ISSN: 2455 - 1597

ijergs
Open Access

International Journal of Engineering Research and Generic Science (IJERGS)
Available online at: https://www.ijergs.in

Volume - 12, Issue - 3, May – June – 2026, Page No. 86 – 98

Query Based Video Summarization Using Machine Learning

[1]Sneha Pasrija, Department of Artificial Intelligence and Data Science (AI&DS), IIMT College of Engineering, Greater Noida, Uttar Pradesh, India

[2]Vaishnavi Pandey, Department of Artificial Intelligence and Data Science (AI&DS), IIMT College of Engineering, Greater Noida, Uttar Pradesh, India

[3]Sneha Sharma, Department of Artificial Intelligence and Data Science (AI&DS), IIMT College of Engineering, Greater Noida, Uttar Pradesh, India

[4]Dr. Shaina, Associate Professor, IIMT College of Engineering, Greater Noida, Uttar Pradesh

Abstract

Video summarization plays a crucial role in efficiently extracting key information from lengthy videos. However, traditional methods often lack consideration for user intent, resulting in summaries that may not align with user preferences. In this paper, we propose a novel approach for video summarization that takes into account the user's query to generate personalized summaries tailored to their interests. Our method leverages a combination of state-of-the-art machine learning techniques. We utilize OpenCV for image processing, EfficientDet for object detection to extract features from each frame, NLTK for word processing, and Spacy for comparing features and user queries to select key frames. The rapid growth of video data across digital platforms has made it challenging for users to efficiently access relevant information. Traditional video summarization techniques often generate generic summaries without considering individual user preferences. This paper presents a query-based video summarization approach that produces personalized summaries based on user input. The proposed system combines computer vision and natural language processing techniques, utilizing OpenCV for frame extraction, EfficientDet for feature detection, and NLP tools such as NLTK and SpaCy for query processing and similarity matching. The user query is analyzed to extract meaningful keywords, which are compared with features identified in video frames to select the most relevant segments. Experiments conducted on datasets such as TVSum, YouTube videos, and a custom real-time dataset demonstrate that the proposed approach improves the relevance and usefulness of generated summaries. This work emphasizes the importance of incorporating user intent into video summarization systems to enhance user experience.

Our experiments demonstrate the effectiveness of our approach in generating user-centric video summaries, outperforming traditional methods by considering user intent. This work contributes to the advancement of video summarization techniques, paving the way for personalized video content consumption experiences.

Keywords: Machine Learning, Open CV, Efficientdet, Spacy, NLTK, Preprocessing.

Introduction

The increasing availability of video content on digital platforms has created significant challenges in efficiently accessing relevant information. Users often need to go through long videos to find specific content, which is time-consuming and inefficient. Video summarization addresses this problem by generating shorter versions of videos that highlight important information.

However, most traditional summarization techniques produce the same output for all users, without considering their individual interests. This limitation reduces the usefulness of the summaries, especially when users are searching for specific information. To overcome this issue, query-based video summarization has emerged as an effective approach, where summaries are generated based on user-defined queries.

Recent advancements in machine learning and deep learning have enabled systems to better understand both visual and textual information. These systems combine feature extraction from video frames with natural language processing techniques to identify relevant content. In this work, we propose a method that integrates object detection and semantic analysis to generate summaries tailored to user queries. The objective is to improve the relevance, accuracy, and efficiency of video summarization while enhancing the overall user experience.

Managing and consuming video data through conventional means is becoming inadequate as its exponential growth continues. Users become overloaded with information due to the constant barrage of content, which lowers engagement. Aware of this difficulty, scientists have set out to create cutting-edge strategies that use machine learning to control the torrent of video data. Of these methods, query-based video summarization stands out as a potentially useful paradigm, providing a way to condense large volumes of footage into succinct, pertinent summaries.

A survey of recent literature reveals a rich tapestry of research endeavors aimed at tackling the video summarization conundrum. Notable contributions include Huang et al. (2023)[1], who propose query-based video summarization with pseudo label supervision, and Messaoud et al. (2021)[2], creators of DeepQAMVS, a query-aware hierarchical pointer network for multi-video summarization. These seminal works lay the foundation for subsequent advancements, setting the stage for the development of more sophisticated techniques.

