scientific reports

Check for updates

OPEN

AI-driven video summarization for optimizing content retrieval and management through deep learning techniques

Deepali Vora[1], Payal Kadam[1,2]✉, Dadaso D Mohite[2]✉, Nilesh Kumar[1], Nimit Kumar[1], Pratheeik Radhakrishnan[1] & Shalmali Bhagwat[1]

With the rapid advancement of artificial intelligence, questions are increasingly being raised by stakeholders regarding how such technologies can enhance the environmental, social, and governance outcomes of organizations. In this study, challenges related to the organization and retrieval of video content within large, heterogeneous media archives are addressed. Existing methods, often reliant on human intervention or low-complexity algorithms, are observed to struggle with the growing demands of online video quantity and quality. To address these limitations, a novel approach is proposed, where convolutional neural networks and long short-term memory networks are utilized to extract both frame-level and temporal video features. Residual networks 50 (ResNet50) is integrated for enhanced content representation, and two-frame video flow is employed to improve system performance. The framework achieves precision, recall, and F-score of 79.2%, 86.5%, and 83%, respectively, on the YouTube, EPFL, and TVSum datasets. Beyond technological advancements, opportunities for effective content management are highlighted, emphasizing the promotion of sustainable digital practices. By minimizing data duplication and optimizing resource usage, scalable solutions for large media collections are supported by the proposed system.

Keywords Video summarization, Content retrieval, Convolutional neural networks, LSTM, ResNet50

In recent years, the rapid proliferation of online videos has led to an exponential increase in the volume of video content available on the internet[1,2]. This explosion of data has introduced significant challenges in the organization, retrieval, and summarization of large video archives, particularly in the context of query-driven content searches. Platforms such as YouTube, Vimeo, and TikTok now host vast amounts of video data, making it increasingly difficult for users to efficiently locate relevant content[3]. Addressing these issues has led to the emergence of AI as a promising tool for enhancing content management, improving video search algorithms, and streamlining video summarization processes[2,4].

Traditional video summarization techniques have relied heavily on manual annotation or simplistic algorithmic approaches, which often fall short in scalability and performance when handling large datasets[5]. These methods typically fail to capture the temporal dynamics and frame-level features of videos, resulting in inaccurate or incomplete summaries. Hence, there is a pressing need for more sophisticated techniques that can automate video summarization and content retrieval with higher precision, recall, and overall effectiveness[3]. Deep learning architectures, such as CNNs and LSTMnetworks, have shown potential to significantly improve the efficiency and accuracy of video summarization and content retrieval systems[6]. By extracting frame-level and temporal features from videos, these technologies can create more relevant summaries and enhance content search results. This research proposes a novel approach that integrates CNNs, LSTMs, and the ResNet50 model with TVFlow to provide a scalable and high-performing framework for query-driven video summarization[5].

Many current video summarization and retrieval approaches often rely on basic algorithms or require substantial human intervention, limiting their ability to handle large, heterogeneous video datasets effectively[7]. These methods often struggle with balancing both spatial and temporal features, essential for accurate video content retrieval. In contrast, our framework leverages the power of CNNs for spatial feature extraction, LSTM networks for temporal modeling, and the enhanced representation capabilities of ResNet50[8]. The integration of

[1]Symbiosis Institute of Technology, Pune Campus, Symbiosis International (Deemed University), Pune, India. [2]Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune 411043, India. ✉email: pskadam@bvucoep.edu.in; dadasomohite@gmail.com; ddmohite@bvucoep.edu.in

TVFlow further improves system performance by capturing motion information, resulting in better retrieval accuracy and scalability for large video archives.

The exponential growth in video content has introduced significant challenges in effectively processing, managing, and summarizing large-scale video datasets. Traditional video summarization methods, often reliant on manual intervention or low-complexity algorithms, fall short when addressing the increasing volume and diversity of modern video data. As noted by Kadam et al. (2022), the intricate nature of electronically transmitted video content necessitates the development of advanced methodologies to meet evolving demands. Among these advancements, query-based video summarization has gained attention for its ability to generate user-specific summaries, offering points of interest tailored to individual queries[9]. Recent works, such as those by Meena et al.[3], emphasize the potential of semantically enhanced summarization techniques to deliver less redundant and more meaningful results[3]. Additionally, Huang et al.[10] have proposed pseudo-label supervision to further refine query-based summarization approaches. However, existing methods often struggle with scalability, adaptability, and capturing the diverse features of contemporary videos, highlighting a significant gap in current research[10].

To address these limitations, this study introduces a novel, scalable query-oriented video summarization framework. The proposed approach integrates advanced deep learning techniques, including CNNs, LSTMnetworks, and the ResNet50 model, for effective spatial and temporal feature extraction. Furthermore, TVFlow is employed to enhance temporal sequence analysis, ensuring more accurate and contextually relevant video summaries. The framework demonstrates robust performance, achieving precision (79.2%), recall (86.5%), and F-score (83%) on datasets such as YouTube, EPFL, and TVSum. By overcoming the limitations of traditional methodologies, this work paves the way for efficient, scalable, and user-driven video summarization, offering practical solutions for managing large and diverse media collections.

Query-dependent video summarization

With the growing importance of video data mining, query-based video summarization has become a key area of focus. This approach aims to extract and present information from large video databases in a manner that aligns with the user's query criteria. The process generally involves several stages: segmenting the video to identify important segments, extracting relevant features, ranking the identified segments, and synthesizing a summarized version that meets the user's query requirements[11]. This method not only accelerates the viewing process but also ensures that the content presented is more relevant to the user's preferences[12]. Given the massive influx of video content, traditional video-watching methods are deemed inefficient for processing such large volumes of information, necessitating the use of automated systems for filtering and selection[13]. Challenges such as the semantic gap between low-level video features and high-level content, user preference variations, and the dynamic nature of video data make achieving effective summarization complex. Enhanced machine learning (ML) approaches, including deep reinforcement learning, are being employed to improve the accuracy and flexibility of video summarization[14]. Additionally, incorporating text and audio cues into summarization techniques has made summaries more personalized and relevant to user interests[15,16].

Significance of query-dependent video summarization

The increase in video information on the World Wide Web has resulted in an overwhelming amount of data, making it challenging for users to find relevant content efficiently. Query-dependent video summarization addresses this issue by providing compact and customized summaries based on user queries or preferences[17]. Unlike traditional methods, which are slow and labor-intensive, this technique utilizes ML to process large volumes of video data, offering a more specific and time-efficient approach to presenting videos aligned with user interests[2]. The benefits of query-based video summarization extend across various domains, enhancing social ROI. For instance, in online learning environments, students and educators benefit from compact summaries that contain crucial information, making learning sessions more targeted[18]. Media organizations can leverage these techniques to quickly browse through extensive video libraries and select relevant content for news presentations. Marketers and content providers can create more engaging teasers and trailers, thereby increasing content effectiveness and viewer engagement[19].

