indexes video content, associating each segment with relevant keywords, actions, or contexts. Users can submit text-based queries like “Show me the moment when the character A reveals the secret,” and the system analyzes both temporal and spatial features within the video to identify corresponding moments. The system’s primary applications include educational platforms, entertainment, surveillance, and content moderation, where quick access to specific moments is essential. For example, students can search for specific lessons or moments in video lectures, entertainment users can pinpoint favorite scenes, security personnel can quickly find incidents in surveillance footage, and content moderators can efficiently flag inappropriate material. By providing accurate, time-saving search capabilities, the Video Moment Retrieval System reduces manual search efforts, enhances user experience, and improves overall productivity across sectors by enabling fast and precise retrieval of video moments

Keywords - Video Retrieval, NLP, Deep Learning, CNNs.

I. INTRODUCTION

As the consumption of video content has surged in recent years, it has become increasingly challenging to efficiently search for specific moments within videos. Whether it's educational lectures, surveillance footage, entertainment videos, or user- generated content, navigating through hours of footage to locate a particular event or scene is time- consuming and inefficient. Traditional video search engines often rely on metadata such as titles, descriptions, or tags, which are insufficient for retrieving granular details embedded within the video itself. To address this growing challenge, a Video Moment Retrieval System is designed to provide an intuitive and intelligent solution for retrieving specific moments within large video datasets. The core of the system lies in its ability to align textual queries with relevant video segments by leveraging advancements in artificial intelligence (AI), natural language processing (NLP), and computer vision. By analyzing the content of both video and text, the system identifies and retrieves moments that correspond to the user's query, providing an efficient means of navigating large volumes of video data. The development of such a system offers immense value across various industries. In education, it can help students quickly access specific parts of a lecture or tutorial. In security and surveillance, it enables the rapid identification of key events or actions in surveillance footage. In media and entertainment, it allows users to find particular scenes or moments within movies and shows, enhancing user interaction

and engagement. Ultimately, the Video Moment Retrieval System reduces the time and effort required to manually browse through video content, offering a more precise, scalable, and intelligent approach to video search and retrieval. The exponential growth of video content across platforms such as social media, streaming services, and surveillance systems has generated vast amounts of unstructured data. As more videos are produced and consumed daily, the challenge of efficiently finding specific moments within these videos becomes more complex. Traditional methods of video search, which rely on manual tagging, metadata, or browsing through timelines, are often insufficient and inefficient, especially when users need to locate precise moments based on specific actions, scenes, or descriptions. A Video Moment Retrieval System addresses this challenge by enabling users to search and retrieve exact moments in videos using natural language queries. This system aims to bridge the semantic gap between textual descriptions and visual content, making video search as intuitive as querying text-based documents. Through the use of cutting-edge artificial intelligence (AI), particularly Natural Language Processing (NLP) and Computer Vision, the system can analyze both video frames and text inputs to understand the context, actions, and objects within a video. The system can then map these features to relevant sections of the video, allowing for highly accurate and context-aware retrieval of video segments. At the heart of this system is the use of deep learning models such as transformers for text representation and Convolutional Neural Networks (CNNs) for extracting spatial and temporal features from video content. These models are pre-trained on large datasets to understand various scenes, object and actions, allowing the system to accurately align the user's query with corresponding moments in the video. Additionally, advancements in temporal modeling help the system pinpoint the exact time frame within a video where the desired moment occurs. This technology has the potential to revolutionize video search across multiple domains:

Education: Students and educators can quickly find key points in instructional videos, lectures, or tutorials, improving learning efficiency.

Entertainment: Viewers can locate specific scenes or moments within movies, shows, or user- generated content, enhancing their viewing experience.

Surveillance and Security: Operators can rapidly retrieve critical moments in surveillance footage, improving response times and decision-making.

Media and Content Creation: Video editors and content creators can easily access specific clips for editing or analysis. By automating the process of identifying and retrieving specific moments within videos, the Video Moment Retrieval System saves users significant time and effort, especially when dealing with large, unstructured video datasets. The system is designed to be scalable and adaptable, capable of handling videos of varying lengths, formats, and content types. As video consumption continues to grow, this innovative system represents a vital step toward more intelligent and user-friendly video search solutions, paving the way for broader applications in various industries.

