Video summarization using deep learning techniques: a detailed analysis and investigation

Article *in* Artificial Intelligence Review · March 2023

DOI: 10.1007/s10462-023-10444-0

Dr Parul Saini
DIT University
17 PUBLICATIONS **368** CITATIONS
SEE PROFILE

Krishan Berwal
88 PUBLICATIONS **1,727** CITATIONS
SEE PROFILE

Shamal Kashid
National Institute of Technology (NIT) Uttarakhand
14 PUBLICATIONS **53** CITATIONS
SEE PROFILE

Alok Negi
National Institute of Technology (NIT) Uttarakhand
35 PUBLICATIONS **355** CITATIONS
SEE PROFILE

Artificial Intelligence Review
https://doi.org/10.1007/s10462-023-10444-0

Video summarization using deep learning techniques: a detailed analysis and investigation

Parul Saini[1] · Krishan Kumar[1] · Shamal Kashid[1] · Ashray Saini[1] · Alok Negi[1]

Abstract

One of the critical multimedia analysis problems in today's digital world is video summarization (VS). Many VS methods have been suggested based on deep learning methods. Nevertheless, These are inefficient in processing, extracting, and deriving information in the minimum amount of time from long-duration videos. Detailed analysis and investigation of numerous deep learning approach accomplished to determine root of problems connected with different deep learning methods in identifying and summarizing the essential activities in such videos. Various deep learning techniques have been investigated and examined to detect the event and summarization capability for detecting and summarizing multiple activities. Keyframe selection Event detection, categorization, and the activity feature summarization correspond to each activity. The limitations related to each category are also discussed in depth. Concerns about detecting low activity using the deep network on various types of public datasets are also discussed. Viable strategies are suggested to evaluate and improve the generated video summaries on such datasets. Moreover, Potential recommended applications based on literature are listed out. Various deep learning tools for experimental analysis have also been discussed in the paper. Future directions are presented for further exploration of research in VS using deep learning strategies.

Keywords Event summarization · Critical information in videos · Surveillance systems · Video analysis · Multimedia analysis

✉ Krishan Kumar
kkberwal@nituk.ac.in

Parul Saini
parulsaini.phd2020@nituk.ac.in

Shamal Kashid
kashid.shamalphd2021@nituk.ac.in

Ashray Saini
mt20cse001@nituk.ac.in

Alok Negi
aloknegi.phd2020@nituk.ac.in

[1] Department of Computer Science and Engineering, National Institute of Technology Uttarakhand, Srinagar Garhwal, Uttarakhand 246174, India

Published online: 15 March 2023

1 Introduction

Videos are the most potent and popular multimedia form as they quickly connect with users. With the arrival of high-speed Internet and low-cost storage, the amount of data has been generated at a rocket pace, most of it in the form of visual or video data (Money and Agius 2008). Video hosting, television show hosting, social media, and online news platforms such as Wistia, SproutVideo, You-tube, Netflix, Amazon Prime, Twitter, LinkedIn, and Facebook have housed a vast amount of video material. YouTube alone produces more than 10 h of video content every second. Video requires more storage and bandwidth to transmit than image and text. Moreover, numerous human resources are necessary to analyze such videos. For such hefty data, effective methods and tools are required to capture videos and present them more compactly and concisely, which may be further used in various applications (Kumar et al. 2018).

The primary objective of VS is to analyze the video by dropping the unnecessary or redundant frames and preserving the keyframes (Kumar et al. 2016). Moreover, it helps to accelerate the browsing of an extensive collection of video data and achieve structured access and representation of the video content. An enormous amount of video recordings are generated and shared on the internet around the clock. In this multimedia era, practical use cases of the videos may be applicable in every corner. Therefore, a video summary can be convenient in any circumstances or situation when a user wants to graze rapidly at video content. Consequently, Automatic VS (AVS) (Binol et al. 2021) is the major trending and growing research area in this field. Artificial Intelligence (AI) enabled software can easily do the task of summarizing lengthy videos.

Various professional and educational applications based on multiple types of VS. These generate or use enormous amounts and volumes of multimedia data, such as monitoring, tracking, diagnosing, identifying, investigating, security analysis, etc. Different media organizations, including sports or entertainment videos, generate teasers or trailers of movies and TV series (Emon et al. 2020) can also use VS to create video highlights for events. Moreover, video search engines can also use VS for video indexing, browsing, retrieval, and recommendation (Emon et al. 2020). In addition, medical video analysis can use VS for complex diagnostics.

Further, VS is also employed to remove frame redundancy, reducing storage requirements and computational time (Xiao et al. 2020b). By choosing the most informative segment of the video, video summarising technologies attempt to provide a brief and complete description. It speeds up the video processing, storage, management, and retrieval of the videos effectively and efficiently, making interpreting and analyzing certain situations or events in long videos easier. Video summary can be static or dynamic. A *static summary* refers to a group or collection of frames called a key-framing or video-storyboard. The static summary is insufficient for users to understand the video, especially in the case of long videos (Emon et al. 2020; Xiao et al. 2020b). However, these techniques can view and index videos and present videos as thumbnails.

Other type, *video skimming* consists of related shots, i.e., a collection of video segments with corresponding audio information, improving the summarization's semantics. Further, seeing a skim or video summary rather than a slide show of frames is generally more entertaining, enjoyable, and fascinating for the users or the audience (Emon et al. 2020; Xiao et al. 2020b); however, time-consuming. While storyboards are not bounded by timing or synchronization issues, they give more flexibility for browsing and navigation of data organization and allow for greater freedom in the data structure for browsing, and navigation (Binol et al. 2021; Emon et al. 2020; Xiao et al. 2020b). Some examples of the VS are generated highlights, video synopsis (Tiwari and Bhatnagar 2021; Ajmal et al. 2012; Sridevi and Kharde 2020). Thumbnail generation domain also considered very close to VS. Conventional thumbnail generation techniques cannot provide meaningful synopsis to the users.

A graph convolved video thumbnail pointer (GTP) can produce a semantically meaningful and coherent video thumbnail from an input video. It also generate the thumbnail semantically related to the natural sentence query (Yuan et al. 2019b). A sentence guided temporal modulation (SGTM) (Rochan et al. 2020) technique uses sentence embedding to control the video thumbnail generating network's normalised temporal activations. These can be coarsely classified into the unsupervised approaches (Mahmoud et al. 2013; Li et al. 2006; Ma et al. 2002; Barbieri et al. 2003) and supervised approaches (Sundaram et al. 2002; Agnihotri et al. 2001; Li et al. 2001). Gesture, audio-visual and objects based detailed framework (Hu et al. 2011) is presented for visual content-based video indexing and retrieval, including structural analysis, feature analysis, video data mining using extracted features and feedback. Barbieri et al. Barbieri et al. (2003) divides video summary into various levels including local (scene level) (Sundaram et al. 2002), global (Agnihotri et al. 2001; Li et al. 2001), and meta-level (Hussain et al. 2019).

Based on the different aspects in the literature, deep learning-based VS (Del Molino et al. 2016; Senthil Murugan et al. 2018; Sreeja and Kovoor 2019; Money and Agius 2008) splits the VS techniques into split subtypes, internal (Khan et al. 2020a; Pereira et al. 2019), external (Sharghi et al. 2017b; Coppola et al. 2020; Lee et al. 2018), and hybrid (Zhu et al. 2016), based on the source of information. The visual surveillance system is proposed to detect a moving object to summarize the videos (Senthil Murugan et al. 2018). VS can also be classified based on the generated summary as generic, object-based, or event-based (Nair and Mohan 2021; Money and Agius 2008; Basavarajaiah and Sharma 2021). A VS technique (Pereira et al. 2019) considered the various standards like the type of source of video, summary or synopsis, preferences, genre, mechanism, and application in different areas or domains. A categorization of various VS focused on compressed domain summarization techniques has been presented (Basavarajaiah and Sharma 2019). Moreover, some state-of-the-art techniques (Hussain et al. 2021) are presented for Multi-View VS (MVS), which poses distinct challenges in summarization than the mono-view videos.

In supervised approaches, training a deep network takes a long time. Thanks to the Graphics processing unit (GPU) for reducing the training time and handling the computational difficulty in deep learning. A large number of Convolutional Neural Networks (CNN) and Deep Convolutional Neural Networks (DCNN), including GoogleNet, Inception V3, AlexNet, variations of ResNet, and variations of VGG Very Deep Convolutional Networks (VGGNet) (Nair and Mohan 2021; Kumar and Shrimankar 2017; Ji et al. 2019; Muhammad et al. 2020; Hussain et al. 2019) have been demonstrated for several applications (Brezeale and Cook 2008). GoogleNet seems to be the most widely used so far. Some key steps on Video Summarization techniques using deep learning are mentioned below:

- *Step1: Analyze Information Sources* Each information source needs to be analyzed, so that the primary information content can be recognized and used further.
- *Step2: Measure of Relevance* The information content based on generic or specialized to a certain issue is generated based on features or semantic approaches.
- *Step3: Synthesize Appropriate Output* The extracted data is structured in an understandable format and represented as accurately as feasible as a output of the model.

The above literature reveals that deep techniques can be more beneficial in solving the video summarization problem. Therefore, it is decided to analyze and investigate the recent developments using State-Of-The-Art (SOTA) deep-learning-based algorithms in the video summarization domain. The significant contributions of the work are as follows:

- In this work, various deep learning frameworks have been analyzed with their pros and cons for video summarization and offer category-wise video summarization techniques using deep learning approaches.
- A detailed exploration of the various existing video summarization techniques is done. It covers the essential aspects, video summarization process, feature-based video summaries, and genre-based summarization.
- An application-based analysis of several video summarization techniques is also presented, along with their limitations and solutions compared with the other existing approaches.
- The details of the existing datasets in the literature have been provided, with the challenges and future directions for future research and video summarization applications.

The remaining article is organized as follows: Sect. 2 presents a detailed comparison of the existing video summarization techniques with their contributions and limitations. Section 3 elaborates on the deep learning techniques-based Video Summarization models and their properties on the basis of supervised, unsupervised, and weakly supervised-based Video Summarization techniques are analyzed. Section 4 presents a detailed and comprehensive overview of several deep learning-based applications of video summarization. Section 5 provides the details of the recent contributions in Video Summarization with the help of Deep learning. Section 6 provides the details of the various datasets and their performance. Section 7 discusses the different evaluation methods for Video Summarization. Section 8 highlights Video Summarization challenges. Section 9 introduced the future directions and the work has been concluded in Sect. 10.

2 Video summarization techniques and their contributions

The video summarization classifications based on their characteristics and properties are shown in Fig. 1.

2.1 Feature based VS techniques

Li et al. (2006) discuss three distinct feature parts of the films, including exposition at the beginning, conflict in the middle, and resolution at the end. In feature-based, the user mainly focuses on the video features like motion, color, gesture, audio-visual, speech,objects, etc. Low-level features such as color and texture are most commonly used to extract the information from the video content because they are easy to compute but not very accurate (Brezeale and Cook 2008; Ajmal et al. 2012).

2.2 Clustering based VS techniques

In Kumar et al. (2016), equal partitions based clustering technique is proposed to detect the key-frames based on the pixel intensity. The research (De Avila et al. 2011; Peker and

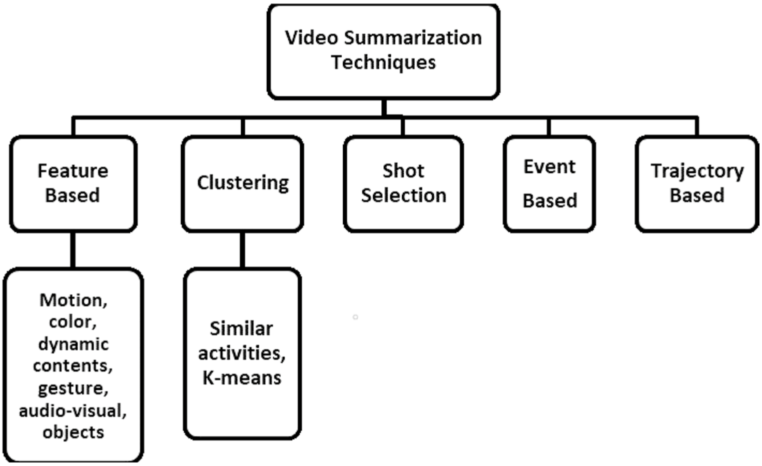
Video Summarization Techniques
Feature Based
Clustering
Shot Selection
Event Based
Trajectory Based
Motion, color, dynamic contents, gesture, audio-visual, objects
Similar activities, K-means

Fig. 1 Primary classification of video summarization

Bashir 2007) revealed that many clustering techniques, including k-means, partitioned, and spectral clustering, have been used for VS. Kumar et al. Kumar et al. (2018) suggested an Eratosthenes sieve based key-frame extraction clustering technique. Summary length is decided based on the inclusion of the content decided on the specific criteria and uses of different evaluation techniques.

2.3 Shot selection based VS techniques

Generic video summaries (Money and Agius 2008; Basavarajaiah and Sharma 2021) are not personalized to the specific user's command or interest but produced by extracting keyframes or shot boundaries detection, scene changes methods, and redundancy reduction (Tiwari and Bhatnagar 2021). In VS, shots are also detected by measuring the transition between the successive frames. VS (Hu et al. 2011) is also classified as static video abstracts, dynamic skims, and hierarchal summarization, where video skimming is achieved by removing redundancy, detecting objects or events, and multimodal integration. Function-based VS methods (Ma et al. 2002) use the attention mechanism to determine the important parts of the video. At the same time, the structure-based VS strategy exploits hierarchical story structure in the form of frames and shots.