Central to the User Interest-Based Video Summarization System is a paradigm shift from one-size-fits-all approaches to personalized content curation. Leveraging insights from Narasimhan et al. (2021)[3], who introduced CLIP- It a language-guided video summarization model, the system employs cutting- edge machine learning algorithms to discern user preferences and tailor summaries accordingly. By harnessing the power of deep reinforcement learning, as advocated by Zhang et al. (2019)[7], the system learns to adapt and refine its summaries over time, ensuring an ever-improving user experience.

Need and Applications of Video Summarization:

Because there is so much video content available these days, video summary is essential in the digital era. Time-saving strategies and effective information extraction are critical as viewers encounter a multitude of movies on many platforms. As we have studied in Query based video summarization with pseudo label supervision Video summarization fills this requirement by dividing lengthy videos into manageable segments that viewers can see quickly and comprehend. Video summarizing is an essential technique for optimizing video consumption in today's fast-paced digital landscape. It is particularly useful for content indexing, fast retrieval, and enhancing user experiences when navigating and comprehending big video datasets.

Applications for video summarizing can be found in many different domains, greatly improving content management and information retrieval. In the context of online learning, it makes learning more effective by offering succinct summaries

of extensive lectures or tutorials. As cited in DeepQAMVS Query- Aware Hierarchical Pointer Networks for Multi-Video Summarization, Video summary helps both journalists and spectators understand important events quickly in news reporting. It helps businesses analyse CCTV footage more quickly, which improves security measures. Furthermore, CLIP-It Language-Guided Video Summarization suggests that video summary plays a crucial role in social media, since it facilitates content surfing and enables users to quickly understand the main points of videos. All things considered; this technology improves user experiences in a variety of industries by reducing the amount of video content to make it easier to consume quickly.

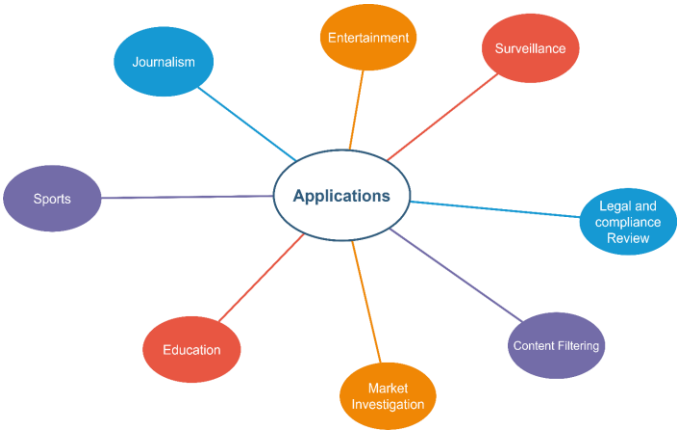


Figure 1: Applications of Video Summarization

Research Objectives:

- Examine different query-dependent video summarizing systems.
- Identify weaknesses in current strategies.
- Clear the path for more desired outcomes in video summarization.
- Develop novel algorithms tailored for providing video summaries based on user preferences.
- Utilize machine learning approaches to customize the summarization process for each user.
- Advance the field of video summarization through algorithmic innovation.
- Validate the developed technology in real-world applications.
- Verify the algorithms' efficiency in processing and summarizing movies in dynamic contexts.
- Use real-time scenarios for validation.
- Conduct a thorough assessment of the suggested user interest-based video summarizing system's effectiveness and efficiency.
- Utilize benchmark datasets for evaluation.

Literature Survey

Video summarization has been an active research area, with various techniques developed to improve efficiency and content relevance. Earlier approaches focused mainly on selecting visually important frames based on low-level features

such as color, motion, and scene changes. While effective to some extent, these methods did not consider user preferences, resulting in generalized summaries.

