IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 6 September 2022, accepted 8 November 2022, date of publication 18 November 2022, date of current version 29 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3223379

RESEARCH ARTICLE

Recent Challenges and Opportunities in Video Summarization With Machine Learning Algorithms

PAYAL KADAM[1,2], DEEPALI VORA[1], (Senior Member, IEEE), SASHIKALA MISHRA[1], SHRUTI PATIL[3], KETAN KOTECHA[3], AJITH ABRAHAM[4], (Senior Member, IEEE), AND LUBNA ABDELKAREEIM GABRALLA[5]

[1]Symbiosis Institute of Technology, Symbiosis International (Deemed University) (SIU), Lavale, Pune 412115, India
[2]Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune 411043, India
[3]Symbiosis Centre for Applied Artificial Intelligence (SCAAI), Symbiosis Institute of Technology, Symbiosis International (Deemed University) (SIU), Lavale, Pune 412115, India
[4]Machine Intelligence Research Labs (MIR Labs), Auburn, WA 98071, USA
[5]Department of Computer Science and Information Technology, College of Applied, Princess Nourah bint Abdulrahman University, Riyadh 11671, Saudi Arabia

Corresponding author: Deepali Vora (deepali.vora@sitpune.edu.in)

This work was supported by the Princess Nourah bint Abdulrahman University Researchers Supporting Project under Grant PNURSP2022R178, through the Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

ABSTRACT The fast progress in digital technology has sparked the generation of the amount of voluminous data from different social media platforms like Instagram, Facebook, YouTube, etc. There are other platforms, as well which generate large data like News, CCTV videos, sports, entertainment, etc. Lengthy Videos typically contain a significant number of duplicate occurrences that are uninteresting to the viewer. Eliminating this unnecessary information and concentra only on the crucial events will be far more advantageous. This produces a summary of lengthy films, which can save viewers time and enable better memory management. The highlights of a lengthy video are condensed into a video summary. Video summarization is an essential topic today, since many industries have CCTV cameras installed for various reasons such as monitoring, security, and tracking. Because surveillance videos are taken 24 hours a day, enormous amounts of memory and time are required if one wishes to trace any incident or person from the full day's video. The summary generated from multiple views is far more challenging, so more study and advancement in MVS is required. The conceptual basis of video summarizing approaches is thoroughly addressed in this paper. This paper addresses applications and technology challenges in Single view and Multi View summarization.

INDEX TERMS Video summarization survey, video sequence, single view summarization (SVS), multi view summarization (MVS), big data.

I. INTRODUCTION

The information generated from different domains like sports, medical, entertainment, surveillance, bulletin, etc. is in digital format and it's difficult to store this large amount of information efficiently. To make video browsing and retrieval easier, video summarization can be of great help. Video summarization is the process of creating and presenting the highlights of a video in a short duration. Using video summarization users will be able to view abstracts of lengthy videos that will make it individual life easy to extract maximum information within less time [1]. An individual can get a visual abstract of a long film using the method known as Video Summarization. The input can be an image or moving images, as we all know that video is nothing, but images are displayed sequentially with some speed that makes them appear moving. Generated video summary can be of mainly two types, one is static, and the other is dynamic. Static summary generation refers to a storyboard that is comprised of related keyframes. These relevant keyframes which are part of the original video sequence are then selected to produce the desired summary of the video. These keyframes are considered highlighted

The associate editor coordinating the review of this manuscript and approving it for publication was Filbert Juwono.

parts of the original input video. Generation of the static summary is divided into four main sections: selecting constitute frames, feature extraction, grouping frames, and skimming of keyframes [1]. Dynamic summary generation refers to video skims that consist of the most relevant parts or events of input video which are termed as segments. These video segments have audio as well as video content. The generation of a dynamic video summary has three main sections: Segmentation of video, score prediction, and selection of segments [4]. While studying and surveying Video summarization some nomenclatures are used which the user needs to keep note of and understand. Table 1 lists the nomenclature that was used in this review work.

TABLE 1. List of nomenclature used in this survey paper.

Nomenclature	Referred to
BIRCH	Balanced Iterative Reducing and Clustering
CCTV	Closed Circuit Television
DB-LSTM	Densely Connected Bidirectional Long Short-Term Memory
ET	Execution Time
FN	False Negative
FOV	Field of Views
FP	False Positive
FS	Feature Selection
GTEA	Georgia Tech Egocentric Activities
KF	Key Frame
LSTM	Long Short-Term Memory
MVS	Multi View Summarization
OMP	Orthogonal Matching Pursuit
SAD	Sum of Absolute Differences
SBOMP	Simultaneous Block Orthogonal Matching Pursuit
SVS	Single View Summarization
TDmap	Temporal Difference Map
TN	True Negative
TP	True Positive
TSCS	Three Step Cross Searching
VS	Video Summarization
YOLO V3	You Only Look Once Version 3

A. MOTIVATION

There is currently no comprehensive evaluation of the literature on single and multi-view video summarization that concentrates on methodology, datasets, application domains, comparative analysis, and future approaches.

- In Single and Multi-view video summarization, current reviews and surveys lack a detailed analysis of application domains, assessment measures, and datasets.
- Traditional methods of low-level feature-dependent grouping or categorization must be replaced with features learned via deep learning models in order to achieve improved efficiency.
- Highlighting current approaches, datasets, applications, issues, and potential future directions in Single and Multiview video summarization is the aim of this literature review.

B. REVIEW STRATEGY

The papers selected for the survey of various video summarization techniques were published from 2005 to 2022. While surveying type of publication, Journals, Conferences were considered. Below figure no. 1 shows result of Co-occurrence analysis on keyword used for re- viewing the papers using VOS Viewer [75]. Keywords occurring more than 20 times are considered for the analysis.

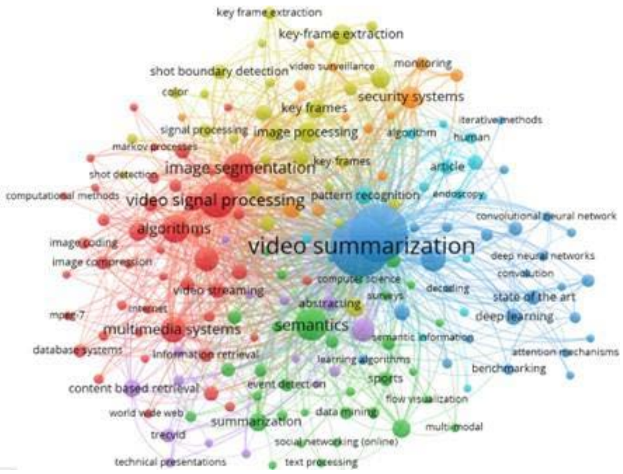

FIGURE 1. Keyword Co-occurrence analysis.

• Inclusive Criteria:

Paper titles, abstracts, contributions, datasets, and evaluation metrics were studied and different algorithms and their performance were discussed for the effective generation of video summaries. A keyword (Video Summarization) search was conducted on papers published in IEEE Xplore, Scopus, SpringerLink, Web of Science, and Science Direct till 2022, and the number of survey papers referred results is shown in figures below 2 and 3. The number of article that come after keyword search is shown in figure 4.

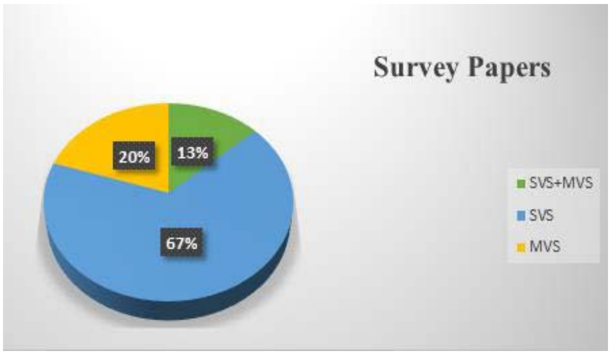

FIGURE 2. Survey Papers referred.

• Exclusive Criteria:

Some papers were duplicated since they were available in more than one database. Many papers were also deemed to be irrelevant after screening titles and abstracts. Finally, a total of 63 papers were selected for systematic review.

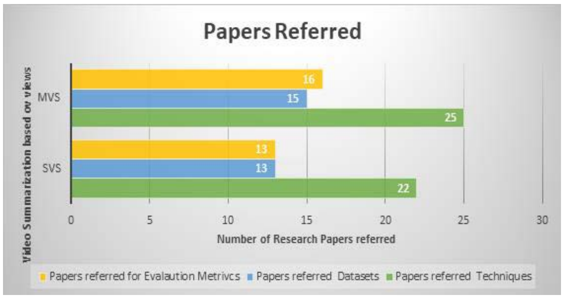

FIGURE 3. Papers referred for MVS & SVS.

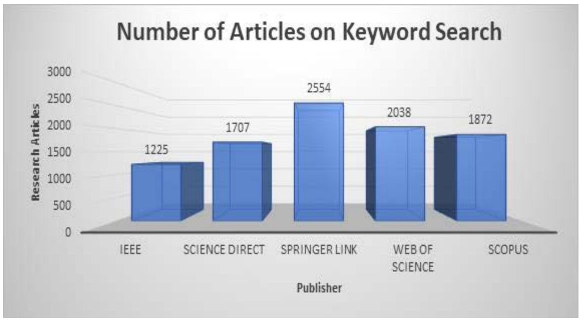

FIGURE 4. Number of articles after keyword search.

C. TERMS AND TERMINOLOGIES

While studying video summarization some concepts are common for both Single view and multi view. These commonly used terms and terminologies are listed below for reference.

1) Video skim- Significant components extracted from the video to generate a video summary.
2) Storyboard - A storyboard is a visual representation of your video's shot-by-shot progression.
3) Ground-truth- The true nature of the problem that is the focus of a machine learning model, as reflected by the relevant data sets connected with the use case in question, is referred to as ground truth.
4) Keyframes- Keyframes are significant frames that reveal the beginning and end of an action. A keyframe informs you with the status of frame at the certain time and also the time at which action has taken place.
5) Video stream- Video streaming is the ongoing transmission of clips from a source to a viewer. Through video streaming, users can view videos online without downloading them like TV serials & YouTube Videos.
6) Features- The components or patterns of an object in a picture that help to identify it are called features.

D. CONTRIBUTION OF WORK

1. This paper reviews existing literature on video summarization, focusing on approaches, datasets, applications, problems, and evaluation measures.

2. In this paper Video summary creation approaches and procedures employed, as well as how they are renovating the research are also discussed.

3. Role of Video summarization in several application domains is also examined in the paper.

4.A list of publicly available datasets is offered to improve domain research in the field of Video summarization.

5. This survey covers trends, application-specific methods, and the taxonomy of single and multi-view summarizing methodologies.

E. ORGANIZATION OF PAPER

In this paper Video summarization, its applications, and classification are explained in detail Different single-view and multi-view summarization techniques are discussed along with datasets and evaluation metrics used.

This paper describes what is video summarization, the necessity of video Summarization and categorization of video summarization techniques. Different algorithms are developed for summarizing videos based on applications, input query, type of video, number of views, etc. Out of these techniques, this paper mainly focuses on summaries generated by single camera and multiple cameras.

The general framework, techniques developed, and datasets used for single view and multiple view Video summarization are discussed in detail. After generating summary, it is important to evaluate the performance of algorithms so different evaluation metrics are discussed and explained which helps the researcher to the select appropriate metric for evaluating the performance of the technique.

F. VIDEO SUMMARIZATION

According to the current scenario sharing, downloading, and uploading of visual data and multimedia, such as images, audio, videos, and movies, is now greatly enhanced. Every person has the ability to produce video content via YouTube, internet, public video sources, and other platforms. In order to organize the growing amount of content on internet and extract useful information from them, much emphasis is paid to video summarization (VS). Summarization of input video is the process of generating a brief video excerpt from the original video that contains some of the source elements and represents the sequence of the original clip. It may be accomplished by selecting the most crucial information from a larger video and presenting it in a concise manner.

Nowadays processing a huge amount of data can be time-consuming, in this case, video summarization can be referred to make these processes efficient in terms of the time factor. The name video summarization itself suggests that the input videos are converted to produce the desired summary that can minimize the consumption of time and make the browsing and retrieving process easier for the user. Many times, a big unedited video is comprised of redundant events which are not of users' interest then discarding the irrelevant frame and only considering the highlighted parts which are of users' interest can be termed as a video summary. The highlighted parts from the input video are referred as events [4], [7]. Referring only to these events we can combine them to form a video summary as shown in figure 5.

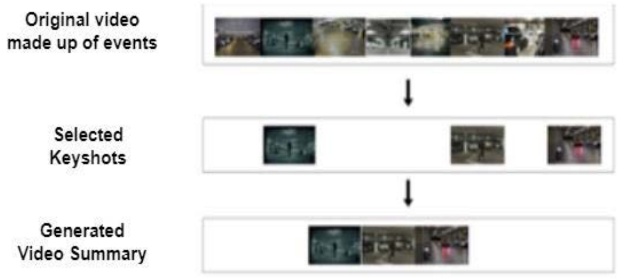

FIGURE 5. Generation of video summary.

Video summarization can be used in different domains as ease of recording and uploading large amount of visual data in day-to-day life. The Surveillance domain generates lengthy videos with a lot of events that are not useful. So, in this case, video summarization can be of great help if a lengthy video captured by CCTV is summarized to focus on the events such as finding suspicious activity, fraud detection, or identification of a person from the input video. This condensed video is storage efficient and easy to share [13].

G. NECESSITY AND APPLICATIONS VIDEO SUMMARIZATION

As we have discussed, earlier large number of videos are generated every day from different domains like surveillance, News, Documentaries, Sports, Education, etc. as shown in figure no. 6. This arises the need for efficient storage, quick sharing, easy retrieval and reduced time consumption technique.

In order to produce a shortened version of the original videos, video summarizing can therefore be used in practically all of the aforementioned fields. [13], [14]. VS can be used in sports to generate a summary of highlights from the match which consist of boundaries, catch, goals, etc. The highlighted version of the game, which is merely a recap of the game, can be watched if a user has limited time and cannot view the entire match. Similarly, original news/documentaries can be of greater length but if a user is interested only in the breaking news, then the summary of that news or documentary can be of great help [6], [7].

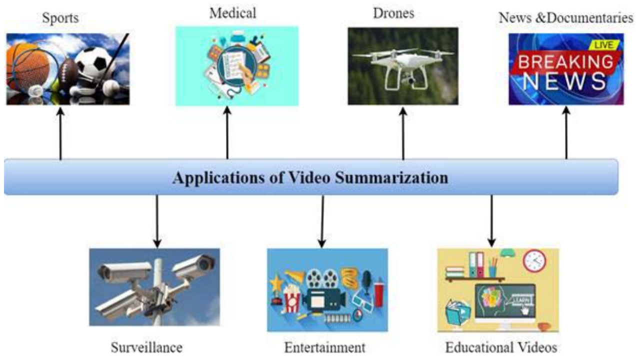

FIGURE 6. Application of video summarization.

