A Systematic Study on Video Summarization: Approaches, Challenges, and Future Directions

Kajal Kansal
National University of Singapore
Singapore
kajal.kansal@nus.edu.sg

Nikita Kansal
MRSPTU, GZSCCET
India
kansalnikita0@gmail.com

Sreevaatsav Bavana
Mahindra University
India
bavanasreevaatsav1@gmail.com

Bodla Krishna Vamshi
Mahindra University
India
bodlavamshi09@gmail.com

Nidhi Goyal
Mahindra University
India
nidhi.goyal@mahindrauniversity.edu.in

ABSTRACT

With the exponential growth of user-generated videos, video summarization has become a prominent research field to quickly understand the essence of video content. The goal is to automate the task of acquiring key segments from the video while retaining the contextual semantics of the video and combining them to generate a summary. The major challenge is to identify important frames or segments corresponding to human perception, which varies from one genre to another. To this end, the survey paper furnishes a thorough panorama encompassing diverse categories of video summarization. In this research work, we investigate, compare, and offer valuable insights into the progress and effectiveness of video summarization techniques. We discuss an end-to-end general pipeline to understand the complexity of the video summarization task. Further, we also discuss several benchmark datasets used to evaluate the performance of video summarization algorithms. Furthermore, this study also explores various challenges specific to video summarization, and potential future directions for further research, and encourages researchers to explore new avenues in video summarization.

CCS CONCEPTS

• **Computing methodologies → Video summarization**;

KEYWORDS

Video Summarization, Deep Learning, Survey, Challenges

ACM Reference Format:
Kajal Kansal, Nikita Kansal, Sreevaatsav Bavana, Bodla Krishna Vamshi, and Nidhi Goyal. 2023. A Systematic Study on Video Summarization: Approaches, Challenges, and Future Directions. In *Proceedings of the 2nd Workshop on User-centric Narrative Summarization of Long Videos (NarSUM '23), October 29, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3607540.3617139

NarSUM '23, October 29, 2023, Ottawa, ON, Canada
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0277-8/23/10.
https://doi.org/10.1145/3607540.3617139

1 INTRODUCTION

The explosion of video content on the internet and the increasing use of surveillance cameras have led to an overwhelming amount of video data that is challenging for users to browse and comprehend effectively. Video summarization techniques [12, 14, 15, 22, 27, 36, 52] have emerged as a solution to this problem by automatically selecting keyframes, shots, or segments that best represent the original video's content and provide a concise overview of the entire video. According to recent statistics by Wyzowl's [1], 75% of viewers watch short-form video content on their mobile devices, and 83% of marketers suggest videos should be under 60 seconds. Video summarization is the process of condensing a video's content into a shorter, more informative representation while maintaining its essential elements and chronological order. The key goals of video summarization are to make video browsing and retrieval more time-efficient, enable effective video analysis, and improve the user experience by offering a concise and meaningful overview of the video's content. Video summarization is a dynamic field with varying objectives depending on the specific application scenarios. Each domain demands a tailored approach to meet the unique requirements of viewers and users. For instance, sports enthusiasts seek video summaries [49] that highlight critical moments impacting game outcomes, while surveillance applications [41] prioritize scenes with unusual or noteworthy events. With the emergence of new video formats like video game live streaming [32], there is a growing need for summarization methods that cater to the preferences of audiences seeking entertainment, education, and information [25]. Additionally, the news, marketing, and educational domains each have their own distinct goals for video summaries.

Traditional video summarization techniques [4, 55] often rely on hand-crafted features, such as color histograms [19], motion vectors [26], and scene transitions [30] combined with clustering or ranking algorithms to identify representative frames or shots. These methods have shown promising results but often struggle with complex and dynamic scenes, as well as handling large-scale video datasets efficiently.

With the rise of deep learning, video summarization has witnessed significant advancements. Deep neural networks [3, 31] have demonstrated remarkable capabilities in learning rich representations directly from raw video frames or feature maps extracted from pre-trained convolutional neural networks (CNNs). Several deep learning-based architectures, including Recurrent Neural Networks

(RNNs), Long Short-Term Memory (LSTM) networks, and attention mechanisms, have been integrated to capture temporal dependencies and highlight salient video segments [54].
In recent years, video summarization has expanded beyond a single modality through methods [44, 56] that can combine different types of modalities (text, audio, and visual) instead of relying on just one modality. These methods leverage not only visual cues but also associated audio and textual data, such as video captions or transcripts, to generate more informative and contextually rich summaries. Moreover, with the increasing popularity of user-generated content on social media platforms [11], there is a growing demand for personalized video summarization. Personalized video summarization [33] aims to create summaries tailored to individual preferences and interests. This area of research involves incorporating user feedback and preferences into the summarization process, making the summaries more relevant and engaging for each user. In addition to personalized video summarization, there has been a growing interest in real-time video summarization. Traditional summarization methods are often computationally intensive and may not be suitable for real-time applications. Real-time video summarization [2, 57] techniques focus on providing efficient and timely summaries without compromising on the quality of the generated summaries. These methods are particularly valuable for applications such as live event coverage, video streaming, and surveillance systems. Another emerging trend in video summarization is the exploration of unsupervised and self-supervised learning approaches. Unsupervised methods aim to learn video representations without relying on annotated data, which can be expensive and time-consuming to obtain. Self-supervised learning leverages the inherent structure and content within the video to guide the learning process, often using pretext tasks such as predicting future frames or reconstructing video segments.

Despite the significant progress, video summarization still faces some challenges and open research questions. One of the key challenges is the trade-off between content preservation and length constraints. Striking the right balance to provide informative yet concise summaries remains an active area of research. Another challenge is the lack of large-scale, diverse, and well-annotated datasets for training and evaluating video summarization algorithms. Creating comprehensive benchmarks that encompass a wide range of video content and user preferences is essential for advancing the field. Additionally, evaluating the quality of video summaries is a non-trivial task, as there is no universally agreed-upon metric that can capture the semantic quality, coherence, and diversity of video summaries.

