Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
DOI : 10.47974/JSMS-1583

Video summarization techniques : A comprehensive review

Charu Kavadia *
Ashutosh Gupta †
Prasun Chakrabarti §
Yashoverdhan Vyas ‡
Department of Computer Science and Engineering
Sir Padampat Singhania University
Udaipur 313601
Rajasthan
India

Abstract

Video summary is becoming a crucial area of research due to the quick growth of video content in many sectors, such as social media, education, entertainment, and surveillance.

This paper presents a comprehensive review of existing methods for video summarization, emphasizing both extractive and abstractive strategies. Using techniques like shot boundary detection and clustering, important frames or segments from the original video are identified as part of the extractive summarizing process. In contrast to extractive methods, Abstractive video summarization aims to interpret and condense the most salient content from a video into a newly generated summary, leveraging advanced machine learning techniques such as deep neural networks, natural language processing, reinforcement learning, attention mechanisms, generative adversarial networks, and multi-modal learning frameworks. We also cover the applications and challenges faced in practical implementations, as well as strategies that combine the two approaches. The datasets used to benchmark these methods are also covered in the publication. This review aims to give readers a comprehensive understanding of the state of video summarization research as well as its future directions.

Subject Classification (2020): *Primary: 68-02, Secondary: 68U10, 68T45, 68T10.*

Keywords: *Event-based summarization, Attention mechanism, Multi-modal learning, Keyframe selection, Supervised methods, Unsupervised methods, Generative adversarial networks.*

* *E-mail:* **charukavadia@gmail.com (Corresponding Author)**
† *E-mail:* **ashu.gupta@spsu.ac.in**
§ *E-mail:* **prasun.chakrabarti@spsu.ac.in**
‡ *E-mail:* **yashoverdhan.vyas@spsu.ac.in**

© Taru Publications

1. Introduction

Rapid technological advancements in recent years have resulted in the integration of camcorders into numerous devices [1]. Individuals frequently record their everyday routines and memorable events, producing vast amounts of video information. The number of videos on the internet has increased dramatically as a result of mobile devices making it simple to produce and share videos on social media [2]. Finding and categorizing certain video footage can take a long time. Although conventional methods that represent videos as a sequence of continuous frames are effective for movie playback, they fall short when applied to contemporary multimedia applications such as video browsing, content-based search, retrieval, and navigation [3]. Techniques for automatically summarizing and indexing video information have been developed in response to the demand for effective time management [1]. These methods aid in finding, looking for, classifying, and identifying actions in videos. The published studies classify video summarizing techniques into a number of important categories. Video summary that is static, chooses a group of key frames that represent important scenes or moments in the video to build a synopsis. In [4], static video summarization is used to identify events by choosing keyframes with k-means clustering. Providing a summary of specific video events is the aim of event-based summarizing approaches. When specific activities are the summary's main objective, several techniques are helpful. The authors in [5] concentrate on providing summaries of particular sporting events. In addition to static and event-driven summarization techniques, personalized video summarization methods have been developed, which tailor the summary output according to individual user preferences and interests. [6] creates topic-related summaries to satisfy the varying subjective requirements of users. The possibilities of video summarization have been substantially improved by recent developments in deep learning. For instance, to produce succinct and insightful summaries, attention processes and reinforcement learning have been used. Models are able to concentrate on pertinent segments of the film thanks to attention processes. These models effectively extract critical spatial and temporal features from video data by dynamically attending to various aspects of the content. Treating the video as an interactive environment, reinforcement learning enables the models to iteratively improve and generate optimal summaries. The quality of the resulting summary determines how the Reinforcement Learning agents learn to maximize the reward [7].

Moreover, multi-modal learning is a potential method for video summarization. Multi-modal models can produce more contextually rich summaries by combining data from other modalities, including text, audio, and visual elements [6].

Researchers have also looked on single-view and multi-view summarizing techniques in addition to these developments. Alongside these developments, scholars have investigated summarization techniques that use both single-view and multi-view perspectives. The goal of single-view summary is to create a succinct depiction of the video from only one viewpoint. Multi-View summarization, on the other hand, incorporates data from several perspectives [8].

Finally, it is increasingly common for video summarizing technologies to incorporate interactive elements. Users can alter and improve summaries according to their requirements and preferences thanks to features like user feedback methods. Users can alter and improve summaries according to their requirements and preferences thanks to features like user feedback methods. Video summarizing systems are more adaptable and user-friendly in a variety of settings and applications thanks to this interaction [9]. The structure of the paper is as follows. An introduction and synopsis of video summary are given in Section 1. The techniques used in video summary are covered in Section 2. The many methods used in the field are examined in Section 3. The datasets used in the publications presented are examined in Section 4. Applications of video summarization are examined in Section 5. Our work is summed up in Section 6, which also offers a look at potential future paths.

2. Video Summarization Methodology

The two primary kinds of video summarizing techniques are extractive and abstractive techniques.

2.1 *Extractive Methods*

Without changing the content, extractive video summarizing techniques pick out significant portions of the original film.