Medical field is also considered as a domain in which lengthy videos are generated and their storage can be one the major issues to be solved. Videos of different surgeries like endoscopy, hysteroscopy, angioplasty, bypass surgery, tooth extraction, etc. are of large size. We can use video summarization, which reduces the original movie to a video storyboard, to facilitate effective analysis of these medical videos. This will help medical practitioners to search for similar set of problem solution and implement procedure for current case [7], [8]. The majority of videos produced in the sports industry involve human involvement in the form of editing tools that can take a lot of time to create a summary video [14]. Video summarization can be useful to track ball as well to track the player for evaluating the result. Also, the matches held are recorded and then are referred while training and deciding the strategies based on previous matches. So, for sports videos we can use video summarization to make it efficient for referring and evaluating. In sports video summarization detects the moments or events of interest based on the defined sports knowledge [6].

Video summarization has huge influence on content-based videos. It also plays major role in indexing, retrieval, and recommendation systems [16]. Video Summarization reduces the size by discarding the irrelevant frame that make video search engines more browsing efficient. Because of this user can get quick access to the videos of interest and content creator gets more views for that video [14]. As due to Covid-19 online teaching has become mandatory and the e-learning resources shared are also in large amount in recent 2 years. So, internet contains large videos having videos containing miscellaneous data. The videos generated in this field, may be recorded from phones that can contain content available on paper or whiteboard. These types of videos have content that is loosely structured and with lot of pause in between and external noise. So, it is important to make video more user specific that highlights the important part or section from video by removing the irrelevant frames from the video. But the summarization of such recorded videos has some challenges like less illumination, background noise, occlusion, etc. which makes it a research topic [13].

CCTV is another area where a lot of video data is produced every day. If a certain parking lot has CCTV camera installed and robbery takes place on some day then instead of wasting time to watch whole video, summarized video containing important events can be refereed to watch the suspicious activity. Nowadays large number of cameras are installed in private as well as public places and the data generated from these cameras per week is approximately of 10 GB. Lot of research is done in summarizing the videos taken from CCTV cameras still in some cases human involvement is necessary to go through the recorded video. But when human intervention is needed for summarizing videos then it requires lot of time and sometimes there is chance that the generated solution is prone to error.

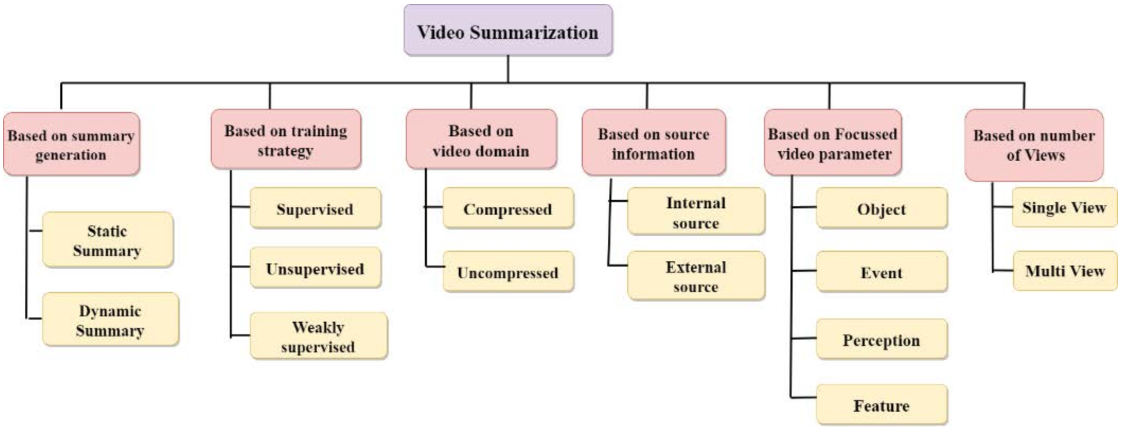

FIGURE 7. Categorization of video summarization.

H. CATEGORIZATION OF VIDEO SUMMARIZATION TECHNIQUES

Several algorithms have been developed with the primary goal of pinpointing video content and generating a video summary. Based on their qualities and characteristics, these techniques are divided into six broad types. Figure 7 depicts the classification of various strategies.

- Based on summary generation

A static storyboard summary that is made up of keyframes and a dynamic video skim that is composed of condensed video segments are the types of video summaries based on generation. Due to their intrinsic characteristics, static summaries can only contain the audio track and potentially some keywords, whereas dynamic summaries can include all three forms of data [68].

1) Static Summary Generation

A static summary is a grouping of the pertinent essential frames that are necessary to produce the desired summary and are chosen in a sequential manner. A video is made up of a number of frames, which are static images that are presented in a specific order and speed to appear to be moving. Key frames are the significant frames of a video's content that contain all the data and are displayed in a chronological order [16]. The general framework of static summary is shown below in figure 8.

A video is made up of several frames, which are still images that are shown in a specific order and at a specific pace to provide the impression of motion. The video clip is first divided into its individual frames, and then, using a feature extraction technique, visual features are retrieved from the frames to provide a static summary. This is a key phase in the process and is based on the kind of video information being described and the context in which it will be used. The features that need to be extracted can primarily be divided into learned and created features. Features that are manually retrieved using any image processing method are referred to as handcrafted features. On the other hand, learnt features are those that a machine learning-based model learns from a large number of video frames and outputs as visual features after substantial training. After that, the relevant frames are grouped using a ML method, either supervised or unsupervised based on the accessibility of data with annotations, and irrelevant frames are discarded. A static summary is

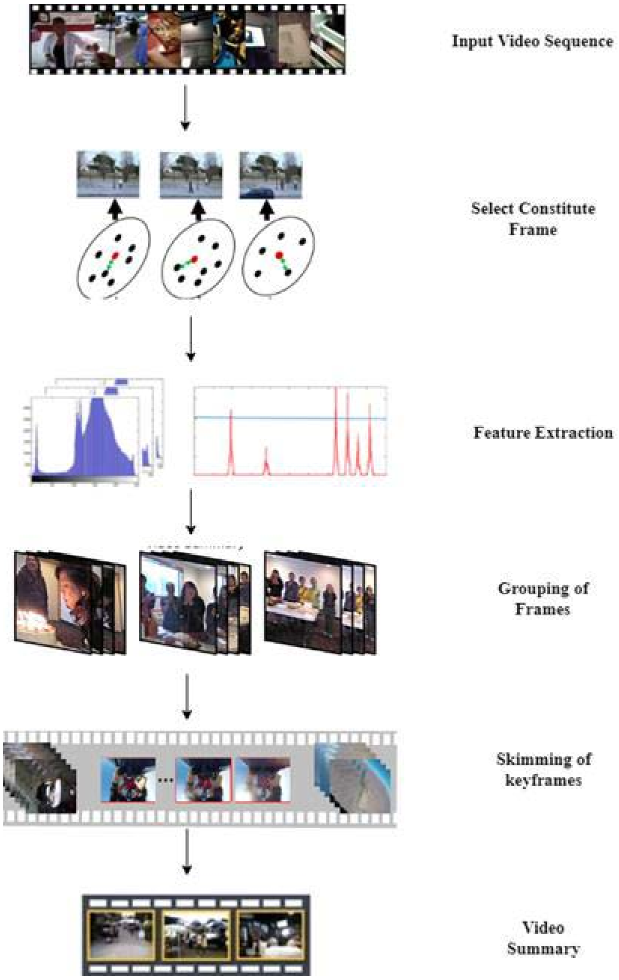

FIGURE 8. Generation of static summary.

created by skimming or extracting the key frames from these groups.

2) Dynamic Summary Generation

A dynamic summary includes key scenes or chunks of a video. Skims comprises of video segments and the related audio, and dynamic video synopses are produced by analyzing visual and aural material of the video stream. in order to create a dynamic synopsis from the current input clip three stages make up the procedure: First by segmenting video followed by predicting the importance score and the selecting the appropriate segments as shown below in figure No. 9

The video is separated into the smallest pieces that can be understood and processed individually in the first phase of video segmentation. These units, also known as skim pieces or skim segments, have a number of frames that adequately describe a certain action [16]. By separating out distinct images and scenes from the video stream, segmentation makes it easier to keep these segments temporally connected.

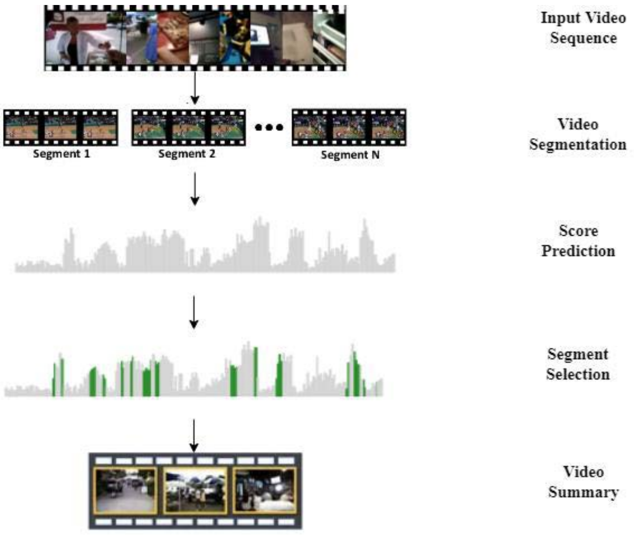

FIGURE 9. Generation of dynamic summary.

Due to the difficulty in identifying shot and scene boundaries in unstructured movies, different segmentation techniques are used for structured and unstructured videos. The important skim units are determined in the second step based on the value calculated. Either all of the frames in a skim unit or the important frames that were taken out of the skim unit can be used to determine the importance score. The second method is more time and money efficient than the first. The calculation of relevance score is an important stage because it tries to draw attention to the key components of a movie. It is equally difficult because the importance criterion varies depending on the application domain, user preferences, requirements, etc. for each video. Certain important criteria, such as aesthetic attraction, compactness, and diversity, have been assigned to some earlier works. To determine the relevance score for segment selection, video diversity data from a video's visual and categorical data is employed [4]. To shorten the overall duration of the video, elimination of unnecessary frames is done in the third phase. Based on the importance scores determined in the earlier step, key segments are determined which makes up the video synopsis [7], [13].

- Based on training strategy

Approaches under supervision that seek to identify the fundamental criteria for choosing video frames and segments and summarizing videos based on datasets with human labelled ground truth. Unsupervised techniques that reduce with the necessity for ground truth data (whose development requires time-consuming and difficult manual annotation procedures), based on learning mechanisms that simply need an acceptable number of original films for training. Weakly-supervised approaches, like unsupervised approaches, seek to reduce the requirement for substantial quantities of hand-labeled data. Although weak labels are less expensive, they are nevertheless capable of producing powerful predictive models despite being inferior to a full set of human annotations [15].

- Based on Domain type

Based on domain video summarization can be categorized as compressed and uncompressed video summarization technique. Uncompressed summarization is also called as pixel video summarization technique. In uncompressed video summarization pixels are taken into consideration for feature extraction and after extraction is used for summary generation. Some videos summarization requires lot of time and space, for such cases compressed summarization can be of great help. This compressed video summarization involves compression of video followed by extraction of features [7].

- Based on the Source of Information

There are three categories of video summarizing approaches and can be categorized based on the information source utilized to summarize videos as mentioned below [4]:

(1) Internal techniques

These techniques make use of data that is found in the video stream and intrinsic part of the video

(2) External techniques

These makes use of data that isn't an intrinsic component of the input clip but is present in its metadata.

(3) Hybrid techniques

These utilize both intrinsic and extrinsic data to solve the summary generation issue.

- Based on focused parameter

1. Object

Whenever the focus while generating a summary is on object like characters, ball, swing, etc. then such summaries are referred as object-based summary. As the focus is on object it may sometimes neglect the text or graphics to focus on objects within the summary. [38], [41], [43]

2. Event

Whenever the focus while generation the summary is on event like boundaries, six, goals, kicks, accidents, etc. then such summaries are referred as event-based summary. Unlike object-based summary event-based summary focus on text and graphics [39], [43], [45].

3. Perception

Perception based summary considers important content of video which is decided by user perspective. It can also be based on the emotions that user may have to the attention user may have towards the content of the input video [18], [40], [49].

4. Feature

Feature based summaries are based on parameter like motion, color or texture in the keyframes of video streams. These attributes are referred as low-level features and are utilized for creating summary in their unprocessed form [47], [50].

- Based on the number of Views

The amount of video data collected by security cameras and regularly recorded by smartphones has greatly increased, making video summarization a popular study area. Single-view video summarization and multi-view video summarization are the two main categories. Single View summarization is the process of constructing a video synopsis with the goal of preserving three properties: low repetition, representativeness, and diversity [11], [13]. The majority of summarizing algorithms are offered for SVS because main goal is to provide a synopsis that is identical to original video while only considering intra-view interrelations into account. Since there is no synchronization and no consistent illumination across all views, SVS is simpler to use than MVS. A problem with video summarizing that is rarely explored is how to summarize videos that have many viewpoints. MVS either creates a collection of illustrative frames (keyframes), a condensed version of a video (video summary), or video skims, like SVS. The primary steps involved are processing the input video followed by extracting the key features then post processing and at the end summary production [14]. The detail description of video summarization based on number of views is explained below with the help of algorithms, datasets and evaluation metrics used.

II. SINGLE VIEW SUMMARIZATION (SVS)

There are many videos that have been produced and disseminated through different media in the digital world of today. Since these video clips are frequently uploaded to the internet or the cloud, viewing them calls for a network with a high bandwidth. Video summaries are becoming more and more popular since they accurately and succinctly offer the best representative description of the original video material. This is a feasible approach to save time, space, and other network and multimedia infrastructure resources. The applications, general framework and different technique used for video summarization are explained below using figure 10.

A. GENERAL FRAMEWORK OF SINGLE VIEW SUMMARIZATION (SVS)

The basic pipeline of single view summarization is shown above in figure 11.