In this survey, we aim to provide a comprehensive overview of video summarization techniques. We discuss the key concepts, methodologies, datasets, and evaluation metrics used in video summarization research. We will also cover emerging trends, challenges, and future directions in video summarization research. We aim to provide a holistic view of the state-of-the-art in this rapidly evolving field and inspire new ideas and approaches to tackle the remaining challenges. Through this comprehensive survey, we hope to facilitate collaboration and knowledge exchange among researchers and practitioners working in video summarization and related areas. Furthermore, we highlight the strengths and limitations of different methods and identify potential research directions for future advancements in this domain.

The remainder of this survey is organized as follows: Section 2 presents an in-depth review of video summarization techniques. Section 3 discusses a generic end-to-end pipeline of video summarization. Section 4 focuses on various benchmark datasets commonly used in video summarization research. Section 4 discusses various applications of video summarization, and Section 5 and 6 provide an analysis of challenges and future research directions. Finally, Section 7 concludes the survey with a summary of the key findings and potential areas for further exploration in video summarization.

2 RELATED WORK

Video summarization has been tackled from various perspectives. Figure 1 shows recent works on video summarization, ordered into several categories such as the nature of video content, applications, and approaches. This comprehensive segregation is identified by analyzing the use cases, nature, and implementation details of the research. In this section, we review the most representative works among all common approaches.

2.1 Single vs Multimodal Video Summarization

Researchers have explored two main approaches to video summarization: single-modal and multimodal video summarization. These approaches differ in how they process and combine information from different sources to generate video summaries. Single-modal video summarization (SVS) focuses on using information from a single source, typically visual content (frames or shots) or audio content (speech or sounds). One such method [37] can be observed from the study that extracts visual features from each frame of the input video and maps them to a visual embedding space, which is learned from image captioning datasets. Jin *et al.* [16] focus on developing an efficient attention mechanism for low-rank high-order cross modalities in a video for the generation of video summaries. This approach is more straightforward and has been extensively explored in the past. In recent years, the focus has shifted towards multimodal video summarization (MVS). MVS [28, 29] involves integrating information from multiple modalities, such as visual content, audio, and text (e.g., subtitles or transcripts). By combining different sources of information, multimodal summarization aims to generate more informative and comprehensive video summaries. MVS is also called query-based video summarization, and its processing method is generally different from that of SVS since it has to handle diverse query-based videos and can take advantage of the query information. Furthermore, there is a direction towards studying multi-view video summarization, which is mainly used in surveillance scenarios to compact the videos captured from different cameras. Often, these approaches utilize multiple sources of input, by leveraging metadata or directly incorporating user preferences through queries, customized summaries, and other interactive means [45, 51]. Some techniques focus on the inclusion of external generic knowledge or take advantage of the metadata of a video to generate many subjective and logically relevant summaries, which would be much more human-sensible [14, 52, 59].

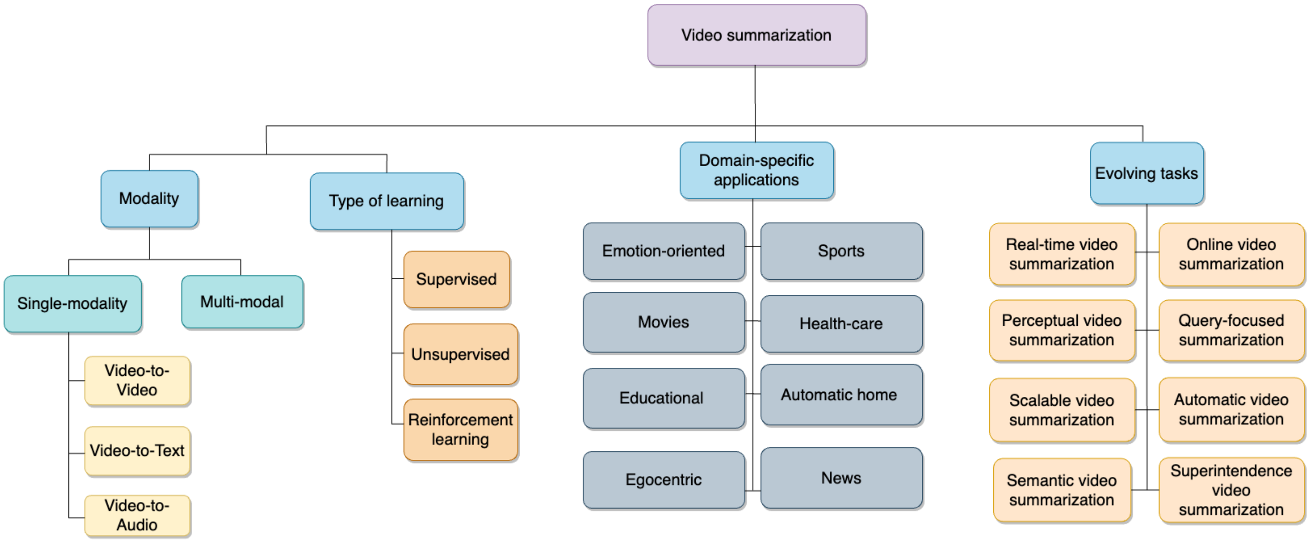
Video summarization
Modality
Type of learning
Domain-specific applications
Evolving tasks
Single-modality
Multi-modal
Video-to-Video
Video-to-Text
Video-to-Audio
Supervised
Unsupervised
Reinforcement learning
Emotion-oriented
Sports
Movies
Health-care
Educational
Automatic home
Egocentric
News
Real-time video summarization
Online video summarization
Perceptual video summarization
Query-focused summarization
Scalable video summarization
Automatic video summarization
Semantic video summarization
Superintendence video summarization

Figure 1: An organized flow diagram for video summarization. Here the boxes at different category of the flow diagram highlight the various types of factors such as "modality" and "type of learning" box for "Approaches", "application" for "use-cases", and "Evolving tasks" for "future directions".