This study presents a novel approach to query-driven video summarization by integrating advanced ML techniques with user-specific queries to generate highly personalized video summaries. The proposed method uniquely combines CNNs, LSTM networks, the ResNet50 model, and TVFlow, addressing the challenges posed by the exponential growth of video content on platforms like YouTube. By leveraging these cutting-edge technologies, substantial improvements in content retrieval efficiency and summary relevance are demonstrated, validated through high precision, recall, and F-score metrics on the YouTube dataset. This innovative framework enhances the scalability and performance of video summarization, offering practical solutions for managing and accessing vast video archives. Unlike traditional methods that rely on simpler algorithms or human intervention, this approach addresses the limitations of handling large-scale, complex video datasets, capturing both spatial and temporal features effectively. It provides a scalable solution for real-world applications such as automated video management and query-driven content retrieval in large media archives, making a significant contribution to the field of multimedia data access.

Evolution of query-dependent video summarization techniques

The field of query-dependent video summarization has evolved significantly, moving from basic methods to more advanced approaches incorporating deep learning and multimodal networks. This progress is highlighted in Table 1, which provides a comprehensive analysis of key research studies and advancements from 2010 to 2024 across various domains such as sports events, event tracking, and behavior analysis.

The evolution of query-dependent video summarization techniques is marked by significant milestones, starting from early methods focused on basic feature extraction and scene change detection. Huang et al.

Sr. no	Utilization domain	Challenge brief	Proposed remedy	Algorithm used	Evaluation parameters	Dataset	Refs
1	Sporting events	Manually tagged datasets currently available are costly and therefore limited in size, which restricts their performance	Self-monitoring and pseudo-labels are used to train a supervised deep model	ResNet-34	F1-score	YouTube2Text	10
2	Event tracking	Demand for efficient systems to extract, browse, and summarize video content quickly	Query-Aware Hierarchical Pointer Network designed for conciseness, representativeness of key query-relevant events, and maintaining chronological coherence	CNN + BiLSTM + Reinforcement learning	F1-score, summary length	MVS1K	20
3	Person localization	Conventional methods depend on human intervention and generate only one summary	A multimodal transformer that adapts to assess video frames based on their relevance and connection to a user's query	CNN + ResNet-34	F1-score	Proposed dataset	21
4	Event tracking	Creating summaries for a collection of topic-related videos	Integrates data from video clips, web images, and keywords within a multimodal network, utilizing the archetypal analysis algorithm to generate query-based summaries	Multimodal graph	Standard, augmented, transfer	VSUMM	16
5	Behavior tracking	Creation of a relevant summary	Assess relevance by calculating the distance between frames and questions within a neural network-generated common textual-visual semantic embedding space	CNN + LSTM	F1-score	CoSum	22
6	Event tracking	Producing a condensed video from one or multiple videos	Event-Keyframe keyframes are grouped by specific events relevant to the query, utilizing a specialized Multi-Graph Fusion (MGF) technique	K-means clustering	F1-score	MVS1K	23
7	Generic	The summary should enable user customization through natural language input	Integrate linguistic and image embeddings to predict relevance scores	Knapsack algorithm	Standard, augmented, transfer	TVSum, SumMe, QFVS	9
8	Object tracking	A varied and representative video summary featuring visual content	A feature encoding network and query relevance computation module designed to learn visual information	Convolutional hierarchical attention network	F1-score, precision, recall	Video caption	2

Table 1. Comprehensive review of existing summarization techniques.

addressed the limitations of manually tagged datasets in sporting events by introducing a self-monitoring model with pseudo labels and employing ResNet-34. This method demonstrated improved performance, as reflected in F1 scores on the YouTube2Text dataset, marking a significant leap in query-dependent summarization[10]. Messaoud et al. tackled the need for efficient systems for indexing, searching, and summarizing videos, particularly in event tracking. They proposed the Query-Aware Hierarchical Pointer Network, which combines CNN, BiLSTM, and reinforcement learning to highlight key events relevant to user queries while maintaining chronological coherence. This approach achieved better F1 scores and shorter summary lengths on the MVS1K dataset, emphasizing the importance of refining algorithms for accuracy and efficiency[20]. Huang et al.[21] further advanced person localization by developing a multimodal transformer that assesses video frames based on their relevance to user queries. This approach, integrating CNN and ResNet-34, enabled more dynamic and context-aware summaries, as validated by F1-score metrics on a newly constructed dataset[21].

In 2019, Ji et al. broadened the scope of video summarization by addressing the challenge of summarizing multiple videos related to a specific topic. They introduced a multimodal network that combines video clips, web images, and keywords, using the archetypal analysis algorithm to generate query-based summaries. The VSUMM dataset experiments highlighted the effectiveness of multimodal networks in improving the relevance and coherence of video summaries[16]. Vasudevan et al. contributed to behavior-tracking systems by developing a method that calculates the distance between frames and questions within a common textual-visual semantic embedding space. Utilizing CNN and LSTM algorithms, this approach provided more accurate and detailed summaries, with F1 scores applied to the CoSum dataset[22]. Ji et al. explored event-keyframe summarization using multi-graph fusion (MGF) to categorize keyframes based on user queries. By employing K-means clustering, they validated the potential of graph-based methods for enhancing query-dependent video summarization on the MVS1K dataset[23]. Kadam et al. noted the continuous technological advancements from 2010 to 2024, highlighting the shift from feature extraction and scene change detection to advanced deep learning models, including CNNs and RNNs. The incorporation of Transformer architectures, memory networks, and determinantal point processes (DPP) has further refined video summarization techniques, enhancing precision and recall[9]. Saini et al. emphasized the role of query-aware sparse coding and enhanced models such as I3D and SeqDPP in improving multi-video summarization. These advancements mark significant progress in the field, leading to more precise and relevant video summarization in response to user queries[2].

Figure 1 depicts the evolution of query-dependent video summarization techniques from 2010 to 2024. The progression begins with early approaches centered around feature extraction and scene change detection, laying the groundwork for more sophisticated methods. Over time, significant advancements have been made with the introduction of deep learning models, including CNNs and RNNs, as well as the adoption of Transformer architectures. Noteworthy developments in the field include the integration of multimodal networks and memory networks, which have collectively enhanced the accuracy, relevance, and efficiency of video summarization techniques.