More recent methods have introduced query-based summarization, where user input is used to guide the summarization process. Some approaches utilize weak supervision or pseudo-labeling techniques to address the challenge of limited labeled data. Other methods employ deep learning models, such as hierarchical networks, to capture relationships between frames and generate more coherent summaries.

Language-guided video summarization techniques have also gained attention, where textual queries are used to identify relevant video segments. These methods combine visual features with semantic representations to improve the alignment between user intent and selected content. Additionally, reinforcement learning-based approaches have been proposed to optimize summary generation by maximizing relevance and diversity.

Despite these advancements, challenges remain in accurately matching user queries with video content and maintaining computational efficiency. Therefore, there is a need for improved systems that effectively integrate feature extraction and semantic understanding to generate personalized video summaries.

The performance of supervised deep video summarization models is constrained by the expensive and consequently small size of the datasets currently available which are manually labeled. By using a pretext task and designing a mechanism to obtain additional data with pseudo labels to pre-train a supervised deep model, self-supervision can address the problem of data sparsity[1]. The demand for systems that can rapidly browse, extract, and summarize video content has grown as a result of the recent expansion of web video-sharing platforms. DeepQAMVS is a new Query-Aware Hierarchical Pointer Network for Multi-Video Summarization that jointly optimizes several criteria, including conciseness, representativeness of significant query-relevant events, and chronological soundness[2]. Users should have the option of customizing the summary by using natural language to specify what is important to them because the significance of scenes in a video is frequently arbitrary[3].

Traditional video summarization approaches generate only a single video summary since it is based on human assistance. The new method models the video summarization as a supervised learning problem[4]. Prior methods of video summarization mainly focused on selecting the most varied and representative visual content for the video summary without taking the user's preferences into account. Convolutional Hierarchical Attention Network (CHAN), which is made up of the query-relevance computing module and the feature encoding network[5]. To develop a novel approach to multi-video summarization that is query-dependent and can generate a summary that is both logical and readable. A Multi-Video Summarization via Multi-modal Weighted Archetypal Analysis (MVS-MWAA) method to extract a concise summarization[6].

Query-conditioned video summarization, which aims to predict a summary of a video that is relevant to a user's query in a concise manner. A deep reinforcement learning approach that consists of a Mapping Network and a Summarization Network[7]. Less research has been done on the challenge of producing a video summary that additionally emphasizes information pertinent to a search query. Measure the distance between frames and queries in a common textual-visual semantic embedding space produced by a neural network extend the model to incorporate properties that aren't affected

by the query, like frame Quality[8]. Because they don't take into account users' search intentions, standard multi-video summarizing techniques frequently don't yield pleasing results. Using an unauthorized multi-graph fusion technique, an event- keyframe presentation structure combines keyframes together for certain events linked to the query[9].

The literature review table of different research papers is displayed in table no. 1 above the evaluation matrix, the algorithm, and the dataset that were used.

Methodology

The increasing availability of video content on digital platforms has created significant challenges in efficiently accessing relevant information. Users often need to go through long videos to find specific content, which is time-consuming and inefficient. Video summarization addresses this problem by generating shorter versions of videos that highlight important information.

However, most traditional summarization techniques produce the same output for all users, without considering their individual interests. This limitation reduces the usefulness of the summaries, especially when users are searching for specific information. To overcome this issue, query-based video summarization has emerged as an effective approach, where summaries are generated based on user-defined queries.

Recent advancements in machine learning and deep learning have enabled systems to better understand both visual and textual information. These systems combine feature extraction from video frames with natural language processing techniques to identify relevant content. In this work, we propose a method that integrates object detection and semantic analysis to generate summaries tailored to user queries. The objective is to improve the relevance, accuracy, and efficiency of video summarization while enhancing the overall user experience.

In the realm of digital multimedia content, the exponential growth of online video platforms has led to an overwhelming abundance of video content. However, the effectiveness of traditional methods for video summarization is challenged by the lack of descriptive metadata and the inability to incorporate user queries into the summarization process. Addressing this challenge, we propose a method as shown in the Fig. 2 for generating video summaries that are tailored to user interests, leveraging machine learning techniques and content analysis.