2.1.1 Keyframe Selection

In extractive video summarization, keyframe selection is often guided by annotated data and specific criteria, including user perception, notable events, and the presence of particular objects. Keyframe selection

strategies can generally be classified into three categories: perception-driven, event-centered, and object-focused approaches. Object-focused techniques emphasize frames containing relevant or significant objects within the video content [28]. Event-based approaches focus on specific times and behaviors. Frames that are deemed significant by the user are used in perception-based approaches [29].

2.1.2 Keyshot Selection

Another important component of extractive summarization is keyshot selection, wherein a video is segmented into discrete shots, and representative segments are chosen to summarize the content. This process can be broadly categorized into four types: area-based, event-driven, object-focused, and activity-oriented approaches. Activity-based methods prioritize selecting shots that emphasize prominent actions or motion patterns within the video. The goal of area-based picture selection is to highlight particular areas of the video [30].

2.2 *Abstractive Methods*

By producing fresh material that encapsulates the spirit of the original video, abstractive video summarizing seeks to produce succinct summaries of videos. Abstractive summarization entails analyzing and synthesizing data to produce new, condensed representations, in contrast to extractive approaches, which choose and combine pre-existing video segments [12].

3. Video Summarization Techniques

This section reviews various video summarization techniques, which can be classified according to the employed learning paradigm, the nature of extracted features, the characteristics of the input video, and the format of the resulting summary.

3.1 *Learning-based Methods*

Techniques for video summarization can be divided into four categories based on the methodology: reinforcement, supervised, unsupervised, and weakly supervised.

3.1.1 Supervised Methods

Supervised video summarizing uses tagged or annotated material in the training phase to learn how to generate summaries. These techniques

use datasets in which every video is accompanied by a synopsis that focuses on the key moments in the video. Supervised video summarization techniques frequently employ deep learning architectures such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), long short-term memory (LSTM) units, and transformer-based models to learn representative features and temporal dependencies [12, 6, 32].

3.1.2 Unsupervised Methods

Unsupervised video summarization has seen notable progress in recent years, driven by the integration of clustering methods and generative adversarial networks (GANs), which enhance summary quality without relying on labeled data. Techniques based on clustering are included in unsupervised methods. Creating distinct and significant clusters from a finite unlabelled dataset is the aim of clustering. The video sequence was segmented using clustering in [32]. The main concept is to improve spatial diversity by including Visual Place Recognition into the clustering procedure. The summary is then created by choosing a representative key frame from each cluster. Using a discriminator and a generator, some methods use adversarial training. The summarizer tries to fool the

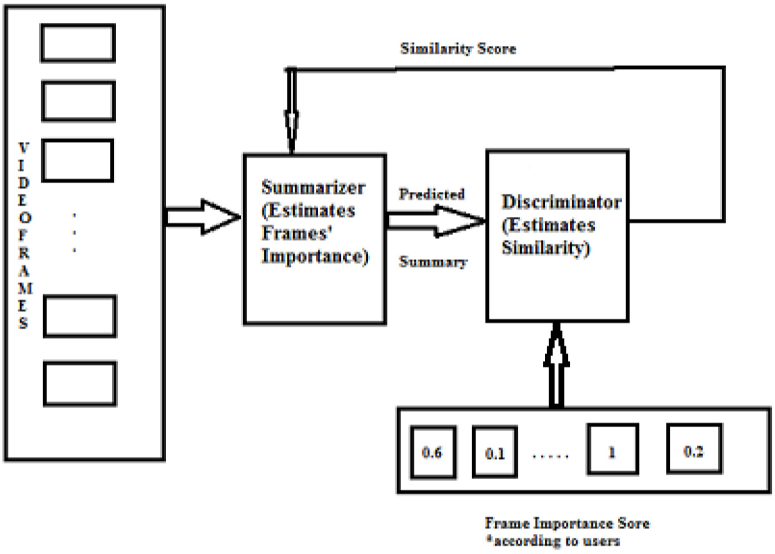


Figure 1

Generator and Discriminator in Video Summarization Process.

discriminator into not differentiating between the predicted and user-generated summaries, while the discriminator aims to learn how to do so.

The Summarizer can produce a summary that closely matches the users' expectations even when the discriminator's confidence level is quite low. An outline of how the discriminator and generator combine to create the video summary is provided in Figure 1. As proposed in [2], adversarial learning is employed wherein a summary discriminator network differentiates between authentic and generated summaries, while a fully convolutional sequence network (FCSN) is responsible for selecting the most relevant frames.

Unsupervised LSTM units are used in the majority of techniques that use GANs to determine the significance of each video frame. Nevertheless, these techniques produced results about the low variance of the frame-level significance scores that LSTMs calculated. Consequently, it appears that the trained LSTM's choices have little effect on summarization. The authors in [34] proposed a technique utilizing Attentive Conditional Generative Adversarial Networks (AC-GANs) for video summarization. This approach makes use of a GAN framework in which a discriminator separates the frame-level significance scores predicted by a BiLSTM from the raw frame characteristics. In order to generate unsupervised video summaries by minimizing the distance between video representation and text representation, [31] proposed a deep semantic and attentive network for Video Summarization (DSAVS). They also incorporated a self-attention mechanism to effectively model long-range temporal dependencies within the video sequences. The work in [33] introduces a framework that models the task of selecting important video segments as a sequence generation problem, integrating an Actor-Critic reinforcement learning model within a Generative Adversarial Network (GAN). In this setup, the Actor and Critic collaboratively determine key video segments over time, while the discriminator provides feedback to guide their decisions at each step.