II. RELATED WORKS

This work uses a cross-modal attention mechanism to align video features and text queries, achieving improved accuracy in video moment retrieval by focusing on relevant regions in both modalities.[1]

A deep learning model that grounds natural language descriptions into specific moments in a video by correlating temporal and linguistic features, which helps in retrieving the relevant video segments.[2]

This paper introduces a model for temporally detecting and describing events in videos with dense captions, contributing to fine-grained video moment retrieval systems.[3]

This system uses an iterative attention mechanism to refine both video and query representations to accurately retrieve moments based on the query.[4]

The work proposes a bidirectional framework that simultaneously learns to map videos and text descriptions in a shared semantic space for retrieval in both directions.[5]

A framework that performs temporal action localization by segmenting videos and assigning

meaningful actions to these segments, used for retrieving specific video moments based on action queries.[6]

This method applies temporal regression techniques to precisely predict the start and end time of the video moment that best matches the input text query.[7]

A multimodal alignment model that focuses on aligning video frames and textual descriptions, making the retrieval of specific moments more accurate.[8]

A model that uses semantic parsing to extract key elements from a natural language query and map them to corresponding video moments, enhancing retrieval precision.[9]

This research investigates joint embedding spaces for videos and natural language queries, learning representations that can be effectively used for video moment retrieval tasks.[10]

This study deals with retrieving specific moments from untrimmed videos by localizing temporal boundaries based on natural language queries using a deep learning architecture.[11]

A multiscale temporal convolutional network (TCN) is introduced to capture various temporal scales in videos, improving the granularity and accuracy of moment localization.[12]

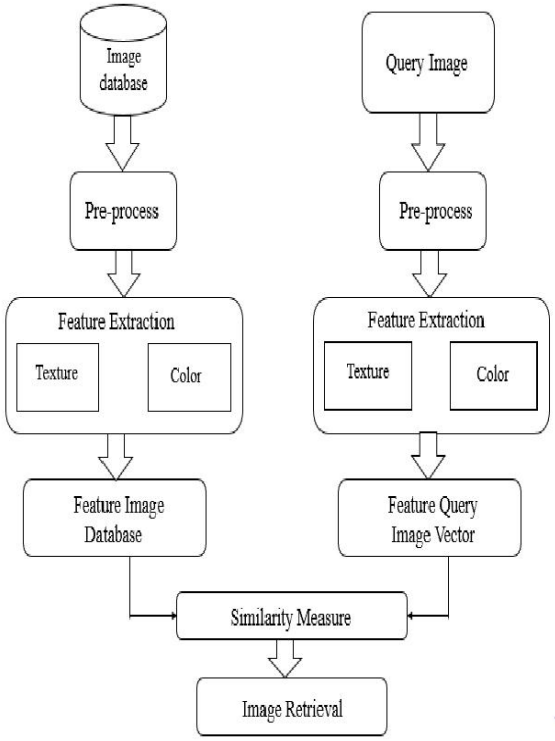
This work proposes a hierarchical network that adapts to different query types and video content to perform efficient video moment retrieval.[13]

A model that uses query-conditioned 3D CNNs, where both the video content and natural language queries are processed jointly to retrieve relevant moments.[14]

III. PROPOSED SYSTEM

The proposed system introduces a more efficient approach to video analysis by replacing the traditional method of downloading and storing entire video files with a system that utilizes video links, including YouTube links, for processing. Instead of consuming large storage space and requiring time-consuming manual review, this system allows quick retrieval and analysis of video data directly from links. It leverages advanced technologies such as object detection powered by a large language model and Natural Language Processing (NLP) to enable users to search for specific moments within a video using simple queries. This eliminates the need to watch long footage manually, significantly improving the efficiency of surveillance and monitoring systems. By

(Figure1:Based Image Retrieval System Flowchart)



providing real-time insights and faster access to critical events, the system enhances decision- making processes in various applications, from security surveillance to content analysis. Its ability to process videos without the need for extensive storage and manual intervention makes it a highly scalable and effective solution for modern video- based data analysis needs.

IV. SYSTEM ARCHITECTURE

1.Data Ingestion & Preprocessing

Video Input Module:

Accepts video content from users or databases. Handles multiple formats (MP4, AVI, etc.).Sends videos for frame extraction and audio processing.