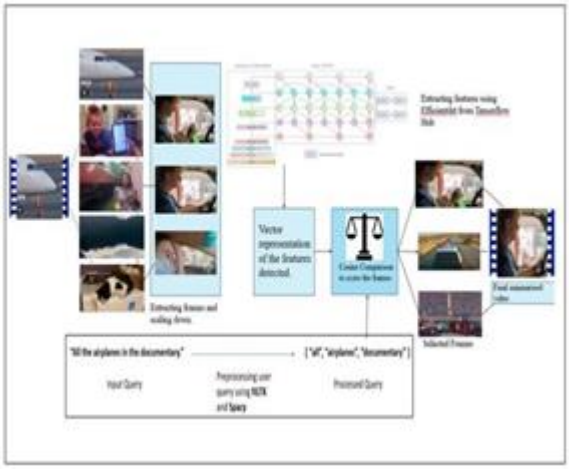


Figure 2: Flowchart

Dataset

To facilitate the development and evaluation of our proposed method, we curated and utilized three distinct datasets:

TVSum Dataset: A widely-adopted benchmark dataset in the field of video summarization, comprising a diverse collection of professionally-produced videos across various domains.

YouTube Dataset: An eclectic selection of videos sourced from the popular online video-sharing platform, encompassing a broad spectrum of content types and genres.

RTV (Realtime Video) Dataset: A novel dataset created specifically for this research, featuring videos captured by smartphone cameras depicting everyday objects and scenarios encountered in real-life situations.

Feature Extraction

Central to our approach is the extraction of salient features from video frames, which serve as the basis for summarization. We employed the state-of-the-art EfficientDet model for feature extraction, retaining features with a confidence level exceeding 0.3. These extracted features are then subjected to further processing to determine their relevance to user queries.

Machine Learning Model

In our pursuit of an effective video summarization model, we conducted comparative experiments with several popular deep learning architectures, including ResNet50 and InceptionV3. However, our empirical findings demonstrated that EfficientDet yielded the most accurate results in terms of relevance to user queries, thus serving as the primary model for our methodology.

Evaluation Metrics

The evaluation of our proposed video summarization method entails the assessment of its performance against established metrics, including Precision, Recall, and F-score. To facilitate the evaluation process, we have initiated a crowdsourcing effort for manual annotation, wherein human annotators provide ground truth summaries for comparison with our generated summaries.

Experimental Setup

The implementation of our methodology was realized within a Python programming environment, leveraging a suite of libraries tailored to image processing, natural language processing (NLP), and video analysis. Notably, OpenCV was utilized for video frame extraction and processing, EfficientDet for feature extraction, and NLTK and Spacy for text preprocessing and semantic analysis.

Methodology Overview

Our methodology comprises several key steps, outlined as follows:

Frame Extraction: In the fig. 3 as we can see video frames are extracted at a predefined frames-per-second (fps) rate using OpenCV, ensuring comprehensive coverage of temporal information.

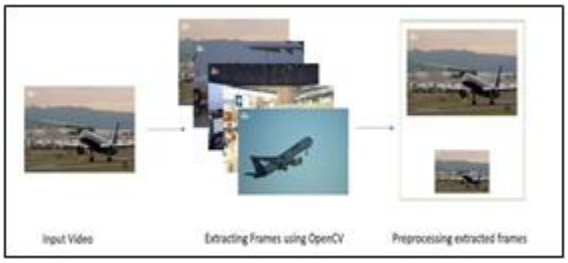


Figure 3: Frame Extraction using OpenCV

Preprocessing: Extracted frames undergo preprocessing, including scaling down and noise reduction, to enhance computational efficiency and feature extraction accuracy. The preprocessing is depicted in the Fig. 4 .