2.4 Event based VS techniques

Agius et al. (Dimitrova et al. 2003) presents the different types of generated video summaries based on the object, event, perception, and feature. High-level features such as events, specific face, motion, gestures, etc., are highly reliable for giving important video content information (Xu et al. 2016a; Wei et al. 2021; Shingrakhia and Patel 2022). In Kumar et al. (2018), events are renovated from the extracted key-frames by fixing the minimum and maximum frame number for the event boundaries. Video events are extracted using graph theory (Kumar 2019) and scale free network (Kumar and Shrimankar 2018a) in mono-view videos and using Basic local alignment searching technique (Kumar 2021) and collections of weak ensembles (Kumar and Shrimankar 2018b) in multi-view videos. Some of the SOTA techniques are proposed for creating video event summary of soccer, cricket, tennis, and basketball games (Vasudevan and Sellappa Gounder 2021). DL is based on an artificial neural network in which the word “deep” reflects the use of multiple hidden layers in a neural network to extract high-level features and can learn vast amounts of data.

2.5 Trajectory based VS techniques

Most researchers initially worked on static VS. A dynamic video summary is generated through a trajectory with stationary backgrounds, which required a lot of computing resources. Deep learning may be the best solutions to detect the important content from video.

3 Deep learning based video summarization

Deep learning (DL) is a dominant branch of machine learning which has been extended with different network structures (Chai et al. 2021). It has been successfully used in various domains, including cybersecurity, natural language processing, bioinformatics, robotics and control, medical information processing, and many more (Alzubaidi et al. 2021). DL has also achieved superior results in video processing, in which VS plays a critical role. DL methods for VS can be supervised, weakly supervised, unsupervised, and Reinforcement learning, as shown in Fig. 2.

3.1 Supervised learning based VS

The supervised techniques learn from the data to predict future outcomes. However, the biggest challenge in supervised learning is to label the data. It requires a high cost to create well-defined datasets as it needs domain knowledge and does not work well with a wide variety of content on the internet. Supervised models are categorized as classification and regression models. Classification models are those where output can be classified as “pass" or “fail" and are used to predict the categories where regression models are used where output is a fundamental value such as sales revenue or weight. Linear classifiers, K-Nearest Neighbors (K-NN), support vector machines, decision trees, and random forests are all standard classification algorithms. Linear, logistic, and polynomial regression are common types of regression algorithms considered machine learning techniques. Table 1 shows comparison of Supervised Learning based DL techniques for VS. Some of the DL techniques are elaborated as follows:

Deep belief network (DBN) is a sophisticated generative model that uses a deep architecture consisting of many stacks of Restricted Boltzmann Machine (RBM) and can be employed in both unsupervised and supervised models. It can extract and classify the features, helpful in many applications. Each RBM’s visible layer is linked to the previous RBM’s hidden layer, and the top two layers are non-directional, as shown in Fig. 3. In Shingrakhia and Patel (2022) DBN, formed by stacking different RBM, is used for cricket video summarization to classify the sports videos into long, medium, close-up, and crowd shots.

Deep neural network (DNN) is the basic architecture of a neural network in which “deep" refers to the multiple hidden layers between the input and output layers of this

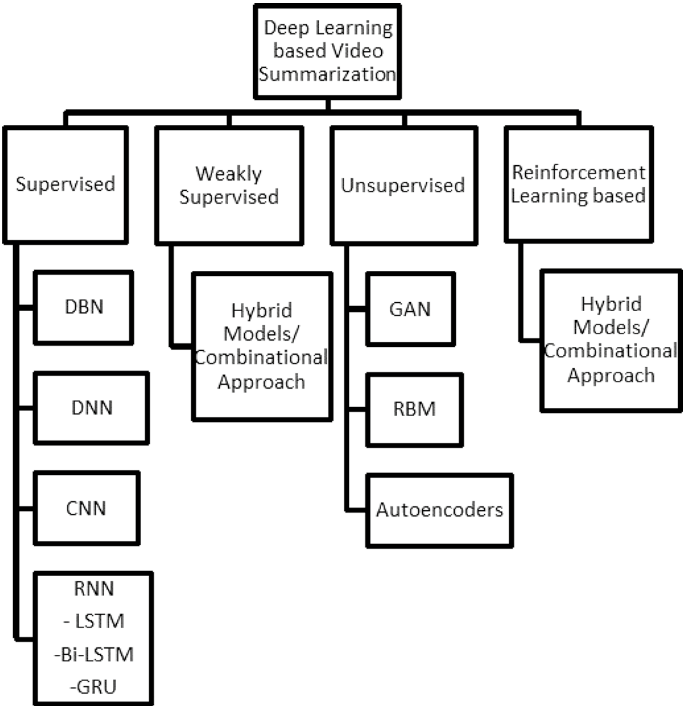
Deep Learning based Video Summarization
Supervised
Weakly Supervised
Unsupervised
Reinforcement Learning based
DBN
DNN
CNN
RNN
- LSTM
-Bi-LSTM
-GRU
Hybrid Models/ Combinational Approach
GAN
RBM
Autoencoders
Hybrid Models/ Combinational Approach

Fig. 2 Deep learning techniques for VS

forward feed network, as shown in Fig. 4. It increases the model's performance accuracy compared to the primary Artificial Neural Network (ANN). In these networks, data flows in the forward direction only, not in the backward direction. Each layer includes some weight, with an activation function that acts as a gateway for passing the signal to the next layer. All the other popular deep learning models have DNN as their primary unit. These include recurrent neural network (Dargan et al. 2020; Pouyanfar et al. 2018; Zhang et al. 2018; Navamani 2019), Autoencoder, long short term memory (Hochreiter and Schmidhuber 1997; Gers et al. 2000, 2002), Gated Recurrent Units (GRUs) (Chung et al. 2014; Zhao et al. 2020; Archana and Malmurugan 2021), CNN (Phung and Rhee 2019), DBN, and RBM and generative adversarial networks (Dargan et al. 2020).

Convolutional neural networks (CNNs) is the most popular and widely used algorithm in deep learning. There are several convolutional layers in a typical CNN, as shown in Fig. 5, followed by pooling layers, and finally, fully connected layers in the final stage. The CNN's job is to minimize the images into a more straightforward process format while preserving essential features for a good prediction. The convolution operation aims to extract high-level features from the input image, such as edges and even more complex features similar to objects, faces, etc. At last, it classifies images using an activation function. Major CNN models used for VS is compared in Table 2.

Some techniques focus on spatial and temporal dependency instead of merely focusing on temporal structures present between frames. CNN is generally used for processing digital images, where images have a grid-like structure that contains some pixel values. A two-stream DCNN for extracting both spatial and temporal information of a video

Table 1 Comparison of supervised learning based DL techniques for VS

SOTA model	DL technique	Contributions	Limitations
TTHRNN (Zhao et al. 2020)	Hierarchical RNN	Contains tensor train embedding layer in RNN to explore Long Range Dependencies (LRD). Reduces training difficulty	Works on defined set of parameters which require uniform sampling for sequences having more than 2000 frames. High computation
Mutiscale RNN (Zhu et al. 2022)	CNN+RNN	Multiscale Hierarchical attention module for structure modelling and LRD to leverage both visual and motion information via intra block and intro block attention	Generation of fixed length blocks which affect the information inside it
OCR Based Model (Guntuboina et al. 2022)	CNN	OCR is used to get the score from cropped scoreboard image detected through YOLO and scoreboard highlights of 30 s is generated	Image preprocessing is required as they may contain noises and false negatives. Only 3 types of scoreboard detectionis applicable
DHAVS (Lin et al. 2022)	Hierarchical LSTM + 3DCNN	3DCNN is used to extract spatio-temporal features and hierarchical LSTM is used to capture LRD. Designed cost sensitive loss function to address unbalanced class distribution	Limited to Short Length Videos (SLV). Improvement to learn from unpaired data is required

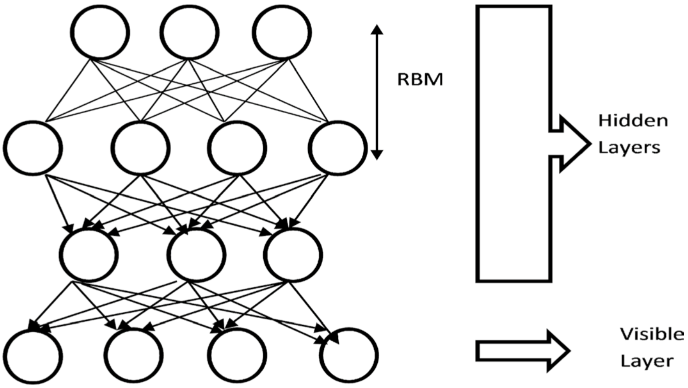

Fig. 3 Basic deep belief network blocks

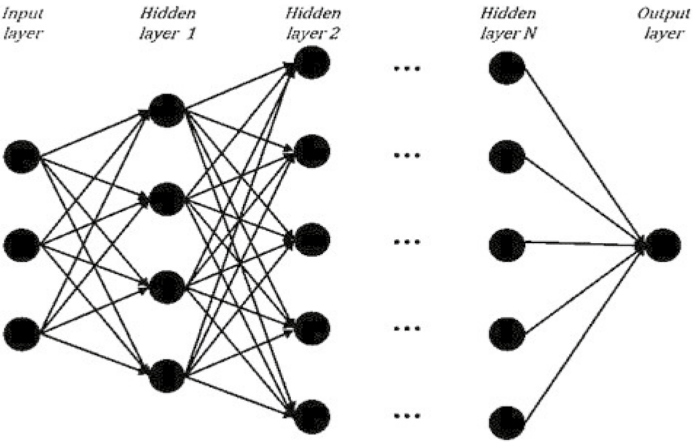

Fig. 4 Deep neural network architecture

employed to utilize parallelization of GPU (Sridevi and Kharde 2020). Experiments done on public datasets shows that the precision with fusion of spatial and temporal DCNNs are higher than their single stream model (Otani et al. 2016). In the cloud-based tier, CNN architecture extracts deep features as input is given to LSTM, which outputs probabilities sequence of frames for the classes, informative and non-informative.

The final summary consists of the sequences with the highest chances of informativeness. Its fusion model is more efficient than single stream DCNNs. After optimizing the architecture, the precision is improved by adding a layer on top called a ranking layer, and the model is retrained to assign higher scores to important segments. A new self-collected dataset was employed to enhance the precision rather than the public dataset, which was not very accurate. Recent applications on CNN are face mask detection (Negi and Kumar 2021b; Negi et al. 2020, 2021b), detection of plant disease (Negi et al. 2021a; Chauhan et al. 2021b), malaria cell detection (Alok et al. 2021), citrus disease detection and classification (Negi and Kumar 2021a), plant seedling image classification (Chauhan et al. 2021a), Covid-19 detection from X-ray images etc.

Table 2 Comparison of CNN models used for VS

Model name	Parameters (In millions)	Depth	Size (MB)	Top 5 error rate
LeNet (Alom et al. 2019)	0.060	005	–	00.95 (MNIST)
AlexNet (Alom et al. 2019)	060	008	–	15.30 (Imagenet)
VGG16 (Simonyan and Zisserman 2014)	138	016	528	08.80 (Imagenet)
VGG19 (Simonyan and Zisserman 2014)	144	019	549	09.00 (Imagenet)
Google V3 (Szegedy et al. 2016)	23.6	159	92	03.50 (Imagenet)
ResNet (He et al. 2016)	25.6	152	98	03.60 (Imagenet)
	25.6	190	–	3.46 (CIFAR10+)
DenseNet (Alom et al. 2019)	25.6	190	–	17.18 (CIFAR100+)
(Huang et al. 2017)	15.3	250	–	05.19 (CIFAR10)
	15.3	250	–	19.64 (CIFAR100)
Xception (Chollet 2017)	23.00	126	088	0.055 (Imagenet)
MobileNetV2 (Howard et al. 2017)	03.50	88	14	–

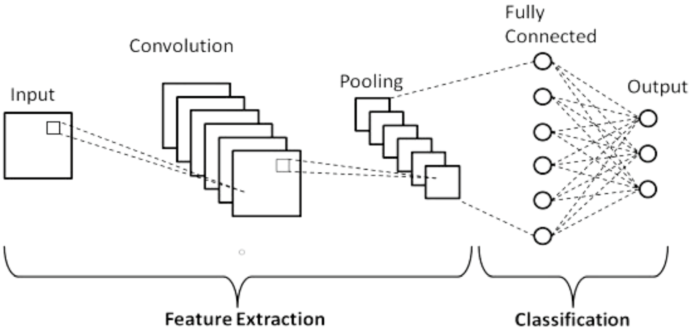

Fig. 5 Basic architecture of CNN

Recurrent neural network (RNN) is a widely used DL algorithm for VS as it utilizes sequential information or considers the time series. In this network, one unit can store information about the previous units (Dargan et al. 2020; Pouyanfar et al. 2018; Zhang et al. 2018). RNNs have a vanishing gradient problem, a long-term dependency problem in which information gets lost over time. RNNs variants were proposed to solve this problem like long LSTMs (Hochreiter and Schmidhuber 1997; Gers et al. 2000, 2002) and GRUs (Chung et al. 2014). Hierarchical Recurrent Neural Network (H-RNN) (Zhao et al. 2017) has two layers. The first layer encodes short video sub shots cropped from the source video. Each sub-shot's hidden state is sent into the second layer to calculate its confidence to be a critical sub-shot.

