International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0326
e ISSN: 2584-2854
Volume: 03
Issue:05 May 2025
Page No: 2080 - 2086

Video Summarization Using Machine Learning Techniques: An Overview

P. Hima Chandana[1], R. Ragupathy[2] and D. Vivekananda Reddy[3]

[1] PhD Scholar, Dept. of CSE, Annamalai University, Annamalainagar, Tamil Nadu, 608002, India.

[2] Professor, Dept. of CSE, Annamalai University, Annamalainagar, Tamil Nadu, 608002, India.

[3] Professor, Dept. of CSE, Sri Venkateswara University College of Engineering, Andhra Pradesh, 517502, India.

Email ID: *hima8022@gmail.com[1], cse_ragu@yahoo.com[2], svuvivek@gmail.com[3]*

Abstract

Technological advancement is a persistent aspect in our lives, and it emerges and grows at an exponential rate. Every day, a large quantity of data is generated in the form of text, image, audio and video. To process large amounts of video data, like movies, social media data and Surveillance data, requires a large amount of storage. Therefore, minimizing these videos by processing only vital content, takes a long processing time. To extract the key content from the video, which is a time-consuming procedure for the viewer, the entire video must be watched to overcome such challenges, video summary can be used to deal with and process lengthy videos. This paper discusses the various machine learning strategies used for summarising videos. Also, this paper presents video processing approaches such as key framing and skimming used for summarization in dynamic environments such as surveillance. Further, this study emphasizes the primary uses of video summarising in both dynamic and static environment. Furthermore, it deals with the datasets used for video summarisation and helps a clear understanding on various techniques applied for video summarisation at both static and dynamic environments.

Keywords: *Video Summarization, Key framing, Video skimming, Recurrent generative adversarial network Self-attention binary neural tree, Global diverse attention, Action ranking, Distinct frame patch index, Domain independent redundancy, Cluster validity index, Discriminative feature learning.*

1. Introduction

Nowadays, digital cameras are utilised for security monitoring, education, news, and entertainment, among other things. These cameras are used in homes, stores, offices, and many other public places to collect data throughout the year. To evaluate these recordings and extract the vital information from them is a hard task that requires more time and storage space, as well as human engagement and concentration. Video summary can be used to solve this problem. Video summarization (VS) is described as the act of computationally shortening a set of data to create a subset or a summary that represents the most significant or relevant information within the original content. In general terms the flow of process of video summarization initially starts with the input of the raw video and it is broken down into frames. Among those frames the feature clip is selected then analysed and processed with the clustering techniques. Later it is summarized as shown in the Figure 1. VS is utilised in a variety of applications, such as the creation of trailers for films and serials, sports and news highlights, and personal video summaries. Static VS / static keyframes and dynamic VS / dynamic video skimming are the two types of VS approaches. Static VS approaches are intended to select the most informative and representative keyframes from the original video sequence and organise them chronologically to represent the video's major content. The keyframe set is not constrained by timing or synchronisation difficulties, allowing for maximum flexibility and adaptability. Dynamic VS approaches concentrate on extracting video clips, which are a collection of frames containing the video's most exciting and relevant

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0326
e ISSN: 2584-2854
Volume: 03
Issue:05 May 2025
Page No: 2080 - 2086

content, including basic audio and visual motion elements. Although this method is more appealing to viewers than simply viewing a sequence of static keyframes, video skimming necessitates advanced semantic analysis. In contrast, static VS approaches based on keyframe extraction have received a lot of attention due to their ease of use, flexibility, and practicality. Summarising news videos will allow us to focus on the most significant components of the news. Some of the applications of video summarising include automatically creating a movie trailer and highlights of sports footage or news video recordings. People can also manage the videos they record on their mobile devices in order to provide access to them by representing them with summaries. The most difficult videos to summarise before displaying to users are those found on the internet. For example, suppose we search for videos on a specific topic, Search engines return a plethora of video results. A summary of the video will be provided as a preview. Egocentric videos are videos taken with a body worn camera. With the recording of these egocentric movies, new computational issues have arisen, ranging from the analysis of socio-behavioural activity to the examination of everyday life occurrences of a human wearing the camera. Wearable cameras are used not just for sports, but also is used for patients with various health conditions to record their daily activities. However, watching such a long recording to discover just an important segment of the video in a specific day may be challenging, such people with unusual health concerns will benefit from a brief description of these video recordings. Closed-circuit television (CCTV) cameras are installed in high-risk areas, and surveillance videos are continuously captured throughout the day and year, resulting in massive amounts of video data being collected. These videos can be easily interpreted using a video synopsis. A summary can be created on surveillance footage that contain thefts in homes, traffic accidents, unusual behaviour of students in exams, polling booths, and hospitals, among other things. Cameras may be integrated into equipment such as robots and drones, allowing them to record films in places that humans cannot reach.