2.2 Learning based methods in Video Summarization

Huang *et al.* [15] proposed a spatio-temporal transformer architecture [10] that is trained in a supervised fashion with the help of attention encoders. The evaluation revealed that this architecture was seen to perform better than most of the SOTA deep learning-based techniques for video summarization. The authors advise incorporating this technique into several applications, such as video saliency analysis. Li *et al.* [22] proposed a weakly supervised deep reinforcement learning (RL)-based architecture, utilizing two sub-networks for video classification and summary generation, working coherently to generate the summaries. The former one is trained on a large-scale video dataset with annotations, and the semantic representations from this sub-model are then used with the RL procedure of summary generation. This implementation outperforms other unsupervised models and is competitive with the supervised ones. Later, Gao *et al.* [12] proposed another RL-based method with label distributions and dual rewards that aimed to address the problem of ambiguity in the generated video summaries. Upon evaluation, this model has achieved good performance compared to other RL-based solutions.

Building upon a different learning nature, a study in 2022 [34] introduced a new method for unsupervised video summarization, utilizing contrastive losses to select the most important video frames. They have proposed new metrics for this experiment, such as local dissimilarity and global consistency among the frames that would be optimized. This work assumes that the videos are restricted to a well-defined central theme, which might not be true with all real-world videos; thus, modifications to the proposed model would be needed. With a slightly different intuition, Maria, Nektaria Minaidi *et al.* [27] have proposed an unsupervised approach based on adversarial learning and self-attention mechanism along with LSTMs for the encoding and decoding parts. The inclusion of an attention mechanism is intuitively made for adapting to new video sequences and for memorizing the long-range dependencies of video frames. This architecture has proven to be more effective than the existing unsupervised techniques for the same task.

On the other hand, a research work published in 2020 [36], named Sum-graph, is based on recursive graph modeling to capture the semantic relations between the frames of a video with the help of deep learning. This method was an improvisation of the existing graph-based strategies for video summarization and thus surpasses many of the existing techniques when trained in a supervised manner; the performance drops slightly compared to the prior method when trained in an unsupervised manner.

With the recent trend of transformers in the domain of machine and deep learning, many researchers have applied the idea of attention mechanism to their work. For instance, in 2020, Yen-Ting Liu *et al.* [24] developed one such work that aimed to enhance interrelations between the video features across the time and space of a video, whereas most previous works primarily focused on applying the attention mechanism to the video frames. This is achieved using a multi-concept video self-attention model, which maps the input video frame into several sub-spaces, allowing for better visualization of several visual features in the summary generation process. This method was generalized to supervised, unsupervised, and semi-supervised learning techniques and was effective against many SOTA methods upon evaluation.

2.3 Domain-specific Video Summarization

Domain-specific Video Summarization includes building different models pertaining to that particular domain [18, 38].

Thiruthuvanathan *et al.* [47] developed an emotional-oriented summarization model that can precisely acquire keyframes through hierarchical summarization and use the keyframes to detect faces and assess the emotional intent of the user. Zhao *et al.* [58] propose human attention based annotation pipeline for movie summarizer that takes advantage of spatio-temporal visual and auditory information. Sahu *et al.* [42] address the problem of egocentric video co-summarization and show how a shot level accurate summary can be obtained in a time-efficient manner using random walk on a constrained graph in transfer learned feature space with label refinement.

Yassir *et al.* [43] proposed a model with multiple pairwise ranking networks, training them based on user preferences. The rankers help in these local summaries, thus enabling specific global summaries that suit preferences. While it was able to perform competitively with the existing methods, the authors also mention the specific need to use this model in sports-video summarization for better understanding and improvisation. Raval *et al.* [40] proposed an approach to detect cricket video shot boundaries and extract replay segments from the cricket video sequence.

Haopeng Li *et al.* [21] addressed some limitations with the existing domain-specific models, specifically the ineffective utilization of high-level information between shots of a video and inter-shot relations. Here, they use similar videos to get the shot-level importance for each of them using cross-video information aggregation with the help of transformer units. Thus, semantically similar videos are encoded, and the generated summaries would be generalized and specific. This implementation was competitive with several of the existing techniques, as evaluated on some benchmark datasets.

Jeiyoon Park *et al.* [35] introduced a transformer-based framework that utilizes a pre-trained video captioning model to produce captions for a video. These captions, combined with encoded audio and video features, form the foundation of their architecture. Another research [14] proposed a new large-scale dataset called YouTube Video-Text Pairs (YTVT) to address the issue of overfitting in existing models caused by limited datasets. The proposed approach uses a multimodal self-supervised framework that captures meaningful relationships between video and text, resulting in a concise and efficient short-video summary.

2.4 Evolving tasks in Video Summarization

Narasimhan *et al.* [29] developed a query-driven transformer for the task. It includes the video captions generated by the pre-trained CLIP model for the text encoding part. They use classification, reconstruction, and diversity losses while training. This model had a significant improvement over the other query-focused techniques when evaluated on multiple datasets. However, the model is biased due to the wide range of data ingested while training. To improve this, Wu *et al.* [51], targets more user-friendliness and allows the users to interactively adjust the output. In this work, the authors have built a new visual-query dataset based on the query-driven video summarization dataset for this purpose. This model achieved comparable performance with the inclusion of other modalities, as user preferences is still being explored.