While generating summaries on focussing the shot segmentation process, such methods and techniques typically create shots of fixed-length segmentation. Multi-edge detection process and multi-edge optimized LSTM RNN (Archana and Malmurugan 2021) are proposed and integrated. Therefore, the proposed model name is Multi-Edge Optimized LSTM RNN for VS(MOLRVS). A combination of a tensor train embedding layer and a multi LSTM (two-layer) forms Tensor-Train Hierarchical Recurrent Neural Network for

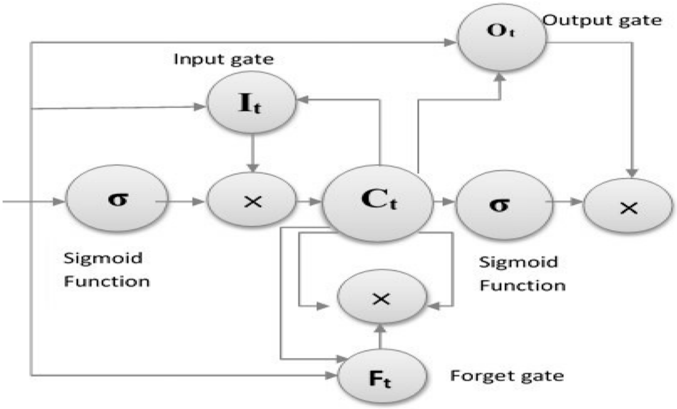

Fig. 6 Basic model of long short term memory

Video Summarization (TTH-RNN) model (Zhao et al. 2020). Tensor train embedding layer prevents the significant feature to hidden mapping matrices caused by the high-dimensional video features.

Long short term memory (LSTM) is an improvement over RNNs. It has a mechanism called gates, as shown in Fig. 6, which controls the flow of information in the network to learn which data is significant to keep or throw away.

Forget gate decides what data should be thrown away or marked with the help of the sigmoid function. The input gate helps to calculate the cell state by updating it from the old state to the new state. Output gate is used to determine the next hidden state, which is also used for predictions. It is also used as Bi-directional LSTM by putting two independent LSTMs together, which allows the model to have both backward and forward information flow about the sequence at every time step (Archana and Malmurugan 2021; Yuan et al. 2019a).

Bidirectional LSTM (BiLSTM) is an encoder and LSTM with an attention mechanism for a summary generation. The encoder in a BiLSTM is responsible for encoding contextual information between video input frames. Furthermore, an attention mechanism on the decoder side of AVS is provided, which explores two techniques to construct video summaries. The proposed approach outruns all the SOTA methods (Ji et al. 2019). Unlike most existing supervised systems, which use BiLSTM, this method uses the underlying hierarchical structure of video sequences. It first divides each video sequence into equal-length blocks and uses intra-block and inter-block attention to understand local and global information. The frame-level, block-level, and video-level representations are combined to predict the frame-level relevance score (Zhu et al. 2022).

3.2 Weakly supervised learning based VS

It is a combination of supervised and unsupervised learning that needs a small number of labeled or annotated data. This weakly labeled or less expensive dataset for learning can create an excellent, predictable model for VS. A weakly-supervised reinforcement learning method for VS proposed by Li and Yang (2021) is based on the combination of two networks. The first is the Video Classification Sub-Network (VCSN) which plays the supervisor role for the second network called the Summary Generation Sub-Network (SGSN).

This enhancement can assist in creating a semantically meaningful summary from its original. Table 3 compares their contribution and limitations of different weakly-supervised learning-based DL techniques for VS.

A unique weakly-supervised way to summarize instructional videos is proposed using text based on a RL concept (Ramos et al. 2022). In this approach, an agent is guided by a novel joint reward function to choose which frames to eliminate and shorten the input video to the desired length without gaps in the output video. Additionally, VDAN+ and a VDAN (Ramos et al. 2020a) modification produce a highly discriminative embedding space to represent both textual and visual input. Experiments on YouCook2 and COIN datasets (Zhou et al. 2018b; Ramos et al. 2020b) show that this approach outperforms the baselines in Precision, Recall, and F1 Score while effectively regulating the video's output duration. A query-adaptive technique is proposed to incorporate saliency maps into a submodular optimization to consider query phrases both in capturing relevant images and representing similarity among shots (Cizmeciler et al. 2022). This study tested the proposed dataset activity-related summaries, as well as a subset of the RAD dataset (Vasudevan et al. 2017).

3.3 Unsupervised learning based VS

In unsupervised techniques, the models learn independently using the data that is neither classified nor labeled. It does not need human supervision to learn. Table 4 compares their contribution and limitations of different unsupervised learning-based DL techniques for VS. It is categorized as clustering, association, and dimensionality reduction such as PCA, K-means, and singular value decomposition (Hatcher and Yu 2018).

Generative adversarial networks (GANs) are robust neural networks used for unsupervised learning and reinforcement learning. In 2014, Goodfellow developed and introduced GAN (Goodfellow et al. 2014) with two neural network models. There is a generator and a discriminator in GANs. The generator and the discriminator are neural networks competing throughout the training phase. The generator creates data samples (images, audio, etc.) to deceive the discriminator. The discriminator seeks to discriminate between actual and fraudulent samples. The procedures are performed multiple times, and with each iteration, the generator and discriminator improve their performance in their respective roles. These networks compete to evaluate and reconstruct the variations within a dataset. In some cases, GAN can be used as supervised learning for video summarization, like in Fu et al. (2019a). It predicts the cutting (beginning and stopping) points for each summarization fragment; the framework uses an attention-based Pointer Network (Ptr-Net) (Vinyals et al. 2015) as the generator as shown in Fig. 7.

In unsupervised learning; GAN has been used in many ways with VS. However, GANs have been widely used for unsupervised VS. Sreeja and Kovoor (2022) uses a GAN to create a VS framework. Summarised movies in a GAN-based technique (Mahasseni et al. 2017) VS used in combination with VAE. Ptr-Net in Lan and Ye (2021) in which the summarizer is the Variational Autoencoder (VAE) based LSTM architecture with Ptr-Net and De-redundancy Mechanism (DM). The discriminator is another LSTM that distinguishes the original and reconstructed video from the summarizer. An actor-critical model was proposed (Apostolidis et al. 2021) with a GAN for treating the selection of key video fragments as a sequence-generating task.

The objective is to minimize the distance between the distributions of the ground truth and generated summaries. Another adversarial learning network (Zhang et al.

Table 3 Comparison of weakly supervised learning based DL techniques for VS

SOTA model	DL technique	Contributions	Limitations
SGSN (Li and Yang 2021)	CNN+LSTM	Employed VCSN to extract video semantic representation to construct meaningful reward	Complex.Limited to SLV
Extended VDAN+ (Ramos et al. 2022)	Hierarchical RNN	Skip Aware Fast Forwarding Agent is used to decide which frames to remove based on textual data as it embeds both textual and visual data	Agent may take incorrect action. Reward is sparse in some scenarions
Query Specific Weakly Supervised Model (Cizmeciler et al. 2022)	CNN	Query focused VS which uses semantic attributes and attention maps to locate related regions. Works for long length videos	It has accuracy vs efficiency tradeoff. More efficient CNN architecture and temporal dependency (LSTM) is not considered

Table 4 Comparison of unsupervised learning based DL techniques for VS

SOTA model	DL technique	Contributions	Limitations
GVSUM (Basavarajaiah and Sharma 2021)	CNN	Generic VS which uses K-Means clustering and sequential keyframe selection algorithm for VS	Summary is not user specific. Limited to visual features
MultiCNN Model (Nair and Mohan 2021)	Multi CNN + AE	It detects keyframes based on feature extracted from four pre-trained CNN models using RFC	Limited to 4 CNN models. Low Sensitivity. No sequence learning from past and future
AC-SUM-GAN (Apostolidis et al. 2021)	GAN	First work which embeds actor-critic model into GAN in the field of VS	Lack of LRD
Adv-Ptr-Der-SUM (Lan and Ye 2021)	LSTM+GAN +AE	Combination of VAE, Ptr-Net and GAN for WCE VS. WCE 2019 dataset is also proposed	Limited to WCE videos. Complex
RSGN (Zhao et al. 2022)	LSTM	Reconstructive Sequence Graph Network which preserves shot level dependencies using LSTM	Limited to SLV. Limited numbers of rewards
CNN Bi-Conv LSTM GAN (Sreeja and Kovoor 2022)	GAN+AE +CNN+ LSTM	Combination of GAN and Knowledge distillation Focusses on diverse and representative elements of video. Generate static and dynamic summary	Complex to some extent. Summary length is fixed

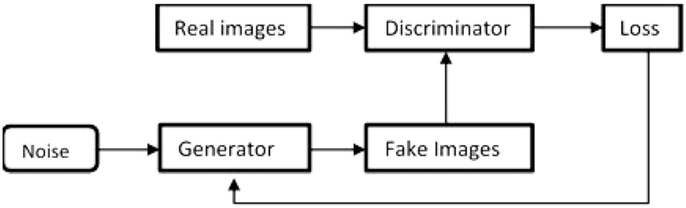

Fig. 7 General working of generative adversarial networks

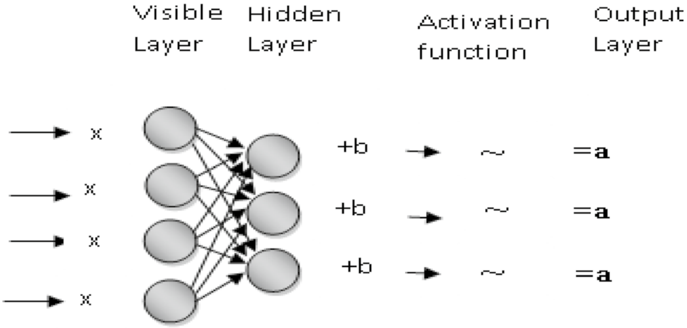

Fig. 8 Restricted Boltzmann machine

2019b) combines LSTM and Dilated Temporal Relational (DTR), which learns to generate a summary by trying the trainable discriminator. GAN-based training framework is proposed attention-based Pointer Network, and the discriminator judges whether a fragment is from a ground-truth or a machine-generated summary (Vinyals et al. 2015). The proposed adversarial learning model's first step ensures that the extracted features contain a wide range of representative video fragments. The adversarial network is followed by a knowledge distillation phase, which uses a primary network with input data acquired from the prior GAN model to operate as a keyframe.

Restricted Boltzmann machines (RBMs) are a variant of Boltzmann machines that restrict the connections between the visible and hidden units. It is a graphical and modeled depiction of a hidden and visible layer, and symmetric relationship between undirected layers. Every visible and hidden layer node is related, but no two nodes in the same layer are connected (Dargan et al. 2020; Navamani 2019). They are a two-layered neural network, with a fully bipartite graph connecting the two layers, as shown in Fig. 8.

AutoEncoders (AE) Automatic AE and neural networks where the output and input layers have the same dimensionality. Output units in the output layer equal the input units in the input layer. An AE reproduces the data from the input to the output in an unsupervised manner, refers as neural replicator network and its types include denoising, sparse, deep autoencoder, VAE, and contractive autoencoder. Mainly, sparse, deep, and VAE have been frequently used for VS. An autoencoder (Dong et al. 2021; Bengio 2009) consists of encoder, decoder, and hidden layer, as shown in Fig. 9. Keyframes based on feature vectors extracted from Multi CNN (Nair and Mohan 2021) is proposed to fed into sparse autoencoder, giving a combined illustration of the input feature vectors. Sparse autoencoder can also generate summaries of high quality from videos of all categories or genres.

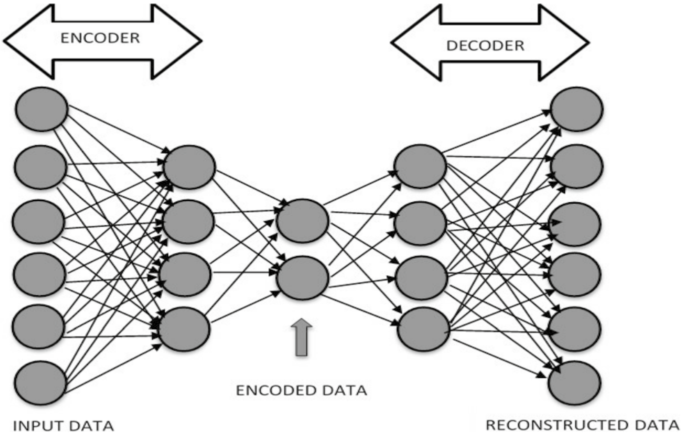

Fig. 9 General working of an Autoencoder

Also, further experiments have been done which prove that this model using a random forest classifier performs much well than any other classifiers. Deep Autoencoders consists of two similar or dissimilar deep networks, one for encoding and another for decoding having four to five layers. An efficient CNN-based summarization (Muhammad et al. 2020) is a shot segmentation method using deep features extracted for surveillance videos in the resource-constrained scenario. VAE learns the input in a compressed form called the latent space and combined with GAN (Lan and Ye 2021).

3.4 RL based VS

RL is a technique based on a series of actions, decisions, and a game-like situation with a trial and error method, learning with its own experiences. It aims to bring the best move or path in a specific domain. The agent gets a positive reward for every correct action and a negative reward for every incorrect action. In this trial and error method, the agent determines the best way to perform the task with maximum rewards. Industrial Automation, Robotics, and Traffic Light Management systems are some applications of RL (Hatcher and Yu 2018). Table 5 compares different RL-based DL techniques for VS with their contribution and limitations.

