Archives available at journals.mriindia.com

International Journal on Advanced Computer Theory and Engineering

ISSN: 2319-2526

Volume 15 Issue 01, 2026

SummarizeX : Intelligent Video Compression Using AI

[1]Suraj R. Nalawade, [2]Praveen R. Barapatre. [3]H. O. Tapase, [4]Suhani Pawar

[1,2,3]Professor, Department of Artificial Intelligence and Data Science Engineering, YSPM's Yashoda Technical Campus, Satara-415001, Affiliated to DBATU University Lonare, Maharashtra, India.)

[4]UG Scholar, Department of Artificial Intelligence and Data Science Engineering, YSPM's Yashoda Technical Campus, Satara-415001, Affiliated to DBATU University Lonare, Maharashtra, India.)

Email: [1]dr.surajsir@gmail.com, [2]dr.p.barapatre@gmail.com, [3]gouribarge.8@gmail.com, [4] suhanipawar3062005@gmail.com

Peer Review Information

Submission: 16 March 2026

Revision: 03 April 2026

Acceptance: 26 April 2026

Keywords

Video Summarization, Deep Learning, CNN, LSTM, Transformers, Attention Mechanism, Keyframe Extraction, Multimedia Data Processing

Abstract

With the rapidly increasing video data, the activity of summarizing videos is becoming very significant. Video summarizing enables the proper storage, retrieval, and utilization of multimedia data by converting long videos into short, meaning-carrying summaries. The domain has seen tremendous change with new developments in the field of deep learning, through which a high-quality automatic summary can be generated with only minimal manual intervention. The paper surveys the designs, challenges, and applications of state-of-the-art deep learning techniques. We examine further directions in this exciting field and provide experimental evaluations.

Introduction

Issues in data management and accessibility arise because of fast spreading video data on YouTube, social media, surveillance systems, and so forth. Video summarization thus tries to find out what are the keyframes or video segments instructive of the video content. Such techniques were bound by limitations of scalability and generality because of the heavy usage of handcrafted characteristics. Video summarization has now advanced significantly with the introduction of deep learning models, making use of transformer topologies, convolutional and recurrent neural networks, and attention processes.

The amount of video data generated daily has increased dramatically in recent years due to the rapid rise of digital technologies and internet usage. Platforms such as YouTube, Instagram, surveillance systems, educational platforms, and entertainment services continuously generate substantial volumes of video content. Consequently, the management, storage, and assessment of these extensive video datasets present significant challenges for both commercial entities and academic investigators. The manual review of lengthy videos to extract pertinent information is both labor-intensive and inefficient, thereby underscoring the importance of automated video summarization systems.

The process of automatically condensing a lengthy film into a brief and educational version while maintaining the most crucial information is known as video summarizing. Reducing repetition and giving viewers a succinct synopsis of the video are the objectives.This enables viewers to rapidly get the key points

without having to watch the full video. Keyframes, highlights, or brief video segments that capture the main ideas of the original video can all be used to create video summaries.

Prior to recent developments, video summarization predominantly utilized rule-based algorithms and manually specified attributes. The necessity for manual feature extraction and domain-specific knowledge constrained the scalability and applicability of these conventional approaches. However, the advent of artificial intelligence has brought about substantial enhancements in the efficacy of video summarization systems, particularly through the application of deep learning techniques. Consequently, models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs), and Transformer architectures are now being utilized to autonomously identify salient patterns within video data.

Deep learning, without needing much manual feature engineering, allows models to learn both temporal and spatial relationships in video data. This has led to the creation of video summarization systems that are more accurate, scalable, and efficient. These systems can identify important scenes, find significant events, and create summaries that are visually coherent and useful.