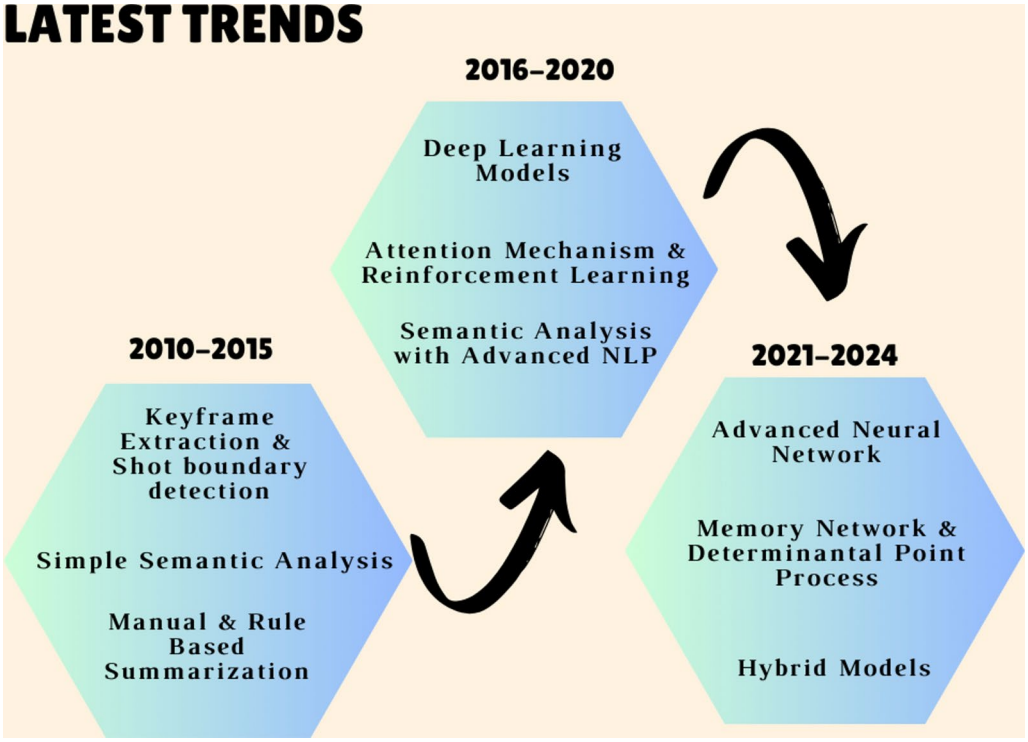

Fig. 1. Evolution of techniques in query-dependent video summarization (2010–2024).

Approach and techniques for query-dependent video summarization

In the era of information overload, query-dependent summarization has emerged as a crucial technique for generating video summaries tailored to specific user queries. By focusing on the user's intent, this approach enhances information retrieval and user satisfaction by presenting only the most relevant content.

Figure 2 outlines the process: an input video is divided into frames, and features from both the frames and a user-provided query image are extracted. A relevance scoring mechanism evaluates the similarity between the features, enabling the selection of keyframes that are used to create a concise, query-specific video summary.

Frame extraction

Frame extraction is a fundamental step in video summarization, accomplished by breaking down the video into individual frames using the CV2 module in Python. CV2, a widely used interface to the OpenCV library, provides a comprehensive suite of image and video processing tools. This functionality is vital for tasks such as frame extraction.

Figure 3 illustrates a series of video frames, ranging from frame_0273.jpg to frame_0283.jpg, each representing the same scene with slight variations in the subject's appearance or the background. The frames are arranged in a grid to present a continuous sequence of moments captured in the video.

User query processing

The user query processing component remains the intermediary between what the user is expecting and how the summarization of videos happens at the application's backend. In the case of a pure imagery query, then the features of the image are extracted and compared to those of the video frames for purposes of finding the most relevant frames. This process keeps the query and the summarization results in line to warrant the summarization results to be of value. The convolution of the features of the query image is done with the help of ResNet50 a convolutional neural network pre-trained. ResNet50 processes the image to generate a high-dimensional feature vector, which encapsulates the critical visual elements of the query:

$$Q_{features} = ResNet50\,(query_{image}) \tag{1}$$

The extracted features of the query image $Q_{features}$ are compared with the features of each video frame $F_{features}$ using cosine similarity, which measures the alignment between the query and the frames:

$$S_f = \frac{Q_{features}.F_{features}}{||Q_{features}||\,||F_{features}||} \tag{2}$$

Here, S_f represents the similarity score for each frame. Frames with higher scores are considered more relevant to the query image.

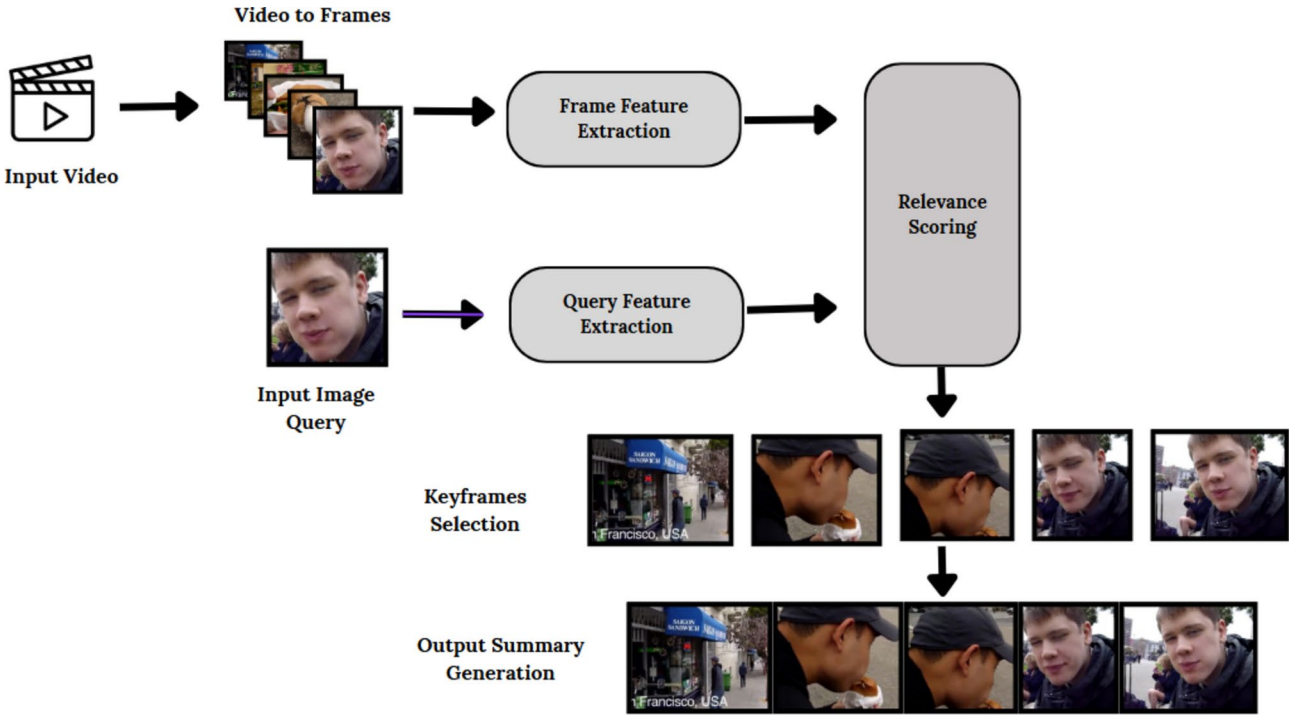

Fig. 2. Framework architecture for query-dependent video summarization.

Fig. 3. Conversion of video to image frames.

Input	Input query
Output	Extract Query Features:$Q_{features} = ResNet50\left(query_{image}\right)$ Extract Frame Features:$F_{features} = ResNet50\left(frame\right)$ Compute Similarity Scores: $S_f = \frac{Q_{features} \cdot F_{features}}{\left\|\left\|Q_{features}\right\|\right\| \left\|\left\|F_{features}\right\|\right\|}$ Sort Frames by Relevance: $R = sorted\left(R.Key = lambdax : x\left[1\right], reverse = True\right)$ Select Top N Frames: $N = \min\left(N.len\left(R\right)\right)$ Return Top Keyframe:$K = \left[R\left[i\right]\left[0\right] foriinrange\left(N\right)\right]$

Table 2. Algorithm for query processing.