Some unsupervised approaches summarize the video by focusing on specific properties for generating the most optimum video summary. The summarizer takes a sequence of video frames as input and produces a summary by predicting importance scores at the frame level.

The calculated results are then joined to get the overall reward value used to train the summarizer. A reinforcement-based unsupervised method (Xu et al. 2021) employed to summarize crowd surveillance videos. A novel crowd location-density reward is proposed to produce high-quality video summaries. A novel reward function (Gonuguntla et al. 2019) is formulated by using the temporal segment networks (Wang et al. 2016) that obtains scores for each frame of the video and seeks to include the temporal and spatial features of the original video in the generated summary. The reward function strives to create an overview in the training phase by having high temporal and spatial scores frames.

Table 5 Comparison of RL based DL techniques for VS

SOTA model	DL technique	Contributions	Limitations
LD-HSN (Xu et al. 2021)	LSTM	Crowd location and density reward function is designed to produce summary of crowd surveillance videos using hierarchical LSTM	Maintains performance for 20 min videos. Limited to two crowd behaviors
3D Spatio-Temporal U-Net RL (Liu et al. 2022)	CNN	Spatio-temporal CNN combined with UNET. Effective in medical videos	Limited to SLV

Table 6 Applications of video summarization techniques

Applications	Techniques
	Deep Cricket Summarization Network (Emon et al. 2020)
	SGRNN-AM (Shingrakhia and Patel 2022)
Sports	YOLO and OCR based sports Video Summarization (Guntuboina et al. 2022)
	Deep action based sports VS (Tejero-de Pablos et al. 2018)
	VS for Query focused (Xiao et al. 2020a)
Query focused	Saliency maps based VS (Cizmeciler et al. 2022)
	Convolutional Hierarchical Attention Network (Xiao et al. 2020b)
	DeepRes (Deep Learning-Based video summarization) (Muhammad et al. 2019)
Surveillance	Convolutional neural network based VS (Muhammad et al. 2020)
	RL based Video Summarization (Xu et al. 2021)
User	DG-CNN (Wu et al. 2020)
Generated	Saliency estimation-based VS (Wei et al. 2021)
Videos	Highlight Detection, Pairwise Model (Sridevi and Kharde 2020)
Movies, News	QuickLook (Najafabadi et al. 2015)
Documentaries	Hierarchical-grained deep Reinforcement Learning model (Ji et al. 2021a)
	Wireless Capsule Endoscopy(WCE) Video Summarization (Lan and Ye 2021)
Medical	3D spatio-temporal -UNet (Liu et al. 2022)
	Ultrasonic imaging Summarization (Liu et al. 2020)
	FCN Network (Davila et al. 2021)
Educational	Text-driven video acceleration (Ramos et al. 2022)
	Multiple Attention Educational VS (Chootong et al. 2021)

In contrast, the enhanced deep summarization network strives to earn higher rewards by learning to produce more comprehensive summaries. VS as a successional stage of decision making is discussed (Zhou et al. 2018a) and trains the summarizer and creates the video summaries using the representative diversity awards. The diversity award quantifies the difference between selected keyframes. while the representativeness award measures the distance of the selected keyframes from the rest of the video frames. The 3D Spatio-temporal-UNet (Ronneberger et al. 2015), RL framework for VS, which efficiently encodes Spatio-temporal information from input movies for downstream RL is proposed (Liu et al. 2022). RL agent learns from spatiotemporal latent scores and anticipates actions for maintaining or rejecting a video frame.

4 DL based VS applications

Since the criteria for keyframe selection are not the same for all genres, the VS technique employed for a particular video material also depends on the video's domain. The type of genre also serves as a criterion for categorizing VS systems. Depending on the characteristics of the video, it can be classed into any genre. The recent VS frameworks investigated for this survey are divided into genres, with approaches associated with each genre. VS method for generic videos named GVSUM (Basavarajaiah and Sharma 2021) applied to many areas like sports, medicine, surveillance, movies, home, etc as shown in Table 6.

Table 7 Different applications of sports based VS

SOTA approach	Sport applications
Deep Cricket Summarization Network (Emon et al. 2020)	Cricket
Deep action based sports video summarization (Tejero-de Pablos et al. 2018)	Kendo, Japanese sport
SGRNN-AM (Shingrakhia and Patel 2022)	Cricket
YOLO & OCR based sports video summarization (Guntuboina et al. 2022)	Football, Cricket & Kabaddi, Kendo

Sports application includes a wide range of sports. The main problem in VS is that important events change between sports. The goal, foul, replay, and so on are momentous occurrences in soccer games, whereas boundary, six, comprehensive, and so on are key events in cricket. Detecting important sporting events is a time-consuming task (Tiwari and Bhatnagar 2021; Sreeja and Kovoor 2019). Deep Cricket Summarization Network (DCSN) (Emon et al. 2020) is used to predict the frame-level probabilities whose allocations are utilized as frame-selection scores to create keyframe selection decisions automatically. To address challenges of DCSN, a new VS method (Tejero-de Pablos et al. 2018) is introduced to access the players conduct as a signal and use that signal to determine the highlights of the source video.

To summarise cricket video, a hybrid approach is also proposed (Shingrakhia and Patel 2022) based on machine learning. Using scoreboard detection, (Guntuboina et al. 2022) provides a very affordable method for automatic key event extraction and summarising of sports movies. YOLO (Redmon and Farhadi 2018) used for soreboard detection which was then clipped out of each frame of the video, followed by noise reduction and false-positive reduction. The different applications of the sports based VS are mentioned in Table 7.

Query-focused video summarization is one of the most significant obstacles to VS research is user subjectivity: users have varying preferences for the summaries they want to view. Two issues arose due to the subjectiveness, as discussed in Tiwari and Bhatnagar (2021), Sreeja and Kovoor (2019), Kumar (2021), Sharghi et al. (2017a). Sharghi et al. (2017a) used memory networks and determinantal point processes to exploit the attention schemes and diversity modelling capabilities. Zhang et al. (2019c) presented MapNet,a Mapping Network, which allows mapping of visual information to query space and SummNet, a deep reinforcement learning-based summarising network, which provides individualised summaries by incorporating relatedness, representativeness, and diversity rewards. These rewards help the agent in selecting the video shots that are most relevant to the user query. The task of query-focused VS is formulated as a scoring problem (Xiao et al. 2020a). The CHAN framework (Xiao et al. 2020b) divides a long video into multiple parts before extracting visual information with a pre-trained network. A query-adaptive technique (Cizmeciler et al. 2022) is proposed to incorporate saliency maps into a submodular optimization to consider query phrases both in capturing relevant images and in representing similarity among shots.

Surveillance Several frameworks have been created that are focused on summarising surveillance videos (Tiwari and Bhatnagar 2021; Sreeja and Kovoor 2019). A DL-based VS Strategy for Resource-Constrained Industrial Surveillance Scenarios system is divided into 4 major parts. Another framework (Muhammad et al. 2020) provides an energy-efficient VS approach based on CNN for surveillance videos acquired by resource-constrained devices, relying on the strength of CNNs for various applications. A novel crowd location-density reward (Xu et al. 2021) is presented to teach a RL-based framework, the summarization network, to produce high-quality summaries.

User generated videos often known as consumer videos, are videos created by ordinary people using ordinary video cameras. As a result, the video is of poor quality and resolution. These videos are categorized as generic since they can come from any domain (Tiwari and Bhatnagar 2021; Sreeja and Kovoor 2019). The goal of VS using highlight detection and pairwise deep ranking model (Sridevi and Kharde 2020) is to create a video summary by modeling a two-stream model with DCNN. An effective saliency estimation-based key frame extraction approach (Wei et al. 2021) is introduced to prevent the influence of emotion-independent frames on the recognition result. A DGCN (Wu et al. 2020) is presented to quantify the importance and relevance of each video shot in its video and the entire video collection.

Movies, news and documenatries Movies are a significant part of the ever-increasing expansion of multimedia information over the internet, and their analysis and summarization are a popular topic among researchers. Because movies are of the entertainment category, the enjoyment component must be kept in mind when summarising them (Tiwari and Bhatnagar 2021; Sreeja and Kovoor 2019). ’QuickLook’ (Najafabadi et al. 2015) is an automatic MS framework that recognizes the main actors and merges several cues retrieved from a movie. Actual demands of news reports VS (Ji et al. 2021a) investigates by focusing on basketball videos of the National Basketball Association.

Medical Endoscopic movies, records of medical operations and surgery, diagnostic hysteroscopy videos, and other medical videos provide a wide range of possibilities for automatic VS (Tiwari and Bhatnagar 2021; Sreeja and Kovoor 2019). A deep RL-based ultrasonic imaging summarization (Liu et al. 2020) approach demonstrates efficacy using the case of fetal ultrasonography screening. WCE (Lan and Ye 2021) generates many duplicate images in a single exam, making it difficult and time-consuming to review by a physician.

Educational Nowadays, educational and instructional videos are valuable learning resources. For a variety of reasons, summarising videos in this genre is difficult. Analyzing the presentation in the video with a focus on slide transitions is one technique to summarise such videos. Other strategies rely on a combination of speech variants, audio content analysis, and speaker gesture analysis (Tiwari and Bhatnagar 2021; Sreeja and Kovoor 2019). A new method for summarising whiteboard/chalkboard lecture recordings is presented as demonstrated in Fully Convolutional Network (Davila et al. 2021) to extract handwritten material from video images with high recall and precision.

5 Recent contributions in DL based VS

This section will discuss the recent contributions made in the VS based on DL. In recent years, much work has been done in considering Spatio-temporal dependencies in the video in the supervised VS. Combination of a tensor train embedding layer and a multi LSTM forms Tensor-Train, TTH-RNN model (Zhao et al. 2020). The tensor embedding layer prevents the significant feature of hidden mapping matrices caused by the high-dimensional video features. The proposed Attentive and Distribution consistent VS (ADSum) (Ji et al. 2021b) approach deals with the two issues of VS. Those are short-term contextual attention insufficiency and distribution inconsistency. It Proposes an encoder self-attention mechanism for Seq2Seq learning-based VS for the former issue and a distribution-based loss function for the latter. The additive and multiplicative attention mechanisms in ADSUM are proposed as ADSUM-A and ADSUM-M (Modified).

Multiscale RNN (Zhu et al. 2022) presents a multiscale hierarchical attention technique. Unlike most existing supervised approaches, which use BiLSTM. This method uses the underlying hierarchical structure of video sequences. It uses intra-block and inter-block attention to learn short- and long-range temporal representations. DHAVS (Lin et al. 2022) uses 3DCNN instead of 2DCNN to provide a more effective and delicate video representation. It has designed cost sensitive loss function to address unbalanced class distribution for creating dynamic summaries.

Weakly supervised video summarization uses only a limited amount of labeled data in the training process according to the requirement of the method or approach. Hence reduce the human intervention for training the model. It is beneficial when a model requires costly labeled data or if there is a limited time for manual annotation. In recent works, a weakly supervised reinforcement learning method for video summarization proposed by Li and Yang (2021) is based on the Combination of networks.

Extended VDAN+ (Ramos et al. 2022) provides a unique, weakly-supervised way to summarize instructional videos using text based on an reinforcement learning concept. In this approach, an agent is guided by a novel joint reward function to choose which frames to eliminate and shorten the input video to the desired length without gaps in the output video.

Query-adaptive video summarization (QSVS) (Cizmeciler et al. 2022) technique incorporates saliency maps into a submodular optimization to consider query phrases. Both in capturing relevant images and representing similarity among shots. This study tested the proposed Activity Related Summaries dataset (ARS), and a subset of the Relevance and Diversity Dataset (RAD) (Vasudevan et al. 2017). Additionally, Unsupervised methods focus on developing heuristic or learning-based criteria for producing summaries of videos as a subset selection issue. The representativeness of frames or shots is frequently assessed using clustering methods. The keyframes or key shots are chosen from the cluster centers formed by the clustering of the video frames (Zhao et al. 2022).

A VS method for generic videos is named GVSUM (Basavarajaiah and Sharma 2021), which can be applied to many areas like sports, medicine, surveillance, movies, home, etc. The video frames are assigned a cluster number based on their visual features. Keyframes are extracted whenever there is a change in the cluster number of the frame. A pre-trained CNN is used to extract visual characteristics of the video on which k-means clustering is applied, followed by a sequential keyframe generation technique is applied to generate the generic summary.

Another proposed method (Nair and Mohan 2021) detects keyframes based on feature vectors extracted from Multi CNN (a combination of four pre-trained Convolutional Neural Network models), fed into Sparse Autoencoder, which in turn gives a combined illustration of the input feature vectors. The keyframes are extracted using Random Forest Classifier. This method performs very well as compared to the other SOTA methods. This model can also generate summaries of high quality from videos of all categories or genres. Also, further experiments have been done which prove that this model using the Random Forest classifier performs much well than any other classifiers.

The AC-SUM-GAN (Apostolidis et al. 2021) structure incorporates the model of an Actor-Critical into a GAN and treats choosing key fragments of input video as a sequence-generating job. The Actor and the Critics participate in a game of selecting key fragments of the source video. Their decisions at each stage of the match result in the Discriminators rewarding them with a set of rewards. The Actor and Critic can use the designed training workflow to locate a space of actions and automatically assess a policy for key fragment selection. Compared to unsupervised approaches, the proposed AC-SUM-GAN model performs well and yields SoTA results.