3.1.3 Weakly Supervised Methods

Given the high cost and effort associated with annotating video data, machine learning approaches that require minimal supervision are increasingly favored for video summarization tasks. Three categories of inadequate supervision exist. The first category is incomplete supervision, in which only a portion of the training dataset is labeled while the remaining data remains unlabeled. The second category is inexact supervision, where annotations are coarse-grained rather than frame- or

segment-specific. For instance, in [22], the model is trained to detect highlights by leveraging high-level video attributes, such as subject tags, which serve as video-level annotations.

Inaccurate supervision is the third kind, meaning that the labels aren't always accurate. This happens, for instance, when the person annotating the photographs is sloppy or exhausted, or when certain images are hard to classify. In [36], segments are ranked according to their significance using a segment score transformer and a contextual temporal video encoder. This method improves scalability for big datasets and eliminates the need for manual annotations.

3.1.4 Reinforcement Learning Methods

As the choice of one video segment influences the relevance and selection of subsequent segments, video summarization is inherently a sequential decision-making task. Reinforcement Learning (RL), well-suited for such sequential problems, enables the model to iteratively refine its selection strategy to generate an optimal summary.

In [7], the authors proposed an end-to-end reinforcement learning framework that introduces a novel reward mechanism independent of labeled data or user feedback. The model is designed to generate summaries that are both diverse and representative. However, there were certain drawbacks to this strategy. Because an agent only receives the reward. By suggesting a Progressive Reinforcement Learning network for Video Summarization (PRLVS) in an unsupervised learning manner, the authors of [38] attempted to solve this issue. The summarizing task is broken down hierarchically in this manner. An agent is trained by the model to gradually alter the summary. At each stage, the agent is rewarded for the entire summary and selects a frame to replace with another frame in the neighbourhood.

To efficiently capture the spatiotemporal features of input videos for reinforcement learning, [39] proposed an unsupervised video summarization approach that integrates a 3D spatiotemporal U-Net with a reinforcement learning framework.

3.2 *Input-Based Methods*

Techniques for video summarizing differ according to the kind and format of the input footage. The number of views in inputs can vary between SingleView and MultiView. Additionally, different input modalities have an impact on how video summarizing is approached.

3.2.1 SingleView Methods

Over the past few decades, there has been a lot of interest in and study on single view video summarization. Numerous methods have been put forth to provide an efficient summary. The authors in [4] suggested a clustering method based on equal partitions for summarizing single-view movies. By using the K-means clustering algorithm to choose key frames, it finds the events.

The authors in [40] suggests a key-frame based single-view video summarizing technique. To eliminate unnecessary frames, the frame difference between neighbouring frames is computed. Key-frames are defined as frames having a difference larger than a predetermined threshold. The authors in [35,57] suggested an abnormality type-based technique that extracts the abnormal occurrences from surveillance recordings. The regular patterns and important areas of the events are initially modelled using a patch-based approach. Blob sequence optimization analyzes spatial, temporal, size, and motion relationships among objects to detect anomalies within a scene.

3.2.2 Multi-View Methods

In order to extract important information from several movies taken at the same time and place, multi-view video summarization is needed. To conform to a single-view input format, single-view summarizers process concurrently captured multi-view videos sequentially, disregarding their temporal synchronization. As a result, explanations become repetitious and redundant, failing to capture the multi-stream character of the video. Two categories of significant events should be covered in the universal overview, as per [13]. The first scenario involves significant overlap among multiple viewpoints, where the summary selects the most representative perspective. The second pertains to spatially independent events, in which each view is analyzed separately without cross-view interaction. The sequential determinantal point process (seqDPP), a probabilistic model for varied sequential subset selection, was proposed by the authors in [41]. The standard DPP interprets video frames as randomly permutable things, which is a shortcoming of the model seqDPP, which takes into account the sequential structures that are naturally present in video data. SeqDPP maintains the capability to model diverse subsets effectively. In [42], the authors introduced a novel video summarization technique that identifies frequently co-occurring shots across videos retrieved using a shared topic keyword, thereby leveraging visual co-occurrence across multiple videos.

The video summarizing task is classified as a multi-view representative selection problem by [43]. Finding visual components that depict the film from various angles is the primary goal. Optimization is performed using the Multi-view Sparse Dictionary Selection with Centroid Co-regularization (MSDS-CC) method. [44] developed another innovative technique for obtaining a summary of configurable length. It creates video abstracts by an accurate frame-adding procedure using a bottom-up method called "video clip growth," giving users the ability to alter the length of the video summaries. To address the challenge of supervised video summarization, [45] proposed a deep multi-view network incorporating an attention mechanism to enhance summary generation. The encoding stage utilizes a Bidirectional Long Short-Term Memory (Bi-LSTM) network to capture contextual information from video frames, while two attention-based LSTM networks function as decoders. The authors in [8] introduced an innovative approach to shot boundary detection by employing frame similarity metrics such as the Jaccard and Dice coefficients. The Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) algorithm is employed to filter out redundant or non-essential frames. The authors of [13] proposed an innovative variant of the well-known Determinantal Point Process (DPP) [41], providing a robust framework that identifies a diverse set of key events from moving cameras—often capturing different scenes—and selects the most representative view(s) of each event for inclusion in a unified summary. It is generalized by Multi-DPP to support multi-stream settings while preserving temporal order.