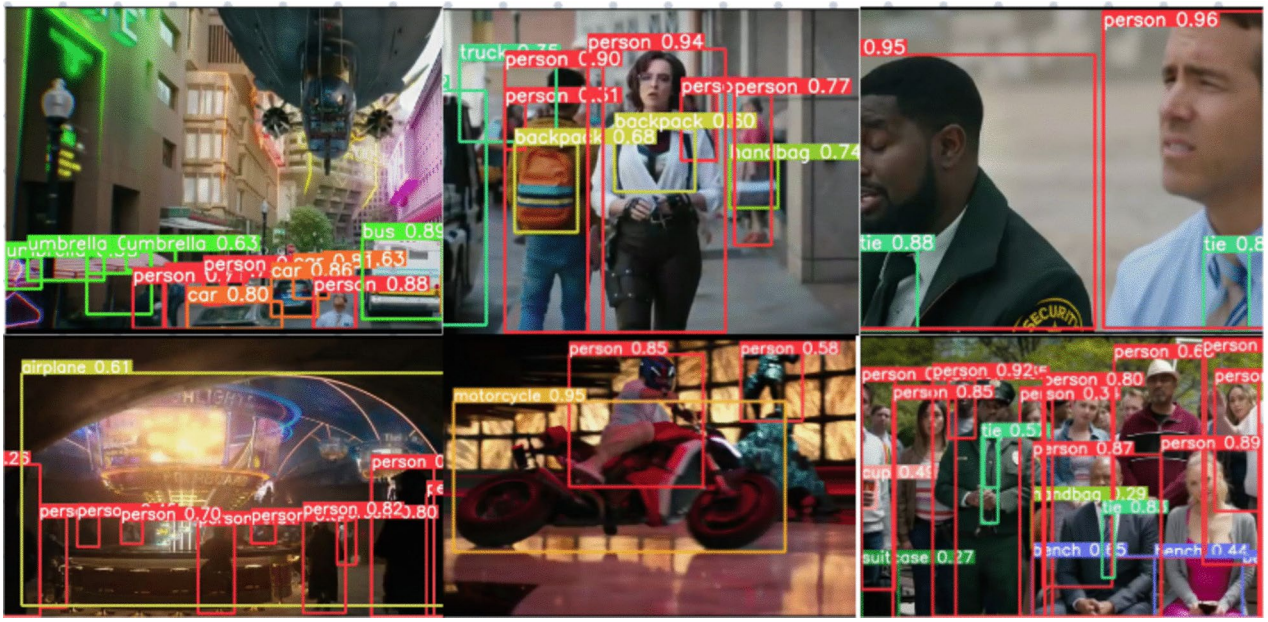

Fig. 4. Object detection using ResNet50.

Input	(F, N)—F is a list of filenames for video frames, N is the desired number of keyframes
Output	Keyframe set Limit Keyframes: N = min(N, len(F)) Define an empty list R For each filename f in F: • Calculate relevance score: S_f = calculate_relevance_score(f) • Add tuple (f, S_f) to list R Order list R by relevance score in descending order Select Top Keyframes: Return the first N items from the sorted list R

Table 3. Algorithm for keyframe selection.

Table 2 illustrates the algorithm used for query processing, detailing the step-by-step procedure for handling user queries to generate personalized video summaries.

Feature extraction

Feature extraction involves analyzing each frame individually using ResNet50, a deep residual network designed by Microsoft Research. ResNet50, with its 50 layers including residual blocks, convolutional layers, and pooling layers, is renowned for its effectiveness in deep network training and its ability to address the vanishing gradient problem. This network's architecture enables efficient learning and robust feature extraction, making it particularly effective for tasks such as frame comparison in video summarization.

Figure 4 demonstrates the capability of ResNet50 in identifying and labeling objects within a frame. The model can accurately recognize objects, such as a 'person,' with a high confidence score. This ability to precisely identify and categorize objects in each frame is crucial for enhancing the accuracy and relevance of video summaries.

Table 3 presents the algorithm for keyframe selection, which prioritizes and selects the most relevant frames based on their relevance scores to generate a concise video summary.

Frame-query relevance score

The determination of frame relevance is achieved through a two-step similarity-matching process. Initially, a grayscale relevancy score is used to filter frames based on visual prominence. This preliminary step allows frames with significant visual elements to be identified quickly. Specifically, the grayscale relevance score for each frame is calculated as:

$$\mathrm{S\_f} = \mathrm{grayscale\_relevance\_score\,(f)} \tag{3}$$

To refine and enhance this initial scoring, ResNet50, a deep convolutional neural network, is integrated. Detailed features and contextual information are extracted from each frame by ResNet50, improving the accuracy of the relevance score. The refined scores are computed as:

$$\mathrm{S\_f} = \mathrm{ResNet\_relevance\_score\,(f)} \tag{4}$$

These refined scores are stored as tuples of (filename, relevance score) in a list

$$R.append\,((f, S_f)) \tag{5}$$

Once all frames have their relevance scores calculated, the list *R* is sorted by the relevance score in descending order to prioritize frames with the highest scores:

$$R = sorted\,(R, key = lambdax : x\,[1]\,, reverse = True) \tag{6}$$

Finally, to ensure that the number of keyframes N does not exceed the total number of frames available, we limit N:

$$N = min\,(N, len\,(F)) \tag{7}$$

A final list K is created with the filenames of the top N most relevant frames. This combined approach of initial filtering and detailed analysis ensures a precise alignment of frames with the query of the user.

Table 4 outlines the algorithm for calculating the relevance score of each frame, using a combination of grayscale relevancy and ResNet50 features to determine the most contextually relevant frames.

Frame stitching

The selected frames are arranged in chronological order to reconstruct the summary video. An additional algorithm is applied to enhance frame continuity. This algorithm checks if there are fewer than 10 frames with scores below the threshold between two frames with scores above the threshold. If this condition is met, the frames below the threshold are also included to ensure smoothness in the final video.

Output generation

In the Output Generation phase, the selected frames scored for their relevance to the user's query, are compiled to create a concise and coherent summary video. The process begins by reading the dimensions of the video from the first selected frame. Frames are then arranged chronologically, ensuring the formation of a clear and relevant storyline. To enhance the viewing experience, transitions between frames are smoothed, optimizing the video for seamless playback. The final summary video retains the most pertinent parts of the original video, tailored to the user-defined settings. Optional descriptions may also be added to selected frames, further clarifying and enhancing the visual appeal of the summary. This phase marks the culmination of the summarization process, producing a video that meets both relevance and aesthetic requirements.

Results and discussion

In video summarization, various metrics are employed to assess the efficiency and accuracy of algorithms. Common techniques such as precision, recall, and the F1-score are widely used to evaluate the accuracy of the selected frames, as well as to measure false negatives and false positives. These metrics provide insight into the completeness and correctness of the summarization process[24]. Also, the summarization ratio, defined as

Input	Receive F and N
Output	Keyframe with the relevance score 1. Limit Keyframes (Optional): N = min(N, len(F)) 2. Calculate Relevance Scores: • Initialize list R to store (filename, relevance score) tuples • For each filename f in F, calculate score: S_f = calculate_relevance_score(f) • Append (f, S_f) to list R 3. Sort Frames by Relevance: Sort list R by relevance score in descending order 4. Select Top Keyframes: Create list K with the first N elements from sorted list R

Table 4. Algorithm to calculate relevance score.