Unsupervised Wireless Capsule Endoscopy(WCE) was employed for video summarization (Lan and Ye 2021). WCE integrates VAE, Ptr-Net, and GAN approaches. At the same time, a discriminator with another LSTM is used to distinguish the original. A reconstructive sequence-graph network (RSGN) (Zhao et al. 2022) encodes frames and shots as sequences and graphs hierarchically, LSTM encoding frame-level dependencies and the GCN capturing shot-level relationships. The sequence-graph summary generator and the graph reconstructors have proven effective.

A framework CNN Bi-ConvLSTM-GAN for summarising videos is created using a generative adversarial network (GAN) (Sreeja and Kovoor 2022). The retrieved characteristics are made sure to encompass a variety of sample video fragments in the first step of the proposed adversarial learning model. The generator model is followed by a discriminator, which seeks to differentiate between the original and reconstructed video samples to improve the generator model's effectiveness. Following the adversarial network distillation phase, knowledge operates as a keyframe or segment selector using a primary network with input data obtained from the prior GAN model. Results from both public and custom datasets are thoroughly evaluated and given different expected summaries.

Certain unsupervised approaches summarize video by focusing on specific properties for generating the most optimum video summary. In this field, RL theory and some reward functions quantify the presence of desired characteristics in the developed outline. The Summarizer takes a sequence of video frames as input and produces a summary by predicting importance scores at the frame level. The generated summary is then sent to the evaluator responsible for quantifying desired features using manually created reward functions. The calculated results are then joined to get the overall reward value used to train the Summarizer (Hatcher and Yu 2018; Xu et al. 2021).

In this context, Xu et al. (2021) an RL-based unsupervised method is employed to summarize crowd surveillance videos. A novel crowd location-density reward is proposed to produce high-quality video summaries. The 3D Spatio-temporal -UNet (Ronneberger et al. 2015; Liu et al. 2022) RL framework for VS encodes Spatio-temporal information from the input movies for downstream RL efficiently. In this method, an RL agent gains knowledge from spatiotemporal latent scores and anticipates whether to keep or discard a video frame. This method can be used for both supervised and unsupervised training. Research demonstrates that 3D CNN features perform better than popular spatial 2D CNN features.

6 Datasets used in VS

The section provides a brief overview of the various relevant datasets available and different evaluation methods for VS. Some of the most commonly used VS datasets are TVSum (Song et al. 2015), SumMe (Gygli et al. 2014), CoSum (Chu et al. 2015), OVP,[1] Youtube (De Avila et al. 2011). Youtube High-light (Sun et al. 2014) as shown in Tables 8 and 9.

[1] https://open-video.org/detailed_search.php, Accessed on 13 March 2022

Table 8 Datasets for evaluating VS performance using DL models

Method	Lol	SumMe	Tvsum	OVP	YouTube	Tour	MED	Cosum	Thumk1K	ADL	VSUMM
DCCN (Sridevi and Kharde 2020)		✓			✓						
Multi-CNN (Nair and Mohan 2021)				✓							✓
GVSUM (Basavarajaiah and Sharma 2021)											✓
TTH-RNN (Zhao et al. 2020)		✓	✓				✓				
MOLRVS (Archana and Malmurugan 2021)					✓		✓				✓
LSTM (Mahasseni et al. 2017)		✓	✓	✓							
Adv-Ptr-Der-SUM (Lan and Ye 2021)		✓	✓								
Cycle-SUM (Yuan et al. 2019a)		✓	✓		✓			✓			
ADSum (Ji et al. 2021b)		✓	✓								
AC-SUM-GAN (Apostolidis et al. 2021)		✓	✓								
HCNN (Zhao et al. 2017)		✓	✓				✓				
DNN (Otani et al. 2016)		✓									
CNN Bi-ConvLSTM GAN (Sreeja and Kovoor 2022)				✓							✓
DHAVS (Lin et al. 2022)		✓	✓								
Multiscale LSTM (Zhao et al. 2022)		✓	✓								
UVS (Zhou et al. 2018a)		✓	✓								
3DST-UNet (Liu et al. 2022)		✓	✓								
SGSN (Li and Yang 2021)		✓	✓								
DeepRes (Muhammad et al. 2019)					✓						
DGCNN (Wu et al. 2020)			✓		✓	✓					
CNN (Purwanto et al. 2018)		✓	✓	✓	✓						
GAZE (Wu et al. 2018)		✓	✓								
Deep (Zhong et al. 2019)		✓	✓								
Multistage (Jappie et al. 2020)		✓	✓								
CNN (Yuan et al. 2019c)		✓	✓								
DSSE (Yuan et al. 2017)									✓		
GAN (Fu et al. 2019b)	✓	✓	✓		✓						

Table 8 (continued)

Method	Lol	SumMe	Tvsum	OVP	YouTube	Tour	MED	Cosum	Thumk1K	ADL	VSUMM
Fully CNN (Rochan et al. 2018)		✓	✓	✓	✓						
DASP (Ji et al. 2020)		✓	✓	✓	✓						
SUM-GAN (Apostolidis et al. 2020)		✓	✓								
PCDL (Zhao et al. 2019)		✓	✓	✓	✓						
SMN (Wang et al. 2019)		✓	✓								
DTR-GAN (Zhang et al. 2019a)		✓	✓	✓		✓					
CSNET (Jung et al. 2019)		✓	✓								
Online Motion-AE (Zhang et al. 2020)		✓	✓		✓			✓			
RNN (Zhong et al. 2018)		✓	✓		✓						
3DCNN (Panda et al. 2017)			✓					✓			
CSM (Sahu and Chowdhury 2020)		✓	✓					✓		✓	
ResNet (Fei et al. 2021)		✓	✓								
Blockchain (Khan et al. 2020b)											✓

Table 9 Datasets details used for VS

Dataset	#Videos	Duration (Min)	Description	Dataset (%) Popularity
Tvsum (Song et al. 2015)	50	02–10	50 Videos of news, documentaries, user-generated content (vlog, egocentric)	20.0
SumMe (Gygli et al. 2014)	25	01–06	Various Events, Holidays, Sports, etc	19.0
OVP (De Avila et al. 2011)	50	01–05	Videos for geners like Documentary, Educational, Historical, Lecture	04.4
VSUMM (De Avila et al. 2011)	50	01–10	Video categories varies from Cartoons, News, Commercials to TV-shows	02.5
MED (Potapov et al. 2014)	160	01–05	Videos of 10 events categories	01.8
Wild-8 (Zhong et al. 2016)	100	–	8 Categories (Scenes: Sky, Tree, Grass, Sand, Water & objects: Bird, Lion)	00.6
SUNY-24 (Zhong et al. 2016)	08	–	24 different categories scenes	00.6
MSR-VTT (Xu et al. 2016b)	10,000	2472	Most different categories and diversified visual content of phrase and vocab	00.6
UCF101 (Ho et al. 2018)	13,320	1620	101 activities grouped into 25 groups individually of 4–7 videotapes	00.6
Thumb1K (Yuan et al. 2017)	1037	–	Query-video pairs collected from Bing	00.6
Lol (Fu et al. 2019b)	218	–	Matches from a League of Legends tournament	00.6
Orange (Zhang et al. 2020)	30	–	Surveillance video sequences	00.6
ADL (Sahu and Chowdhury 2020)	20	120	Regular actions (hand-washing, face-washing, coffee making, Viewing TV)	00.6

- *TVSum* contains 50 videos from 10 categories: News, tutorials, user-generated content, and documentaries from TRECVid Multimedia Event Detection (MED) task (Smeaton et al. 2006). The range of videos is from 2 to 11 min, and each has been annotated by 20 people based on frame relevance and frame-level ratings of one to five.

- *SumMe* is a collection of 25 movies ranging in length from 1 to 6 min, and also it covers various topics like holidays, events, and different games. In numerous critical portions of the film, annotations are received from 15 to 18 users.
- *CoSum* Each topic contains videos with a total length of 4444 films ranging from 11 to 25 min. There are 51 films available, and each video spans around 147 min.
- *Thumk1K* Contains 10 various topics. Sky diving, Velopolo, Tower Bridge, Jcb Bridge Crossing, Youngsters Dancing in Trees, MLB, NFL, Notre Dame Cathedral, Browsing, and the National Historic Site were all videos collected from YouTube.
- *Open video project (OVP)* Contains 50 videos annotated with five different user keyframe sets. The videos range in length from just 1 to 4 min. Educational, transitory, scientific, comics, media, events, advertising, TV shows, and home movies are all part of OVP.
- *YouTube* Platform offers a video collection of 50, and the YouTube Highlight dataset includes a video collection of 100. Raw video annotations are made using AMT technology, with individual selections lasting around 5 s as actual value.
- *UCF101* is a package of real-time activity videos obtained from a Platform named YouTube with action consisting of 101 categories for action recognition. This data set supplements the UCF50 (Reddy and Shah 2013) data collection, which includes 50 different activities. The 101 activity types videos were grouped into 25 groups, each with four to seven action videos. Videos from the same group may have certain similarities, such as a similar setting, a similar point of view, and so on.
- *MSR-VTT* combining different 257 popular video search queries with 118 videos from a commercial video search engine. MSR-VTT now offers 10K online video clips with a total duration of 41.2 h and 200K clip-sentence pairings, encompassing the most extensive categories and visual content and the massive amount of data based on sentence structure and vocabulary.
- *LoL* is composed of 218 long videos where the time for each video is between 30 and 50 min. The annotation originates from a Channel on YouTube that includes 5 to 7-min videos of society highlight points.

Some other datasets are also exist which are rarely employed: Videos in the Wild (VTW) (Zeng et al. 2016), FVPSum (Ho et al. 2018), UCLA (Pei et al. 2011), MED-summaries (Potapov et al. 2014), YouCook2 (Zhou et al. 2018b) and COIN datasets (Ramos et al. 2020b).

- *MED-Summaries* consists of 160 annotated videos from 10 different event categories collected from TRECVID 2011 MED Dataset (Smeaton et al. 2006).
- *VTW* consists of about 18,100 videos, with 2000 rated at the sub-short level to summarize videos. The average duration of the video is 1.5 min.
- *FPVSum*is a dataset for a first-person video summary that has 98 videos in total from 14 categories of easy-to-watch GoPro videos of various lengths over 7 h in total. The

Table 10 Static summary on TVSum datasets based on F-Score (%)

VS model	F-score (%)
3DST-Unet-Reinforcement Learning (Liu et al. 2022)	58.1
ADV-PTR-DER-SUM (Lan and Ye 2021)	58.3
TTHRNN:Tensor-Train Hierarchical Recurrent Neural Network for Video Summarization (Zhao et al. 2020)	62.3
Deep Attentive Preserving (Ji et al. 2020)	63.6
CNN Bi-convolutional Long short-term memory GAN (Sreeja and Kovoor 2022)	**69.0**

The best results are depicted in bold

remaining videos are untagged. Approximately 35% of video sequences within each category are annotated with ground-truth ratings by at least 10 individuals.

- *UCLA* is made up of three single and two-person activity surveillance recordings. These three video episodes total about 35 min in length. The remaining 60 movies are validation sets from 15 various types, with the majority lasting between 1 and 5 min. Annotations are a collection of critical scores averaged across one to four annotators.
- *YouCook2* is an extensive dataset made up of unconstrained YouTube cooking videos that cover a wide range of cuisines and cooking methods. It contains 2,000 videos spread among 89 recipes, totaling 176 h.
- *COIN*is a large-scale dataset comprising 11,827 instructional videos split among 180 tasks and a total length of 476 h, arranged hierarchically into 12 categories.

7 Performance measures

The following sections provides a brief overview of different evaluation methods for VS.

7.1 Static VS evaluation

The evaluation was founded on detailed standards, like the applicability of individual keyframes to the videotape content, repetitious or missing knowledge (de Avila et al. 2008), or the instructive importance and comfort of the conceptual (Ejaz et al. 2014). While this can exist as a proper evaluation procedure, the evaluation procedure for individually resulting summary can be time-consuming for the user. The evaluation outcomes cannot be readily duplicated or used for coming comparisons. The current assignment considers its keyframe-based outlines with ground-truth overviews or employs some objective standards to overwhelm the drawbacks of user analyses. In this context, (Chasanis et al. 2008) assesses the rate of the created outlines using the standard of commitment (Chang et al. 1999) and the measure of the capacity of reconstruction of the image (Liu et al. 2004), comparison of user summaries is presented in (De Avila et al. 2011). The F- score evaluation of static summary on dataset TVSum, SumMe, OVP, YouTube, VSUMM are shown in Tables 10, 11, 12, 13 and 14 respectively.