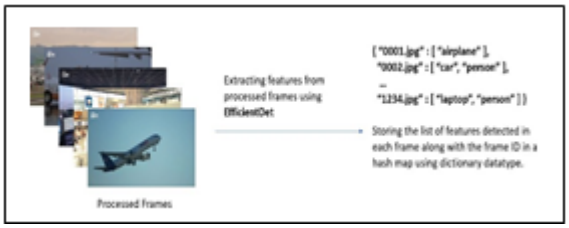


Figure 4: Preprocessing the extracted frames

Query Processing: User queries are preprocessed using NLTK to remove stop words and tokenize the query into meaningful components, facilitating semantic analysis and comparison with extracted features as shown in Fig. 5.

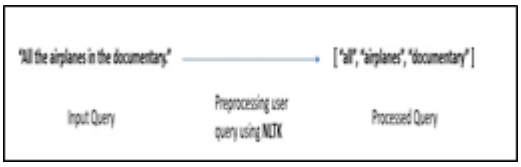


Figure 5: Query preprocessing using NLTK

Feature Comparison: In the fig. 6 as we can see the processed user queries are compared with extracted video features using Spacy, a natural language processing library, to compute relevance scores indicative of feature-query alignment. The Fig. 6 shows comparing the similarities between lists of objects.

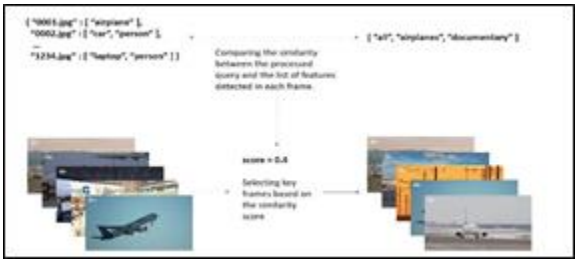


Figure 6: Comparing the similarities between lists of object

Key Frame Selection: Key frames, representing the most relevant segments of the video content with respect to the user query, are selected based on the computed relevance scores as shown in Fig. 7. These key frames are subsequently stitched together to form the final video summary using OpenCV's image manipulation capabilities.

Figure 7: Key frame Selection

Algorithm

- Frames are extracted from the input video by specifying the FPS required and stored in a folder named frames. frame_interval = int(fps / desired_fps)
- Frames are processed using OpenCV and loaded into a folder named train.
- The images in train are loaded and stored in list images=[0001.jpg,0002.jpg, ... N.jpg]
- Features are extracted from each image in images using EfficentDet and stored in a hashmap using Dictionary datatype.
  feature_list={ “0001.jpg” : [fx1],
  “0002.jpg” : [fx2, fx3],
  “N.jpg” : [fxN] }
- Input query is processed for stopwords removal and tokenization using NLTK.
  query=[w1, w2 ... wN]
- The processd query is now compared with the list of features detected in each frame and a smiliarity score is given using Spacy.
  score = doc1.similarity(doc2)
- Based on this relevance score key frames are selected.

if relevance_score(query,fx)>0.3: selected_frames.append(fr)

- All the selected key frames are stitched together to generate the summary video.

Datasets And Evaluation Matrix

Datasets: Datasets are essential for the development and assessment of the algorithms that are trained. These datasets have to be varied, encompassing a variety of video genres, to guarantee adaptability. Annotating datasets with user interest information, such ratings or feedback, allows for the simulation of real-world scenarios.By enabling the machine learning model to generalize well across a range of user preferences, this variation increases the resilience of the system. Moreover, benchmark datasets allow for comparative analyses with existing methods, ensuring the effectiveness of the proposed video summarizing system.

TVSum: TVSum (TVSum50) is a dataset widely used in the domain of video summarization research, and it can be particularly relevant to the investigation of "User Interest-Based Video Summarization Using Machine Learning." TVSum consists of 50 videos, each associated with multiple user-generated summaries, as cited in [4]. These summaries

reflect the diverse perspectives and interests of users, making the dataset valuable for training and evaluating machine learning algorithms aimed at personalized video summarization.