Mohamed Elfeki *et al.* [9] explored the DPP method on single-view tasks in videos with multiple tasks. There were no prior publicly available datasets for this specific task. Therefore, the research work created a new dataset named Multi-Ego. This work has no bias about the perspective of a video and finds utility in many real-world applications, such as video surveillance and sports summarization. Jingyang Lin *et al.* [23] curated a new dataset called VideoXum and proposed a novel system architecture for generating textual and video summaries for a given video. The approach leverages a state-of-the-art pre-trained model for visual language understanding and employs an encoder-decoder architecture to improve the performance of both text and video outputs.

3 A GENERAL PIPELINE

Given an input sequence of frames $x = x_1, x_2, ..., x_t, ..., x_T$ in a long video to be summarized while x_t is the visual features extracted at the $t^t h$ frame. The output of the video summarization algorithm can take one of two forms. First, keyframes selection, where a subset of isolated frames can be selected. Second, interval-based keyshots, where the summary is a set of short intervals along the time axis. The general pipeline of video summarization involves several stages and steps to process the input video and generate a concise and informative summary. Here's a high-level overview of the typical video summarization pipeline:

3.1 Sampling and Pre-processing

3.1.1 Frame Extraction. The video is divided into frames or shots, which are individual image frames representing different segments of the video.

3.1.2 Feature Extraction. Features are extracted from each frame, representing visual, audio, or textual information. These features will serve as input for the summarization framework.

3.1.3 External Data. The inclusion of external knowledge is suggested to improve the video summary. One such technique that makes use of extra knowledge is mentioned in [52], where Shaoping Lu *et al.*designed a hybrid multi-fusion model in which both the image and audio features of a video frame are used in the process. The proposed multi-fusion model maps the inter- and intra-level features across the modalities given, and it outperforms the SOTA methods with a significant difference, especially in a canonical training approach. This approach can be extended by implying the principles of cinematography and visual storytelling for the generation of summaries, as mentioned by the authors.

Other research works [10, 18, 30, 36, 44, 49, 50] that use the integration of external data are designed with multimodal solutions, while the rest of them [23, 24, 29, 43] are mainly focused on query-driven techniques.

3.2 Modality Integration for Video Representation

Using a multimodal approach, the features from different modalities (e.g., visual, audio, and text) are combined to create a unified representation of the video.

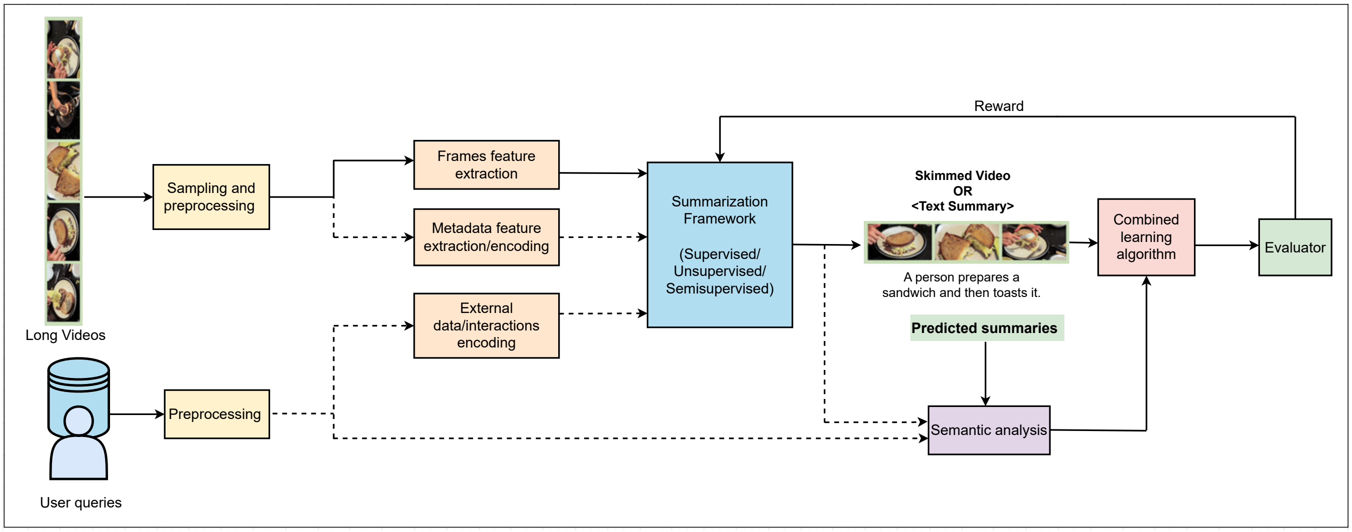
Long Videos
Sampling and preprocessing
Frames feature extraction
Metadata feature extraction/encoding
External data/interactions encoding
User queries
Preprocessing
Reward
Summarization Framework
(Supervised/ Unsupervised/ Semisupervised)
Skimmed Video
OR
<Text Summary>
A person prepares a sandwich and then toasts it.
Predicted summaries
Combined learning algorithm
Evaluator
Semantic analysis

Figure 2: A generic pipeline portraying the different stages involved in Video Summarization.

3.3 Summarizer framework

3.3.1 Keyframe Extraction. Keyframes are selected to represent the most salient and informative frames in the video. These keyframes may be chosen based on visual or audio features, or a combination of both.

3.3.2 Shot Segmentation. The video may be divided into shots (short consecutive sequences of frames) to further analyze and summarize each shot independently.

3.4 Summary Generation

3.4.1 Summary Length. The desired length of the video summary is determined, either as a fixed number of keyframes/shots or as a percentage of the original video length.