7.2 Dynamic VS evaluation

Initially, user-created datasets are employed for an accurate estimate. Later, F-score is used in VS (Gygli et al. 2014; Song et al. 2015). Based on a set of predefined criteria as in BBC

Table 11 Static summary performance on SumMe dataset based on F-Score (%)

VS model	F-score (%)
ADV-PTR-DER-SUM (Lan and Ye 2021)	43.6
3D Spatio-Temporal U-Net Reinforcement Learning (Liu et al. 2022)	44.6
Tensor-Train Hierarchical Recurrent Neural Network for Video Summarization (Zhao et al. 2020)	45.0
Deep Attentive Preserving (Ji et al. 2020)	**45.5**

The best results are depicted in bold

Table 12 Static summary of recent work in VS on OVP dataset based on F-Score (%)

VS model	F-score (%)
PCDL:Property-Constrained Dual Learning for Video Summarization (Zhao et al. 2019)	78.3
Convolution Neural Network (Purwanto et al. 2018)	78.7
Multi Convolutional Neural Networks (Nair and Mohan 2021)	**82.0**

The best results are depicted in bold

Table 13 Static summary of recent work in VS on Youtube dataset based on F-Score (%)

VS model	F-score (%)
Convolution Neural Network (Purwanto et al. 2018)	63.2
PCDL:Property-Constrained Dual Learning for Video Summarization (Zhao et al. 2019)	63.4
GAN:Generative Adversarial Network (Fu et al. 2019b)	69.7
DeepRes(Deep Learning-Based video summarization) (Muhammad et al. 2019)	79.0
Multi-Edge Optimized LSTM RNN for VS (Archana and Malmurugan 2021)	**85.8**

The best results are depicted in bold

Table 14 Static summary of recent work in VS on VSUMM dataset based on F-Score (%)

VS model	F-score (%)
Blockchain (Khan et al. 2020b)	53.7
Convolutional Neural Network Bi-Convolutional LSTM GAN (Sreeja and Kovoor 2022)	68.0
GVSUM:Generic Video Summarization (Basavarajaiah and Sharma 2021)	78.0
Multi Convolutional Neural Networks (Nair and Mohan 2021)	83.0
MOLRVS(Multi-Edge Optimized LSTM RNN for VS(MOLRVS)) (Chai et al. 2021)	**92.4**

The best results are depicted in bold

rushes VS evaluation is done in TRECVID workshop (Smeaton et al. 2006; Over et al. 2008). An evaluation process was introduced concurrently with the SumMe dataset for a videotape summary (Song et al. 2015).

Table 15 Dynamic summary performance on TVSum datasets based on F-Score (%)

VS model	F-score (%)
Online Motion-AE (Zhang et al. 2020)	51.5
Summary Generation Sub-Network (Li and Yang 2021)	55.7
Cycle-SUM (Yuan et al. 2019a)	57.6
Reconstructive Sequence-Graph Network (Zhao et al. 2022)	58.0
AC-SUM-generative adversarial network (Apostolidis et al. 2021)	60.6
DHAVS (Lin et al. 2022)	60.8
Dilated Temporal Relational Generative Adversarial Network (Zhang et al. 2019a)	61.5
Deep Attentive Video Summarization with Distribution Consistency Learning-A -M (Ji et al. 2021b)	64.3
Deep Attentive Video Summarization with Distribution Consistency Learning-A -A (Ji et al. 2021b)	64.5
Convolutional Neural Network Bi-Convolutional LSTM GAN (Sreeja and Kovoor 2022)	**72.0**

The best results are depicted in bold

Table 16 Dynamic summary on SumMe dataset based on F-Score (%)

VS model	F-score (%)
Online Motion-Auto Encoder (Zhang et al. 2020)	37.7
DHAVS (Lin et al. 2022)	40.6
Summary Generation Sub-Network (SGSN) (Li and Yang 2021)	41.5
Cycle-SUM (Yuan et al. 2019a)	41.9
Reconstructive Sequence-Graph Network (Zhao et al. 2022)	42.3
Deep Attentive Video Summarization with Distribution Consistency Learning-A (Ji et al. 2021b)	45.9
Deep Attentive Video Summarization with Distribution Consistency Learning-M (Ji et al. 2021b)	46.1
AC-SUM-GAN (Apostolidis et al. 2021)	50.8
Dilated Temporal Relational Adversarial Network for Video Summarization (Zhang et al. 2019a)	**51.4**

The best results are depicted in bold

Table 17 Dynamic summary on Youtube dataset based on F-Score (%)

VS model	F-score (%)
Cycle-SUM(Supervised) (Yuan et al. 2019a)	64.2
Cycle-SUMmarization: Cycle-SUM (Unsupervised) (Yuan et al. 2019a)	**77.3**

The best results are depicted in bold

An evaluation process was introduced concurrently with the SumMe dataset for a videotape summary (Apostolidis et al. 2021). Another method utilize the Mathews (Matthews 1975) correlation coefficient to evaluate implementation. A different approach was used in (Mahasseni et al. 2017; Yuan et al. 2019a) to estimate submission using a single ground-truth summary instead of considerable user summaries. The F-score evaluation of dynamic summary on dataset TVSum, SumMe, YouTube, are shown in Tables 15, 16 and 17 respectively.

The following observations are made from Tables 10, 11, 12, 13, 14, 15, 16 and 17 based on the F-score evaluation of the static and dynamic video summaries.

- The TVSum dataset is the maximum used dataset by many of the existing video summarization techniques during the last decade, where the Convolutional Neural Network Bi-Convolutional Long Short Term Memory Generative Adversarial Network method (Sreeja and Kovoor 2022) is outperformed on TVSum dataset with an F-score of 69.0% on the static summary as mentioned in Table 10.
- From Table 11, it is observed that the SumMe dataset is the second most highly used dataset by the exiting video summarization techniques during the last decade, where Deep Attentive Preserving (Ji et al. 2020) is outperformed with an F-score of 45.5% to generate the static summary on SumMe dataset.
- Multi Convolutional Neural Network outperformed on Open Video Project dataset in terms of F-score of 82.0% to generate the static summary over Convolutional Neural Network (Purwanto et al. 2018), and Property Constrained Dual Learning (Zhao et al. 2019) as shown in Table 12.
- From Table 13, it is observed that the Multi-edge optimized LSTM RNN for video summarization (Chai et al. 2021) approach outperformed the recent video summarization techniques on the YouTube dataset in terms of F-score of 85.8% for generating the static summary.
- On a static summary using the VSUMM dataset, the Multi-edge optimized LSTM RNN for video summarization (Chai et al. 2021) approach delivered the best results compared to other video summarization approaches employed recently where the F-score value for the static summary is 92.4%, as shown in Table 14.
- From Table 15, it is noticed that the F-score value for the Convolutional Neural Network Bi-Convolutional Long Short Term Memory Generative Adversarial Network method (Sreeja and Kovoor 2022) for generating the dynamic summary on the Tvsum dataset is 72.0%, better than the other exiting techniques.
- The Dilated Temporal Relational-Generative Adversarial Network (Zhang et al. 2019a) method performed well on the dynamic summary using SumMe dataset with the F-score value of 51.4 % as shown in Table 16.
- From Table 17, it is observed that the unsupervised learning-based Cycle-SUM (Yuan et al. 2019a) method outperformed with the F-score value of 77.3 % to generate the dynamic summary.

8 Challenges

A video captures spatial and temporal data comprising frames, shots, and scenes has a hierarchical structure. The final video includes sets that combine images and a grouping of frames, which are the elementary unit of the video. Hence VS presents the following challenges:

- Multimodal: A video stream's multimodal nature, which may include pictures, audio, videotape, rotating images, and documents (text), is more complex than summarising any other type of content. It is challenging to generate summaries using high-level features from different video categories, as discussed in (Nair and Mohan 2021). 3D CNN (Liu et al. 2022; Lin et al. 2022) is used for high-level features but faces complexity and computation challenges.
- Spatio-temporal dependencies: The design of the potential architecture of VS model to capture Spatio-temporal dependencies is a complex task. Exploiting high-dimensional video features, whether shallow or deep, to represent such a massive amount of information is challenging, as stated in (Zhao et al. 2020; Zhu et al. 2022; Apostolidis et al. 2021; Lin et al. 2022).
- User Subjectivity: The fact that a single video may create numerous summaries based on the customer's preferences (Matthews 1975). As a result, no video summarizer can meet the needs of all users unless it interacts with them and adjusts to their needs. As in query-specific VS (Cizmeciler et al. 2022), it requires understanding the present visual data and the given textual queries, which is user specific.
- Generating importance scores: Importance scores are used to determine whether frames or segments of a video are significant or not (Matthews 1975). However, (Li and Yang 2021) discusses that procedures for significance vary for various peoples and are based on multiple aspects such as summary classification, people conditions, and video type.
- Evaluation of summaries: It is primarily due to the lack of a single qualitative and quantitative evaluation metric. F-score, precision, recall, and accuracy are standard quantitative evaluation metrics. One method is requesting viewers to evaluate the rate of quality of the summary. Informativeness, coverage, and rank are some of the other criteria (Ramos et al. 2022; Xu et al. 2021; Matthews 1975).
- Application based: Range of Applications is vast, but mostly work in VS is based on sports, query-focused, or surveillance genres. Sports videos are complex and differ in their rules, statistics, directions, and several exciting events (Guntuboina et al. 2022; Tejero-de Pablos et al. 2018). Other applications like user-generated videos, education, and documentaries are quite challenging (Sridevi and Kharde 2020; Wu et al. 2020; Ji et al. 2021a; Davila et al. 2021). User-generated videos can be of inferior quality and resolution, whereas in VS, educational videos rely on the combination of speech, audio, and gesture analysis.
- Storage and computation: VS based on DL required a massive amount of annotated data for learning which requires more Storage and more complexity. However, most datasets are small, and hence scope of VS techniques on these datasets is minimal. Annotating large-scale BIG datasets, especially surveillance videos, is very challenging considering Storage and computation. These data are more diverse, complex, unstructured, and high-dimensional. Also, Application of VS based on DL is unexplored mainly and ineffective when data comes to as streaming data (Muhammad et al. 2019; Xu et al. 2021).
- Data mining and information retrieval: Extracting meaningful information from the huge amount of generated raw videos is complicated. The data mining process also comes with its challenges since these videos are unstructured and may have quality issues. It is challenging in video data to consider perceptual content like color, intensity, and so on, and semantic content includes visual objects, events, and their relationships (Najafabadi et al. 2015; Muhammad et al. 2019; Xu et al. 2021) for information retrieval.

9 Future direction

DL represents a significant advancement in neural networking technologies. It uses several stages of processing information in a network to categorize features extracted and comprises an input layer containing essential data. Numerous hidden layers examine the data and produce results. In recent years, it has increased in popularity. The current VS can enhance by applying the most recent technique. DL algorithms are divided into three groups based on their architecture: unsupervised, supervised, and hybrid.

AE and RBM are two approaches used in unsupervised DL. ANN and AE utilize hidden layers the same way, but the autoencoder consists of only three hidden layers. The input and output levels have almost the same nodes. Hidden layer nodes are utilized to lower the convolution neural network, one kind of supervised learning that allows quick understanding. Both strategies are used in a hybrid approach. DNNs are an example of hybrid architecture. DNNs create cascaded multilayer networks by providing a fully connected hidden layer.

Deep Learning has been used to classify images. Adopting Deep Learning for event recognition in video summarization is a challenge. Deep Learning for event detection is based on converting code to images and then using CNN to learn features using images as input. Traditional detection outcomes are improved when combined with supervised and unsupervised Deep Learning algorithms. The generation of new event detection systems is more reliable than existing machine learning methods. As it does the extensive data analysis, Deep Learning adapts to the changing context of the data.

On the other hand, DL for analysis remains an unexplored and challenging field for researchers. Videos are multi-dimensional and include voice, motion, and a time-series dimension in addition to being simply a collection of numerous frames or images. Each of these factors is essential to understanding a video, and different elements may be necessary depending on the target audience for the summarization.

Longformer EncoderDecoder (LRD) and low computation should be considered in supervised VS while using RNN or LSTM. There is also much scope for improving learning from unpaired data. If we believe in unsupervised VS, the summaries often lack context and appear disjointed. In weakly supervised and reinforcement VS, different reward functions can be defined, which play an essential role in selecting keyframes or shots. GAN creates unique summaries for videos while keeping the context and significance of the original videos intact. Since it is limited to SLV, it should be designed with advanced mechanisms to maintain LRD. Concerning VS, these techniques are just the beginning of a new era in deep learning technology. Shortly, many advancements will be made to produce and optimize the best summaries based on the intended audience, delivery method, and summary goal.

10 Conclusion

This paper discusses procedures, challenges, and applications for VS. Additionally reviews DL-based video summarising methods, algorithms, techniques, and approaches. The most promising method for unsupervised VS appears to be utilizing GANs to create a representative video description. There is some variation in the evaluation protocols used for VS techniques, which is related to how the employed data are divided for training and testing purposes and the number of experiments conducted using various randomly generated splits of the data of the SumMe and TVSum datasets. The TVSum dataset is mainly employed by most of the existing video summarization techniques during the last decade.

The Multi-edge optimized LSTM RNN for VS approach beat the best F-score 92.4% to generate the static summary on the VSUMM dataset over other recent VS techniques. In addition, to solve the shortage of information in conveying various VS efforts regarding the procedure used for deciding which trained model to keep and which to discard. To solve this problem for each newly proposed approach, the relevant community should be aware of these concerns and take the appropriate steps.

The advanced techniques for VS could be based on Feature-Based VS, Clustering Based VS, Shot Selection, Event-Based, Trajectory Based VS. Deep learning-based methods for VS can be supervised, weakly supervised, unsupervised, and reinforcement learning. Supervised VS techniques are DBN, DNN, CNN, RNN, LSTM, and Bi-LSTM. Unsupervised VS includes DL models like GAN, RBM, and AE. Deep Learning based VS applications consists of sports, query-focused, surveillance, movies, news documentaries, medical, and educational.

Recent Contributions in VS using DL includes AC-SUM-GAN structure incorporating the actor-critical model into a GAN and treating choosing key fragments of input video as a sequence-generating job. Different challenges are present in VS like multimodal, user subjectivity, spatio-temporal dependencies, generating importance scores, evaluation of summaries, Storage and computation, Data mining, and information retrieval. User Subjectivity is the biggest challenge in VS. A single video may create numerous summaries based on the customer's preferences. As a result, no video summarizer can meet the needs of all users unless it interacts with them and adjusts to their needs.