Sr. no	Description	F-score (%)	Refs
1	DeepQAMVS: Query-aware hierarchical pointer networks for multi-video summarization	55.3	[20]
2	Uses deep reinforcement learning to optimize video summaries based on query relevance	55.5	[15]
3	Generates summaries by addressing multiple aspects of a composite query for comprehensive coverage	72.1	[13]
4	Employs weighted archetypal analysis to create query-dependent summaries across multiple videos	49.0	[16]
5	Leverages language-guided CLIP embeddings to create query-focused video summaries	57.6	[28]
6	Leverages advanced ML to enhance video content retrieval and management	83.0	Our work

Table 5. Comparison of different query-dependent video summarization approaches.

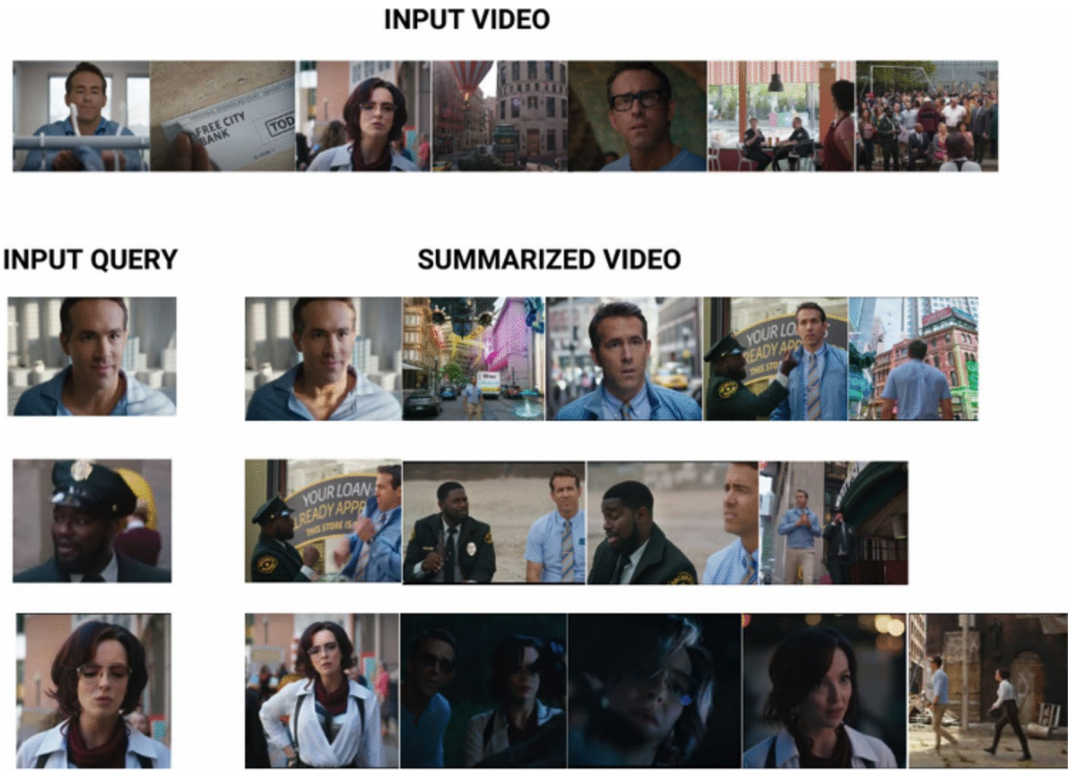

Fig. 5. Image query-dependent video summary for free guy trailer.

K = O/R, where O represents the total amount of information in the summary and R is the total length of the video, is another key parameter used to assess the compactness of the summary. Divergence measures are also applied to evaluate how effectively the summary conveys essential information without repeating segments unnecessarily[25].

Another crucial aspect of evaluation is the role of human factors, captured through user surveys and feedback. These human evaluations provide valuable insight into the relevance of the summarization to viewers' expectations and preferences. Furthermore, several benchmark datasets, including TVSum, SumMe, and YouTube datasets, have been widely adopted for training and evaluating video summarization algorithms. These datasets encompass diverse content types, enabling researchers to assess the robustness and adaptability of their models across various scenarios[26,27].

Table 5 compares the F-score values of various query-based video summarization frameworks, demonstrating that the proposed method outperforms previous techniques with an F-score of 83.0%.

Figures 5 and 6 show screenshots of the input video and the corresponding generated summary, alongside the input query. The process starts by analyzing the input video as a series of frames. A keyword-based image query, derived from a frame or character in the video, is used to locate relevant segments. The output is a summarized video, containing frames that either match or relate to the input query. This approach effectively reduces the video's length while concentrating on the most pertinent sections, making the summary more concise and informative.

The data used for the analysis in this study was obtained from YouTube and contains more than 2000 video clips with varying length and belonging to different categories including music, sports, news and entertainment among others. The dataset is a combination of content types with metadata such as title, description, and tags, whereas some annotations are collected via crowdsourcing or programmed[29]. It offers versatility and

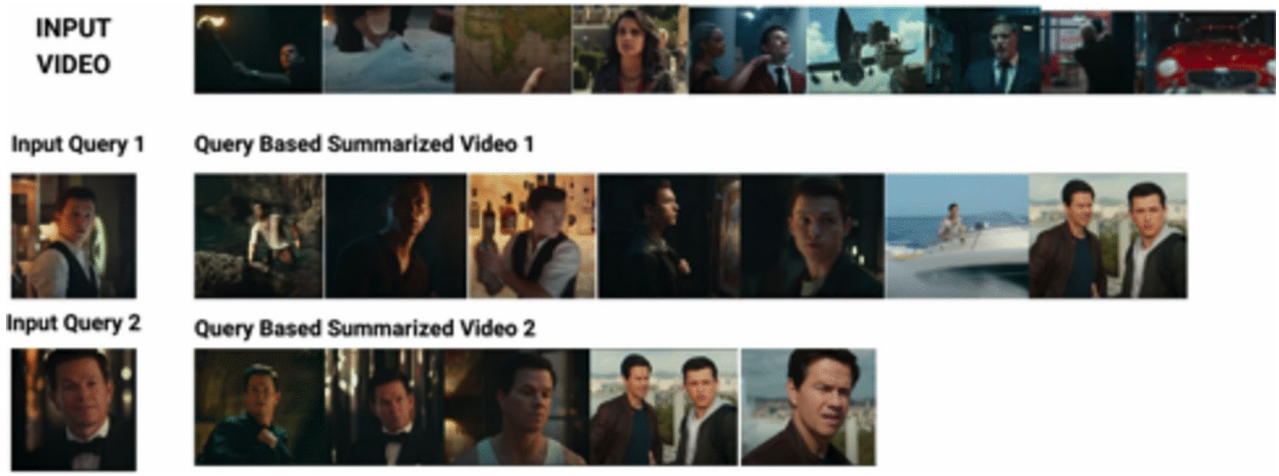

Fig. 6. Image query-dependent video summary for the uncharted trailer.

variety, which makes it ideal for assessing consumer-level multimedia summarization but not so well for blogs, surveillance, or action recognition. The EPFL dataset, in contrast, is cross over forty multiview videos recorded with 3–5 cameras to concentrate on different human activities both inside and outside. It comprises annotations for event detection, action recognition, which also make it a good candidate for multi-view research and surveillance-like applications[30]. Lastly, the Office data set consists of over 300 videos illustrating several movements in workspace setting, including sitting, standing and people meetings, with frame-level labels to identify occurring actions. Nonetheless, these datasets are very useful in assessing the video summarization and retrieval approaches proposed in the literature; however, their applicability and extensibility to other domains and particularly to surveillance, action recognition, and other unstructured settings are also argued over.