SumMe: SumMe can be used as a benchmark dataset to evaluate how well algorithms designed to produce customized video summaries perform. SumMe can be used by researchers developing user-centric methods to assess how well their algorithms identify and prioritize content that is in line with the interests of specific users. By comparing the generated summaries with human comments or ground truth summaries, this evaluation approach enables researchers to measure the efficacy of their methods statistically.

QueryVS: QueryVS refers to Query-Dependent Video Summarization. This concept involves tailoring the video summarization process based on specific user queries or interests. Unlike generic summarization methods, which aim to condense videos without considering individual preferences, QueryVS aims to personalize the summarization output according to what a user is specifically interested in.

Real Time Videos (RTV): Within the dynamic field of video summarization research, the Real Time Videos (RTV) dataset is a crucial tool that provides a new angle on unaltered, real-life footage shot by cell phone cameras in typical home environments. With over 30 videos at this point, the RTV dataset is constantly growing as efforts are made to broaden its scope and diversify its content. Although the dataset is currently only accessible for research purposes, efforts are being made to expand its availability upon request, which should encourage more cooperation and creativity in the field of user interest-based video summarization.

The Real Time Videos (RTV) dataset constitutes a collection of non- professional videos captured using mobile phone cameras in everyday household settings, portraying raw footage of common scenarios. Unlike conventional datasets in video summarization research, RTV emphasizes unedited content, offering a unique glimpse into real-life environments. Access to the dataset is currently restricted to research purposes, with plans for future availability upon request. The significance of RTV lies in its provision of authentic, unfiltered content, enabling exploration of summarization techniques tailored to non-professional video contexts. Efforts are underway to expand the dataset and facilitate broader research collaboration in the field of user interest- based video summarization.

Evaluation Matrix

An evaluation matrix helps check how well the system is working. It looks at things like how much the video summaries match what users are interested in, if the system can make personalized summaries, and if it covers all the important parts of the videos. The matrix also checks if the summaries make sense in terms of timing and if the system can do this in real-time. User feedback, like what people say in surveys, is considered, and the system's performance is tested against existing standards. This evaluation matrix is like a checklist to make sure the system is doing a good job of summarizing videos based on what users like. system's effectiveness against established standards.

F1 Score: When discussing machine learning, the F1 score is a statistic that is frequently employed in applications such as user interest-based video summarization. It is a metric for evaluating a model's accuracy that takes recall and precision into account. When used in the context of video summarization, precision refers to how well the algorithm chooses pertinent video parts depending on user interest. A high precision means that a significant amount of the content in the

created video summary is in line with the user's preferences. Conversely, recall gauges how well the system can include all pertinent parts into the summary, guaranteeing that no important information is missed. The F1 Score can be calculated using the eq. (1) as shown:

The traditional F-measure or balanced F-score (**F_1 score**) is the harmonic mean of precision and recall:[2]

$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2\frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} = \frac{2\mathrm{tp}}{2\mathrm{tp} + \mathrm{fp} + \mathrm{fn}}. \quad \text{…….(1)}$$

Precision: Precision is the term used to describe the relevance and correctness of the video summaries that the system generates depending on user interests. Precision is the ratio of pertinent content to overall content offered in the video summaries as shown in eq. (2). A greater degree of precision guarantees that the information gathered is useful and relevant because the generated summaries closely match the user's interests. Achieving high precision is critical in this case because it immediately impacts the user experience by providing user- tailored, succinct video summaries that closely align with their preferences.

$$\text{Precision} = \frac{\text{Relevant retrieved instances}}{\text{All retrieved instances}}$$

$$\text{Precision} = \frac{tp}{tp + fp} \quad \text{….(2)}$$

Recall: Recall is the system's capacity to find and play appropriate videos that match the user's interests. More specifically, recall quantifies the percentage of pertinent videos that the system correctly recognizes from all of the relevant videos in the collection. Achieving high recall is very important in machine learning-based video summarization to make sure the system records and incorporates all relevant information that corresponds with the user's preferences. By providing a more thorough and customized video summary, a better recall shows that the system is good at remembering important information and improving the user experience. The formula to calculate recall is given below in eq. (3).