3.2.3 Modality-Based Methods

It is challenging for one modality to give enough information for complex video sequences, which could result in inaccurate prediction findings. Combining the expertise of several modalities compensates for the shortcomings of a single modality, which helps with future decision-making and enhances the caliber of summaries. This method increases the efficacy and precision of the summary process by utilizing the rich, complimentary information that each modality offers. A range of modalities, such as text, audio, and visual, can be included in videos.

- Visual Modality: The visual modality identifies different scenes and their transitions, facilitating the segmentation of the video into manageable segments that can be summarized individually. Furthermore, identifying objects, persons, and activities in the

movie highlights visually significant times and offers insightful information about the context and substance of scenes.

- Audio Modality: The audio modality leverages automatic speech recognition (ASR) to convert spoken dialogue into text, facilitating the detection of key phrases, conversations, and narrative elements within the video content.
- Textual Modality: This modality, which includes video-related text overlays and subtitles, uses text to represent spoken content. It enhances understanding of the video's content and provides context.
- Multi-modal learning has recently demonstrated that there are semantic links between the audio and visual modalities and that they share a consistency space. The audio modality complements the visual modality by enhancing the understanding of the video's structure and content. In practical terms, audio and visual aids complement each other to deliver activities in many modalities. The authors of [46] attempt to use LSTM to encode both visual and aural data. Initially, the audio and visual features are encoded in sequence, and their temporal dependencies are modeled using a two-stream LSTM architecture. The audio-visual fusion follows. LSTM was created to use an adaptive gating method to integrate auditory and visual data, taking advantage of the consistency space between them. To model the complex interactions between audio and visual modalities, [22] proposed a low-rank tensor fusion approach that efficiently generates informative fused features. This audio-visual tensor fusion method addresses the limitations associated with traditional bilinear pooling and linear fusion techniques. Because the feature distributions of audio and visual modalities differ—audio cues may not consistently align with corresponding video frames—linear fusion techniques fail to effectively capture the complex inter-modal relationships. The authors in [22] presented the Audio-Visual Recurrent Network (AVRN), a cutting-edge technique that improves video summarization by utilizing both audio and visual data. AVRN enhances the understanding of video content and structure by integrating both audio and visual modalities, unlike traditional approaches that rely solely on visual information. Evaluation of AVRN on the SumMe and TVSum datasets revealed that incorporating both audio and visual information led to superior video summarization performance compared to using visual features alone.

In [47], the authors generate video summaries by aligning visual features with semantic information—such as titles, user queries, descriptions, and comments—within a shared latent space to enhance relevance. In [48], the authors utilize semantic attention networks and human-annotated video descriptions to select representative shots by minimizing the discrepancy between the generated summary sentences and the original annotations. In order to produce thorough summaries, [16] use both textual and visual data. They do this by employing a pre-trained action recognition algorithm to extract visual features from films, which capture important elements and actions.

Textual information is simultaneously obtained through transcripts, either generated manually by users or automatically via a pre-trained Automatic Speech Recognition (ASR) system. [37] proposed a language-guided multimodal transformer that evaluates video frames based on their importance and relevance to either auto-generated detailed captions or user-specified queries. [12] presented an alignment-guided self-attention module to fuse cross-modal information in a cohesive way and align the temporal correlation between text and video modalities. In order to simulate the cross-modal correlation at various granularities, it also develops two new contrastive losses to model intra-sample and inter-sample correlations.

In [14], the authors presented computational models that forecast the significance of certain (long) video parts. Data from both modalities are integrated using bi-LSTM layers. [6] presented an innovative multimodal Transformer architecture capable of dynamically integrating features from multiple modalities, thereby addressing the limitations of individual modalities and improving the overall quality of the video summaries. The approach is structured into four key modules: feature extraction, feature representation learning, subject classification, and frame selection.

3.3 *Features-Based Methods*

This section categorizes video summarizing methods based on the kinds of features that are part of the employed strategy.

3.3.1 Spatial-Features-Based Methods

Spatial features in video summarization involve analyzing and extracting information based on the spatial characteristics of individual video frames. Each frame's visual content is captured by spatial

characteristics, which yield useful information for locating crucial scenes and frames.

[49] employed a modified nonparametric Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm to cluster color and texture features extracted from video frames, aiming to generate a summarized representation of the video content [50]. Pre-sampling is done on the original video. After obtaining color features through a histogram in the HSV color space, texture characteristics are derived using a 2D HAAR wavelet transform, also within the HSV space. Then, using a modified DBSCAN algorithm to cluster video frames, important frames are selected. Lastly, to facilitate visual comprehension, these important frames are rearranged in their original sequence.