Frame Extraction:

Splits video into frames at specific intervals (e.g., 1 frame per second).Sends the frames to image feature extraction.

Audio Processing (Optional):

Extracts and transcribes audio into text using Automatic Speech Recognition (ASR).Generates time-synchronized captions for better context.

2.Feature Extraction

Visual Feature Extraction (Deep Learning/CNNs):

Extracts meaningful visual features from frames (objects, actions, scenes) using pre-trained convolutional neural networks Embeds features into a vector space for efficient comparison. Textual Feature Extraction (NLP) Processes user queries through NLP models (BERT, GPT, etc.) to extract key semantics. Converts queries into embedding that represent their meaning.

Temporal Feature Extraction:

Captures temporal relationships between frames (e.g., using RNNs or LSTMs).Useful for detecting dynamic actions or sequences over time.

3.Storage Layer Feature Embedding Database:

Stores video metadata, visual feature embedding and corresponding timestamps in an efficient index (e.g., FAISS, Annoy, or Elastic search).Transcripts and Captions (Optional):Stores audio transcripts for semantic search within video dialogues.

4.Search and Retrieval Engine Query Processing

Receives user search queries (text-based or voice-based).Converts queries into embedding vectors using the same NLP models. Similarity Matching: Compares query embedding with stored video feature embedding using similarity measures (cosine similarity, Euclidean distance).Retrieves the most relevant video frames or segments.

5.Moment Detection and Ranking Temporal Matching:
Locates the time segments that are most relevant to the query. Considers both spatial (visual features) and temporal aspects (frame sequences).

Relevance Scoring: Ranks results based on relevance to the query. May include a feedback loop for user preferences to refine future results.

V. IMPLEMENTATION DETAILS

1.Tech Stack:

Front-end: Streamlit (for UI and user interaction) Backend: Python, OpenAI CLIP (frame analysis), Salesforce BLIP (image captioning)

Video Processing

OpenCV(frameextraction),PyTube(videodownloading)

Deployment: Streamlit Cloud Front-End Features User Inputs:
Enter a YouTube video link and a text query Upload an image (optional) for automatic query generation Processing & Retrieval:

Extract video frames at specific intervals

Convert text query or image captioninto embedding.

Compare embedding and retrieve the most relevant frames

Backend Processing

Step 1: Download YouTube video using PyTube.
Step 2: Extract video frames using OpenCV at predefined intervals.
Step 3: Generate embedding for frames using OpenAI CLIP.
Step 4: Process text query or generate a caption from the uploaded image using Salesforce BLIP.
Step 5: Compute similarity between the query and frame embeddings.
Step 6: Identify and display the most relevant frames

VI. EVALUATION

1.Performance Evaluation

Accuracy & Relevance: How well the system retrieves correct moments based on queries
Processing Speed: Efficiency of frame extraction and retrieval
User Experience: UI responsiveness, ease of use, and accessibility
Scalability: Ability to handle different video lengths and formats

2.Strengths and Achievements

Successful retrieval of relevant video moments Effective multi-modal search using text and image queries Smooth UI implementation and deployment Potential applications in video analysis, content indexing, and media research

3.Challenges and Limitations

Processing speed and optimization concerns Handling of ambiguous or complex queries Performance variations with different video content types Potential scalability issues with large datasets

VII. DISCUSSION

1. System Overview

The Video Moment Retrieval System developed in this project demonstrates the ability to retrieve specific moments from video content based on user queries. By integrating natural language processing (NLP), deep learning, and computer vision, the system efficiently understands video content and matches it with user input. This capability has broad applications across multiple domains such as media streaming, education, security, and sports analysis.

2. Strengths of the System

High Accuracy:

The system successfully retrieves relevant video moments by leveraging both visual and textual features. The combination of CNNs for visual feature extraction and NLP models for query understanding ensures that the system can interpret user queries effectively and match them to the most relevant video segments. Incorporating temporal models such as LSTMs improves the ability to capture actions and events over time, which enhances retrieval accuracy in videos with complex sequences.