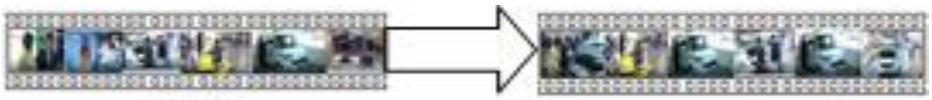
$$\text{Recall} = \frac{\text{Relevant retrieved instances}}{\text{All relevant instances}}$$

$$\text{Recall} = \frac{tp}{tp + fn} \quad \text{……(3)}$$

Results and Discussions

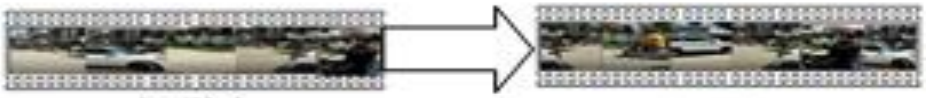
Table 3: Same Video Different Queries

		TVSum			YouTube				RTV			
SR.NO	TITLE	QUERY	ORIGNAL VIDEO LENGTH	SUMARIZED LENGTH	TITLE	QUERY	ORIGNAL VIDEO LENGTH	SUMARIZED LENGTH	TITLE	QUERY	ORIGNAL VIDEO LENGTH	SUMARIZED LENGTH
1	37rzWOQsNIw	birds	3m 11 s	38s	Crazy Passenger Refuses to Miss Flight	typing on laptop	2m 36s	4s	college video	cellphone	4m 30s	20s
2	37rzWOQsNIw	cars	3m 11 s	12s	Crazy Passenger Refuses to Miss Flight	train arriving	2m 36s	2s	college video	person	4m 30s	10s
3	37rzWOQsNIw	food	3m 11 s	1m 15s	Crazy Passenger Refuses to Miss Flight	cell phone	2m 36s	5s	college video	food	4m 30s	5s
4	37rzWOQsNIw	a cup of coffee	3m 11 s	1s	Crazy Passenger Refuses to Miss Flight	cat	2m 36s	7s	college video	car	4m 30s	40s
5	37rzWOQsNIw	burger	3m 11 s	23s	Crazy Passenger Refuses to Miss Flight	dog	2m 36s	9s	college video	plants	4m 30s	30s



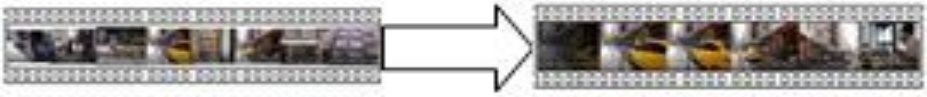
Input Video 1

Output Video 1



Input Video 2

Output Video 2



Input Video 3

Output Video 3

Table 4: Different Query Different Video

		TVSum			YouTube				RTV			
SR.NO	TITLE	QUERY	ORIGNAL VIDEO LENGTH	SUMARIZED LENGTH	TITLE	QUERY	ORIGNAL VIDEO LENGTH	SUMARIZED LENGTH	TITLE	QUERY	ORIGNAL VIDEO LENGTH	SUMARIZED LENGTH
1	AV6I7DC8FM0	bicycles in the video	1m 44s	20s	Inside Virat Kohli's Spacious Nature Inspired Holiday Home | AD India	furniture	3m 53s	35s	potted plant	potted plant	3m 44s	26s
2	37rzWOQsNIw	birds and dogs	3m 11s	25s	The True Scale of the World's Largest Airports	airplanes	6m 44s	1m 48s	dog	dog	28s	1s
3	LRtw_obCPUY	burgers and salads	4m 24s	20s	Despicable Me 4 | Official Trailer	car scenes	2m 26s	2s	scissor	scissor	2m 50s	44s
4	xTI2EL5N-vY30	train accident	2m 28s	21s	Harley Davidson Street 750 Accident With Cow	cow accident	1m 30s	5s	knife	knife	2m 44s	1s
5	WG0MBPgPC6I	chopping vegetables	6m 37s	1s	HGV almost crushes car in shocking footage!	traffic lights	24s	1s	refrigerator	refrigerator	3m 1s	36s

Future Scope

The proposed system for user interest-based video summarization demonstrates promising results; however, there are several areas where further improvements and advancements can be made:

Integration of Deep Learning Models: Future work can incorporate advanced deep learning architectures such as Transformers and attention-based models to improve accuracy and contextual understanding of video content.