3.4.2 Selection Strategy. A selection strategy is applied to choose the top-ranked keyframes or shots to form the final summary. This could be a simple selection of the top N keyframes/shots or more sophisticated methods that consider diversity and coverage of the content.

3.5 Evaluator

3.5.1 Frame/Shot Importance Scoring. Each keyframe or shot is assigned a score based on its importance and relevance to the overall content of the video. Various criteria may be used, such as visual uniqueness, audio importance, or text relevance.

3.5.2 Ranking. The keyframes or shots are ranked based on their scores in descending order to prioritize the most important ones for the summary.

3.6 Post-processing

3.6.1 Smooth Transitions. Transitions between selected keyframes or shots can be smoothed out to create a more visually appealing and coherent summary.

3.6.2 Summary Refinement. Optional refinement steps can be applied to improve the quality and coherence of the generated summary.

3.7 Predicted Summary

The final output is a condensed video summary containing selected keyframes or shots that best represent the content of the original video. It is important to note that the specific algorithms and techniques used in each step of the pipeline may vary depending on the video summarization approach (single-modal, multimodal, supervised, unsupervised, etc.). Additionally, advances in deep learning and reinforcement learning have also been incorporated into video summarization pipelines, leading to more sophisticated and context-aware summarization methods.

4 DATASETS

Below are the details of the datasets which are commonly used for video summarization research.

4.1 Video to Video datasets

- **SumMe[13]:** The dataset contains 25 videos from YouTube, covering different domains such as sports, cooking, and music. Each video is associated with multiple summaries created by human annotators.
- **TVSum[46]:** The dataset consists of 50 videos of 1 to 11 minutes duration from diverse TV shows, with each video accompanied by human-generated summaries. The dataset covers various domains, including news, documentaries, and talk shows.
- **MED summaries[39]:** MED summaries dataset for evaluation of dynamic video summaries. It contains annotations of 160 videos: a validation set of 60 videos and a test set of 100 videos.

Table 1: A study on datasets available for video summarization for Video-to-Video summarization and Video-to-Text summarization. The '-' denotes the unavailability of details.

Dataset	Dataset type	Number of videos	Duration (min)	Type of content	Annotation type	Number of annotators per video
SumMe [13]	Video to Video	25	1 - 6	holidays, events, sports	multiple sets of key-fragments	15-18
TVSum [46]		50	2-10	news, how-to's, user-generated, documentaries (10 categories - 5 videos each)	multiple fragment-level scores	20
MED [39]		160	1 - 5	15 categories of various genres	one set of imp. scores	1 - 4
CoSum [5]		51	1-4	multi modal summarization	multiple set of key frames	3
Youtube [6]	Video to Text	50	1 - 10	cartoons, sports, tv-shows, commercial, home videos	multiple sets of key-frames	5
MSR-VTT [53]		10000	10 (avg)	Various web videos Human-written captions	sentence level scores	20
UET-Surveillance [8]		1200	5-15	surveillance videos	key frames	4
ActivityNet [20]		20000	120 sec	annotated events	multiple key frames	3

- **CoSum[5]:** This dataset is designed for multi-modal summarization and serves as a benchmark to validate video co-summarization techniques, where the goal is to create video summaries given a video collection of the same topic. It is collected from YouTube using 10 queries, in total 51 videos of 147m 40s.

4.2 Video to Text Datasets

- **YouTube[6]:** Youtube dataset contain 50 videos, whose annotations are sets of key-frames, produced by 5 users. The video duration ranges from 1 to 10 minutes. This dataset comprised of videos with diverse video content, such as cartoons, news, sports, commercials, TV-shows and home videos.
- **MSR-VTT (Microsoft Research-Video to Text)[53]:** The MSR-VTT dataset is a large collection of videos with captions. It has 10,000 web video clips, and each video comes with 20 captions written by humans. These captions give detailed descriptions of what happens in the videos.
- **UET Surveillance[8]:** This dataset comprises 1200 surveillance videos from the University of Engineering and Technology (UET), Lahore. The videos are gathered from four different locations: Girls Student Service Center (GSSC), Boys Student Service Center (BSSC), UET Bus Stand, and opposite CS department. Each location has approximately 300 videos, with each video's duration ranging from 5 to 15 seconds. The videos are accompanied by detailed descriptions in multi-line English sentences, typically consisting of 4 to 6 sentences. Additionally, each video has an abstractive textual summary provided, which is a shorter version of the description, containing 1 to 2 sentences. To facilitate training and evaluation, the dataset is divided into three subsets: 800 videos for training, 200 videos for testing, and another 200 videos for validation.
- **ActivityNet[20]:** The dataset is based on ActivityNet v1.3 and consists of 20,000 untrimmed YouTube videos with 100,000 caption annotations. On average, the videos are around 120 seconds long. Each video contains more than 3 annotated events, and each event is associated with corresponding start and end times, along with human-written sentences that describe the events. These sentences contain an average of 13.5 words. The dataset is split into three subsets for training, validation, and testing, with 10,024 videos in the training set, 4,926 videos in the validation set, and 5,044 videos in the test set. This dataset is widely used for tasks such as video captioning, event recognition, and temporal localization, as it provides a rich source of video content with detailed annotations.

5 CHALLENGES

Despite the growing research on related areas, video summarization poses several challenges that require further research and development. Understanding and addressing these challenges are crucial for advancing the field and improving the quality and effectiveness of video summarization techniques.

5.1 Extracting Semantically Meaningful Content

Identifying and extracting semantically meaningful content from videos is a challenging problem. Simply selecting keyframes or shots based on visual saliency may not always capture the most important events, actions, or semantic concepts in the video. Advancements are required to develop techniques that can understand the semantic context of video content and summarize videos based on high-level semantic information.