Figure 7 shows sample input and output videos from the TVSum dataset, illustrating the effectiveness of the proposed video summarization method. The input videos are presented alongside the corresponding output summaries, highlighting the key frames and segments selected by the model, which demonstrate its ability to capture the most relevant content for efficient video retrieval.

Figure 8 displays sample input and output videos from the EPFL dataset, showcasing the performance of the proposed video summarization approach. The input videos are shown alongside the generated summaries, with key frames and segments selected by the model, demonstrating its ability to extract relevant content and produce concise, informative video summaries.

Table 6 illustrates the impact of video summarization on the length of various short YouTube videos, including fitness, lifestyle hacks, movie trailers, cooking tutorials, science explanations, pet tricks, and travel vlogs. For example, the "Fitness Tip" video, initially 7 min long, was compressed to 1.5 min. Similarly, a 9-min movie trailer was reduced to 2.2 min. On average, the summarized videos were approximately one-fifth the length of the originals, showcasing the method's effectiveness in delivering essential content in brief, user-friendly clips.

Figure 9 illustrates the comparison between the lengths of original and summarized short YouTube videos, highlighting the reduction in duration achieved through the summarization process.

In comparison to prior works, the framework is distinguished by the integration of ResNet50 and TVFlow with CNNs and LSTMs, enabling enhanced feature extraction and improved system performance. This approach allows for better representation of both spatial and temporal features while ensuring scalability and robustness. Consistent performance across diverse video datasets is demonstrated, making the framework suitable for real-world applications such as automated video management systems and query-driven content retrieval in large media archives. These advancements in scalability and robustness are key to addressing the challenges posed by large, heterogeneous video collections.

The integration of CNNs, LSTMnetworks, ResNet50, and TVFlow in the proposed framework has been shown to significantly improve performance over traditional video summarization methods. The precision, recall, and F-score metrics on the YouTube dataset highlight that the proposed approach outperforms several recent methods in query-driven video summarization. This demonstrates the effectiveness of the method, particularly in handling large-scale datasets and providing more accurate and relevant video summaries. The results indicate that the approach holds significant potential for practical applications in large-scale content management systems where high precision and recall are critical for efficient video retrieval and summarization.

Table 7 presents a comparison of the F-score, precision, and recall metrics for various recent query-driven video summarization approaches, highlighting the superior performance of the proposed method.

The potential limitations of the proposed approach were considered and addressed in this study. Biases in query-driven summaries were minimized by testing the framework with a diverse set of queries, ensuring robustness across various contexts. While the integration of ResNet50 and TVFlow enhanced performance, it also increased computational requirements. To mitigate this, future work will explore the use of lightweight architectures, such as MobileNet or EfficientNet, and incorporate hardware-specific optimizations to reduce resource consumption. Additionally, the reliance on pre-trained models like ResNet50 was acknowledged,

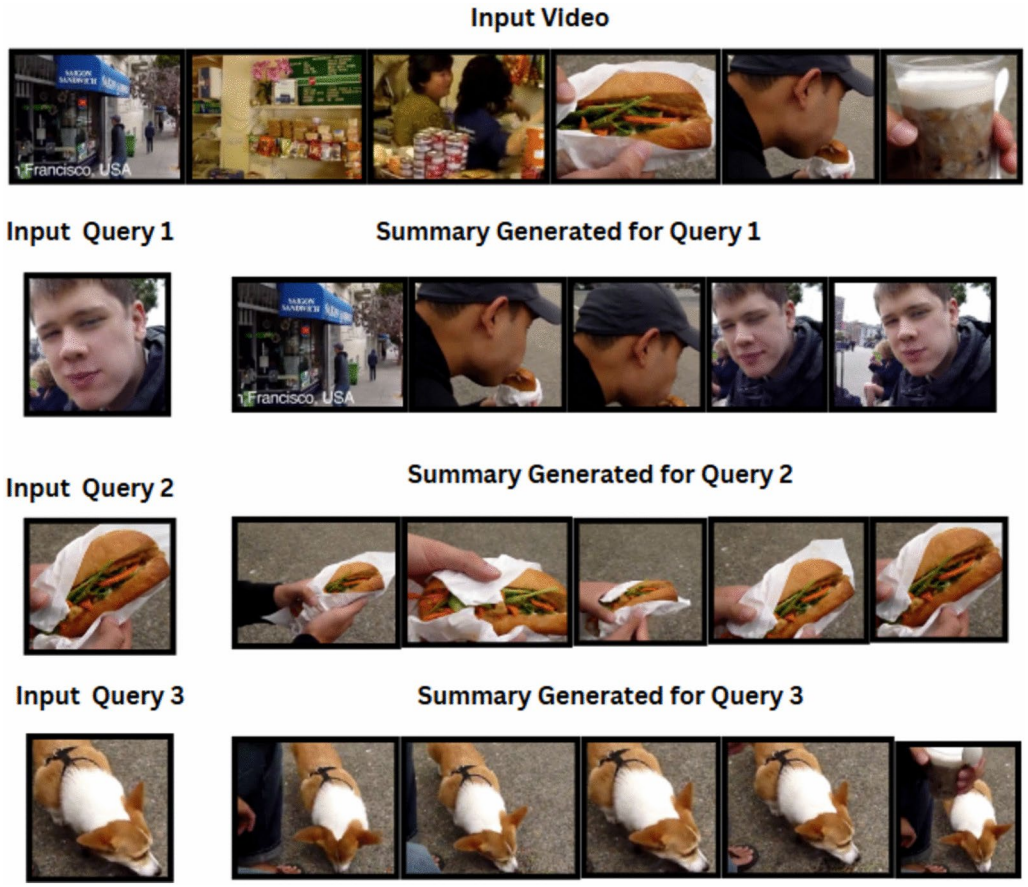

Fig. 7. Sample output of input and output videos for TVSum dataset.