3.3.2 Temporal-Features-Based Summarization

Segmenting the video into shots or temporal segments is known as temporal-based summarizing.This can be accomplished in a number of ways, including scene change detection, which uses sudden frame shifts to signal the start or finish of a section. To divide up temporal occurrences, the authors of [51] employed clustering of frame colour histograms. The authors in [52] employs temporal video segmentation to identify shot or scene boundaries. This method considers how each pair of frames differs from the others. In order to incorporate local temporal hints between adjacent segments, a new hierarchical temporal context encoder was presented in [22].

3.3.3 Spatio-temporal-Features-Based Summarization

In order to provide a succinct summary, spatiotemporal summarizing entails extracting and displaying spatial and temporal information from films. It plays a crucial part in managing the enormous volumes of video data produced in different industries, making it possible to extract the most important content without having to go through all the footage by hand. In [53], a novel method known as SUM-GDA (SUMmarization by Global Diverse Attention) is presented. It creates an attention mechanism in a convolutional neural network to produce globally diverse and educational video summaries. A novel framework named Deep Hierarchical LSTM Networks with Attention (DHAVS) was proposed in [31] for the task of video summarization. The approach leverages 3D ResNeXt-101 to extract spatiotemporal features and employs a deep hierarchical LSTM architecture with an attention mechanism to model

long-term dependencies and highlight essential keyframes. A cost-sensitive loss function is also introduced to improve the selection of significant frames, particularly under conditions of class imbalance. Testing DHAVS on the SumMe and TVSum datasets revealed that it performs better than the most advanced techniques, indicating notable progress in video summarizing.

3.4 *Output-Based Methods*

The video summarization techniques are covered in this part based on the output summary type.

3.4.1 Generic Video Summarization

A general-purpose video summary can be produced through techniques such as storyboard summarization or video skimming. By selecting the segments of the video that are thought to best represent its content in a synopsis format, a general summary is produced in each of these approaches. When a consumer wants to quickly review the video's highlights and lacks intuition about its substance, this kind of summary is helpful. In generic video summarizing, there is disagreement over what aspects of a video qualify to be included in the summary. It is possible to generate a basic general summary by randomly or uniformly sampling the frames.

However, a sophisticated method of choosing the frames that best capture the film's semantics can be created by utilizing computer vision and video processing techniques.

A generic video summarizing method known as GVSUM was proposed by the authors of [54] and may be used for a range of films, including home videos, sports videos, movies, and surveillance videos. It starts by separating the video's individual frames. Rather than decoding the complete video, a partial decoding approach is employed to extract I-Frames — intra-coded frames that independently contain all visual information for a specific video frame. Unsupervised grouping of the frames is used to assign each frame a unique cluster number based on the visual attributes that have been collected from them. A frame is selected for inclusion in the summary whenever there is a change in its assigned cluster label. A change in the cluster assignment indicates a transition in the visual content of the video, as frames are grouped according to their visual characteristics.

3.4.2 Personalized Video Summarization

Personalized video summary is a new technique that creates brief videos tailored to each viewer's preferences, requirements, and interests. In order to address the subjective needs, the writers of [6] concentrate on creating different topic-related video summaries. The new visual features are intended to be categorized into topic classes using the topic classification module. The topic classification problem is actually a multi-label classification challenge because a single video frame may show numerous topics. To assess whether a given frame aligns with a specific topic, the topic classification module is decomposed into multiple binary classification tasks.

Personalized video summarization enables the generation of sports highlights tailored to individual viewer preferences, ensuring that each user sees the segments of a sporting event that align most with their interests. Any sport where a game involves a series of activities can benefit from the architecture that the authors in [55] developed, which employs neural networks to create a tailored video summary. They utilize the players' actions as a contract to choose the best parts of the original video. By identifying the enthusiasm score in the audio stream, [5] describes an efficient summarization technique for sports films based on user-defined summary length.

3.4.3 Input-Query-Based Video Summarization

Query-based video summarization aims to generate concise and informative summaries that align with specific user queries or areas of interest. Unlike conventional video summarization methods that depend on predefined rules or algorithms, query-based summarization incorporates user-specified criteria to tailor the summary according to individual preferences.

To capture the visual characteristics of each shot, [56] formulate the task as a similarity estimation problem between video shots and the user query. Subsequently, they introduce a convolutional architecture that incorporates both query-aware global attention and localized self-attention to enhance relevance and contextual understanding. [10] introduced 'IntentVizor,' an interactive video summarization framework that effectively integrates both visual and textual query inputs to tailor the generated summary. The system comprises two primary components: the intent module, responsible for interpreting the user's objectives, and the summary module, which generates the video summary. Both modules

utilize a shared architecture called the Granularity-Scalable Ego-Graph Convolutional Network (GSE-GCN) to effectively align video features with user-defined intents or generic queries.

4. Datasets

The selection of datasets is essential for training and assessing the effectiveness of different approaches in video summarizing research. Table 1 provides an overview of the most commonly utilized datasets in video summarization research, highlighting their characteristics, content types, and relevance to different summarization approaches.