The input video is obtained via CCTV and then bifurcated into frames, with a greater number of frames being acquired. This part reduces the number of unnecessary frames by using the Three Step Cross Searching Algorithm techniques, also known as the motion estimation algorithm. Due of its effectiveness and simplicity, the three-step algorithm has become a very common search mechanism. It looks for a better motion vector to make the search pattern more exact and the algorithm, however, is based on SAD. Following pre-sampling, the ID value is assigned to each frame. The primary rationale for providing an ID number is to sort the frames in an orderly fashion, resulting in an efficient summarization and reduced execution time. At this stage, the frames with allocated Unique ID numbers are processed to extract the Haralick, contrast, edge feature, and blocks correlation. Following the feature selection procedure, the N-K means clustering technique is used to group the selected features. According to preliminary cluster centroids, similar data are also divided into clusters using the efficient clustering method known as normal K means, and these values are directly grouped. The rapid sort method, which uses a Conquer and Divide algorithm, summarizes the movie depending on the id value. It divides the specified array across an element that serves as a pivot. Any element in the array, including the first, last, or any other arbitrary member, could serve as the pivot element [24].

B. SINGLE VIEW SUMMARIZATION TECHNIQUES

Zhong Ji et al. offers an encoder self-attention mechanism that complements the attention by assigning weights to encoder outputs based on their relevance in the short-term environment. This satisfies both immediate and long-term context - specific awareness, which is a necessary condition for efficient video summary. It proposes a distribution-based loss function to get over the problem of using MSE alone. It learns the distribution consistency from the series of estimated score and actual data, based on the assumption of least distance between them, ensuring better use of human annotations [18]. The Detect-to-Summarize network [47] approach presents video summarizing as forecasts importance ratings and fragment placements simultaneously while solving a temporal interest identification challenge.

The anchor-based technique provides temporal proposals to address length fluctuations of interest, and the anchor-free approach learns importance score and temporal location directly. A TTH-RNN [50] can avoid the huge feature-to-hidden mapping matrices created by high-dimensional video features by adding the tensor-train embedding layer. The bulk of training parameters are removed, making training easier. A TTH-RNN may capture long-range temporal dependencies between frames by creating the hierarchical structure of an RNN. As a result, the maximum length of sequence that an RNN can handle is increased, while the nonlinear fitting ability is improved by modelling the frame sequence hierarchically A TTH-RNN hierarchically examines intrasubshot and intrasubshot temporal dependencies. Because video data is layered as frames and subshots, it adheres to the temporal structure of video data. Different techniques, evaluation

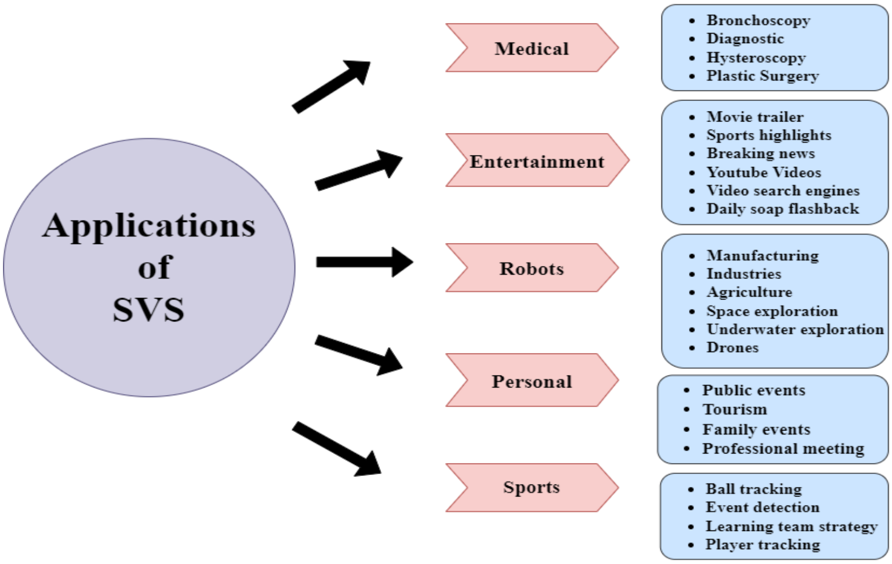

FIGURE 10. Applications of SVS.

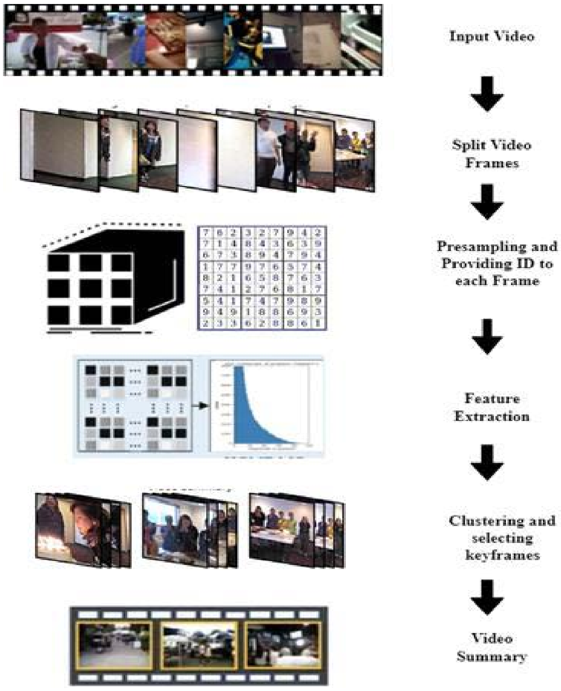
Input Video
Split Video Frames
Presampling and Providing ID to each Frame
Feature Extraction
Clustering and selecting keyframes
Video Summary

FIGURE 11. General framework of SVS.

metrics used for evaluating performance and datasets used are surveyed in Table No. 2

Meta learning is used in the MetaL-TDVS [55] approach to perform video summarization. To the best of the research, meta learning has never been used to the field of video summarization. Focusing on the challenge of VS rather than only temporal or architectural video data, MetaL-TDVS will specifically look at the underlying strategy of video summarization. MetaLTDVS' usefulness is demonstrated using both substantial and conventional aspects. To produce better video summaries, Optimum-Path Forest includes temporal information alongside OPF throughout the clustering computation procedure [57]. In a special graph-based architectural variance model analysis to define the VS problem, the structural information in the features for every video frame is examined and presented in graphs to fill the differences between the existing semantic structural features and the raw characteristics of video frames.

We create a graph-based unit of measurement to assess the difference between frames. It is possible to employ median plots as keyframes to reflect the general trend of the video, and this difference in graphical form might draw attention to any discrepancies between consecutive frames [52]. The block sparsity property of applicant keyframes produced from the regional similarity among neighboring frames are taken into consideration, and the VS problem is defined as a block sparse dictionary selection-based model in order to directly and effectively address the concerns of redundancy and outliers. The OMP algorithm's simultaneous block variation, known as SBOMP, is designed to retrieve keyframes by first choosing keyframe units, and then extracting keyframes from the keyframe blocks generated using two techniques dependent on block centroid and accurate representation [51].

The figure 12 below displays the most recent developments in Single view summarization, giving us a quick overview of algorithms used.

C. DATASETS

Appropriate datasets are needed for the training, testing, and comparative analysis of different video summarizing algorithms. When it comes to summarizing, evaluating, and analyzing data, there are some application fields where the datasets are still in the growth stage. Tables No. 3 and 4 below

TABLE 2. Comprehensive information of research Single View Summarization techniques.

Sr. No	Methodology /Algorithm	Problem Statement	Proposed Solution	Evaluation Metric					Limitation\Future Scope
				F-Score	Precision	Recall	Canonical	Augmented	
1.	ADSum [18]	short-term contextual attention insufficiency and distribution inconsistency	encoder self-attention mechanism for Seq2Seq learning-based video summarization	64.3	-	-	64.3	65.7	Deficient in modelling very long-term contextual attention
2	DSNet [47]	regression problem without temporal consistency and integrity constraints	generates temporal interest proposals to handle length variations of interest & learns importance scores and temporal locations	62.1	83.14	91.63	-	-	Interest proposals are required for generating importance score and segment boundaries
3	FCN-LectureNet [49]	extraction and summarization of the handwritten content	produce a spatial-temporal index of handwritten content	87.12	83.4	91.63	-	-	Improvement in temporal segmentation and content-based indexing
4	TTH-RNN [50]	large feature-to-hidden matrices & deficiency in long-range temporal dependence exploration.	tensor-train embedding layer to avert the large feature-to-hidden matrices, together with a hierarchical structure of an RNN	46.6	40.5	53.2	-	-	Modify a TTH-RNN to a general architecture and provide more insights
5	SBOMP [51]	redundancy among selected keyframes and poor robustness to outlier frames.	candidate keyframes derived from the local correlation among adjacent frames	48.54	41.41	64.49	-	-	No guarantee of uniform partitioning
6	Graph-based structural difference analysis [52]	detection of shot transitions like hard cuts, dissolves, wipes, and fade-ins/fade-outs	graphs to bridge the gap between the actual semantic structural information and the raw features	67.5	54.6	88.4	-	-	improve the effectiveness by introducing more robust features.
7	DASP [53]	inherent relations between the original video and its summary,	encoder-decoder attention and semantic preserving loss in a deep Seq2Seq framework	63.6	-	-	63.6	64.5	Computational time can be reduced
8	Adv-Ptr-Der-SUM [54]	shortened and informative WCE	generative adversarial framework, consisting of a summarizer and a discriminator	58.3	-	-	-	-	Performance can be improved
9	MetaL-TDVS [55]	focus more on video summarization problem itself instead of only on sequential or structural video data.	reformulating video summarization as a meta learning problem and promote generalization ability of the trained model	58.2	-	-	-	-	Foreground superior meta learners to explore the mechanism for summarizing video extraction
10	OPFSumm [57]	compact representation for a better storage and retrieval purposes	temporal information to obtain better video representations.	72.8	-	-	-	-	Evaluate OPFSumm under features obtained through deep learning techniques, &study other distance functions

list the datasets used in various techniques along with their specifics. The most popular datasets, which we can refer to as benchmark datasets, are TVSum and SumMe, as seen in the table. SumMe consists of 25 videos of maximum 6 min, in 25 different themes, including sports, events, and holidays. In the form of several sets of important pieces, the annotations are collected from 15 to 18 people for each movie. Each summary is between 5 and 15 percent of the length of the original video. TVSum includes fifty videos of 10 categories, with 5 videos included in each. These categories include news, user-generated content, and documentaries. Each movie contains 20 users who have annotated it with shot- and frame-level relevance values. The videos range in length from 2 to 11 minutes (ranging from 1 to 5).

- NII TV-RECS (NII TV Broadcast Video Research Corpus) [73]

The NII TV-RECS is an experimental prototype of a research-purpose broadcast video corpus being developed by a group of NII researchers with the goal of creating a useful resource for video processing research. A corpus, in general, is a large collection of study materials. Text corpora (e.g., language corpus, online corpus, news corpus, etc.) are useful for natural-language processing research, and video corpus are projected to be useful for video processing research.

- Cross Task [70]

The Cross Task dataset is made up of instructional films for 83 different tasks. It provides an orderly set of stages with manual descriptions for each task. There are two pieces to the dataset: 18 major tasks and 65 associated tasks. Manually captured videos for the core tasks are provided with annotations for temporal step boundaries. The videos for the relevant tasks are automatically collected and do not include annotations.

- MED [76]

MED is a novel evaluation dataset that covers a wide spectrum of monotonicity reasoning and was compiled from linguistics publications via crowdsourcing. The collection was created by crowdsourcing spontaneously occurring cases and well-designed examples from linguistics journals. There are 5,382 cases in total.

- VSUMM [67]

This dataset contains 50 Open videos and each video is 352 × 240 pixels in size, in MPEG-1 format, with sound and color. The lengths of these films range from one to four minutes, with a total runtime of about 75 minutes. They are split into numerous genres (documentary, educational, ephemeral, historical, lecture). Along with this, 250 user summaries are also available in this dataset. These summaries were manually generated by 50 individuals, each of whom dealt with five videos, resulting in five video summaries created by five separate users for each film.

- GTEA-gaze+ [72]

Seven different daily tasks, such as preparing burgers, beverages, are included in the Georgia Tech Egocentric Activities dataset. There are a total of 28 videos, with four different people performing each action. There are roughly 20 minute-long, fine-grained action instances in each video.

To help the user select the most effective method, various tactics have been discussed and contrasted. Some feature-based techniques perform badly in long films for object detection and demand expensive system specifications. When compared to other methods, clustering-based methods summarized the movie more precisely. A tracking technique is also suitable for various environments, such CCTV Videos with a stationary camera. The most effective supervised methods to date for summarizing have learned the importance of frames by simulating the varying range temporal dependency of video segments using RNN and specific attention techniques [6]. Due to increase in security demand need for multiple view videos has gained more interest and need which single view videos fail to do. Below figure 13 shows some of the advantages and disadvantages of single view summarization from which we can conclude that is why multi view summarization has gain attention recently.

III. MULTI VIEW SUMMARIZATION

Since a lot of data is collected from surveillance cameras every day in daily life, video summarization is a very recent study area. Surveillance video can be categorized into single view and multi view. Video taken from single camera is called as Single View whereas video captured by multiple cameras of same location is called as Multi View. The summarization of multiple view videos is more difficult to address due to factors such as different angles, unaligned videos, different light intensity, and synchronization scarcity [30]. Lot of work is done on single view summarization as compared to multi view summarization. Recently for security and tracking purpose CCTV with multiple views are used which makes MVS an crucial topic in today's world. The video summary of videos taken from multiple cameras of same or different locations can come under multi view summarization. Different areas where Multiview summarization is gaining interest is listed below in figure 14.

A. IMPORTANCE OF CCTV CAMERAS

Closed Circuit Television (CCTV) is used to monitor interior and exterior area of any property. The recorded video of the specific location is made available on limited number of monitors. Currently in industrial sector CCTV plays very important role for maintaining the safety and privacy of the company which has increased the use of CCTV to huge amount. Nowadays, CCTV cameras are installed near every public as well as private places like grocery store, private companies, schools, malls, offices, parking lot, private lobby, and many other places [58]. Whenever there is any suspicious activity, or any crime reported then police officers and crime investigators will take help of these installed CCTV cameras to solve the case. In many private companies CCTV are installed to monitor employee activities, visitors and various other activities that can provide useful information to the concerned authorities. According to the studies due to installation of CCTV cameras in private and public premises has reduced the rate of crime by almost 51%. Whenever any illegal activity has taken place then crime investigators take help of recorded videos to gather crime evidence which can make tricky case easier. In many workplaces CCTV is used to provide employee a safer environment preventing sexual harassment as CCTV gives 24∗7 real time data [59].

According to a survey, China is the country with large number of CCTV cameras. The nation has been featured

TABLE 3. Datasets used for different Single View Summarization techniques.