5.2 Handling Noisy and Redundant Content

Videos often contain noisy or redundant segments that may negatively impact the quality of the generated summaries. Filtering out irrelevant or redundant content is essential to produce concise and informative summaries. Developing robust algorithms that can effectively identify and remove noisy or redundant segments while preserving important information remains a challenging task.

5.3 Handling Diverse Video Content

Videos exhibit significant diversity in terms of content, genre, length, and quality. Existing video summarization techniques may not be equally effective for different video types such as sports videos, movies, documentaries, or user-generated content. It is essential to develop techniques that can handle the diverse nature of video content and adapt to different video genres and characteristics.

5.4 Temporal Coherence and Context

Preserving the temporal coherence and context of the original video while generating a summary is a challenging task. Maintaining a narrative flow and smooth transitions between selected keyframes or shots is crucial for creating coherent and meaningful summaries. Developing algorithms that can effectively capture and represent the temporal dynamics and contextual information of videos remains an open problem.

5.5 Scalability and Efficiency

Dealing with large-scale video datasets presents significant challenges in terms of computational efficiency and scalability. Current video summarization algorithms often struggle to process extensive collections of videos efficiently. There is a need to develop algorithms that can handle the ever-increasing volume of video data while maintaining real-time or near-real-time processing capabilities.

5.6 Subjectivity in Evaluation

Evaluating the quality of video summaries is subjective, as it depends on individual preferences and information needs. While existing evaluation metrics such as F-measure, precision, and recall provide some insights, there is a need for more comprehensive and objective evaluation methodologies. Developing evaluation metrics that can capture the effectiveness, representativeness, and informativeness of video summaries in a consistent and reliable manner is an ongoing research challenge.

6 FUTURE RESEARCH DIRECTIONS

The field of video summarization offers numerous open problems and avenues for future research. Some potential open problems include the following:

6.1 Personalized Video Summarization

Exploring the potential for personalized video summarization techniques that can adapt to individual user preferences, information needs, and interaction patterns. Recent efforts have focused on personalized video summarization to cater to individual user preferences. Mujtaba *et al.* [28] focus on a client-driven technique for generating personalized summaries with significantly lower computational expenses compared to many existing methods [2, 3]. The authors proposed a novel framework using an online transfer of data from the server to the client. This process includes the extraction of video thumbnails according to the user's preference and uses a lightweight architecture to be used efficiently on the client's end. The research work could be explored and extended to serve many online streaming applications where the content is focused more on user behavior. Sharghi et al [45] able to address the subjectivity of users to personalize the generated video summaries. A memory-based network was designed with a query module that captures the user's preferences and a summary module that generates the relevant video frames as per the query vector constructed. For this task, they have built a new dataset based on UT-Egocentric data. However, adaptation of models based on user preferences over time and the improvement of existing architecture for memory efficiency still need to be explored. Also, investigating techniques that can understand and summarize videos based on high-level semantic concepts and contextual understanding. Different users or applications may have specific preferences for the summaries, and the system should be able to adapt accordingly.

6.2 Real-time Video Summarization

Real-time summarization involves processing large volumes of video data, especially in surveillance systems [17] and live event monitoring [7]. Deep learning-based summarization approaches [25, 26] pose challenges for real-time implementation due to high computational complexity. Recent research [23, 50] mainly focused on the new state-of-the-art architectures and their adaptation to real-time systems. However, these frameworks lack generalizability across diverse tasks for summarizing a video in terms of scalability. Many of these research works [48] have left some challenges and adaptations to be delved into further. Scalable solutions that can handle varying video lengths and diverse content are essential for real-time deployment. Addressing these challenges in real-time (live streaming, surveillance systems, and event monitoring) is an exciting area of research that offers numerous opportunities for video summarization. Developing efficient and adaptive algorithms, incorporating online learning techniques, and considering user preferences will be essential for advancing this field and making real-time summarization a reality in various applications.

6.3 Benchmarking and Evaluation

Despite the growing research in related areas, benchmarking and evaluation metrics still need to be discussed. Most notably, no uniform method exists to evaluate video data summarization quality or quality criteria. A high-quality video summarization dataset is highly subjective, and every dataset may be appropriate for specific applications, but its quality may be insufficient for other purposes. Therefore, establishing standardized benchmarks, comprehensive datasets, and evaluation metrics that capture various aspects of summarization quality will enable fair comparison and progress measurement. Such quality parameters would enable robust and generalizable evaluation of video data pipelines. Furthermore, these quality aspects will promote the development of more effective and reliable summarization methods to cater to the diverse needs of various real-world applications.

7 CONCLUSION

This paper provides a survey for video summarization based on different perspectives and also compares different literature works related to video summarization. The paper also study a general end-to-end framework and challenges in the area of video summarization to present a comprehensive review. A general pipeline for an end-to-end video summarization system is discussed. Different techniques, perspectives, and modalities are considered to preserve the diversity of the survey. This paper also talks about the open research problems and future directions of the video summarization.

REFERENCES

[1] [n. d.]. How Video consumption is changing in 2023 [New Research]. https://blog.hubspot.com/marketing/how-video-consumption-is-changing.

[2] Muhammad Ajmal, Muhammad Husnain Ashraf, Muhammad Shakir, Yasir Abbas, and Faiz Ali Shah. 2012. Video summarization: techniques and classification. In *Computer Vision and Graphics: International Conference, ICCVG 2012, Warsaw, Poland, September 24-26, 2012. Proceedings*. Springer, 1–13.

[3] Evlampios Apostolidis, Eleni Adamantidou, Alexandros I Metsai, Vasileios Mezaris, and Ioannis Patras. 2021. Video summarization using deep neural networks: A survey. *Proc. IEEE* 109, 11 (2021), 1838–1863.