Real-Time Video Processing: The system can be enhanced to support real-time video summarization for live streaming platforms, enabling instant generation of personalized summaries.

Multimodal Analysis: Currently, the system mainly relies on visual and textual features. Future improvements can include audio analysis (speech recognition, sentiment analysis) to generate more comprehensive summaries.

Improved User Personalization: User profiling techniques and recommendation systems can be integrated to learn user preferences over time and provide more accurate and adaptive summaries.

Scalability and Cloud Deployment: The model can be deployed on cloud platforms to handle large-scale video datasets efficiently and make the system accessible as a web or mobile application.

Support for Multiple Languages: Enhancing the NLP module to process queries in multiple languages will make the system more globally applicable.

Higher Accuracy with Larger Datasets: Training the model on larger and more diverse datasets can improve generalization and performance across different types of videos.

Interactive User Interface: Developing an intuitive UI where users can refine queries or adjust summary length will improve user experience.

Conclusion

Query-based video summarization, extracting important details from videos, is a powerful technique with many useful applications, promising efficiency and better decision-making. Despite challenges like accuracy and privacy concerns, effective solutions can address these issues, ensuring continuous progress in the field. Our project has made significant strides by introducing personalized video summaries. Using machine learning, we customize summaries based on each user's interests, enhancing the user experience and making content more relevant. We improve the quality of our video summaries through thorough content analysis, and the inclusion of recommendation algorithms encourages further exploration, enhancing viewers' overall interaction with video content.

References

1. Huang, J.-H., Murn, L., Mrak, M., & Worring, M. (2023). Query- based Video Summarization with Pseudo Label Supervision (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2307.01945
2. Messaoud, S., Lourentzou, I., Boughoula, A., Zehni, M., Zhao, Z., Zhai, C., & Schwing, A. G. (2021). DeepQAMVS: Query-Aware Hierarchical Pointer Networks for Multi-Video Summarization (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2105.06441
3. Narasimhan, M., Rohrbach, A., & Darrell, T. (2021). CLIP-It! Language-Guided Video Summarization. arXiv. https://doi.org/10.48550/ARXIV.2107.00650

4. Huang, J.-H., & Worring, M. (2020). Query-controllable Video Summarization (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2004.03661
5. Xiao, S., Zhao, Z., Zhang, Z., Yan, X., & Yang, M. (2020). Convolutional Hierarchical Attention Network for Query-Focused Video Summarization. arXiv. https://doi.org/10.48550/ARXIV.2002.03740
6. Ji, Z., Zhang, Y., Pang, Y., Li, X., & Pan, J. (2019). Multi-video summarization with query-dependent weighted archetypal analysis. In Neurocomputing (Vol. 332, pp. 406–416). Elsevier BV. https://doi.org/10.1016/j.neucom.2018.12.038
7. Zhang, Y., Kampffmeyer, M., Zhao, X., & Tan, M. (2019). Deep Reinforcement Learning for Query-Conditioned Video Summarization. In Applied Sciences (Vol. 9, Issue 4, p. 750). MDPI AG. https://doi.org/10.3390/app9040750
8. Vasudevan, A. B., Gygli, M., Volokitin, A., & Van Gool, L. (2017). Query-adaptive Video Summarization via Quality-aware Relevance Estimation (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1705.00581
9. Ji, Z., Ma, Y., Pang, Y., & Li, X. (2017). Query-Aware Sparse Coding for Multi-Video Summarization (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1707.04021
10. Sharghi, A., Laurel, J. S., & Gong, B. (2017). Query-Focused Video Summarization: Dataset, Evaluation, and A Memory Network Based Approach (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1707.04960