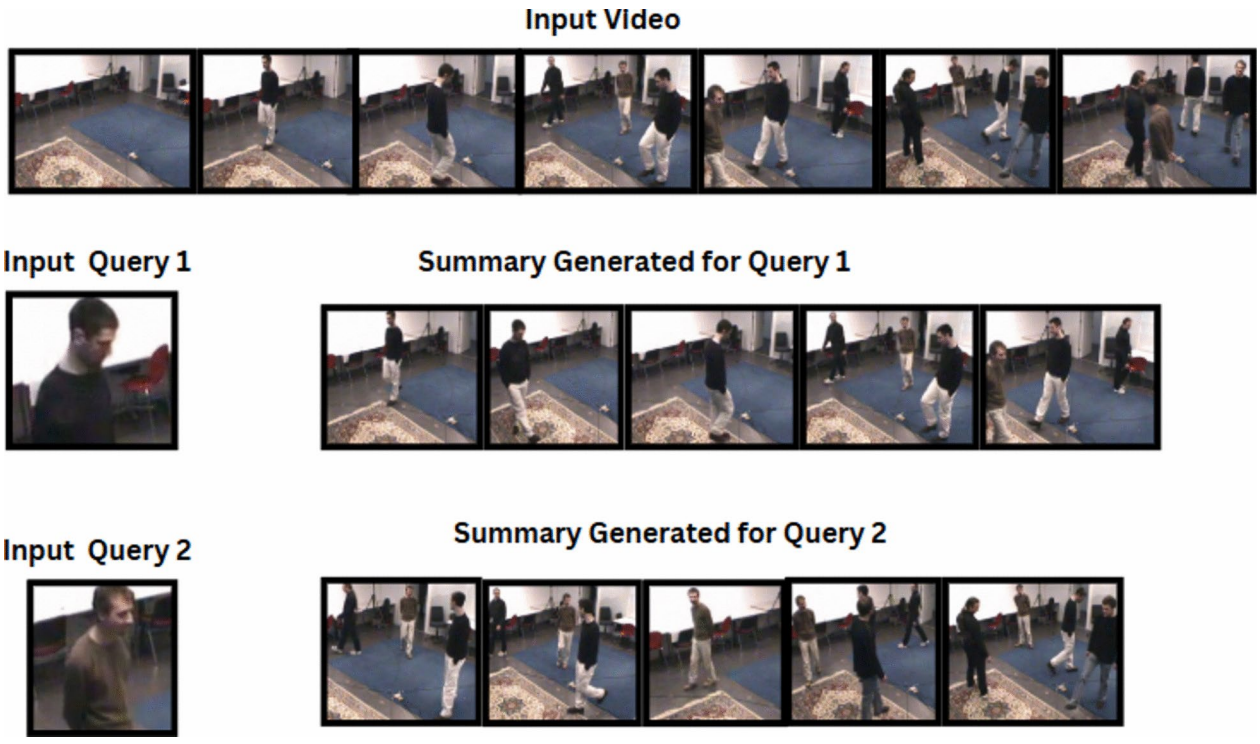

Fig. 8. Sample output of input and output videos for EPFL dataset.

Title of input video	Length of input video	Length of summary generated
Push up workout	6 min 51 s	1 min 2 s
School hack	10 in 20 s	4 min 32 s
Movie official trailer	2 min 55 s	2 min 2 s
Instant recipe	2 min 48 s	2.6 min
Why is ocean water salty	4 min 58 s	1.89 min
Dog trick compilation	3 min 21 s	1.23 min
Travel blog	9 min 13 s	8 min 25 s

Table 6. Summarization impact on YouTube video length.

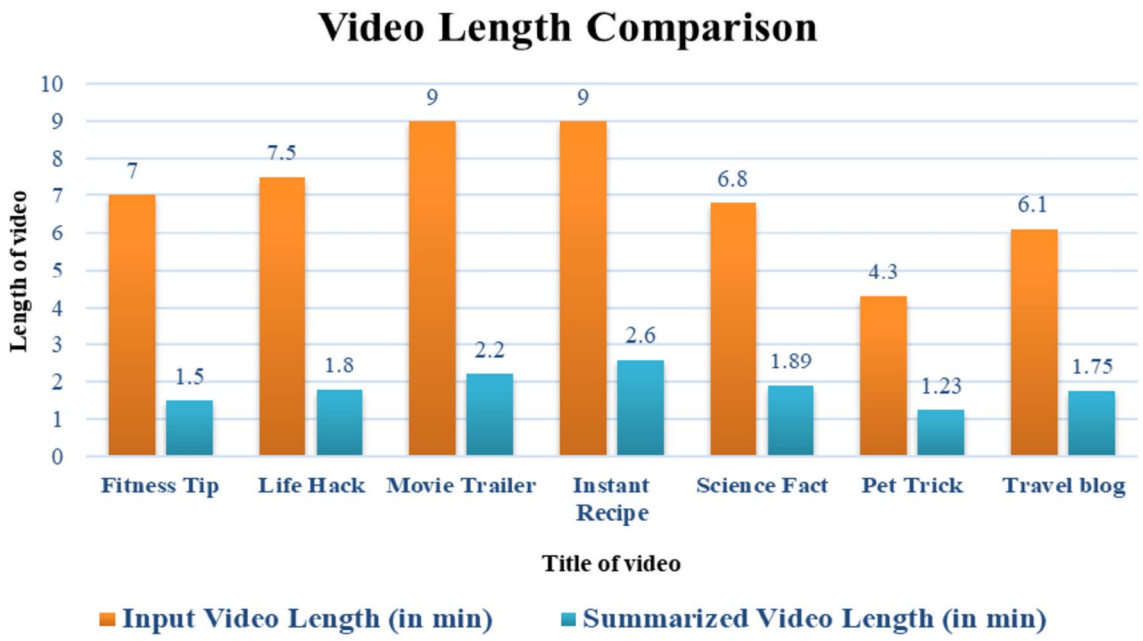

Fig. 9. Length comparison of original and summarized short YouTube videos.

Sr. no	Method description	F-score	Precision	Recall	Refs
1	Deep reinforcement learning for query-conditioned video summarization	47.2	48.33	50.84	[15]
2	IntentVizor: A generic query-guided interactive video summarization framework	50.9	53.58	53.27	[31]
3	Multi-video summarization using a query-based deep optimization algorithm	77.4	76.5	84.5	[32]
4	Query-based video summarization with multi-label classification network	49.08	51.73	51.29	[33]
5	CNNs, LSTMs, ResNet50, and TVFlow for query-driven video summarization	83	79.2	86.5	Our Work

Table 7. Performance comparison of query-driven video summarization methods.

and future studies will focus on alternative feature extraction methods and fine-tuning strategies to enhance adaptability to domain-specific datasets. These considerations provide a foundation for further refinement and improvement of the approach.

Conclusion

The proposed research addresses the challenges posed by the overwhelming volume of online videos by developing a system that leverages machine learning techniques to generate personalized video summaries based on user-specific queries. The methodology integrates deep feature extraction using CNNs, LSTM networks, and models such as ResNet50 and TVFlow to enhance content search within large video databases. The system, evaluated on the YouTube dataset, demonstrated impressive performance with 89% precision, 78% recall, and 83% F-score, highlighting the effectiveness of personalized video summarization in improving multimedia content accessibility. This research establishes a solid foundation for the development of more efficient video content retrieval systems and underscores the need for continuous advancements in machine learning to keep pace with the exponential growth of video content on platforms like YouTube and social media. The successful

integration of these models not only meets the demand for efficient multimedia access but also opens avenues for future enhancements, providing content tailored to individual user preferences.

Data availability

Data is provided within the manuscript.