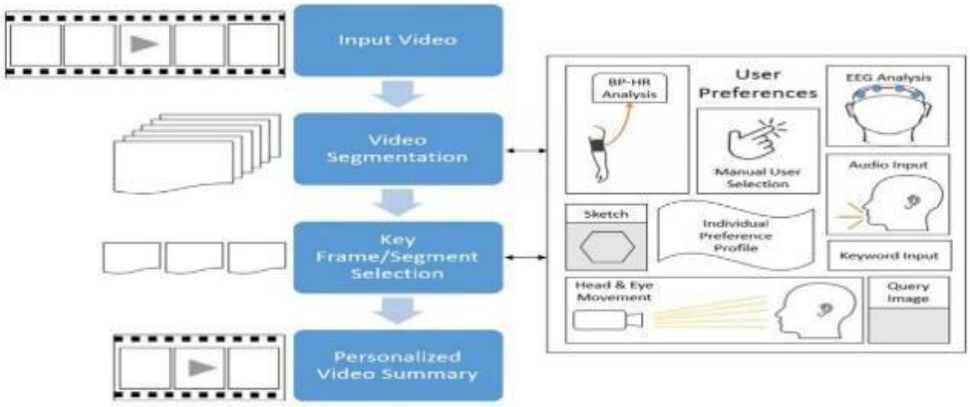

Fig 1: Video summarization using deep learning (Source: Internet)

Objective: Based on the review of prominent approaches, datasets, challenges, and applications, this paper investigates the use of deep learning algorithms for video summarization.

Background And Related Work

1. Traditional Methods for Video Summarization

Prior to the development of deep learning, video summarization was accomplished using traditional computer vision methods. These traditional methods were centered on the extraction of low-level features from video images. Some of the traditional methods used for video summarization included shot boundary detection, where a video was broken down into smaller segments called shots based on the visual differences between consecutive video frames.

After the detection of the shot boundaries, clustering algorithms such as K-means clustering or hierarchical clustering were used to cluster similar video frames. From the clusters, a representative frame called a keyframe was chosen. These keyframes made up the video summary.

Another traditional method used for video summarization included the computation of the importance scores of video frames based on predefined rules. Video frames with higher importance scores were chosen as keyframes for video summarization. However, these traditional methods had some drawbacks. These methods were dependent on handcrafted features, which were not capable of extracting high-level semantic information from videos.

Additionally, these methods were not effective when used in different videos because the handcrafted features were domain-specific. Therefore, traditional video summarization methods were not scalable.

2. Video Summarization with Deep Learning

The application of deep learning algorithms has brought a remarkable change in the area of video summarization. Unlike conventional approaches, deep learning algorithms have the capability to learn significant features automatically from the raw video data itself. This makes the system more robust and efficient in generalizing various types of videos.

Deep learning algorithms have the ability to learn complex spatial and temporal patterns from video data. For example, Convolutional Neural

Networks are efficient in learning spatial features from individual frames of the video, whereas Recurrent Neural Networks and LSTM networks are capable of learning temporal dependencies between the frames.

Another benefit of using deep learning algorithms for video summarization is the end-to-end learning process. In this process, the entire system can be trained together using large amounts of data, and the algorithm can learn optimal representations and summarization strategies automatically. This makes deep learning algorithms more efficient than conventional approaches in terms of accuracy and efficiency.

Recent studies have also focused on more advanced architectures like attention models and transformer models, which have the capability to learn long-range dependencies in videos. These models have further improved the quality of the generated video summaries.

3. Summarization Types

Video summarization methods can be broadly classified into two categories: keyframe-based summarization and video skimming.

Keyframe-based summarization is a method that concentrates on identifying a set of key frames that represent the most significant events in a video. These frames give a visual summary of the video content and are widely used in video browsing and indexing applications.

Video skimming, on the other hand, is a method that concentrates on identifying a set of short video clips that represent the most significant events in a video. Video skimming is different from keyframe summarization because it preserves the temporal continuity of the original video. This makes video skimming more suitable for applications that require motion and context.

Both methods have their own strengths and weaknesses depending on the application. Keyframe-based summarization is easier and more efficient, while video skimming is more informative and engaging.

Video Summarization Techniques with Deep Learning

1. Convolutional Neural Networks (CNN)

Due to the power of CNNs to fool low-level and high-level information from images and videos, Convolutional Neural Networks have been implemented for many computer vision problems. CNNs have also been used for video summarization in which they retrieve such information as objects, scenes and texture from the videos.

Long short-term memory network (LSTM) modules have explored relationships across the frames of video. However, the knowledge fetched had been used to get iconic highlights of a video to be filled.