Sr. No .	Single View Summarization techniques	Algorithm	Dataset
1.	FCN-Lecture Net: Extractive Summarization of Whiteboard and Chalkboard Lecture Videos [49] 2021	FCN-Lecture Net	Access Math dataset
2.	Deep Attentive Video Summarization with Distribution Consistency Learning [18] 2021	ADSum	SumMe, TVSum
3.	Recurrent generative adversarial networks for unsupervised WCE video summarization [54] 2021	Adv-Ptr-Der-SUM	SumMe, TVSum
4.	DSNet: A Flexible Detect-to-Summarize Network for Video Summarization 2021[47]	DSNet	SumMe, TVSum
5.	TTH-RNN: Tensor-Train Hierarchical Recurrent Neural Network for Video Summarization [50] 2021	TTH-RNN	SumMe, TVSum, MED, and VTW
6.	Graph-based structural difference analysis for video summarization [52] 2021	Graph-based structural difference analysis model	Open video project dataset, VSUMM, YouTube
7.	Meta Learning for Task-Driven Video Summarization [55] 2020	MetaL-TDVS	SumMe, TVSum
8.	Dynamic graph convolutional network for multi-video summarization [56] 2020	dynamic graph deep learning model	Tour 20, TVSum
9.	A Novel Key-frames Selection Framework for Comprehensive Video Summarization [48] 2019	Caps Net to extract spatiotemporal features and generate inter-frames motion curve.	VSumm, SumMe, TVSum
10.	Video summarization via block sparse dictionary selection [51] 2019	SBOMP	VSumm, TVSum
11.	Deep Attentive and Semantic Preserving Video Summarization [53] 2019	Encoder-decoder attention and semantic preserving loss in a deep Seq2Seq framework	SumMe, TVSum, YouTube
12.	OPFSumm: on the video summarization using Optimum-Path Forest [57] 2018	OPFSumm	SumMe, Open Video Dataset

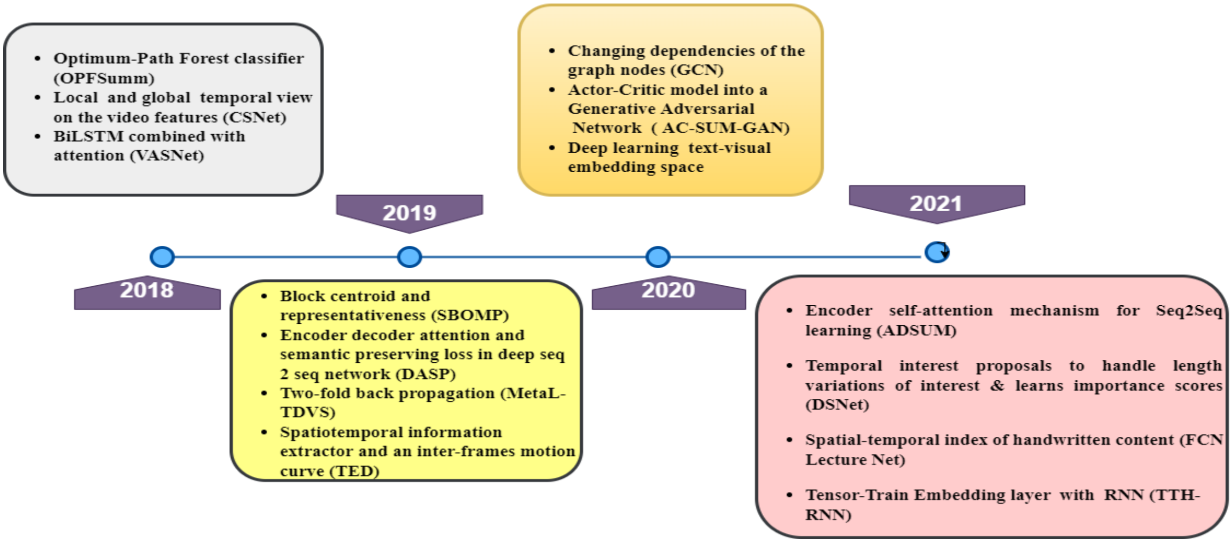
• Optimum-Path Forest classifier (OPFSumm)
• Local and global temporal view on the video features (CSNet)
• BiLSTM combined with attention (VASNet)
• Changing dependencies of the graph nodes (GCN)
• Actor-Critic model into a Generative Adversarial Network (AC-SUM-GAN)
• Deep learning text-visual embedding space
2019
2021
2018
2020
• Block centroid and representativeness (SBOMP)
• Encoder decoder attention and semantic preserving loss in deep seq 2 seq network (DASP)
• Two-fold back propagation (MetaL-TDVS)
• Spatiotemporal information extractor and an inter-frames motion curve (TED)
• Encoder self-attention mechanism for Seq2Seq learning (ADSUM)
• Temporal interest proposals to handle length variations of interest & learns importance scores (DSNet)
• Spatial-temporal index of handwritten content (FCN Lecture Net)
• Tensor-Train Embedding layer with RNN (TTH-RNN)

FIGURE 12. Latest trends in SVS.

prominently across the entire ranking, which examined the 150 most dynamic metropolises worldwide and excluded those for which insufficient data was provided. The nation has been making waves for its liberal use of monitoring technology. However, India is becoming more competitive in the top 20, which is dominated by China, with Indore

TABLE 4. Details of commonly used datasets for video summarization.

Sr. No.	Name of Dataset	Number of Videos	Duration	Domain
1.	SumMe [63] https://gyglim.github.io/me/vsum/index.html	50	1 to 6 min	Holiday, Events, Sports
2.	MED [65] https://github.com/veryplumbing/MED	160	1 to 5 min	15 categories of various fields
3.	TVSum [61] https://github.com/yalesong/tvsum	50	2 to 10 min	Headlines, documentary, and user created videos
4.	YouTube [64] https://research.google.com/youtube8m/	50	1 to 4 min	Cartoon, Sports, TV shows, Commercial, Home

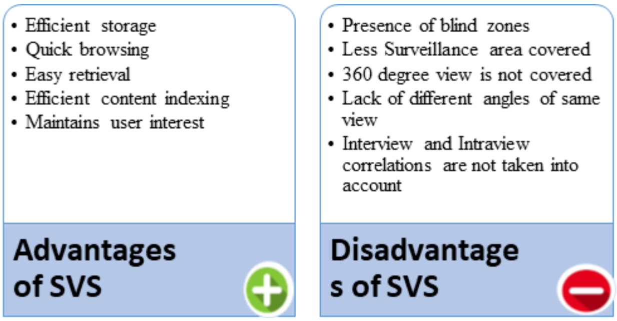
• Efficient storage
• Quick browsing
• Easy retrieval
• Efficient content indexing
• Maintains user interest
Advantages of SVS
• Presence of blind zones
• Less Surveillance area covered
• 360 degree view is not covered
• Lack of different angles of same view
• Interview and Intraview correlations are not taken into account
Disadvantages of SVS

FIGURE 13. Advantages & Disadvantages of SVS.

entering at position 4, Hyderabad moving up from position 16 in 2020 to position 12, and Delhi entering at position 16 in 2021, as illustrated in figures 15, 16, and 17.

Multi view VS and single view VS generation are different from each other in following ways. First, multi-view has architecture, and there are various correlations and FOV(Field of View). For an effective summary, content correlations as well as inconsistencies among distinct movies must be accurately modelled. Second, many unaligned clips were produced because of using various view angles and depths of field to capture the same image. As a result, differences in lighting, position, view angle, and synchronization concerns make summing these movies difficult. As a result, approaches for extracting summary from single-view films rarely generate the best set of representatives when summarizing multi-view videos [25]. Below figure 18 shows Pros and Cons of MVS which gives idea while selecting the techniques suitable for the summarization as per the need.

B. GENERAL FRAMEWORK OF MULTI VIEW SUMMARIZATION (MVS)

The MVS summary challenge seeks to construct a video synopsis or key-frame sequence that highlights the most crucial elements of the input videos in a brief amount of time [25]. It takes a series of input films recorded from several cameras concentrating on about the same FOV from various angles. The general framework of generation summary if multi view video is shown in figure 19.

The multi view videos streams are taken as an input and then pre-processing done on the respective videos. The processing of these inputs includes shot segmentation, shot boundary detections and removal of redundant frames. Pre-processing is followed by several ways for extracting information from the input video. Features extracted can be of two types of hand crafted (SIFT, edge and color histogram) or learned features (C3D and deep features). Post processing is done on the features extracted which includes computation of inter and intra view correlations, Spatio temporal graph generation and histogram computation. Finally, the summary of video is generated using different machine learning or deep learning algorithms [19].

C. MULTI VIEW SUMMARIZATION TECHNIQUES

A summary of the activities in the scenario that is taken at the end of the film depending on its actions and highlights plays important role to make browsing easy. In other words, it's a method for human activity recognition and summarization that begins by sensing human physical movements before identifying them through the use of a background-subtraction-based method that has been proposed. The recommended method allows for the detection and recognition of many human body activities as opposed to present systems, which only record the action of one individual in this case. For this, techniques are suggested: the one which makes use of the TDMap's HOGs' Cosine Similarity value, while other uses TDMap images to classify activities using CNN. [20]. This article describes a novel deep learning approach for multi-modal video summary generation that is built on context - specific word vectors and a specialized attention network. To more adequately assess the underlying interactional data between the input and the query, an interactive attention network built on Convolutional Neural Network is used. Experiments on the current multi-modal video summarizing dataset have thoroughly verified our strategy [21]. The current VS solutions have attempted to generate the VS; however, the technologies have processing time concerns and have trouble adequately compressing the video content for each domain. To overcome these drawbacks, an approach is developed which provides effective VS for surveillance systems based on normalized k-means and a rapid sort approach. The eight steps of this technique include pre-sampling, providing a unique number, feature selection and extraction,

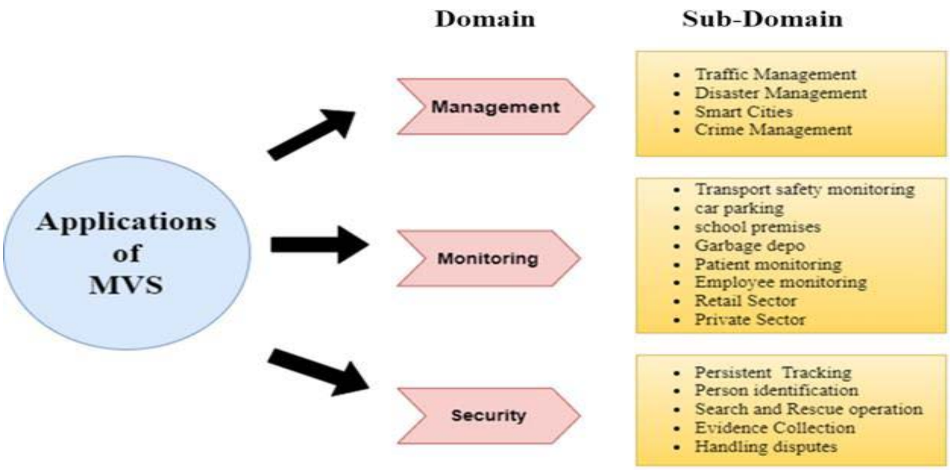

FIGURE 14. Applications of MVS.

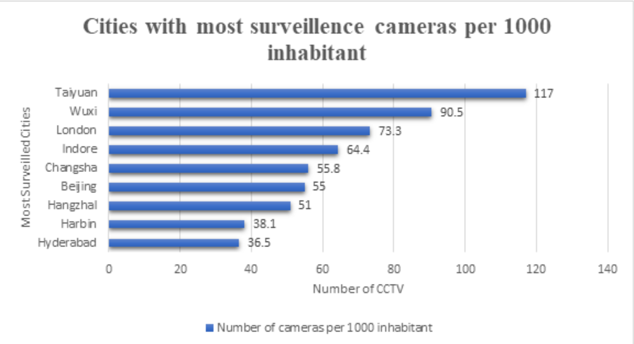

FIGURE 15. Cities with most surveillance cameras.

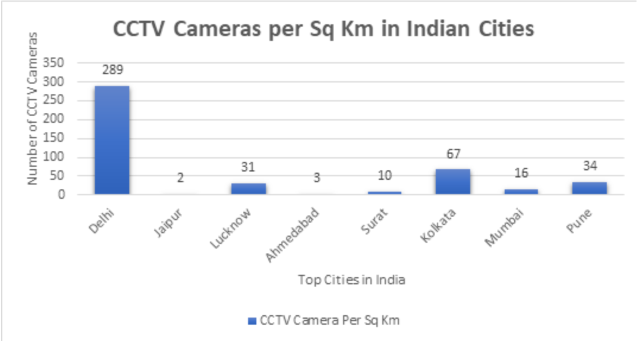

FIGURE 17. Number of CCTV cameras per Sq Km in India among the top cities.

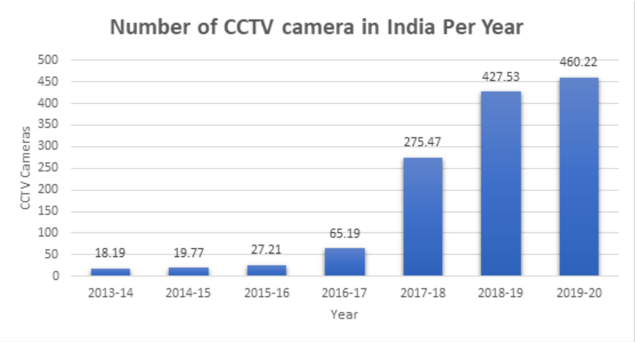

FIGURE 16. Number of CCTV cameras in India from 2013-20.

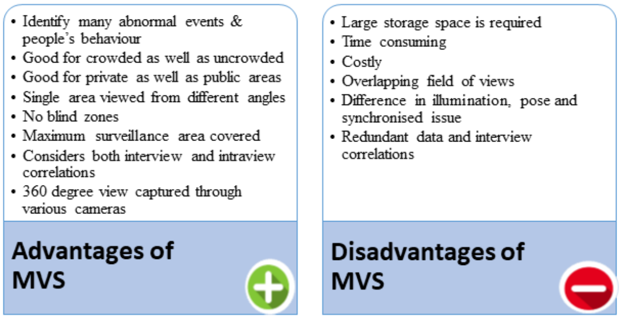
• Identify many abnormal events & people's behaviour
• Good for crowded as well as uncrowded
• Good for private as well as public areas
• Single area viewed from different angles
• No blind zones
• Maximum surveillance area covered
• Considers both interview and intraview correlations
• 360 degree view captured through various cameras
Advantages of MVS
• Large storage space is required
• Time consuming
• Costly
• Overlapping field of views
• Difference in illumination, pose and synchronised issue
• Redundant data and interview correlations
Disadvantages of MVS

FIGURE 18. Advantages and disadvantages of MVS over SVS.

clustering, and video synopsis. This technique yields pronounced falls where an event shift occurs using Jaccard and Dice similarity measures. Because of this strategy, the amount of missed shot cuts is significantly reduced, and the final summary film has practically all the crucial events. This presents a fresh boundary selection method based on sliding windows. This digital technique permits precise event boundary detection. All critical events gathered by either view is included thanks to a synchronized merging and partitioning approach. While merging shot boundaries obtained for each view, the temporal order and a minimum gap are preserved [30].