Table 1
Datasets for Video Summarization

Dataset	Description	Size	Duration (min)	Modality
TVSum [10]	Title-based, with 1,000 annotations of shot-level relevance scores gathered through crowdsourcing, and categories include news, how-to, documentaries, vlogs, and egotistical content.	50	2-10	Uni
SumMe [11]	There are at least fifteen human summaries annotated on each video.	25	1-6	Uni
BLiSS [12]	Includes transcripts with multimodal summaries and live stream videos (674 videos, 12,629 text).	13303	5	Multi
TopicSum [6]	Includes frame-level importance scores along with annotations of topic labels.	136	5	Multi
Multi-Ego [13]	Includes annotations for subject labels and significance scores at the frame level.	41	3-7	Uni
EDUVSUM [14]	Edx, YouTube, and TIB AVPortal are three well-known e-learning portals that offer instructional films with subtitles.	98	10-60	Multi

Contd...

FineGym [15]	A hierarchical video dataset for action interpretation at the fine level.	156	10	Multi
How2 [16]	Each of the brief educational movies from various subjects comes with a transcript.	80000	1-2	Multi
COIN [17]	Includes 180 projects and 11,000 instructional videos.Used to generate pseudo-summaries during training.	8521	3	Uni
CrossTask [18]	Includes 4,700 training movies that cover 83 jobs; it is used for training and creating fake summaries.	3675	3	Uni
HowTo100M [19]	A large-scale collection comprising over 100 million YouTube video clips, each accompanied by textual annotations.	136M	6-7	Multi
WikiHow Summaries [20]	High-quality test set produced by harvesting WikiHow articles with step visuals and video demos.	2106	2-7	Multi
VTW [21]	A big dataset of YouTube videos with annotations and key shots highlighted.	2529	2-5	Uni

5. Applications

Video summarization has a wide range of applications across diverse industries, revolutionizing how video content is analyzed, utilized, and managed. This section explores the various domains where video summarization techniques are applied.

5.1 *Video Highlight Detection*

Summarization techniques are used by streaming providers to produce highlight reels and trailers that give viewers a brief synopsis of the content. This may persuade prospective viewers to watch entire series or films. The emergence of programs for sharing short-form videos on the Internet has garnered international interest. To find highlight segments in lengthy, uncut videos, an automated technique is badly needed [12,58].

5.2 *Personalized Recommendations*

Recommendation systems are able to customize recommendations to each user's viewing preferences by examining the themes and subjects found in video summaries. For example, the platform may give priority to suggesting new culinary videos or similar material to a user who often watches summaries of cooking videos, offering a more tailored experience.

In [6], the authors proposed a method for generating video summaries across diverse topics, aiming to cater to the varying interests of different viewers. Instead of producing a single summary of an input video as is the case with current methods, topic-aware video summarizing considers the subjectivity of video summarization and gives consumers more options.

5.3 *Security and Monitoring*

In the domains of security and monitoring, video summarization is an essential technique that greatly improves the efficacy and efficiency of surveillance systems. The authors of [9] suggested IntentVizor, which makes video summarizing more engaging and streamlines the monitoring procedure.

5.4 *Video Indexing*

In the field of multimedia management, video indexing is a critical process that involves organizing and tagging video content to facilitate efficient search, retrieval, and analysis. It is important for many purposes, ranging from security and education to media and entertainment. Condensed versions of the content are provided by summarized videos, which facilitate the processing and indexing of important themes and subjects by search algorithms [23]. More accurate search results can be obtained by using summarization to help identify pertinent keywords and phrases that are essential for indexing [24]. For indexing systems, summarized films use less bandwidth and storage space, which can result in lower costs and quicker processing times.

5.5 *Anomaly Detection*

Prominent techniques for visual anomaly identification frequently depend on extensive training datasets and lengthy training periods. Video summarization allows for quicker model deployment while preserving similar detection performance by utilizing pre-trained deep models and denoising auto encoders (DAEs) [25]. The framework proposed in [9] not

only detects anomalies in surveillance footage but also provides explanations for each detection, offering valuable insights into the underlying decision-making process.

5.6 *Education and E-Learning*

Summaries of lengthy lectures, tutorials, or instructional videos are provided by video summarizing, which can aid in learning. The learning process is improved since students may rapidly understand important ideas, go over important details, and concentrate on topics of interest [26].

5.7 *Social Media and Marketing*

Video summary can be used by social media sites to provide product demos or promotional content. These summaries aid in drawing in viewers and efficiently conveying important points in a little amount of time [27].

6. Conclusions

Research on video summarization is essential since it supports numerous useful applications, including surveillance systems, tailored video recommendations, and video retrieval.

There are numerous ways to classify video summarizing techniques. Different strategies are adopted based on the nature of the input video or the desired format of the output summary. This work aimed to provide a comprehensive overview of the existing methodologies, benchmark datasets, and video summarization techniques currently employed in the field.

References

[1] Y. Pritch, A. Rav-Acha, and S. Peleg, "Non-chronological video synopsis and indexing," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 30, no. 11, pp. 1971–1984 (2008).