For future research, it is suggested that additional efforts be put into VS algorithms for real-world applications by incorporating such techniques to support the demands of the current multimedia management system for quick adaptation, process, storage, retrieval, and reuse of video content. The potential applications could be intelligence transport systems, educational videos, entertainment, movies, video games, security surveillance etc.

Acknowledgements The authors would like to thank to the DST GoI for sponsoring the work under DST/ICPS/General/2018.

Data availability NA.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Agnihotri L, Devara KV, McGee T, Dimitrova N (2001) Summarization of video programs based on closed captions. In: Storage and retrieval for media databases 2001, vol 4315, SPIE, Bellingham, pp 599–607

Ajmal M, Ashraf MH, Shakir M, Abbas Y, Shah FA (2012) Video summarization: techniques and classification. In: International conference on computer vision and graphics, Springer, pp 1–13

Alok N, Krishan K, Chauhan P (2021) Deep learning-based image classifier for malaria cell detection. In: Machine learning for healthcare applications, pp 187–197

Alom MZ, Taha TM, Yakopcic C, Westberg S, Sidike P, Nasrin MS, Hasan M, Van Essen BC, Awwal AA, Asari VK (2019) A state-of-the-art survey on deep learning theory and architectures. Electronics 8(3):292

Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, Santamaría J, Fadhel MA, Al-Amidie M, Farhan L (2021) Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 8(1):1–74

Apostolidis E, Adamantidou E, Metsai AI, Mezaris V, Patras I (2020) Unsupervised video summarization via attention-driven adversarial learning. In: International conference on multimedia modeling, Springer, pp 492–504

Apostolidis E, Adamantidou E, Metsai AI, Mezaris V, Patras I (2021) Ac-sum-gan: connecting actor-critic and generative adversarial networks for unsupervised video summarization. IEEE Trans Circuits Syst Video Technol 31(8):3278–3292

Archana N, Malmurugan N (2021) Multi-edge optimized lstm rnn for video summarization. J Ambient Intell Humaniz Comput 12(5):5381–5395

Barbieri M, Agnihotri L, Dimitrova N (2003) Video summarization: methods and landscape. In: Internet multimedia management systems IV, vol 5242, SPIE, pp 1–13

Basavarajaiah M, Sharma P (2019) Survey of compressed domain video summarization techniques. ACM Comput Surv (CSUR) 52(6):1–29

Basavarajaiah M, Sharma P (2021) Gvsum: generic video summarization using deep visual features. Multimed Tools Appl 80(9):14459–14476

Bengio Y (2009) Learning deep architectures for AI. Found Trends Mach Learn 2(1):1–127. https://doi.org/10.1561/2200000006

Binol H, Niazi MK, Elmaraghy C, Moberly AC, Gurcan MN (2021) Automated video summarization and label assignment for otoscopy videos using deep learning and natural language processing. In: Medical imaging 2021: imaging informatics for healthcare, research, and applications, vol 11601, SPIE, pp 153–158

Brezeale D, Cook DJ (2008) Automatic video classification: a survey of the literature. IEEE Trans Syst Man Cybern C 38(3):416–430

Chai J, Zeng H, Li A, Ngai EW (2021) Deep learning in computer vision: a critical review of emerging techniques and application scenarios. Mach Learn Appl 6:100134

Chang HS, Sull S, Lee SU (1999) Efficient video indexing scheme for content-based retrieval. IEEE Trans Circuits Syst Video Technol 9(8):1269–1279

Chasanis V, Likas A, Galatsanos N (2008) Efficient video shot summarization using an enhanced spectral clustering approach. In: International conference on artificial neural networks, Springer, pp 847–856

Chauhan P, Mandoria HL, Negi A (2021a) Deep residual neural network for plant seedling image classification. In: Agricultural informatics: automation using the IoT and machine learning, pp 131–146

Chauhan P, Mandoria HL, Negi A, Rajput RS (2021b) Plant diseases concept in smart agriculture using deep learning. In: Smart agricultural services using deep learning, big data, and IoT. IGI Global, Hershey, pp 139–153

Chollet F (2017) Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1251–1258

Chootong C, Shih TK, Ochirbat A, Sommool W, Zhuang YY (2021) An attention enhanced sentence feature network for subtitle extraction and summarization. Expert Syst Appl 178:114946

Chu WS, Song Y, Jaimes A (2015) Video co-summarization: video summarization by visual co-occurrence. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3584–3592

Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555

Cizmeciler K, Erdem E, Erdem A (2022) Leveraging semantic saliency maps for query-specific video summarization. Multimed Tools Appl 81(12):17457–17482

Coppola C, Cosar S, Faria DR, Bellotto N (2020) Social activity recognition on continuous rgb-d video sequences. Int J Soc Robot 12(1):201–215

Dargan S, Kumar M, Ayyagari MR, Kumar G (2020) A survey of deep learning and its applications: a new paradigm to machine learning. Arch Comput Methods Eng 27(4):1071–1092

Davila K, Fei X, Setlur S, Govindaraju V (2021) Fcn-lecturenet: extractive summarization of whiteboard and chalkboard lecture videos. IEEE Access 9:104469–104484

de Avila SE, da_Luz Jr A, Araújo AD, Cord M (2008) VSUMM: an approach for automatic video summarization and quantitative evaluation. In: 2008 XXI Brazilian symposium on computer graphics and image processing, IEEE, pp 103–110

De Avila SE, Lopes AP, da Luz Jr A, de Albuquerque Araújo A (2011) Vsumm: a mechanism designed to produce static video summaries and a novel evaluation method. Pattern Recognit Lett 32(1):56–68

Del Molino AG, Tan C, Lim JH, Tan AH (2016) Summarization of egocentric videos: a comprehensive survey. IEEE Trans Hum-Mach Syst 47(1):65–76

Dimitrova N, Zimmerman J, Janevski A, Agnihotri L, Haas N, Bolle R (2003) Content augmentation aspects of personalized entertainment experience. In: Proceedings of the third workshop on personalization in future TV, pp 42–51

Dong S, Wang P, Abbas K (2021) A survey on deep learning and its applications. Comput Sci Rev 40:100379. https://doi.org/10.1016/j.cosrev.2021.100379

Ejaz N, Mehmood I, Baik SW (2014) Feature aggregation based visual attention model for video summarization. Comput Electr Eng 40(3):993–1005

Emon SH, Annur AH, Xian AH, Sultana KM, Shahriar SM (2020) Automatic video summarization from cricket videos using deep learning. In: 2020 23rd international conference on computer and information technology (ICCIT), IEEE, pp 1–6

Fei M, Jiang W, Mao W (2021) Learning user interest with improved triplet deep ranking and web-image priors for topic-related video summarization. Expert Syst Appl 166:114036

Fu T, Tai S, Chen H-T (2019a) Attentive and adversarial learning for video summarization. In: 2019 IEEE winter conference on applications of computer vision (WACV), IEEE, pp 1579–1587

Fu T-J, Tai S-H, Chen H-T (2019b) Attentive and adversarial learning for video summarization. In: 2019 IEEE winter conference on applications of computer vision (WACV), IEEE, pp 1579–1587

Gers FA, Schmidhuber J, Cummins F (2000) Learning to forget: continual prediction with lstm. Neural Comput 12(10):2451–2471

Gers FA, Schraudolph NN, Schmidhuber J (2002) Learning precise timing with lstm recurrent networks. J Mach Learn Res 3:115–143

Gonuguntla N, Mandal B, Puhan NB (2019) Enhanced deep video summarization network. In: 30th British Machine Vision Conference, 9–12 Sep 2019, Cardiff

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial networks. arXiv preprint arXiv:1406.2661

Guntuboina C, Porwal A, Jain P, Shingrakhia H (2022) Video summarization for multiple sports using deep learning. In: Proceedings of the international e-conference on intelligent systems and signal processing, Springer, pp 643–656

Gygli M, Grabner H, Riemenschneider H, Gool L (2014) Creating summaries from user videos. In: European conference on computer vision, Springer, pp 505–520

Hatcher WG, Yu W (2018) A survey of deep learning: platforms, applications and emerging research trends. IEEE Access 6:24411–24432

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778

Ho H-I, Chiu W-C, Wang Y-CF (2018) Summarizing first-person videos from third persons' points of view. In: Proceedings of the European conference on computer vision (ECCV), pp 70–85

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780

Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861

Hu W, Xie N, Li L, Zeng X, Maybank S (2011) A survey on visual content-based video indexing and retrieval. IEEE Trans Syst Man Cybern C 41(6):797–819

Huang G, Liu Z, van der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708

Hussain T, Muhammad K, Ullah A, Cao Z, Baik SW, de Albuquerque VHC (2019) Cloud-assisted multiview video summarization using cnn and bidirectional lstm. IEEE Trans Ind Inform 16(1):77–86

Hussain T, Muhammad K, Ding W, Lloret J, Baik SW, de Albuquerque VHC (2021) A comprehensive survey of multi-view video summarization. Pattern Recognit 109:107567

Jappie Z, Torpey D, Celik T (2020) Summarynet: a multi-stage deep learning model for automatic video summarisation. arXiv preprint arXiv:2002.09424

Ji Z, Xiong K, Pang Y, Li X (2019) Video summarization with attention-based encoder-decoder networks. IEEE Trans Circuits Syst Video Technol 30(6):1709–1717

Ji Z, Jiao F, Pang Y, Shao L (2020) Deep attentive and semantic preserving video summarization. Neurocomputing 405:200–207

Ji N, Zhao S, Lin Q, Yu D, Zhao Y (2021a) NBA basketball video summarization for news report via hierarchical-grained deep reinforcement learning. In: International conference on image and graphics, Springer, pp 712–728

Ji Z, Zhao Y, Pang Y, Li X, Han J (2021b) Deep attentive video summarization with distribution consistency learning. IEEE Trans Neural Netw Learn Syst 32(4):1765–1775. https://doi.org/10.1109/TNNLS.2020.2991083

Jung Y, Cho D, Kim D, Woo S, Kweon IS (2019) Discriminative feature learning for unsupervised video summarization. In: Proceedings of the AAAI Conference on artificial intelligence, vol 33, pp 8537–8544

Khan AA, Shao J, Ali W, Tumrani S (2020a) Content-aware summarization of broadcast sports videos: an audio-visual feature extraction approach. Neural Process Lett 52(3):1945–1968

Khan G, Jabeen S, Khan MZ, Khan MUG, Iqbal R (2020b) Blockchain-enabled deep semantic video-to-video summarization for iot devices. Comput Electr Eng 81:106524

Kumar K (2019) Evs-dk: event video skimming using deep keyframe. J Vis Commun Image Represent 58:345–352

Kumar K (2021) Text query based summarized event searching interface system using deep learning over cloud. Multimed Tools Appl 80(7):11079–11094

Kumar K, Shrimankar DD (2017) F-des: fast and deep event summarization. IEEE Trans Multimed 20(2):323–334

Kumar K, Shrimankar DD (2018a) ESUMM: event summarization on scale-free networks. IETE Technical Review

Kumar K, Shrimankar DD (2018b) Deep event learning boost-up approach: delta. Multimed Tools Appl 77(20):26635–26655

Kumar K, Shrimankar DD, Singh N (2016) Equal partition based clustering approach for event summarization in videos. In: 2016 12th international conference on signal-image technology & internet-based systems (SITIS), IEEE, pp 119–126

Kumar K, Shrimankar DD, Singh N (2018) Eratosthenes sieve based key-frame extraction technique for event summarization in videos. Multimed Tools Appl 77(6):7383–7404

Lan L, Ye C (2021) Recurrent generative adversarial networks for unsupervised wce video summarization. Knowl-Based Syst 222:106971

Lee S, Sung J, Yu Y, Kim G (2018) A memory network approach for story-based temporal summarization of 360 videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1410–1419

Li Z, Yang L (2021) Weakly supervised deep reinforcement learning for video summarization with semantically meaningful reward. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 3239–3247

Li Y, Ming W, Kuo CCJ (2001) Semantic video content abstraction based on multiple cues. In: IEEE international conference on multimedia and expo, 2001. ICME 2001, IEEE Computer Society, pp 159–159

Li Y, Lee S-H, Yeh C-H, Kuo CCJ (2006) Techniques for movie content analysis and skimming: tutorial and overview on video abstraction techniques. IEEE Signal Process Mag 23(2):79–89

Lin J, Zhong S, Fares A (2022) Deep hierarchical lstm networks with attention for video summarization. Comput Electr Eng 97:107618

Liu T, Zhang X, Feng J, Lo K-T (2004) Shot reconstruction degree: a novel criterion for key frame selection. Pattern Recognit Lett 25(12):1451–1457

Liu T, Meng Q, Vlontzos A, Tan J, Rueckert D, Kainz B (2020) Ultrasound video summarization using deep reinforcement learning. In: International conference on medical image computing and computer-assisted intervention, Springer, pp 483–492

Liu T, Meng Q, Huang J-J, Vlontzos A, Rueckert D, Kainz B (2022) Video summarization through reinforcement learning with a 3d spatio-temporal u-net. IEEE Trans Image Process 31:1573–1586

Ma YF, Lu L, Zhang HJ, Li M (2002) A user attention model for video summarization. In: Proceedings of the tenth ACM international conference on Multimedia, pp 533–542

Mahasseni B, Lam M, Todorovic S (2017) Unsupervised video summarization with adversarial lstm networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 202–211

Mahmoud KM, Ghanem NM, Ismail MA (2013) Unsupervised video summarization via dynamic modeling-based hierarchical clustering. In: 2013 12th international conference on machine learning and applications, vol 2, IEEE, pp 303–308