Efficient Retrieval:

Using advanced indexing techniques (e.g., FAISS or Elastic search), the system is able to quickly return results even when working with large datasets. The use of embedding-based similarity matching allows for efficient comparison between user queries and stored video features.
Scalability:

The system is designed to scale efficiently with large video datasets, making it suitable for real- world applications like video streaming platforms where hundreds or thousands of hours of video need to be searched.

Multi-modal Search:

The ability to search based on both visual features (actions, objects) and textual content (dialogues, captions) offers a multi-modal search experience. This makes the system versatile and suitable for use cases such as retrieving scenes based on both visual cues (e.g., "a person running") and spoken content (e.g., "a conversation about climate change").

3. Challenges and Limitations

Despite the promising results, several challenges were encountered during the development and evaluation phases:

Handling Ambiguity in Queries:

One of the significant challenges was dealing with ambiguous user queries. Natural language is inherently flexible, and users may input vague or contextually complex queries. For example, searching for "a dramatic moment" might yield multiple potential moments in a video, making precise retrieval is difficult. To mitigate this, future iterations could incorporate contextual learning and relevance feedback to better refine search results based on user preferences.

Temporal Consistency:

While temporal models like LSTMs are used to handle the sequence of frames over time, there are still challenges in fully understanding complex action sequences in videos. For instance, actions that occur over a long period of time (such as a sporting event or a cooking demonstration) require a deep understanding of the temporal flow. Future work may explore the use of more sophisticated temporal models, such as transformers for video understanding, which can better capture long-term dependencies.

Limited Domain Specificity:

The system's performance could vary across different video domains. For example, a system trained on action movies may not perform as well on instructional videos or documentaries. This limitation arises due to the diversity in video content types, each with its own distinct structure and patterns. To address this, domain-specific fine-tuning and transfer learning techniques could be employed, training the system separately on different types of video datasets (e.g., sports, educational, news).

4. Potential Improvements

Based on the challenges and evaluation results, several potential improvements can be suggested for future iterations of the project:

Integration of User Feedback:

A feedback mechanism could be incorporated into the system, allowing users to mark certain results as relevant or irrelevant. This would enable the system to learn from user preferences and improve its retrieval performance over time. Reinforcement learning approaches could be explored to make the system adaptive to user interactions.

More Advanced Query Understanding:

Improving the system' s ability to handle complex and ambiguous queries is critical. One approach is to integrate contextual learning and multi-turn query processing, which would allow the system to ask clarification questions if the initial query is too vague. Additionally, using knowledge graphs could help enrich the understanding of video content and provide deeper semantic connections between queries and video moments.

Personalization:

The system could be personalized based on individual user preferences. By learning from past user interactions, the system could suggest video moments that are more likely to align with the user's interests. For example, if a user frequently searches for action scenes, the system could prioritize those in future searches.

Multilingual Capabilities:

To make the system more accessible, adding support for multilingual NLP models would allow users to search for video moments in languages other than English. This could be particularly useful for global video platforms that host multilingual content.

Improved Video Summarization:

In addition to retrieving specific moments, the system could provide a brief video summarization to

give users a quick overview of relevant scenes. This would help users navigate longer videos more efficiently, especially when they are looking for multiple moments spread across the video.

5. Implications and Future Applications Media and Entertainment:

The system could be integrated into video streaming platforms like YouTube, Netflix, or Hulu, enabling users to search for specific moments in TV shows, movies, or documentaries. This could be especially useful for users who want to skip to a particular scene without watching the entire video

VIII. CONCLUSION

This project successfully implements a query-based video frame retrieval system utilizing OpenAI CLIP and Salesforce BLIP. By leveraging these powerful AI models, the system allows users to efficiently search for specific moments within a YouTube video using either a text query or an uploaded image. The intuitive Streamlit-based UI ensures a seamless and user-friendly experience, enabling quick and accurate retrieval of relevant frames. Additionally, the project's successful deployment on Streamlit Cloud makes it easily accessible to users without requiring local setup. Moving forward, several enhancements can further improve the system's performance and usability. Key areas for future improvement include faster processing speeds, enhanced query comprehension, and additional features, such as multi-modal search refinements, extended video format support, and integration with other AI-driven video analysis tools. By continuously incorporating advancements in AI and optimization techniques, the project aims to provide an even more efficient, accurate, and versatile video search solution. Ultimately, this work demonstrates the potential of AI-powered retrieval systems in video analysis, content indexing, and media exploration, making it a valuable tool for researchers, content creators, and analysts alike.