[2] M. Rochan and Y. Wang, "Video summarization by learning from unpaired data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7902–7911 (2019).

[3] C. Panagiotakis, A. Doulamis, and G. Tziritas, "Equivalent key frames selection based on iso-content principles," I*EEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 3, pp. 447–451 (2009).

[4] K. Kumar, D. Shrimankar, and N. Singh, "Equal partition based clustering approach for event summarization in videos," in *2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pp. 119–126 (2016).

[5] A. Banjar, H. Dawood, A. Javed, and B. Zeb, "Sports Video Summarization using Acoustic Symmetric Ternary Codes and SVM," *Applied Acoustics*, vol. 216, p. 109795 (2024).

[6] Y. Zhu, W. Zhao, R. Hua, and X. Wu, "Topic-aware Video Summarization using Multimodal Transformer," *Pattern Recognition*, vol. 140, p. 109578 (2023).

[7] K. Zhou, Y. Qiao, and T. Xiang, "Deep Reinforcement Learning for Unsupervised Video Summarization with Diversity-Representativeness Reward," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32 (2018).

[8] A. S. Parihar, J. Pal, and I. Sharma, "Multi-view video summarization using video partitioning and clustering," *Journal of Visual Communication and Image Representation*, vol. 74, p. 102991 (2021).

[9] G. Wu, J. Lin, and C. T. Silva, "Intentvizor: Towards generic query guided interactive video summarization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* pp. 10503–10512 (2022).

[10] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "TVsum: Summarizing web videos using titles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5179–5187 (2015).

[11] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in *Computer Vision – ECCV 2014,* pp. 505–520.

[12] B. He, J. Wang, J. Qiu, T. Bui, A. Shrivastava, and Z. Wang, "Align and attend: Multimodal summarization with dual contrastive losses," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14867–14878 (2023).

[13] M. Elfeki, L. Wang, and A. Borji, "Multi-stream dynamic video summarization," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 339–349 (2022).

[14] J. A. Ghauri, S. Hakimov, and R. Ewerth, "Classification of important segments in educational videos using multimodal features," arXiv:2010.13626 (2020).

[15] D. Shao, Y. Zhao, B. Dai, and D. Lin, "Finegym: A hierarchical video dataset for fine-grained action understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2616–2625 (2020).

[16] S. Palaskar, J. Libovický, S. Gella, and F. Metze, "Multimodal abstractive summarization for how2 videos," arXiv:1906.07901 (2019).

[17] Z. Tang, Y. Xiong, Y. Xu, W. Wang, X.-S. Hua, and J. Zhang, "Coin: A large-scale dataset for comprehensive instructional video analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1207–1216 (2019).

[18] Z. Xu, S. Buch, D. Ramanan, and J. C. Niebles, "Cross-task weakly supervised learning from instructional videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3537–3545 (2019).

[19] A. Miech, I. Laptev, and J. Sivic, "Howto100m: Learning a text-video embedding by watching hundred million video clips," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2630–2640 (2019).

[20] S. Hassan, M. Saleh, and M. Kubba, "Wikihow: A large scale text summarization dataset," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3243–3252.

[21] K.-H. Zeng, T.-H. Chen, J. C. Niebles, and M. Sun, "Title generation for user generated videos," in *Computer Vision – ECCV 2016*, pp. 609–625.

[22] Q. Ye, X. Shen, Y. Gao, Z. Wang, Q. Bi, P. Li, and G. Yang, "Temporal cue guided video highlight detection with low-rank audio-visual fusion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7950–7959 (2021).

[23] V. Tiwari and C. Bhatnagar, "A survey of recent work on video summarization: approaches and techniques," *Multimedia Tools and Applications,* vol. 80, no. 18, pp. 27187–27221 (2021).

[24] E. Apostolidis , E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "Video summarization using deep neural networks: A survey," *Proceedings of the IEEE*, vol. 109, no. 11, pp. 1838–1863 (2021).

[25] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018).

[26] I. Benedetto, A. Tedeschi, M. Piazza, and G. Sansonetti, "Abstractive video lecture summarization: applications and future prospects," *Education and Information Technologies,* vol. 29, no. 3, pp. 2951–2971 (2023).

[27] A. Sabha and A. Selwal, "Data-driven enabled approaches for criteria-based video summarization: a comprehensive survey, taxonomy, and future directions," *Multimedia Tools and Applications*, vol. 82, no. 21, pp. 32635–32709 (2023).

[28] P. Meena, H. Kumar, and S. K. Yadav, "A review on video summarization techniques," *Engineering Applications of Artificial Intelligence,* vol. 118, p. 105667 (2023).

[29] B. Köprü and E. Erzin, "Use of affective visual information for summarization of human-centric videos," I*EEE Transactions on Affective Computing* (2022).

[30] G. Mujtaba, A. Malik, and E.-S. Ryu, "LTCSUM: Lightweight client-driven personalized video summarization framework using 2D CNN," *IEEE Access,* vol. 10, pp. 103041–103055 (2022).