The YOLO version 3 and Deep SORT tracking method are combined in a deep learning- detection technique. To evaluate YOLOv3's performance, it was first trained on a collection of front view sample images before being put to the test on a batch of numerous people data obtained via an IP camera that was entirely unrelated to the training data

TABLE 5. Comprehensive information of researched Multi View Summarization techniques.

Sr. No	Paper No.	Application /Domain	Algorithm /Methodology/Approach	Problem Statement	Problem Solution	Feature Used	Evaluation Metrics Used				Technique Used
							Fscore	Precision	Recall	other	
1.	DiMS [29] 2017	Tourist video	DiMS	diverse informative summary	minimizes the overall objective function	RGB and HSV color spaces	88.26	100	86	-	CNN
2.	Long term Identity Aware [31] 2017	Person tracking	identity-aware multi-object tracking	computational analysis of surveillance video	sparse label information in a manifold learning framework	Motion Features, histograms	66.3	69.2	63.5	-	-
3.	F-DES [41] 2018	Motion detection, Indoor security & monitoring, Employee monitoring	F-DES	difficult in real-time to manage and access the huge amount of video-content	Interview dependencies among multiple views of video are then captured via the FASTA algorithm	Visual features	89.6	85.4	94.3	-	CNN
4.	Energy Efficient CNN [43] 2018	Smart cities, news, home, entertainment	shot segmentation using deep features	indexing, retrieval and management of surveillance videos	efficient convolutional neural network-based summarization	salient Features	79	-	-	-	CNN
5.	DELTA [44] 2018	Indoor security & monitoring, Employee monitoring	DELTA	difficulty in fast browsing, retrieval, and analysis of surveillance videos	AdaBoosting approach captures the inter-view dependencies	-	89.6	91.3	87.9	-	CNN
6.	Video Clip Growth [45] 2018	Motion detection	Video Clip Growth	customize the length of the video summaries	The average energy of the video clip can then be determined using the energy of each frame.	Spatio-temporal features	89.2	82.12	97.62	-	Energy Function
7.	Cloud Assisted [33] 2019	Indoor security & monitoring, Employee monitoring & industries	CNN and Bi-Directional LSTM	enormous data amount, redundancy, view overlap, light fluctuations, and correlations between viewpoints	deep bi-directional long short-term memory	Visual frame-level deep features	89	94	85	-	CNN
8.	Multiple Action Recognition [20] 2020	Action Recognition	CS-HOG-TDMap & CNN-TDMap	Actual information retrieval	a correlation of the produced and current HOGs	motion, appearance, space-time	-	-	-	Accuracy-98.9	CNN
9.	Deep SORT & YOLOv3 [32] 2020	Person tracking	deep SORT and YOLOv3	Tracking person in surveillance videos	tracking using the Deep SORT technique and YOLO version 3	Visual features	95	85	82	-	Transfer Learning

set. Using a transfer learning technique, the model's detection accuracy is increased. The top view dataset, which is already a part of the trained model, is used for the extra training [32]. Table no. 3 helps to survey different Multi View summarization techniques along with their evaluation metrics and datasets.

By continuously observing changes, video summarization reduces the number of frames that must be stored. With the

TABLE 5. *(Continued.)* Comprehensive information of researched Multi View Summarization techniques.

Sr. No	Paper No.	Application /Domain	Algorithm /Methodology/Approach	Problem Statement	Problem Solution	Feature Used	Evaluation Metrics Used				Technique Used
							Fscore	Precision	Recall	other	
11.	GPT2MVS [21] 2021	Event detection, Indoor security & monitoring	GPT2MVS	video representations regardless of user interest	specialized attention network and contextualized word representations	Visual & Text feature	50.78	-	-	Accuracy-70.37	CNN
12.	Multi Ego [26] 2021	Employee monitoring	Multi-DPP	Summary of same Event taken from different devices	accommodate multi-stream setting while maintaining the temporal	Spatio-temporal features	34.33	34.27	35.03	-	Reinforcement Learning
13.	VQBMVS [28] 2021	Person detection & recognition	VQBMVS	effective and steady methods of information retrieval of videos for analysis	keyframe importance & maximum frame coverage	Spatio-Temporal	88.26	100	79	-	CNN
14	BIRCH clustering [30] 2021	Indoor security & monitoring, Employee monitoring	video partitioning and clustering	extensive surveillance footage with an overlapping FOV	BIRCH clustering algorithm	histograms	-	-	-	C-71.4. (Q)-96	Deep Learning

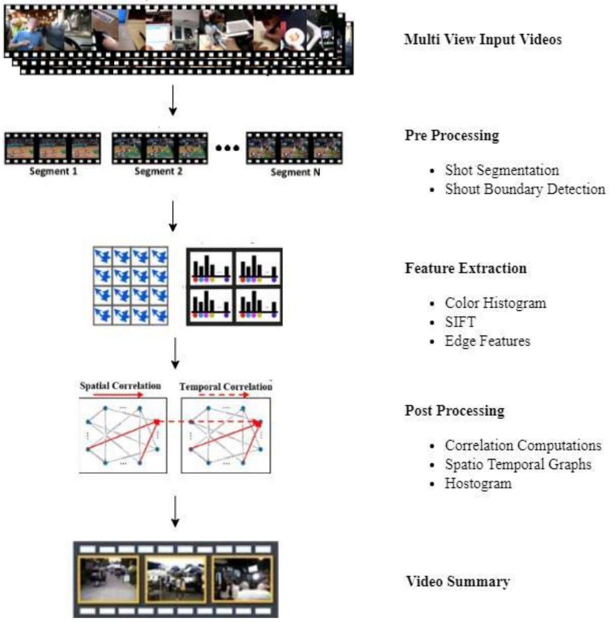

FIGURE 19. General Framework of MVS.

use of a threshold-dependent algorithm, this work enables summarization by selecting the key frames that are most suited for storage and further analysis. All of the frames in a surveillance video are subjected to a global threshold dependent on the Otsus methodology in order to identify the crucial frames, after which each frame is subjected to a retrospective statistical comparison based on the threshold. A similarity index is produced by comparing frames repeatedly using both global and local threshold comparison [22]. Calculating inter-view correlations in this summarization is the most difficult and crucial pre-processing phase, but only with the person holding the camera. Alex Net Convolutional Neural Network along with drop out regularization is used to select key-frames and reduce over fitting problems [27].

A new diversity-aware technique for multi-video summarization is developed after researching the compatibility between the videos. This method generates a multi-movie summary that is both engaging and representational of the entire video library. To effectively address our optimization problem, a method of coping approach is developed that fixes the other videos while ensuring minimal objective function with regard to one video clip at a time. A new benchmark dataset, Tour20, is also introduced [29]. Because a person's look and clothing fluctuate significantly over time, and because they frequently leave and return to a place, long-term monitoring is challenging. We use face recognition data, which is immune to changes in clothing, to immediately initialize tracker after several days of recordings in order to solve these problems. Regrettably, familiar faces are frequently absent. As a result, our tracker spreads identification data to frames without facial features by displaying the look and spatial manifold formed by person detections [31]. Without assuming any existing connection or orientations between the multi-view movies, such as those obtained in an uncalibrated camera network, the data is placed into a contextual integrating to describe the multi-view structure, which preserves both the correlations in this central principle, called "subspace learning," aims to find a concealed subspace that many viewpoints share by making the assumption that these viewpoints are derivations of this short subspace. Over the trained embedding, we applied a sparse representative method of selecting to condense the multi-view recordings. The work of creating summaries is described as a sparse coding problem,

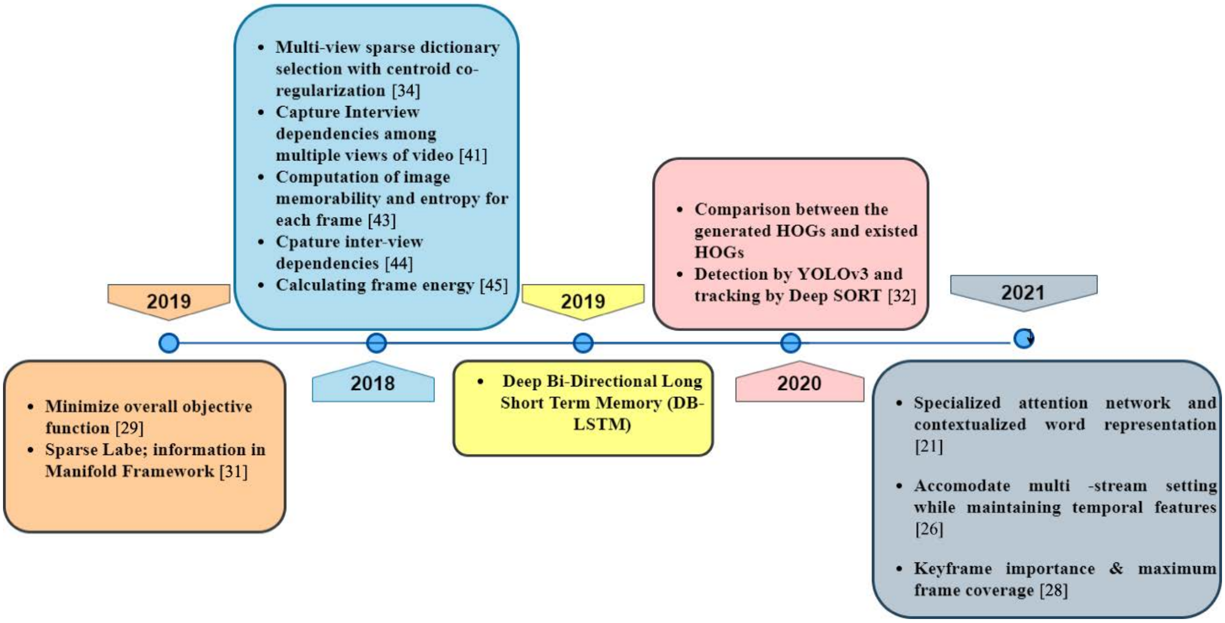
2019
Minimize overall objective function [29]
Sparse Labe; information in Manifold Framework [31]
2018
Multi-view sparse dictionary selection with centroid co-regularization [34]
Capture Interview dependencies among multiple views of video [41]
Computation of image memorability and entropy for each frame [43]
Cpature inter-view dependencies [44]
Calculating frame energy [45]
2019
Deep Bi-Directional Long Short Term Memory (DB-LSTM)
2020
Comparison between the generated HOGs and existed HOGs
Detection by YOLOv3 and tracking by Deep SORT [32]
2021
Specialized attention network and contextualized word representation [21]
Accomodate multi -stream setting while maintaining temporal features [26]
Keyframe importance & maximum frame coverage [28]

FIGURE 20. MVS trends.

where the glossary must have a stable basis (i.e., it must be a matrix of the same data points) and the multi-view synopsis are represented by nonzero rows of the sparse coefficient matrix. Finally, this method jointly train the embedding and optimal representations in order to improve the efficiency of the multi-view embedding and selection technique. By adaptively updating and integrating, it is possible to combine these two objectives into a single optimization issue rather than only using the embedding mechanism for explaining multi-view interrelations and selecting a proper strategy [25].

D. DATASETS

The datasets used for different Multi View Summarization techniques are listed below in Table No.2. According to the survey the most used datasets for MVS are Office, Lobby, VISIOCITY, Campus, TVSum, SumMe, Tour 20, UT Ego and YouTube 8M Segments which are described below table. Commonly used datasets are shown in figure no. 21.

• Office [30]

It consists of TV series that has daily routine of office employees of Fictional Dunder Miffin Paper Company consisting of videos taken by 4 stable cameras. Office datasets consist of 188 rows and 12 columns which has various episodes like Pilot, Diversity, Health Care, Basketball, Olympics, The fire, Halloween, The flight, etc.

• Lobby [28]

Lobby dataset contains videos recorded by 3 cameras installed in certain area. The cameras are in sync with each other, and the cameras are not stable.

• VISIOCITY [77]

VISIOCITY has 67 videos that are of different categories with average duration of 55 mins. Videos in the dataset are from different areas like TV shows, Sports, educational and surveillance. It is comprised of annotations for supervised learning and evaluation and has ground truth summaries. It is used for event localization or action recognition due to its rich annotations.

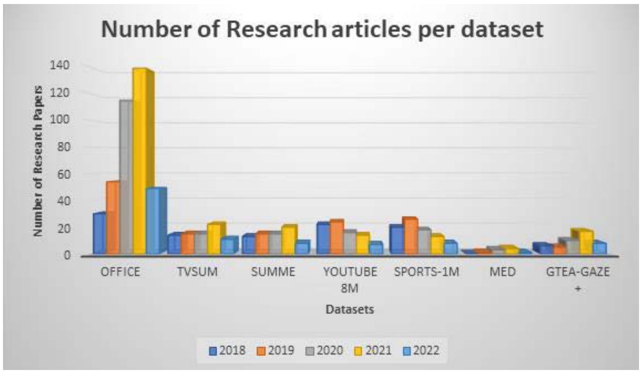

FIGURE 21. Number of articles per dataset.

• Campus [26]

Campus dataset is recorded by non-specialist at the university campus with 180-degree angle view using four cameras. The camera used are not stable and are not synchronized so the recorded videos are not stable.

• TVSum [64]

It consists of 50 videos and has 1000 annotations that permits an evaluation for various video summarization methods. TVSum data set obtains shot level importance score using crowdsourcing.

• SumMe [63]

It consists of 25 videos with 15 human annotations. It comprises of annotations evaluation code and videos which allows a consistent automatic evaluation

- Tour20 [29]

It is used primarily for Multi View Summarization (MVS) as it contains 140 videos of total 6 hours. It consists of 3 human created ground truth summaries and provides the shot segmentation

- UT Ego [27]

It contains 4 videos captured at University of Texas at Austin. It is taken from head mounted cameras that produced 3.5 hours long video. The cameras are worn by four subjects while doing different activities lie attending lecture, cooking, eating, etc.

- YouTube 8M Segments [64]

It consists of 237k segments on 1000 classes large scale labelled videos. It has high quality machine generated annotations and enables deep exploration of complex audio video visual models.