CNN architectures have found a successful application on different problems in video summarization capitalizing on the representations learned by CNNs.

2. Recurrent Neural Networks and LSTM

It has been able to achieve such a high level of success in managing information pertaining to space, but not time on frames of videos. But this issue has been addressed by Recurrent Neural Networks (RNNs).

In fact with RNN, memory has been sequentially stored. Thus the time relation between different frames of videos is taken care of. But its effectiveness in processing the data for a long chain of videos is limited due to the 'vanishing gradient' problem.

To solve this problem, Long Short-Term Memory (LSTM) was proposed. Herein is memory stored in memory cells. Thus now, the data for a huge series of videos has been processed. In video summarization, this model had been applied to find the information in terms of time-wise related information.

3. Attention Mechanisms and Transformers

Attention mechanisms have been used a lot in recent deep learning models. Attention essentially allows the model to pay more attention to specific parts of the input while ignoring others.

Attention mechanisms aid in identifying the most significant frames for video summarization as well. These architectures are further enhanced by using self-attention mechanisms that can handle long-range dependencies between frames.

Transformers have achieved amazing results in various sequence modeling tasks and are being incorporated into video analysis applications.

4. Reinforcement Learning

Another area that can be applied effectively to video summarization is reinforcement learning. Under this approach, the summarization process is considered an agent that interacts with the environment and learns decisions based on rewards.

The model chooses frames or video segments, and a reward function is used to determine the quality of the summary. The objective of the model is to maximize the reward by choosing the most informative frames and minimizing redundancy.

Reinforcement learning enables the model to learn the best strategies for summarization without the need for explicit supervision for

every frame.

4. Graph Neural Networks (GNNs)

The purpose of Graph Neural Networks (GNNs) is to handle the relationships that exist between different entities in a graph structure. In video summarization, a video can be modeled as a graph, where the nodes are frames or shots in a video, and the edges are the relationships between the nodes. GNNs can handle complex structural relationships that exist in a video and aid in the identification of important segments of a video based on their relationships with other frames in the video. This is especially useful when dealing with complex scene transitions and interactions in a video.

5. Efficient learning without labeled data

One of the biggest problems in video summarization is the absence of labeled datasets. It is a time-consuming task to annotate the summaries, and this makes it difficult to obtain training data.

The problem of the absence of labeled datasets can be overcome by unsupervised and self-supervised learning techniques. These techniques learn patterns from the unlabeled video data.

These techniques are gaining popularity because they decrease the reliance on expensive labeled datasets.

Evaluation Metrics and Datasets

1. Datasets for benchmarking

There are some benchmark datasets that are used to test the performance of video summarization models. The SumMe dataset includes user videos along with summaries created by humans. These summaries are used as ground truth. The SumMe dataset is widely used to test the performance of summarization models.

The TVSum dataset includes video clips from different TV shows, and importance scores are assigned by different human annotators. The dataset is used to test the performance of models to see how well they align with human perception.

The YouTube Highlights dataset includes videos of different categories like sports and cooking. The dataset is used to identify highlight moments in videos.

2. Metrics for Evaluation

To evaluate the performance of video summarization systems, various metrics have been employed. The most popular metric is the F-measure, which calculates the similarity between the resultant summary and the ground truth summary based on precision and recall.

Coverage is another critical metric that measures the amount of significant content from the original video that is preserved in the resultant summary.

Diversity is used to ensure that the selected frames or clips are not redundant and cover different portions of the video.

The above metrics enable researchers to make an objective comparison of the performance of various video summarization methods.

Challenges With Video Summarization Via Deep Learning

Despite the progress made, there are still some challenges in video summarization. One of the biggest challenges is the lack of annotated datasets. The process of creating high-quality labeled summaries is time-consuming and requires human judgment.

Another challenge in video summarization is that it is a subjective task. Different people may have different opinions about what is important in a video. Developing models that can be tailored to different user preferences is still an open research challenge.

Scalability is also a challenge when working with long videos such as movies, surveillance videos, or recorded lectures. Handling such a large amount of data requires a lot of computational power.

In addition, developing models that can generalize to different video genres is still a challenge. A model that is trained on sports videos may not work well on educational or surveillance videos.