[31] J. Lin, S.-H. Zhong, and A. Fares, "Deep hierarchical LSTM networks with attention for video summarization," *Computers & Electrical Engineering*, vol. 97, p. 107618 (2022).

[32] C. Chen, Y. Zhang, X. Liu, H. Wang, and J. Wu, "Scene summarization: Clustering scene videos into spatially diverse frames," arXiv:2311.17940 (2023).

[33] E. Apostolidis, A. I. Metsai, E. Adamantidou, V. Mezaris, and I. Patras, "AC-SUM-GAN: Connecting actor-critic and generative adversarial networks for unsupervised video summarization," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 31, no. 8, pp. 3278–3292 (2020).

[34] X. He, Y. Hua, T. Song, Z. Zhang, Z. Xue, R. Ma, N. Robertson, and H. Guan, "Unsupervised video summarization with attentive conditional generative adversarial networks," in *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 2296–2304 (2019).

[35] W. Lin, Y. Zhang, J. Lu, B. Zhou, J. Wang, and Y. Zhou, "Summarizing surveillance videos with local patch-learning-based abnormality detection," *Neurocomputing*, vol. 155, pp. 84–98 (2015).

[36] M. Narasimhan, A. Rohrbach, and T. Darrell, "TL;DW? Summarizing Instructional Videos with Task Relevance & Cross-Modal Saliency," in *European Conference on Computer Vision*, pp. 540–557 (2022).

[37] M. Narasimhan, A. Rohrbach, and T. Darrell, "Clip-it! Language-Guided Video Summarization," *Advances in Neural Information Processing Systems*, vol. 34, pp. 13988–14000 (2021).

[38] G. Wang, X. Wu, and J. Yan, "Progressive Reinforcement Learning for Video Summarization," *Information Sciences*, vol. 655, p. 119888 (2024).

[39] T. Liu, H. Meng, M. Vlachos, and J. Zhao, "Video summarization through reinforcement learning with a 3D spatio-temporal U-Net," *IEEE Transactions on Image Processing*, vol. 31 (2022).

[40] S. Tirupathamma, "Key frame based video summarization using frame difference," *International Journal of Innovative Computer Science & Engineering*, vol. 4, no. 3 (2017).

[41] B. Gong, W.-L. Chao, K. Grauman, and F. Sha, "Diverse sequential subset selection for supervised video summarization," *Advances in Neural Information Processing Systems,* vol. 27 (2014).

[42] W.-S. Chu, Y. Song, and A. Jaimes, "Video summarization by visual cooccurrence," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3584–3592 (2015).

[43] J. Meng, H. Wang, J. Yuan, and Y.-P. Tan, "Video summarization via multi-view representative selection," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 1189–1198 (2017).

[44] G. Pan, Y. Zhu, and Q. M. J. Wu, "Video clip growth: A general algorithm for multi-view video summarization," in *PCM 2018*, pp. 112–122.

[45] Z. Ji, K. Xiong, Y. Pang, and X. Li, "Video summarization with attention-based encoder–decoder networks," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 30, no. 6, pp. 1709–1717 (2019).

[46] B. Zhao, X. Li, and X. Lu, "Audiovisual video summarization," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 34, no. 8, pp. 5181–5188 (2021).

[47] Y. Yuan, T. Mei, and W. Zhu, "Video summarization by learning deep side semantic embedding," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 29, no. 1, pp. 226–237 (2017).

[48] H. Wei, B. Ni, Y. Yan, H. Yu, X. Yang, and C. Yao, "Video summarization via semantic attended networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32 (2018).

[49] K. M. Mahmoud, M. A. Ismail, and M. S. Donia, "VSCAN: An enhanced video summarization using density-based spatial clustering," Springer, pp. 733–742 (2013).

[50] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proceedings of KDD (1996).

[51] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 1346–1353 (2012).

[52] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-Specific Video Summarization," Springer, pp. 540–555 (2014).

[53] P. Li, Q. Ye, L. Yuan, X. Wang, and J. Shao, "Exploring global diverse attention via pairwise temporal relation for video summarization," *Pattern Recognition*, vol. 111, p. 107677 (2021).

[54] M. Basavarajaiah and P. Sharma, "Gvsum: generic video summarization using deep visual features," *Multimedia Tools and Applications,* vol. 80, no. 9, pp. 14459–14476 (2021).

[55] A. Tejero-de Pablos, Y. Nakashima, T. Sato, N. Yokoya, and Y. Lin, "Summarization of user-generated sports video by using deep action recognition features," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2000–2011 (2018).

[56] S. Xiao, Z. Zhao, Z. Zhang, X. Yan, and M. Yang, "Convolutional hierarchical attention network for query-focused video summarization," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7 (2020).

[57] S. Sharma and D. Goyal, "Enhanced security using video summarization for surveillance system using deep LSTM model with K-means clustering technique," *Journal of Discrete Mathematical Sciences & Cryptography,* vol. 26, pp. 913–925 (2023).

[58] S. Sharma, A. Khunteta, and D. Goyal, "A deep learning-based highlight detection with pairwise segment ranking for effective video summarization," *Journal of Information and Optimization Sciences,* vol. 45 (2024).

Received July, 2025