Office and campus are clearly the most frequently used datasets where monitoring and security are needed from the preceding table. The overview of domain wise used datasets shown below in figure no. 22

E. CHALLENGES IN MVS DATASETS

Multiple view datasets are complicated to work with than single view datasets because they have more difficulties, such as camera instability when recording video, lack of synchronization among multiple cameras, and cluttered situations. The key issues for these datasets are covered in detail in the following sections, along with references to relevant works [2].

1) CROWDED SCENES

Lobby and Soccer datasets in MVS literature are highly congested and contain more activities than Office dataset, which is captured in indoor situations. Working with busy situations and strong traffic is more difficult than dealing with simple scenes with a few people doing some tasks. The difficulty in the Lobby dataset is crowd density, in terms of crowd frequency, The simplest datasets in the existing research are Office and BL-7F. In addition to these issues, the Office dataset has a variable frame rate and extremely variable light variations.

2) LACK OF SYNCHRONIZATION

Mainstream videos in multiple view databases from various perspectives are not synchronized. For example, the Office dataset videos are extremely disjointed. The video clips in the Soccer dataset are not synchronized during field recording, but they are manually synchronized later. Synchronization issues make it difficult to calculate inter-view correlations across numerous video clips which are not properly synchronized, which makes creating summaries difficult. To address manually aligned the video films before conducting summary generating experiments. Other technique used various feature matching algorithms to intelligently find correlations for summary production.

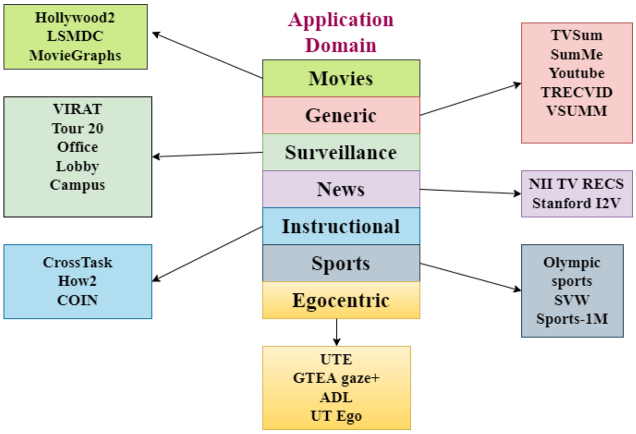

FIGURE 22. Domain wise utilization of dataset.

IV. EVALUATION METRICS

To build an effective machine learning model evaluating of model plays a very crucial step. When we must evaluate any model, we need to first understand what parameters are needed to consider while calculating specific metric.

A. CLASSIFICATION OF EVALUATION METRICS

Different evaluation metrics are present to evaluate model performance such as accuracy, precision, and F1 score as shown in figure no. 26. Depending upon the parameters and need we must select appropriate evaluation metric for any model. Base to many evaluation metrics is confusion matrix which is basically used for 2 class classification problem. In binary classification task we get two results, either correct or incorrect.

For example, if we are classifying input image as cat or dog then in supervised learning if output matches with the predicted output, then output is considered as correct output. While evaluating performance of any model some errors are made that are needed to be considered which are explained with the matrix shown below.

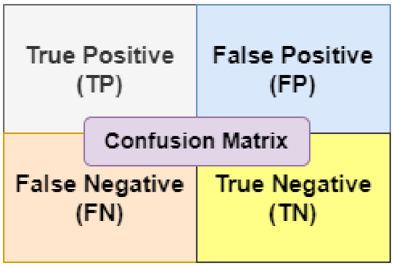

FIGURE 23. Confusion metrics.

True positive (TP): A chosen frame during the time period of the noted significant occurrence

False positive (FP): A frame that is chosen by the algorithm as consecutive frames despite not being present within the

TABLE 6. Datasets used in MVS research papers with references.

Sr.No.	Multi View Summarization Techniques	Data sets
1.	A combined multiple action recognition and summarization for surveillance video sequences [20] 2021	Weizmann, KTH, UCF-ARG, UT-Interaction, IXMAS and MHAD
2.	Generative Pre-trained Transformer-2 for Multi-modal Video Summarization [21] 2021	Query VS
3.	An efficient video summarization for surveillance system using normalized k-means and quick sort method [24] 2021	YouTube.
4.	Multi-stream dynamic video Summarization [26] 2021	Office, Campus, Lobby
5.	Intelligent Video Analytic Based Framework for Multi-view Video Summarization [28] 2021	Office, Lobby
6.	Multiview video summarization using video partitioning and clustering [30] 2021	Office, Lobby
7.	Top view multiple people tracking by detection using deep SORT and YOLOv3 with transfer learning: within 5G infrastructure [32] 2020	captured in an indoor unconstrained university campus environment (Institute of Management Sciences, Pakistan.).
8.	Cloud-Assisted Multi-View Video Summarization using CNN and Bi-Directional LSTM [33] 2019	Office
9.	Egocentric Video Summarization Based on People Interaction Using Deep Learning [27] 2018	UT Ego
10.	Video Summarization via Multi-View Representative Selection [36] 2018	TVSum and SumMe
11.	F-DES: Fast and Deep Event Summarization [41] 2018	Office, Lobby, BL-7F
12.	Efficient CNN based summarization of surveillance videos for resource-constrained devices [43] 2018	open video (OV), VSUMM,
13.	Diversity-aware Multi-Video Summarization [29] 2017	Tour20, YouTube.
14.	Long-Term Identity-Aware Multi-Person Tracking for Surveillance Video Summarization [31] 2017	PETS 2009 sequences, Virat, TRECVID 2008 and Town Centre
15.	Multi-View Surveillance Video Summarization via Joint Embedding and Sparse Optimization [25] 2017	Office, Campus, Lobby, Badminton, BL-7F, Road

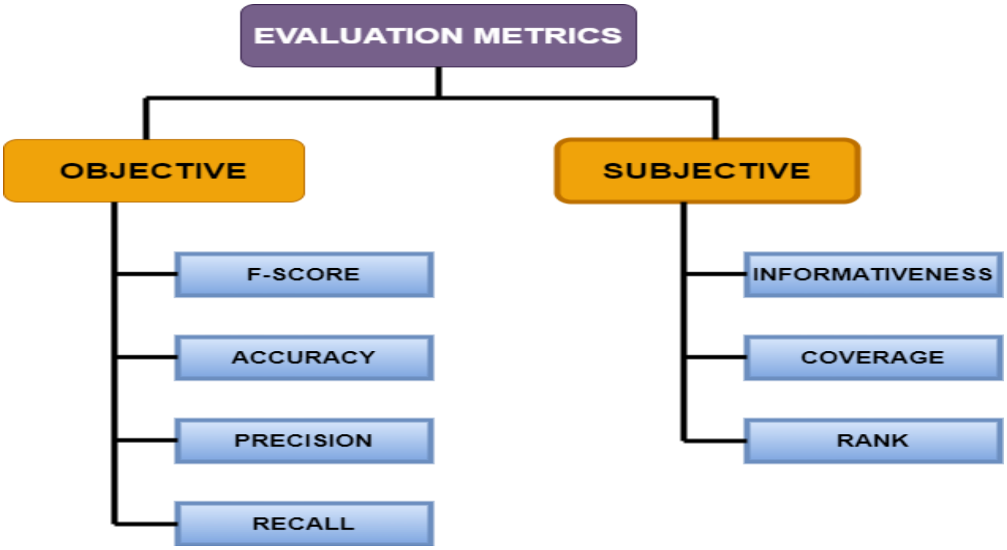

FIGURE 24. Classification of evaluation metrics.

duration of the prominent event. True negative (TN): A frame that neither falls within the prominent event interval nor is chosen by the strategies

False negative (FN): A frame between important events that the technique does not choose.

The evaluation metrics can be categorized as follows:

1) Objective metrics:

F-score, precision, recall, and accuracy are often used metrics for statistical or analyzation. Additional measures are also employed, including computing time, compression ratio, key frame frequency, training time, area under the curve, and redundancies. Calculating the F-score is the most often employed statistic in the most of methods.

1) Subjective metrics:

The majority of qualitative evaluation is subjective because it is based on user replies. Users rate the summary based on factors like informativeness, how effectively the semantics is expressed, uniqueness, representativeness, readability, scope, score, etc. Most frequently, a questionnaire is used to collect user responses in the form of scores, which are then examined for assessment.

- Accuracy (A)

Accuracy is used to determine the overall number of accurate estimates.

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

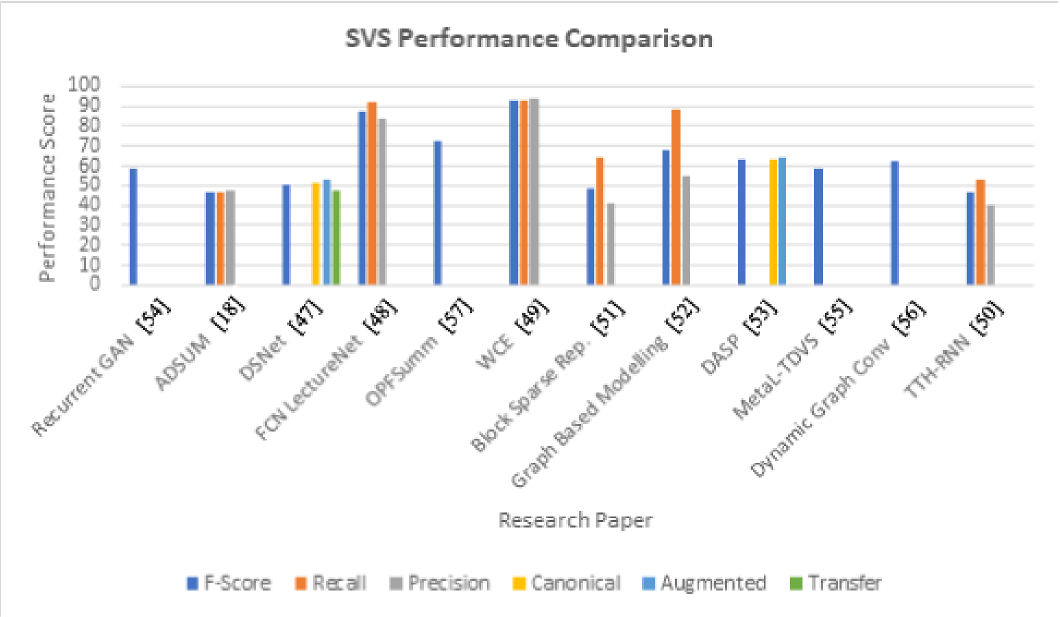

FIGURE 25. Comparison of various SVS algorithms' performance with respect to F-Score, Recall, Precision, Canonical, Augmented, Transfer.

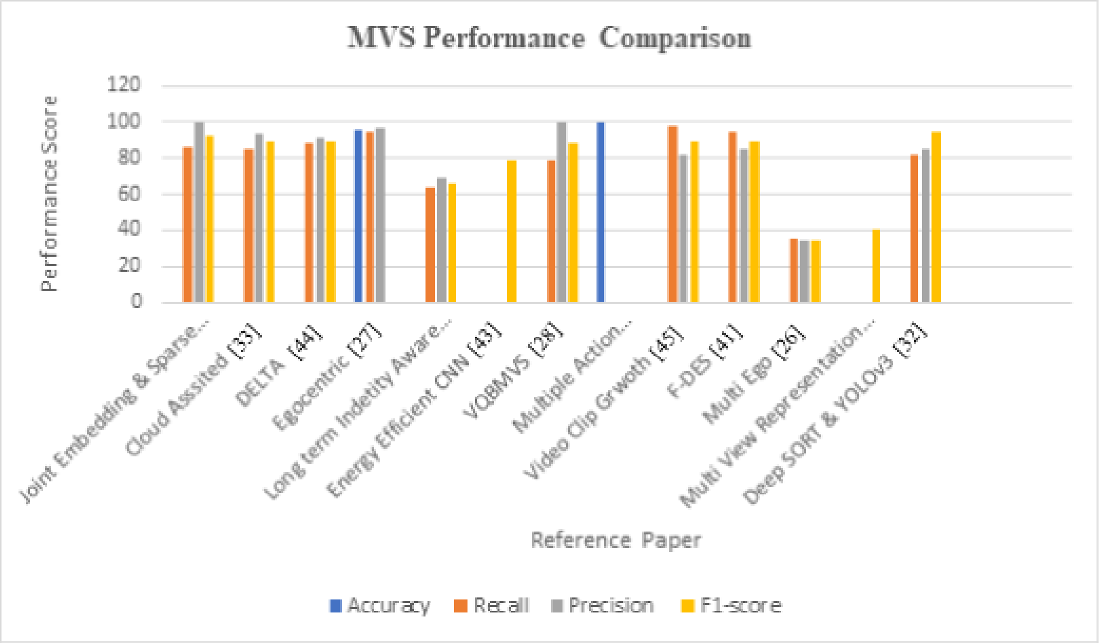

FIGURE 26. Comparison of various MVS algorithms' performance with respect to Accuracy, Recall, Precision and F-score.

- Precision (P)

Precision is mainly focused on evaluating how model predicts positive labels. It is the ratio of predicted positive output and actual positive output.

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- Recall (R)

It is the ratio of TP to all positives in dataset

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- F-Score

F-score is mean of P & R and measure tests accuracy. When we have perfect precision, we get 1.00 F-Score is used measuring search and classification performance.

$$F1 = \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

B. EVALUATION METRICS IN SINGLE VIEW SUMMARIZATION

We may deduce from the table above that F-score, precision, and recall are the most often utilized assessment metrics for

TABLE 7. Evaluation Metrics used for evaluating different SVS techniques.

Sr. No.	Paper No.	F-Score	Precision	Recall	Canonical	Augmented	Transfer	Other Metric
1.	[49]	✓	✓	✓	-	-	-	-
2.	[18]	✓	-	-	✓	✓	-	-
3.	[54]	✓	-	-	-	-	-	-
4.	[47]	✓	-	-	✓	✓	✓	-
5.	[50]	✓	✓	✓	-	-	-	-
6.	[52]	✓	✓	✓	-	-	-	-
7.	[55]	✓	-	-	-	-	-	-
8.	[56]	✓	-	-	-	-	-	✓
9.	[49]	✓	✓	✓		--	-	✓
10.	[48]	✓	-	-	-	-	-	-
11.	[51]	✓	✓	✓	-	-	-	✓
12.	[53]	✓	-	-	✓	✓	-	-
13.	[57]	✓	-	-	-	-	-	-

TABLE 8. Evaluation Metrics used for evaluating different MVS techniques.