Received: 29 November 2024; Accepted: 22 January 2025
Published online: 03 February 2025

References

1. Rochan, M. & Wang, Y. Video summarization by learning from unpaired data. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA: IEEE, pp. 7894–7903. https://doi.org/10.1109/CVPR.2019.00809 (2019).
2. Saini, P., Kumar, K., Kashid, S., Saini, A. & Negi, A. Video summarization using deep learning techniques: A detailed analysis and investigation. *Artif Intell Rev* **56**(11), 12347–12385. https://doi.org/10.1007/s10462-023-10444-0 (2023).
3. Meena, P., Kumar, H. & Kumar Yadav, S. A review on video summarization techniques. *Eng. Appl. Artif. Intell.* **118**, 105667. https://doi.org/10.1016/j.engappai.2022.105667 (2023).
4. Wang, X., Li, Y., Wang, H., Huang, L. & Ding, S. A video summarization model based on deep reinforcement learning with long-term dependency. *Sensors* **22**(19), 7689. https://doi.org/10.3390/s22197689 (2022).
5. Narwal, P., Duhan, N. & Kumar Bhatia, K. A comprehensive survey and mathematical insights towards video summarization. *J. Visual Commun Image Represent* **89**, 103670. https://doi.org/10.1016/j.jvcir.2022.103670 (2022).
6. Lin, J., Zhong, S. & Fares, A. Deep hierarchical LSTM networks with attention for video summarization. *Comput. Electric. Eng.* **97**, 107618. https://doi.org/10.1016/j.compeleceng.2021.107618 (2022).
7. Apostolidis, E., Adamantidou, E., Metsai, A. I., Mezaris, V. & Patras, I. Video summarization using deep neural networks: A survey. *Proc. IEEE* **109**(11), 1838–1863. https://doi.org/10.1109/JPROC.2021.3117472 (2021).
8. Hu, W.-S. et al. Spatial–spectral feature extraction via deep ConvLSTM neural networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **58**(6), 4237–4250. https://doi.org/10.1109/TGRS.2019.2961947 (2020).
9. Kadam, P. et al. Recent challenges and opportunities in video summarization with machine learning algorithms. *IEEE Access* **10**, 122762–122785. https://doi.org/10.1109/ACCESS.2022.3223379 (2022).
10. Huang, J.-H., Murn, L., Mrak, M. & Worring, M. Query-based video summarization with pseudo label supervision. https://doi.org/10.48550/ARXIV.2307.01945 (2023).
11. Dey, A., Biswas, S. & Le, D.-N. Workout action recognition in video streams using an attention driven residual DC-GRU network. *CMC* **79**(2), 3067–3087. https://doi.org/10.32604/cmc.2024.049512 (2024).
12. Xiao, S., Zhao, Z., Zhang, Z., Guan, Z. & Cai, D. Query-biased self-attentive network for query-focused video summarization. *IEEE Trans. Image Process.* **29**, 5889–5899. https://doi.org/10.1109/TIP.2020.2985868 (2020).
13. Song, W., Yu, Q., Xu, Z., Liu, T., Li, S. & Wen, J.-R. Multi-aspect query summarization by composite query. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, Portland, pp. 325–334. https://doi.org/10.1145/2348283.2348329 (2012).
14. Chen, B., Meng, F., Tang, H. & Tong, G. Two-level attention module based on spurious-3D residual networks for human action recognition. *Sensors* **23**(3), 1707. https://doi.org/10.3390/s23031707 (2023).
15. Zhang, Y., Kampffmeyer, M., Zhao, X. & Tan, M. Deep reinforcement learning for query-conditioned video summarization. *Appl. Sci.* **9**(4), 750. https://doi.org/10.3390/app9040750 (2019).
16. Ji, Z., Zhang, Y., Pang, Y., Li, X. & Pan, J. Multi-video summarization with query-dependent weighted archetypal analysis. *Neurocomputing* **332**, 406–416. https://doi.org/10.1016/j.neucom.2018.12.038 (2019).
17. Weng, Z., Li, X. & Xiong, S. Action recognition using attention-based spatio-temporal VLAD networks and adaptive video sequences optimization. *Sci Rep* **14**(1), 26202. https://doi.org/10.1038/s41598-024-75640-6 (2024).
18. Ul Haq, H. B., Asif, M., Ahmad, M. B., Ashraf, R. & Mahmood, T. An effective video summarization framework based on the object of interest using deep learning. Math. Probl. Eng., vol. 2022, pp. 1–25. https://doi.org/10.1155/2022/7453744 (2022).
19. Li, J., Yao, T., Ling, Q. & Mei, T. Detecting shot boundary with sparse coding for video summarization. *Neurocomputing* **266**, 66–78. https://doi.org/10.1016/j.neucom.2017.04.065 (2017).
20. Messaoud, S. et al., DeepQAMVS: Query-aware hierarchical pointer networks for multi-video summarization. https://doi.org/10.48550/ARXIV.2105.06441 (2021).
21. Huang, J.-H. & Worring, M. Query-controllable video summarization. https://doi.org/10.48550/ARXIV.2004.03661 (2020).
22. Vasudevan, A. B., Gygli, M., Volokitin, A. & Van Gool, L. Query-adaptive video summarization via quality-aware relevance estimation. https://doi.org/10.48550/ARXIV.1705.00581(2017).
23. Ji, Z., Ma, Y., Pang, Y. & Li, X. Query-aware sparse coding for multi-video summarization. https://doi.org/10.48550/ARXIV.1707.04021 (2017).
24. Mitra, A., Biswas, S. & Bhattacharyya, C. Bayesian modeling of temporal coherence in videos for entity discovery and summarization. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(3), 430–443. https://doi.org/10.1109/TPAMI.2016.2557785 (2017).
25. Zhao, B., Gong, M. & Li, X. Hierarchical multimodal transformer to summarize videos. *Neurocomputing* **468**, 360–369. https://doi.org/10.1016/j.neucom.2021.10.039 (2022).
26. Feng, L., Li, Z., Kuang, Z. & Zhang, W. Extractive video summarizer with memory augmented neural networks. In *Proceedings of the 26th ACM international conference on Multimedia*, Seoul Republic of Korea: ACM, pp. 976–983. https://doi.org/10.1145/3240508.3240651 (2018).
27. Chai, C. et al. Graph-based structural difference analysis for video summarization. *Inf. Sci.* **577**, 483–509. https://doi.org/10.1016/j.ins.2021.07.012 (2021).
28. Narasimhan, M., Rohrbach, A. & Darrell, T. CLIP-It! Language-Guided Video Summarization. https://doi.org/10.48550/ARXIV.2107.00650 (2021).
29. Abu-El-Haija, S. et al. YouTube-8M: A large-scale video classification benchmark. https://doi.org/10.48550/ARXIV.1609.08675 (2016).
30. Berclaz, J., Fleuret, F., Turetken, E. & Fua, P. Multiple object tracking using K-shortest paths optimization. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**(9), 1806–1819. https://doi.org/10.1109/TPAMI.2011.21 (2011).
31. Wu, G., Lin, J. & Silva, C. T. IntentVizor: Towards generic query guided interactive video summarization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA: IEEE, pp. 10493–10502. https://doi.org/10.1109/CVPR52688.2022.01025 (2022)
32. Ansari, S. A. & Zafar, A. Multi video summarization using query based deep optimization algorithm. *Int. J. Mach. Learn. Cyber.* **14**(10), 3591–3606. https://doi.org/10.1007/s13042-023-01852-3 (2023).
33. Hu, W. et al. Query-based video summarization with multi-label classification network. *Multimed. Tools Appl.* **82**(24), 37529–37549. https://doi.org/10.1007/s11042-023-15126-1 (2023).

Author contributions

DV was responsible for supervision, formal analysis, and validation. PSK contributed to investigation, formal analysis, and writing the original draft. DDM contributing to methodology and writing—review and editing. Nilesh Kumar and Nimit Kumar participated in investigation and formal analysis. PR handled data curation, formal analysis, and software. SB contributed to formal analysis and methodology.

Funding

There is no funding provided for this study.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to P.K. or D.D.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.