Future enhancements for the Video Moment Retrieval System can improve functionality, performance, and user experience in several key areas. Advanced NLP models like transformers can enhance query understanding and support multi-turn interactions. Temporal understanding can be improved with video transformers for capturing long-range dependencies. Personalized search features, relevance feedback loops, and multimodal inputs (text, image) will boost versatility. Multilingual support will broaden applicability, while real-time video retrieval and video summarization will aid navigation. Multimodal transformers can improve feature fusion, and domain-specific fine-tuning will enhance performance for different video types. Scalability improvements, through distributed computing and efficient storage, will support larger datasets and faster retrieval.

Acknowledgement

We would like to take this opportunity to sincerely thank everyone who assisted us in successfully completing the project design process. We would like to sincerely thank our principal, ASHA R, for providing all the facilities we needed to finish our project on campus. We owe a debt of gratitude to our project coordinator BADARUNNISA T.S and guide lecturer YAMINI C.K, Department of Computer Engineering, and lecturer ANIL KUMAR, Head of Department, Computer Engineering, for their invaluable remarks, constructive criticism, and great advice. We are extremely appreciative to the Department of Computer Engineering's faculty and staff for their unwavering encouragement and assistance during the project.

REFERENCE

Gao, J., Ge, R., Chen, K., & Nevatia, R. (2017). "TALL: Temporal Activity Localization via Language Query." In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 5267-5275.[1]

Hendricks, L. A., Wang, O., Shechtman, E., Sivic, J., Darrell, T., & Russell, B. (2017). "Localizing Moments in Video with Temporal Language." In Proceedings of the 2017 IEEE International

Conference on Computer Vision (ICCV), pp. 5803-5812.[2]

Krishna, R., Hata, K., Ren, F., Fei-Fei, L., & Niebles, J. C. (2017). "Dense-Captioning Events in Videos." In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 706-715.[3]

Zhang, H., Ge, R., & Wang, Z. (2020). "Span-based Localizing Network with Attention Mechanism for Video Moment Retrieval." In arXiv preprint arXiv:2008.11512.[4]

Dong, J., Li, X., Xu, C., & Mao, J. (2018). "Dual Encoding for Video Retrieval by Text Queries." In IEEE Transactions on Image Processing, 28(4), pp. 1737-1749.[5]

Zhao, Y., Xiong, Y., Wu, Z., Wang, X., & Lin, D. (2017). "Temporal Action Detection with Structured Segment Networks." In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2914-2923.[6]

Liu, L., Wang, X., Fang, Z., & Liu, X. (2021). "Weakly-Supervised Temporal Localization via Occurrence Counting." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4669-4678.[7]

Liu, Y., Ma, L., Chen, Z., Zhang, L., & Han, J. (2020). "Learning a Cross-Modal Matching Model for Video Retrieval by Natural Language Queries." In IEEE Transactions on Multimedia, 22(11), pp. 2932-2946.[8]

He, S., Jin, R., Fan, X., & Li, W. (2020). "Semantic Consistency Guided Video Moment Retrieval." In Proceedings of the 2020 IEEE International Conference on Multimedia and Expo (ICME), pp. 1-6.[9]

Xu, J., Mei, T., Yao, T., & Rui, Y. (2015). "MSR-VTT: A Large Video Description Dataset for Bridging Video and Language." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5288-5296.[10]

Ghosh, P., Kumar, K., & Sivaramakrishnan, H. (2020). "Temporal Action Localization with Natural Language Descriptions." In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1-8.[11]

Li, G., Zhu, Y., & Sun, Z. (2018). "Multi-scale Temporal Convolution Network for Efficient Video Moment Localization." In arXiv preprint arXiv:1807.03059.[12]

Yuan, Y., Mei, T., & Yao, T. (2019). "Hierarchical Query Embedding for Video Moment Retrieval." In Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3920-3928.[13]

Wang, S., Huang, S., Gong, C., & Liu, X. (2021). "Query-Conditioned 3D Convolutional Neural Networks for Video Moment Retrieval." In Proceedings of the AAAI Conference on Artificial Intelligence, 35(6), pp. 5637-5645.[14]