Sr.No.	Paper No.	Accuracy	Recall	Precision	F-Score	Other Metric
1.	[20]	✓	-	-	-	-
2.	[21]	✓	-	-	✓	-
3.	[26]	-	✓	✓	✓	-
4.	[28]	-	✓	✓	✓	-
5.	[30]	-	-	-	-	✓
6.	[32]	-	✓	✓	✓	-
7.	[33]	-	✓	✓	✓	-
8.	[27]	✓	✓	✓		-
9.	[34]	-	-	-	✓	-
10.	[41]		✓	✓	✓	-
11.	[45]	✓	✓	✓	✓	-
12.	[44]	✓	✓	✓	✓	-
13.	[43]	-	-	-	-	✓
14.	[29]	-	✓	✓	✓	-
15.	[31]	-	✓	✓	✓	-
16.	[25]	-	✓	✓	✓	-

summarization techniques. Along with these regularly used metrics, other metrics like as Transfer, Canonical and Augmented are also used to assess the technique performance.

- Canonical-This is the standard supervised learning setting where the training, validation, and testing sets are from the same dataset, though they are disjoint.
- Augmented-In this setting, for a given dataset, we randomly leave 20% of it for testing, and augment the remaining 80% with the other three datasets to form an augmented training and validation dataset.
- Transfer: In this setting, for a given dataset, we use the other three datasets for training and validation and test the learned models on the dataset

C. EVALUATION METRICS IN MULTI VIEW VIDEO SUMMARIZATION

As per table no. 5, F-Score is most used evaluation metric used for Multi View Video summarization technique along with precision, recall and accuracy. F-score was employed by 12 of the 16 journal papers to evaluate their MVS approach.

As demonstrated in Fig. 21, the majority of MVS methods achieve the maximum accuracy value possible, demonstrating that all keyframes reconstructed by these approaches are fully aligned with the actual truth. Due to the mismatch between the matching important frames and the number of frames used in the actual data, the recall rate for the majority of MVS algorithms varies and encounters abrupt spikes.

Previous Multi view summarized systems used input film that weren't properly ordered for the final keyframe, which caused a poor recall score selections [2]. Aside from using low-level statistical features, algorithms that use learnt features give convincing results in form of precision and recall, resulting in a higher F1 score.

V. CONCLUSION

There are increasing number of security cameras that offer single- or multi-view coverage. In comparison to single-view cameras, dispersed video cameras provide superior scene courage and generate a disproportionately large amount of video data. Rare events are contained in this Big Data, however most of them are redundant frames with no useful information. The requirement for MVS approaches arises from the necessity to extract prominent elements from such Big Data. We provided a summary of VS methods, their datasets, and the performance criteria that were used to analyze them. The application-specific distribution, the generic flow and concise discussion of datasets can guide researchers to various single & multi view video summarization deployment directions, such as industry sectors, law and order, healthcare, education, surveillance, and entertainment, etc. It has been discovered that the Multi View Summarization (MVS) problem is not being sufficiently addressed, even though it is required for many applications. The examined literature reveals that most of the study to date has focused on artisanal or mid-level features. This study makes some recommendations for further research into how to make video summarizing systems more efficient and effective. In addition to these suggestions for upcoming investigation, we think that more work can be done on the real-time application of summarization methodologies by embedding such cutting-edge technology into corpus that facilitate the needs of contemporary news agencies for making it time-effective & reusable.

FUTURE RESEARCH

The advancement in unsupervised VS algorithms that integrate the benefits of adversarial and reinforcement learning, as well as the use of latest multi-head techniques for more precise estimation of variable-range temporal relationships between video segments, should all be priorities for future SVS research. Also, the examination of how 3D-CNNs and convolutional LSTMs might be used in architectures to represent the video's Spatio temporal architecture.

Future MVS research should focus on creating auto summaries utilizing devices with limited resources as these devices have capability that can be applied to a variety of different scenarios. Modern MVS algorithms offer experimental results that are exclusively accuracy-focused, with no consideration for their execution time or computational viability for use in actual surveillance networks. In the Multi view summarization survey, there is no evaluation of the algorithms' execution time given their system setups. In terms of effectiveness or efficiency, novel methods contribute to ongoing research in SVS and other related fields of video analysis. It is advised that the recently developed Multi View video summarizing methods be tested for running time and that complete implementation information to be shared. For upcoming study in the Multiple view field, embedded programs can be used with good resolution cameras for Multiple view Summarization at the site autonomously. Delivering only the produced summary, which can speed up video analysis while also conserving bandwidth and the valuable time of CCTV analysts, is preferable to sending all of the video clips across wireless or local networks.

REFERENCES

[1] M. Kini and K. Pai, "A survey on video summarization techniques," *Innov. Power Adv. Comput. Technol. (i-PACT)*, vol. 1, pp. 1–5, Mar. 2019.

[2] A. G. D. Molino, C. Tan, J.-H. Lim, and A.-H. Tan, "Summarization of egocentric videos: A comprehensive survey," *IEEE Trans. Hum.-Mach. Syst.*, vol. 47, no. 1, pp. 65–76, Feb. 2017.

[3] N. Dilshad, J. Hwang, J. Song, and N. Sung, "Applications and challenges in video surveillance via drone: A brief survey," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, Oct. 2020, pp. 728–732, doi: 10.1109/ICTC49870.2020.9289536.

[4] A. G. Money and H. Agius, "Video summarisation: A conceptual framework and survey of the state of the art," *J. Vis. Commun. Image Represent.*, vol. 19, no. 2, pp. 121–143, Feb. 2008.

[5] T. Hussain, K. Muhammad, W. Ding, J. Lloret, S. W. Baik, and V. H. C. de Albuquerque, "A comprehensive survey of multi-view video summarization," *Pattern Recognit.*, vol. 109, Jan. 2021, Art. no. 107567.

[6] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "Video summarization using deep neural networks: A survey," *Proc. IEEE*, vol. 109, no. 11, pp. 1838–1863, Nov. 2021.

[7] M. Basavarajaiah and P. Sharma, "Survey of compressed domain video summarization techniques," *ACM Comput. Surveys*, vol. 52, no. 6, pp. 1–29, Nov. 2020.

[8] M. U. Sreeja and B. C. Kovoor, "Towards genre-specific frameworks for video summarisation: A survey," *J. Vis. Commun. Image Represent.*, vol. 62, pp. 340–358, Jul. 2019.

[9] S. A. Ahmed, D. P. Dogra, S. Kar, and P. P. Roy, "Trajectory-based surveillance analysis: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 7, pp. 1985–1997, Jul. 2019.

[10] K. Schoeffmann, M. A. Hudelist, and J. Huber, "Video interaction tools: A survey of recent work," *ACM Comput. Surveys*, vol. 48, no. 1, pp. 1–34, Sep. 2015.

[11] D. Sen and B. Raman, "Video skimming: Taxonomy and comprehensive survey," *ACM Comput. Surveys*, vol. 52, no. 5, pp. 1–38, Sep. 2020.

[12] K. B. Baskurt and R. Samet, "Video synopsis: A survey," *Comput. Vis. Image Understand.*, vol. 181, pp. 26–38, Apr. 2019.

[13] M. R. Suguna and A. Kalaivani, "A research on multi-view video summarization techniques," *Int. J. Eng. Adv. Technol.*, vol. 9, no. 1, pp. 6837–6846, Oct. 2019.

[14] P. Kalaivani and S. M. M. Roomi, "Towards comprehensive understanding of event detection and video summarization approaches," in *Proc. 2nd Int. Conf. Recent Trends Challenges Comput. Models (ICRTCCM)*, Feb. 2017, pp. 61–66, doi: 10.1109/ICRTCCM.2017.84.

[15] S. V. Bhagwat and S. S. Thokal, "A survey on automatic summarization using multi-modal summarization system for asynchronous collections," *Int. J. Innov. Res. Sci., Eng. Technol.*, vol. 8, no. 2, pp. 2347–6710, Feb. 2019.

[16] H. B. U. Haq, M. Asif, and M. B. Ahmad, "Video summarization techniques: A review," *Int. J. Sci. Technol. Res.*, vol. 9, no. 11, pp. 146–153, Nov. 2020.

[17] B. Bineesh and S. Shunmugan, ''Comparative work on video summarization methods,'' *J. Inf. Comput. Sci.*, vol. 10, no. 6, pp. 1548–7741, 2020.

[18] Z. Ji, Y. Zhao, Y. Pang, X. Li, and J. Han, ''Deep attentive video summarization with distribution consistency learning,'' *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 4, pp. 1765–1775, Apr. 2021.

[19] H. Fu and H. Wang, ''Self-attention binary neural tree for video summarization,'' *Pattern Recognit. Lett.*, vol. 143, pp. 19–26, Mar. 2021.

[20] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, A. Bouridane, and A. Beghdadi, ''A combined multiple action recognition and summarization for surveillance video sequences,'' *Int. J. Speech Technol.*, vol. 51, no. 2, pp. 690–712, Feb. 2021.

[21] J. H. Huang, L. Murn, M. Mrak, and M. Worring, ''GPT2MVS: Generative pre-trained transformer-2 for multi-modal video summarization,'' in *Proc. Comput. Vis. Pattern Recognit.*, Apr. 2021, pp. 580–589.

[22] B. Balasubramanian, P. Diwan, and D. Vora, ''Deep learning based approaches for recommendation systems,'' in *Intelligent Data Communication Technologies and Internet of Things* (Lecture Notes on Data Engineering and Communications Technologies), vol. 38, D. Hemanth, S. Shakya, and Z. Baig, Eds. Cham, Switzerland: Springer, 2020, doi: 10.1007/978-3-030-34080-3_58.

[23] M. U. Shaikh, D. Vora, and A. Anurag, ''Surveillance system for intruder detection using facial recognition,'' in *Intelligent Computing and Networking* (Lecture Notes in Networks and Systems), vol. 146, V. E. Balas, V. B. Semwal, A. Khandare, and M. Patil, Eds. Singapore: Springer, 2021, doi: 10.1007/978-981-15-7421-4_18.

[24] D. M. Davids and C. S. Christopher, ''An efficient video summarization for surveillance system using normalized K-means and quick sort method,'' *Microprocessors Microsyst.*, vol. 83, Jun. 2021, Art. no. 103960.

[25] R. Panda and A. Roy-Chowdhury, ''Multi-view surveillance video summarization via joint embedding and sparse optimization,'' *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 2010–2021, Sep. 2017, doi: 10.1109/TMM.2017.2708981.

[26] M. Elfeki, L. Wang, and A. Borji, ''Multi-stream dynamic video summarization,'' in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Dec. 2018, pp. 339–349.

[27] H. A. Ghafoor, A. Javed, A. Irtaza, H. Dawood, H. Dawood, and A. Banjar, ''Egocentric video summarization based on people interaction using deep learning,'' *Math. Problems Eng.*, vol. 2018, pp. 1–12, Nov. 2018.

[28] V. Parikh and P. Sharma, ''Intelligent video analytic based framework for multi-view video summarization,'' *Int. J. Comput. Digit. Syst.*, vol. 12, no. 1, pp. 619–628, Aug. 2018.

[29] R. Panda, N. C. Mithun, and A. K. Roy-Chowdhury, ''Diversity-aware multi-video summarization,'' *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4712–4724, Oct. 2017.

[30] A. S. Parihar, J. Pal, and I. Sharma, ''Multiview video summarization using video partitioning and clustering,'' *J. Vis. Commun. Image Represent.*, vol. 74, Jan. 2021, Art. no. 102991, doi: 10.1016/j.jvcir.2020.102991.

[31] S. I. Yu, Y. Yang, X. Li, and A. G. Hauptmann, ''Long-term identity-aware multi-person tracking for surveillance video summarization,'' in *Proc. Comput. Vis. Pattern Recognit.*, Apr. 2016, pp. 1–14.

[32] I. Ahmed, M. Ahmad, A. Ahmad, and G. Jeon, ''Top view multiple people tracking by detection using deep SORT and YOLOv3 with transfer learning: Within 5G infrastructure,'' *Int. J. Mach. Learn. Cybern.*, vol. 12, no. 11, pp. 3053–3067, 2020, doi: 10.1007/s13042-020-01220-5.

[33] T. Hussain, K. Muhammad, A. Ullah, Z. Cao, S. W. Baik, and V. H. C. de Albuquerque, ''Cloud-assisted multiview video summarization using CNN and bidirectional LSTM,'' *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 77–86, Jan. 2020, doi: 10.1109/TII.2019.2929228.

[34] R. V. Bidwe, S. Mishra, S. Patil, K. Shaw, D. R. Vora, K. Kotecha, and B. Zope, ''Learning approaches for video compression: A bibliometric analysis,'' *Big Data Cogn. Comput.*, vol. 6, no. 2, p. 44, 2022, doi: 10.3390/bdcc6020044.

[35] S. Patil, L. Chavan, J. Mukane, D. Vora, and V. Chitre, ''State-of-the-art approach to E-learning with cutting edge NLP transformers: Implementing text summarization, question and distractor generation, question answering,'' *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 1, pp. 1–9, 2022, doi: 10.14569/IJACSA.2022.0130155.

[36] J. Meng, S. Wang, H. Wang, J. Yuan, and Y.-P. Tan, ''Video summarization via multiview representative selection,'' *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2134–2145, May 2018, doi: 10.1109/TIP.2017.2789332.

[37] A. Moawad, E. Islam, N. Kim, R. Vijayagopal, A. Rousseau, and W. B. Wu, ''Explainable AI for a no-teardown vehicle component cost estimation: A top-down approach,'' *IEEE Trans. Artif. Intell.*, vol. 2, no. 2, pp. 185–199, Apr. 2021.

[38] R. Confalonieri, L. Coba, B. Wagner, and T. R. Besold, ''A historical perspective of explainable artificial intelligence,'' *WIREs Data Mining Knowl. Discovery*, vol. 11, no. 1, p. e1391, Jan. 2021.

[39] M. Hudec, E. Mináriková, R. Mesiar, A. Saranti, and A. Holzinger, ''Classification by ordinal sums of conjunctive and disjunctive functions for explainable AI and interpretable machine learning solutions,'' *Knowl.-Based Syst.*, vol. 220, May 2021, Art. no. 106916.

[40] L. A. de Souza, R. Mendel, S. Strasser, A. Ebigbo, A. Probst, H. Messmann, J. P. Papa, and C. Palm, ''Convolutional neural networks for the evaluation of cancer in Barrett's esophagus: Explainable AI to lighten up the black-box,'' *Comput. Biol. Med.*, vol. 135, Aug. 2021, Art. no. 104578.

[41] K. Kumar and D. D. Shrimankar, ''F-DES: Fast and deep event summarization,'' *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 323–334, Feb. 2018, doi: 10.1109/TMM.2017.2741423.