Video Summarization Applications

Video summarization has numerous applications in different fields. In video sharing platforms like YouTube and Netflix, video summarization algorithms can be used to create highlight reels that help viewers understand the content of the video.

In security systems, video summarization can assist security personnel in viewing hours of recorded video in a very short period of time by pointing out key events. Learning platforms can apply video summarization to identify key points in lengthy lectures, making it easier for viewers to refer to them.

In the medical field, video summarization algorithms can be applied to record lengthy medical procedures or surgeries by pointing out key points for analysis.

Results Of the Experiment

Experimental evaluation is an important aspect in understanding the efficacy of video summarization models. In most cases, the

models are trained and tested on benchmark datasets like SumMe and TVSum.
Recent experiments have proved that transformer models and attention mechanisms can greatly enhance the performance of summarization models. These models perform better in understanding long-term dependencies and key frames.
The performance of the models is measured using metrics such as F-measure, coverage, and diversity. The higher the score, the better the summary's alignment with the ground truth summary.

Prospective Paths
Future work in video summarization is likely to concentrate on multimodal learning, where data from various sources like audio, text, and images is integrated to produce more informative summaries.
Another area of interest in video summarization is real-time video summarization, which focuses on producing summaries simultaneously while the video is being captured.
Personalized video summarization is also an emerging area, where summaries are produced according to the user's preferences and interests.
In addition, there is interest in applying Explainable AI (XAI) methods to enhance the transparency of deep learning models by providing an explanation for why particular frames were chosen for the summary.

Conclusion
Deep learning-based video summarization has utterly changed the utility and availability of video content. Despite significant improvement, challenges such as subjectivity, scalability, and availability of datasets still remain. Future research will most likely focus on multimodal, real-time, and adaptive summarization techniques to enhance the field's applicability even further.

References
Satya Prasad, K., and Munagala, V. (2019). Clustered entropy computing: A Holoentropy-based encoding approach for very efficient computing systems. Computing Clusters 22, 1429-1441.

Magnenat-Thalmann, N. (Vol 37, issue 8) Preface. 2051–2052 in The Visual Computer, 37(8). 2021.

Tariq, J., Ashraf, I., Rahman, H., Ijaz, A., Ali, H., Alfalou, A., & Rehman, S. (2022). HEVC: Intraphysical mode selection using the statistical approach can be executed within a fast intra mode decision. 3903–3918 in Computers, Materials and Continua, 70(2).

Verbist, F., Deligiannis, N., Slowack, J., Van de Walle, R., Schelkens, P., & Munteanu, A. (2012). Iossifides, A. C. Video coding with Wyner-Ziv for wireless, low-power multimedia applications. EURASIP Journal of Wireless Networking and Communications, 2012, 1–20.

Sadagopan, P., and Sharma, S. N. (2022). Impact of feature selection based on conditional holoentropy on automatic recommendation systems in the e-commerce industry. Computer and Information Sciences Journal of King Saud University, 34(8), 5564-5577.

In 2019, Liu, H., Li, J., Wu, Y., and Fu, Y.

clustering while eliminating outliers. IEEE Knowledge and Data Engineering Transactions, 33(6), 2369-2379.

Prasad, K. S., and Munagala, V. (2019). A holoentropy-based encoding approach for very efficient computing systems is called clustered entropy computing. Networks Software Tools and Applications Journal Jan.

Mishra, N., and B. Gupta (2022). enhanced attack detection system based on deep learning for safe cloud virtualized environments. Electronic Networks, Devices, and Fields: International Journal of Numerical Modelling, 35(1), e2945.

Ravi, J., Dabbu, M., Karuppusamy, L., & Lakshmanan, S. (2022). Fuzzy entropy-based deep belief network for cloud intrusion detection based on the chronological salp swarm algorithm. Electronic Networks, Devices, and Fields: International Journal of Numerical Modelling, 35(1), e2948. 18

Timmerer, C., Ilangovan, A., Zabrovskiy, A., Agrawal, P., & Prodan, R. FastTTPS: Fast Scheduling and Transcoding Time Prediction for HTTP Adaptive Streaming. Cluster Comput, 24(3), 1605-1621 (2021).