[4] Bo-Wei Chen, Jia-Ching Wang, and Jhing-Fa Wang. 2009. A novel video summarization based on mining the story-structure and semantic relations among concept entities. *IEEE Transactions on Multimedia* 11, 2 (2009), 295–312.

[5] Wen-Sheng Chu, Yale Song, and Alejandro Jaimes. 2015. Video co-summarization: Video summarization by visual co-occurrence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3584–3592.

[6] Sandra Eliza Fontes De Avila, Ana Paula Brandao Lopes, Antonio da Luz Jr, and Arnaldo de Albuquerque Araújo. 2011. VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern recognition letters* 32, 1 (2011), 56–68.

[7] Alan J Demers, Johannes Gehrke, Biswanath Panda, Mirek Riedewald, Varun Sharma, Walker M White, et al. 2007. Cayuga: A General Purpose Event Monitoring System.. In *Cidr*, Vol. 7. 412–422.

[8] Aniqa Dilawari and Muhammad Usman Ghani Khan. 2019. ASoVS: abstractive summarization of video sequences. *IEEE Access* 7 (2019), 29253–29263.

[9] Mohamed Elfeki, Liqiang Wang, and Ali Borji. 2021. Multi-Stream Dynamic Video Summarization. arXiv:1812.00108 [cs.CV]

[10] Antonio H de O Fonseca, Emanuele Zappala, Josue Ortega Caro, and David van Dijk. 2023. Continuous spatiotemporal transformers. *arXiv preprint arXiv:2301.13338* (2023).

[11] Tsu-Jui Fu, Shao-Heng Tai, and Hwann-Tzong Chen. 2019. Attentive and adversarial learning for video summarization. In *2019 IEEE Winter Conference on applications of computer vision (WACV)*. IEEE, 1579–1587.

[12] Yongbiao Gao, Ning Xu, and Xin Geng. 2021. Video Summarization via Label Distributions Dual-Reward. In *International Joint Conference on Artificial Intelligence*. https://api.semanticscholar.org/CorpusID:237101049

[13] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. 2014. Creating summaries from user videos. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*. Springer, 505–520.

[14] Li Haopeng, Ke Qiuhong, Gong Mingming, and Tom Drummond. 2022. Progressive Video Summarization via Multimodal Self-supervised Learning. arXiv:2201.02494 [cs.CV]

[15] Tzu-Chun Hsu, Yi-Sheng Liao, and Chun-Rong Huang. 2023. Video Summarization With Spatiotemporal Vision Transformer. *IEEE Transactions on Image Processing* 32 (2023), 3013–3026. https://doi.org/10.1109/TIP.2023.3275069

[16] Tao Jin, Siyu Huang, Yingming Li, and Zhongfei Zhang. 2019. Low-Rank HOCA: Efficient High-Order Cross-Modal Attention for Video Captioning. arXiv:1911.00212 [cs.LG]

[17] Kajal Kansal, Yongkang Wong, Wei Jian Peh, Hui Lam Ong, and Mohan Kankanhalli. 2023. Handling Privacy Regulations in Video Surveillance Systems. (2023).

[18] Vishal Kaushal, Sandeep Subramanian, Suraj Kothawade, Rishabh Iyer, and Ganesh Ramakrishnan. 2019. A framework towards domain specific video summarization. In *2019 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 666–675.

[19] Irena Koprinska, James Clark, Sergio Carrato, et al. 2004. VideoGCS-a clustering-based system for video summarization and browsing. In *6th COST 276 Workshop on Information and Knowledge Management for Integrated Media Communication*. Citeseer, 34–40.

[20] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*. 706–715.

[21] Haopeng Li, Qiuhong Ke, Mingming Gong, and Rui Zhang. 2023. Video Joint Modelling Based on Hierarchical Transformer for Co-Summarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 3 (2023), 3904–3917. https://doi.org/10.1109/TPAMI.2022.3186506

[22] Zutong Li and Lei Yang. 2021. Weakly Supervised Deep Reinforcement Learning for Video Summarization With Semantically Meaningful Reward. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 3238–3246. https://doi.org/10.1109/WACV48630.2021.00328

[23] Jingyang Lin, Hang Hua, Ming Chen, Yikang Li, Jenhao Hsiao, Chiuman Ho, and Jiebo Luo. 2023. VideoXum: Cross-modal Visual and Textural Summarization of Videos. arXiv:2303.12060 [cs.CV]

[24] Yen-Ting Liu, Yu-Jhe Li, and Yu-Chiang Frank Wang. 2020. Transforming Multi-Concept Attention into Video Summarization. arXiv:2006.01410 [cs.CV]

[25] Yu-Fei Ma, Xian-Sheng Hua, Lie Lu, and Hong-Jiang Zhang. 2005. A generic framework of user attention model and its application in video summarization. *IEEE transactions on multimedia* 7, 5 (2005), 907–919.

[26] Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, and Mingjing Li. 2002. A user attention model for video summarization. In *Proceedings of the tenth ACM international conference on Multimedia*. 533–542.

[27] Maria Nektaria Minaidi, Charilaos Papaioannou, and Alexandros Potamianos. 2023. Self-Attention Based Generative Adversarial Networks For Unsupervised Video Summarization. arXiv:2307.08145 [cs.CV]

[28] Ghulam Mujtaba, Adeel Malik, and Eun-Seok Ryu. 2022. LTC-SUM: Lightweight Client-Driven Personalized Video Summarization Framework Using 2D CNN. *IEEE Access* 10 (2022), 103041–103055. https://doi.org/10.1109/access.2022.3209275

[29] Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. 2021. CLIP-It! Language-Guided Video Summarization. arXiv:2107.00650 [cs.CV]

[30] Chong-Wah Ngo, Yu-Fei Ma, and Hong-Jiang Zhang. 2005. Video summarization and scene detection by graph modeling. *IEEE Transactions on circuits and systems for video technology* 15, 2 (2005), 296–305.