[42] A. Dilawari and M. Khan, ''ASoVS: Abstractive summarization of video sequences,'' *IEEE Access*, vol. 7, pp. 29253–29263, 2019, doi: 10.1109/ACCESS.2019.2902507.

[43] K. Muhammad, T. Hussain, and S. W. Baik, ''Efficient CNN based summarization of surveillance videos for resource-constrained devices,'' *Pattern Recognit. Lett.*, vol. 130, pp. 370–375, Feb. 2020.

[44] K. Kumar and D. D. Shrimankar, ''Deep event learning boosT-up approach: DELTA,'' *Multimedia Tools Appl.*, vol. 77, no. 20, pp. 26635–26655, Oct. 2018, doi: 10.1007/s11042-018-5882-z.

[45] G. Pan, X. Qu, L. Lv, S. Guo, and D. Sun, ''Video clip growth: A general algorithm for multi-view video summarization,'' in *Advances in Multimedia Information Processing* (Lecture Notes in Computer Science), vol. 11166, R. Hong, W. H. Cheng, T. Yamasaki, M. Wang, and C. W. Ngo, Eds. Cham, Switzerland: Springer, 2018, doi: 10.1007/978-3-030-00764-5_11.

[46] D. Tank, ''A survey on sport video summarization,'' *IJSART*, vol. 2, no. 10, pp. 1–5, Oct. 2016.

[47] W. Zhu, J. Lu, J. Li, and J. Zhou, ''DSNet: A flexible detect-to-summarize network for video summarization,'' *IEEE Trans. Image Process.*, vol. 30, pp. 948–962, 2021, doi: 10.1109/TIP.2020.3039886.

[48] C. Huang and H. Wang, ''A novel key-frames selection framework for comprehensive video summarization,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 2, pp. 577–589, Feb. 2020, doi: 10.1109/TCSVT.2019.2890899.

[49] K. Davila, F. Xu, S. Setlur, and V. Govindaraju, ''FCN-LectureNet: Extractive summarization of whiteboard and chalkboard lecture videos,'' *IEEE Access*, vol. 9, pp. 104469–104484, 2021.

[50] B. Zhao, X. Li, and X. Lu, ''TTH-RNN: Tensor-train hierarchical recurrent neural network for video summarization,'' *IEEE Trans. Ind. Electron.*, vol. 68, no. 4, pp. 3629–3637, Apr. 2021, doi: 10.1109/TIE.2020.2979573.

[51] M. Ma, S. Mei, S. Wan, J. Hou, Z. Wang, and D. D. Feng, ''Video summarization via block sparse dictionary selection,'' *Neurocomputing*, vol. 378, pp. 197–209, Feb. 2020, doi: 10.1016/j.neucom.2019.07.108.

[52] C. Chai, G. Lu, R. Wang, C. Lyu, L. Lyu, P. Zhang, and H. Liu, ''Graph-based structural difference analysis for video summarization,'' *Inf. Sci.*, vol. 577, pp. 483–509, Oct. 2021.

[53] Z. Ji, F. Jiao, Y. Pang, and L. Shao, ''Deep attentive and semantic preserving video summarization,'' *Neurocomputing*, vol. 405, pp. 200–207, Sep. 2020, doi: 10.1016/j.neucom.2020.04.132.

[54] L. Lan and C. Ye, ''Recurrent generative adversarial networks for unsupervised WCE video summarization,'' *Knowl.-Based Syst.*, vol. 222, Jun. 2021, Art. no. 106971, doi: 10.1016/j.knosys.2021.106971.

[55] X. Li, H. Li, and Y. Dong, ''Meta learning for task-driven video summarization,'' *IEEE Trans. Ind. Electron.*, vol. 67, no. 7, pp. 5778–5786, Jul. 2020, doi: 10.1109/TIE.2019.2931283.

[56] J. Wu, S. H. Zhong, and Y. Liu, ''Dynamic graph convolutional network for multi-video summarization,'' *Pattern Recognit.*, vol. 107, Nov. 2020, Art. no. 107382, doi: 10.1016/j.patcog.2020.107382.

[57] G. B. Martins, D. R. Pereira, J. G. Almeida, V. H. C. de Albuquerque, and J. P. Papa, ''OPFSumm: On the video summarization using optimum-path forest,'' *Multimedia Tools Appl.*, vol. 79, nos. 15–16, pp. 11195–11211, Apr. 2020, doi: 10.1007/s11042-018-5874-z.

[58] M. Ma, S. Mei, S. Wan, Z. Wang, and D. Feng, "Video summarization via nonlinear sparse dictionary selection," *IEEE Access*, vol. 7, pp. 11763–11774, 2019, doi: 10.1109/ACCESS.2019.2891834.

[59] Y. Jiang, K. Cui, B. Peng, and C. Xu, "Comprehensive video understanding: Video summarization with content-based video recommender design," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1562–1569, doi: 10.1109/ICCVW.2019.00195.

[60] J. Basak, V. Luthra, and S. Chaudhury, "Video summarization with supervised learning," in *Proc. 19th Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4, doi: 10.1109/ICPR.2008.4761475.

[61] M. P. J. Ashby, "The value of CCTV surveillance cameras as an investigative tool: An empirical analysis," *Eur. J. Criminal Policy Res.*, vol. 23, no. 3, pp. 441–459, Sep. 2017, doi: 10.1007/s10610-017-9341-6.

[62] G. Van Voorthuijsen, H. A. J. M. Van Hoof, M. Klima, K. Roubik, M. Bernas, and P. Pata, "CCTV effectiveness study," in *Proc. 39th Annu. Int. Carnahan Conf. Secur. Technol.*, Oct. 2005, pp. 105–108.

[63] M. Gygli, H. Grabner, H. Riemenschneider, and L. V. Gool, "Creating summaries from user videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014, pp. 505–520. [Online]. Available: https://gyglim.github.io/me/

[64] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "TVSum: Summarizing web videos using titles," in *Proc. CVPR*, Jun. 2015, pp. 5179–5187. [Online]. Available: https://github.com/yalesong/tvsum

[65] Y. Fu, Y. Guo, Y. Zhu, F. Liu, C. Song, and Z.-H. Zhou, "Multi-view video summarization," *IEEE Trans. Multimedia*, vol. 12, no. 7, pp. 717–729, Nov. 2010.

[66] S.-H. Ou, C.-H. Lee, V. S. Somayazulu, Y.-K. Chen, and S.-Y. Chien, "On-line multi-view video summarization for wireless video sensor network," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 1, pp. 165–179, Feb. 2015.

[67] S. E. F. De Avila, A. P. B. Lopes, A. Da Luz, and A. De A. Araújo, "VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognit. Lett.*, vol. 32, no. 1, pp. 56–68, Jan. 2011.

[68] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," in *Proc. Eur. Conf. Comput. Vis.*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014, pp. 540–555. [Online]. Available: http://lear.inrialpes.fr/people/potapov/medsummaries

[69] K.-H. Zeng, T.-H. Chen, J. C. Niebles, and M. Sun, "Title generation for user generated videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 609–625. [Online]. Available: http://aliensunmin.github.io/project/video-language/

[70] D. Zhukov, J. B. Alayrac, R. G. Cinbis, D. Fouhey, I. Laptev, and J. Sivic, "Cross-task weakly supervised learning from instructional videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3537–3545.

[71] Y. Li, B. Merialdo, M. Rouvier, and G. Linares, "Static and dynamic video summaries," in *Proc. 19th ACM Int. Conf. Multimedia*, 2011, pp. 1573–1576.

[72] J. Xu, L. Mukherjee, Y. Li, J. Warner, J. M. Rehg, and V. Singh, "Gaze-enabled egocentric video summarization via constrained submodular maximization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2235–2244, doi: 10.1109/CVPR.2015.7298836.

[73] I. Ide, "Report on the analyses and the applications of a large-scale news video archive: NII TV-RECS," *IEICE Electron. Exp.*, pp. 9–17, Mar. 2014, doi: 10.2201/NiiPi.2014.11.3.

[74] V. Kaushal, S. Kothawade, A. Tomar, R. Iyer, and G. Ramakrishnan, "How good is a video summary? A new benchmarking dataset and evaluation framework towards realistic video summarization," in *Proc. Comput. Vis. Pattern Recognit.*, vol. 26 Jan. 2021, pp. 1–19.

[75] *Visualizing Scientific Landscapes*. Accessed: Jan. 2022. [Online]. Available: https://www.vosviewer.com/

[76] H. Yanaka, K. Mineshima, D. Bekki, K. Inui, S. Sekine, L. Abzianidze, and J. Bos, "Can neural networks understand monotonicity reasoning," in *Proc. ACL Workshop BlackboxNLP, Analyzing Interpreting Neural Netw. (NLP)*, Florence, Italy, 2019, pp. 31–40.

[77] V. Kaushal, S. Kothawade, R. Iyer, and G. Ramakrishnan, "Realistic video summarization through VISIOCITY: A new benchmark and evaluation framework," in *Proc. 2nd Int. Workshop AI Smart TV Content Prod., Access Del.*, Oct. 2020, pp. 37–44, doi: 10.1145/3422839.3423064.

PAYAL KADAM received the Bachelor of Engineering degree in electronics and telecommunication from Shivaji University, and the Master of Technology degree in electronics (VLSI) from Bharati Vidyapeeth (Deemed to be University), Pune. She is currently a Research Scholar at Symbiosis International University, Pune. Her research interests include image processing and deep learning.

DEEPALI VORA (Senior Member, IEEE) received the Ph.D. degree in computer science and engineering from Amity University, Mumbai.

She worked as the Head of the Information Technology Department, Vidyalankar Institute of Technology, Mumbai. She is currently working as an Associate Professor of Computer Science and Engineering, and engineering with the Symbiosis Institute of Technology Pune, Symbiosis International University (Deemed), Pune, India. She has more than 20 years of experience in total in teaching, research, and industry. She has published more than 50 research papers in reputed national, international conferences and journals. She has coauthored three books and two book chapters and delivered various talks in data science and machine learning. She received grants from government bodies, such as AICTE, ISTE, and the industry. She has organized many value-added courses for the benefit of the students. More than 18 students have completed and currently, and three students are pursuing their post-graduate studies under her guidance from Mumbai University. In addition to that, three students are pursuing research (Ph.D.) under her guidance with Symbiosis International University, Pune. Her course developed on deep learning is currently available on the Unschool platform for all, and two technical blogs are available on the KnowledgeHut.com website. She is acting as a Reviewer for many international conferences and journals, such as IEEE Access, *Journal of Big Data Analytics* (IGI Global), *Journal of Intelligent Systems*, and Inderscience.

SASHIKALA MISHRA received the Ph.D. degree in the field of bioinformatics and data mining from Siksha 'O' Anusandhan University, Bhubaneswar, Odisha, in 2015. She has authored more than 63 international journals in almost all publishing houses, including IEEE, Elsevier, Inderscience, Springer, and ACM, with citations of more than 50 every year. She also filed three patents and published two IPRs. She has also contributed book chapters in *Biological Knowledge Discovery Handbook* by Wiley Publisher. Her paper has also received the Best Paper Award in International Conference on Modeling, Optimization and Computing 2012 held at Noorul Islam University, Kanyakumari, India, by Elsevier. Her research interests include artificial intelligence, conservation biology, pricing theory, bioinformatics, data mining, image processing, and networking. Her main research interests lay in the design and implementation of algorithms related to prediction, classification, and clustering.

SHRUTI PATIL received the M.Tech. degree in computer science and the Ph.D. degree in data privacy from Pune University. She is currently an industry professional currently associated with the Symbiosis Institute of Technology as a Professor and with the SCAAI as a Research Associate, Pune, Maharashtra. She has three years of industry experience and ten years of academic experience. She has expertise in applying innovative technology solutions to real world problems. She is also working in the application domains of healthcare, sentiment analysis, emotion detection, and machine simulation in which she is also guiding several UG, PG, and Ph.D. students as a domain expert. She has published more than 30 research articles in reputed international conferences and Scopus and Web of Science indexed journals, as well as books with more than 90 citations. Her research interests include applied artificial intelligence, natural language processing, acoustic AI, adversarial machine learning, data privacy, digital twin applications, GANS, and multimodal data analysis.

KETAN KOTECHA received the Ph.D. and M.Tech. degrees from the IIT Bombay, India. He currently holds position as the Head of the Symbiosis Center for Applied AI (SCAAI), the Director of the Symbiosis Institute of Technology, and the Dean of the Faculty of Engineering, Symbiosis International (Deemed University). He has gained expertise and experience in cutting-edge research and projects in AI and deep learning over the last 25 years.

AJITH ABRAHAM (Senior Member, IEEE) received the Master of Science degree from Nanyang Technological University, Singapore, in 1998, and the Ph.D. degree in computer science from Monash University, Melbourne, VIC, Australia, in 2001. He is currently the Director of the Machine Intelligence Research Laboratories (MIR Laboratories), a Not-for-Profit Scientific Network for Innovation and Research Excellence Connecting Industry and Academia. The Network with HQ in Seattle, WA, USA, is also more than 1,500 scientific members from over 105 countries. He also works as a Professor of artificial intelligence at Innopolis University, Russia, and the Yayasan Tun Ismail Mohamed Ali Professorial Chair in Artificial Intelligence at UCSI, Malaysia. As an Investigator/a Co-Investigator, he has won research grants worth over more than 100 Million U.S. dollars. He holds two university professorial appointments. He works in a multi-disciplinary environment. He has authored/coauthored more than 1,400 research publications out of which there are more than 100 books covering various aspects of computer science. One of his books was translated into Japanese and a few other articles were translated into Russian and Chinese. He has more than 49,000 academic citations (H-index of more than 104 as per Google Scholar). He has given more than 150 plenary lectures and conference tutorials (in more than 20 countries). He was the Chair of the IEEE Systems Man and Cybernetics Society Technical Committee on Soft Computing (which has over more than 200 members), from 2008 to 2021, and served as a Distinguished Lecturer for the IEEE Computer Society representing Europe (2011–2013). He was the Editor-in-Chief of *Engineering Applications of Artificial Intelligence* (EAAI), from 2016 to 2021, and serves/served on the editorial board for over 15 international journals indexed by Thomson ISI.

LUBNA ABDELKAREIM GABRALLA received the B.Sc. and M.Sc. degrees in computer science from the University of Khartoum and the Ph.D. degree in computer science from the Sudan University of Science and Technology, Khartoum, Sudan. She is also an Associate Professor with the Department of Computer Science and Information Technology, Princess Nourah bint Abdulrahman University, Saudi Arabia. Her current research interests include soft computing, machine learning, and deep learning. She became a Senior Fellow (SFHEA), in 2021.