Matthews BW (1975) Comparison of the predicted and observed secondary structure of t4 phage lysozyme. Biochim Biophys Acta (BBA)-Protein Struct 405(2):442–451

Money AG, Agius H (2008) Video summarisation: a conceptual framework and survey of the state of the art. J Vis Commun Image Represent 19(2):121–143

Muhammad K, Hussain T, Del Ser J, Palade V, De Albuquerque VH (2019) Deepres: a deep learning-based video summarization strategy for resource-constrained industrial surveillance scenarios. IEEE Trans Ind Inform 16(9):5938–5947

Muhammad K, Hussain T, Baik SW (2020) Efficient cnn based summarization of surveillance videos for resource-constrained devices. Pattern Recognit Lett 130:370–375

Nair MS, Mohan J (2021) Static video summarization using multi-cnn with sparse autoencoder and random forest classifier. Signal Image Video Process 15(4):735–742

Najafabadi Maryam M, Flavio V, Khoshgoftaar Taghi M, Naeem S, Randall W, Edin M (2015) Deep learning applications and challenges in big data analytics. J Big Data 2(1):1–21

Navamani TM (2019) Efficient deep learning approaches for health informatics. In: Deep learning and parallel computing environment for bioengineering systems. Elsevier, Amsterdam, pp 123–137

Negi A, Kumar K (2021a) Classification and detection of citrus diseases using deep learning. In: Data science and its applications, Chapman and Hall/CRC, Boca Raton, pp 63–85

Negi A, Kumar K (2021b) Face mask detection in real-time video stream using deep learning. In: Computational intelligence and healthcare informatics, pp 255–268

Negi A, Chauhan P, Kumar K, Rajput RS (2020) Face mask detection classifier and model pruning with keras-surgeon. In: 2020 5th IEEE international conference on recent advances and innovations in engineering (ICRAIE), IEEE, pp 1–6

Negi A, Kumar K, Chauhan P (2021a) Deep neural network-based multi-class image classification for plant diseases. In: Agricultural informatics: automation using the IoT and machine learning, pp 117–129

Negi A, Kumar K, Chauhan P, Rajput RS (2021b) Deep neural architecture for face mask detection on simulated masked face dataset against covid-19 pandemic. In: 2021 international conference on computing, communication, and intelligent systems (ICCCIS), . IEEE, pp 595–600

Otani M, Nakashima Y, Rahtu E, Heikkilä J, Yokoya N (2016) Video summarization using deep semantic features. In: Asian conference on computer vision, Springer, pp 361–377

Over P, Smeaton AF, Awad G (2008) The trecvid 2008 bbc rushes summarization evaluation. In: Proceedings of the 2nd ACM TRECVid video summarization workshop, pp 1–20

Panda R, Das A, Wu Z, Ernst J, Roy-Chowdhury AK (2017) Weakly supervised summarization of web videos. In: Proceedings of the IEEE international conference on computer vision, pp 3657–3666

Pei M, Jia Y, Zhu S-C (2011) Parsing video events with goal inference and intent prediction. In: 2011 international conference on computer vision, IEEE, pp 487–494

Peker K, Bashir F (2007) Content-based video summarization using spectral clustering, September 27 (2007). US Patent App. 11/361,829

Pereira MH, Pádua FL, Dalip DH, Benevenuto F, Pereira AC, Lacerda AM (2019) Multimodal approach for tension levels estimation in news videos. Multimed Tools Appl 78(16):23783–23808

Phung VH, Rhee EJ (2019) A high-accuracy model average ensemble of convolutional neural networks for classification of cloud image patches on small datasets. Appl Sci 9(21):4500

Potapov D, Douze M, Harchaoui Z, Schmid C (2014) Category-specific video summarization. In: European conference on computer vision, Springer, pp 540–555

Pouyanfar S, Sadiq S, Yan Y, Tian H, Tao Y, Reyes MP, Shyu ML, Chen SC, Iyengar SS (2018) A survey on deep learning: algorithms, techniques, and applications. ACM Comput Surv (CSUR) 51(5):1–36

Purwanto D, Chen YT, Fang WH, Wu WC (2018) Video summarization: how to use deep-learned features without a large-scale dataset. In: 2018 9th international conference on awareness science and technology (iCAST), IEEE, pp 220–225

Ramos W, Silva M, Araujo E, Marcolino LS, Nascimento E (2020a) Straight to the point: fast-forwarding videos via reinforcement learning using textual data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10931–10940

Ramos W, Silva M, Araujo E, Marcolino LS, Nascimento E (2020b) Straight to the point: fast-forwarding videos via reinforcement learning using textual data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10931–10940

Ramos W, Silva M, Araujo E, Moura V, Oliveira K, Marcolino LS, Nascimento ER (2022) Text-driven video acceleration: a weakly-supervised reinforcement learning method. IEEE Trans Pattern Anal Mach Intell 45(2):2492–2504

Reddy KK, Shah M (2013) Recognizing 50 human action categories of web videos. Mach Vis Appl 24(5):971–981

Redmon J, Farhadi A (2018) Yolov3: an incremental improvement. arXiv preprint arXiv:1804.02767

Rochan M, Ye L, Wang Y (2018) Video summarization using fully convolutional sequence networks. In: Proceedings of the European conference on computer vision (ECCV), pp 347–363

Mrigank R, Mahesh KKR, Yang W (2020) Sentence guided temporal modulation for dynamic video thumbnail generation. arXiv preprint arXiv:2008.13362

Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention, Springer, pp 234–241

Sahu A, Chowdhury AS (2020) Summarizing egocentric videos using deep features and optimal clustering. Neurocomputing 398:209–221

Senthil Murugan A, Suganya Devi K, Sivaranjani A, Srinivasan P (2018) A study on various methods used for video summarization and moving object detection for video surveillance applications. Multimed Tools Appl 77(18):23273–23290

Sharghi A, Laurel JS, Gong B (2017a) Query-focused video summarization: dataset, evaluation, and a memory network based approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4788–4797

Sharghi A, Laurel JS, Gong B (2017b) Query-focused video summarization: Dataset, evaluation, and a memory network based approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4788–4797

Shingrakhia H, Patel H (2022) Sgrnn-am and hrf-dbn: a hybrid machine learning model for cricket video summarization. Vis Comput 38(7):2285–2301

Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556

Smeaton AF, Over P, Kraaij W (2006) Evaluation campaigns and trecvid. In: Proceedings of the 8th ACM international workshop on multimedia information retrieval, pp 321–330

Song Y, Vallmitjana J, Stent A, Jaimes A (2015) Tvsum: summarizing web videos using titles. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5179–5187

Sreeja MU, Kovoor BC (2019) Towards genre-specific frameworks for video summarisation: a survey. J Vis Commun Image Represent 62:340–358

Sreeja MU, Kovoor BC (2022) A multi-stage deep adversarial network for video summarization with knowledge distillation. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-021-03641-8

Sridevi M, Kharde M (2020) Video summarization using highlight detection and pairwise deep ranking model. Procedia Comput Sci 167:1839–1848

Sun M, Farhadi A, Seitz S (2014) Ranking domain-specific highlights by analyzing edited videos. In: European conference on computer vision, Springer, pp 787–802

Sundaram H, Xie L, Chang SF (2002) A utility framework for the automatic generation of audio-visual skims. In: Proceedings of the tenth ACM international conference on Multimedia, pp 189–198

Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826

Tejero-de-Pablos A, Nakashima Y, Sato T, Yokoya N, Linna M, Rahtu E (2018) Summarization of user-generated sports video by using deep action recognition features. IEEE Trans Multimed 20(8):2000–2011

Tiwari V, Bhatnagar C (2021) A survey of recent work on video summarization: approaches and techniques. Multimed Tools Appl 80(18):27187–27221

Vasudevan AB, Gygli M, Volokitin A, Van Gool L (2017) Query-adaptive video summarization via quality-aware relevance estimation. In: Proceedings of the 25th ACM international conference on Multimedia, pp 582–590

Vasudevan V, Sellappa Gounder M (2021) Advances in sports video summarization–a review based on cricket videos. In: International conference on industrial, engineering and other applications of applied intelligent systems, Springer, pp 347–359

Vinyals O, Fortunato M, Jaitly N (2015) Pointer networks. Adv Neural Inf Process Syst 28

Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, Tang X, Van Gool L (2016) Temporal segment networks: towards good practices for deep action recognition. In: Leibe B, Matas J, Sebe N, Welling M (eds) Computer vision—ECCV 2016, Springer, Cham, pp 20–36

Wang J, Wang W, Wang Z, Wang L, Feng D, Tan T (2019) Stacked memory network for video summarization. In: Proceedings of the 27th ACM international conference on multimedia, pp 836–844

Wei J, Yang X, Dong Y (2021) User-generated video emotion recognition based on key frames. Multimed Tools Appl 80(9):14343–14361

Wu J, Zhong SH, Ma Z, Heinen SJ, Jiang J (2018) Gaze aware deep learning model for video summarization. In: Pacific rim conference on multimedia, Springer, pp 285–295

Wu J, Zhong SH, Liu Y (2020) Dynamic graph convolutional network for multi-video summarization. Pattern Recognit 107:107382

Xiao S, Zhao Z, Zhang Z, Guan Z, Cai D (2020a) Query-biased self-attentive network for query-focused video summarization. IEEE Trans Image Process 29:5889–5899

Xiao S, Zhao Z, Zhang Z, Yan X, Yang M (2020b) Convolutional hierarchical attention network for query-focused video summarization. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, pp 12426–12433

Xu B, Wang X, Jiang YG (2016a) Fast summarization of user-generated videos: exploiting semantic, emotional, and quality clues. IEEE MultiMedia 23(3):23–33

Xu J, Mei T, Yao T, Rui Y (2016b) Msr-vtt: a large video description dataset for bridging video and language. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5288–5296

Xu J, Sun Z, Ma C (2021) Crowd aware summarization of surveillance videos by deep reinforcement learning. Multimed Tools Appl 80:6121–6141. https://doi.org/10.1007/s11042-020-09888-1

Yuan Y, Mei T, Cui P, Zhu W (2017) Video summarization by learning deep side semantic embedding. IEEE Trans Circuits Syst Video Technol 29(1):226–237

Yuan L, Tay FEH, Li P, Feng J (2019a) Unsupervised video summarization with cycle-consistent adversarial lstm networks. IEEE Trans Multimed 22(10):2711–2722

Yuan Y, Ma L, Zhu W (2019b) Sentence specified dynamic video thumbnail generation. In: Proceedings of the 27th ACM international conference on multimedia, pp 2332–2340

Yuan Y, Li H, Wang Q (2019c) Spatiotemporal modeling for video summarization using convolutional recurrent neural network. IEEE Access 7:64676–64685

Zeng KH, Chen TH, Niebles JC, Sun M (2016) Generation for user generated videos. In: European conference on computer vision, Springer, pp 609–625

Zhang Q, Yang LT, Chen Z, Li P (2018) A survey on deep learning for big data. Inf Fusion 42:146–157

Zhang Y, Kampffmeyer M, Liang X, Zhang D, Tan M, Xing EP (2019a) Dilated temporal relational adversarial network for generic video summarization. Multimed Tools Appl 78(24):35237–35261

Zhang Y, Kampffmeyer M, Zhao X, Tan M (2019b) Dtr-gan: dilated temporal relational adversarial network for video summarization. In: Proceedings of the ACM turing celebration conference—China, ACM TURC '19, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3321408.3322622

Zhang Y, Kampffmeyer M, Zhao X, Tan M (2019c) Deep reinforcement learning for query-conditioned video summarization. Appl Sci 9(4):750

Zhang Y, Liang X, Zhang D, Tan M, Xing EP (2020) Unsupervised object-level video summarization with online motion auto-encoder. Pattern Recognit Lett 130:376–385

Zhao B, Li X, Lu X (2017) Hierarchical recurrent neural network for video summarization. In: Proceedings of the 25th ACM international conference on Multimedia, pp 863–871

Zhao B, Li X, Xiaoqiang L (2019) Property-constrained dual learning for video summarization. IEEE Trans Neural Netw Learn Syst 31(10):3989–4000

Zhao B, Li X, Xiaoqiang L (2020) Tth-rnn: tensor-train hierarchical recurrent neural network for video summarization. IEEE Trans Ind Electron 68(4):3629–3637

Zhao B, Li H, Xiaoqiang L, Li X (2022) Reconstructive sequence-graph network for video summarization. IEEE Trans Pattern Anal Mach Intell 44(5):2793–2801

Zhong G, Tsai Y-H, Yang M-H (2016) Weakly-supervised video scene co-parsing. In: Asian conference on computer vision, Springer, pp 20–36

Zhong G, Tsai Y-H, Liu S, Su Z, Yang M-H (2018) Learning video-story composition via recurrent neural network. In: 2018 IEEE winter conference on applications of computer vision (WACV), IEEE, pp 1727–1735

Zhong S, Jiaxin W, Jiang J (2019) Video summarization via spatio-temporal deep architecture. Neurocomputing 332:224–235

Zhou K, Qiao Y, Xiang T (2018a) Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In: Proceedings of the AAAI conference on artificial intelligence. https://doi.org/10.1609/aaai.v32i1.12255

Zhou L, Xu C, Corso JJ (2018b) Towards automatic learning of procedures from web instructional videos. In: Thirty-second AAAI conference on artificial intelligence

Zhu X, Loy CC, Gong S (2016) Learning from multiple sources for video summarisation. Int J Comput Vis 117(3):247–268

Zhu W, Jiwen L, Han Y, Zhou J (2022) Learning multiscale hierarchical attention for video summarization. Pattern Recognit 122:108312

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