[31] Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya. 2017. Video summarization using deep semantic features. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part V 13*. Springer, 361–377.

[32] Mayu Otani, Yale Song, Yang Wang, et al. 2022. Video Summarization Overview. *Foundations and Trends® in Computer Graphics and Vision* 13, 4 (2022), 284–335.

[33] Costas Panagiotakis, Harris Papadakis, and Paraskevi Fragopoulou. 2020. Personalized video summarization based exclusively on user preferences. In *European conference on information retrieval*. Springer, 305–311.

[34] Zongshang Pang, Yuta Nakashima, Mayu Otani, and Hajime Nagahara. 2022. Contrastive Losses Are Natural Criteria for Unsupervised Video Summarization. arXiv:2211.10056 [cs.CV]

[35] Jeiyoon Park, Kiho Kwoun, Chanhee Lee, and Heuiseok Lim. 2023. Multimodal Frame-Scoring Transformer for Video Summarization. arXiv:2207.01814 [cs.LG]

[36] Jungin Park, Jiyoung Lee, Ig-Jae Kim, and Kwanghoon Sohn. 2020. SumGraph: Video Summarization via Recursive Graph Modeling. arXiv:2007.08809 [cs.CV]

[37] Bryan A. Plummer, Matthew Brown, and Svetlana Lazebnik. 2017. Enhancing Video Summarization via Vision-Language Embedding. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1052–1060. https://doi.org/10.1109/CVPR.2017.118

[38] Danila Potapov, Matthijs Douze, Zaid Harchaoui, and Cordelia Schmid. 2014. Category-Specific Video Summarization. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 540–555.

[39] Danila Potapov, Matthijs Douze, Zaid Harchaoui, and Cordelia Schmid. 2014. Category-specific video summarization. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*. Springer, 540–555.

[40] Khushali R Raval and Mahesh M Goyani. 2023. Shot Segmentation and Replay Detection for Cricket Video Summarization. In *2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)*. IEEE, 933–938.

[41] Paolo Remagnino, Graeme A Jones, Nikos Paragios, and Carlo S Regazzoni. 2002. Video-based surveillance systems: computer vision and distributed processing. (2002).

[42] Abhimanyu Sahu and Ananda S Chowdhury. 2023. Egocentric video co-summarization using transfer learning and refined random walk on a constrained graph. *Pattern Recognition* 134 (2023), 109128.

[43] Yassir Saquil, Da Chen, Yuan He, Chuan Li, and Yong-Liang Yang. 2021. Multiple Pairwise Ranking Networks for Personalized Video Summarization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 1698–1707. https://doi.org/10.1109/ICCV48922.2021.00174

[44] Xindi Shang, Zehuan Yuan, Anran Wang, and Changhu Wang. 2021. Multimodal video summarization via time-aware transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1756–1765.

[45] Aidean Sharghi, Jacob S. Laurel, and Boqing Gong. 2017. Query-Focused Video Summarization: Dataset, Evaluation, and A Memory Network Based Approach. arXiv:1707.04960 [cs.CV]

[46] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. 2015. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5179–5187.

[47] Michael Moses Thiruthuvanathan and Balachandran Krishnan. 2022. Multimodal emotional analysis through hierarchical video summarization and face tracking. *Multimedia Tools and Applications* 81, 25 (2022), 35535–35554.

[48] Vasudha Tiwari and Charul Bhatnagar. 2021. A survey of recent work on video summarization: approaches and techniques. *Multimedia Tools and Applications* 80, 18 (2021), 27187–27221.

[49] Dian Tjondronegoro, Yi-Ping Phoebe Chen, and Binh Pham. 2004. Highlights for more complete sports video summarization. *IEEE multimedia* 11, 4 (2004), 22–37.

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. arXiv:1706.03762 [cs.CL]

[51] Guande Wu, Jianzhe Lin, and Claudio T. Silva. 2022. IntentVizor: Towards Generic Query Guided Interactive Video Summarization. arXiv:2109.14834 [cs.CV]

[52] Jiehang Xie, Xuanbai Chen, Shao-Ping Lu, and Yulu Yang. 2022. A Knowledge Augmented and Multimodal-Based Framework for Video Summarization. In *Proceedings of the 30th ACM International Conference on Multimedia* (Lisboa, Portugal) *(MM '22)*. Association for Computing Machinery, New York, NY, USA, 740–749. https://doi.org/10.1145/3503161.3548089

[53] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5288–5296.

[54] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. 2021. Deep learning for person re-identification: A survey and outlook. *IEEE TPAMI* 44, 6 (2021), 2872–2893.

[55] Bin Yu, Wei-Ying Ma, Klara Nahrstedt, and Hong-Jiang Zhang. 2003. Video summarization based on user log enhanced link analysis. In *Proceedings of the eleventh ACM international conference on Multimedia*. 382–391.

[56] Bin Zhao, Maoguo Gong, and Xuelong Li. 2021. Audiovisual video summarization. *IEEE Transactions on Neural Networks and Learning Systems* (2021).

[57] Bin Zhao and Eric P Xing. 2014. Quasi real-time summarization for consumer videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2513–2520.

[58] Defang Zhao, Dandan Zhu, Xiongkuo Min, Jiaomin Yue, Kaiwei Zhang, Qiangqiang Zhou, Guangtao Zhai, and Xiaokang Yang. 2023. Human attention based movie summarization: Dataset and baseline model. *Neurocomputing* 534 (2023), 106–118.

[59] Xiatian Zhu, Chen Change Loy, and Shaogang Gong. 2015. Learning from Multiple Sources for Video Summarisation. arXiv:1501.03069 [cs.CV]