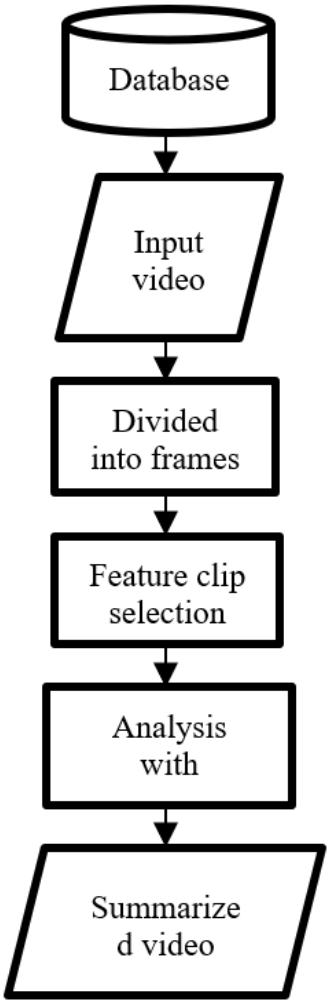

Figure 1 Flow of Process of Video Summarization

These videos will be easier to interpret if they are summarised. Medical movies such as diagnostic hysteroscopy videos, endoscopic films, surgical videos, and so on result in a range of automatic video summaries. The length of these medical movies might vary greatly. Summarising lengthy medical movies can help medical practitioners perform a thorough review of medical procedures, as well as teach those processes to medical students. This will allow medical practitioners to quickly look for comparable instances or go over the details of an earlier case. The rest of the paper is organized as follows. In section 2 various machine learning techniques used for video summarization are presented in Section 3. Conclusion is made in Section 4.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0326
e ISSN: 2584-2854
Volume: 03
Issue:05 May 2025
Page No: 2080 - 2086

2. Machine Learning Techniques for Video Summarization

Graph-based structural technique

A graph-based structural difference analysis model has been proposed by Chunlei Chai et al. [1] to formulate the Video summarization problem, in which the structural information in each video frame feature is considered and modelled in graphs to bridge the gap between the actual semantic structural information and the raw video frame features. A graph-based metric is used for measuring frame dissimilarity. This structural difference in the graph might show potential disparities between continuous frames, and median graphs which can be created as keyframes to reflect the overall trend of the video.

2.1 Recurrent generative adversarial network

Libin Lan and Chunxiao Ye [2] developed a novel approach to unsupervised Wireless capsule endoscopy (WCE) video summarization by merging variational auto encoder (VAE), pointer network (Ptr-Net), and generative adversarial network (GAN) approaches, which are commonly utilised in multimedia video processing. These approaches, particularly Ptr-Net, are initially used to summarise WCE videos. It employs a de-redundancy method (DM) to reduce redundant frames in both user and medical movies. The concept is inspired by the coverage method, which is used to solve the repetition problem in text summary. They provided WCE-2019-Video, a novel dataset that enables a repeatable evaluation of WCE video summarising methods. It is, to the best of the knowledge, the first dataset that can be utilised for future WCE.

2.2 Self-attention binary neural tree

The primary purpose of this model is to predict a score ranging from 0 to 1 for each shot. The greater a shot's score, the more likely it will be included in the final report. In this regard, Hao Fu and Hongxing [3] Wang proposed the self-attention binary tree neural network (SABTNet) model, which consists of the backbone network, shot encoding, branch routing, self-attention, and score prediction modules. Shot segmentation algorithms or subsequent shot summarising should first divide it into non-overlapping shots. For shot segmentation, it was used like kernel temporal segmentation (KTS). In this it is obtained like Nv shots for video v in this manner, where the nth shot ranges from the nstth frame to the nedth frame. It made use of a pre-configured backbone network. To extract and re-encode visual characteristics from individual video frames, a Bidirectional long short-term memory (Bi-LSTM) converts features inside the same shot into shot-level features. Following that, a full binary tree is used to expose each shot to several assessment paths for comprehensive scoring. Each non-leaf node in the tree is accompanied by a branch routing module, which determines the likelihood of which path the shot will be delivered to for review. It has anticipated scores and proportions based on different root-to-leaf paths thanks to branch routing can be anticipated. Also, an integrated module along with each edge of the tree to optimise shot attributes by factoring self-attention among video frames to ensure a more accurate forecast.

2.3 Global diverse attention

A global diversified attention mechanism is designed to describe the temporal dependency of video by leveraging pairwise relations between every two frames regardless of stride magnitude, which aids in the handling of recurrent neural network (RNN) models, long-range dependency problem. By SUM-global diverse attention (GDA) is considerably more efficient than competing systems since it immediately calculates the pairwise similarity matrix that depicts the relationships between the source and target frames. SUM-GDA model suggested by Ping Li et al. [4]is investigated in supervised, unsupervised, and semi-supervised settings. Both the variety of generated summaries and the impact of optical flow features were explored.

2.4 Egocentric based action ranking

In this technique, Abhimanyu Sahu and Ananda S. Chowdhury [5] claimed that keyframes from a series of clusters are picked to obtain a summary of a specific length. These summaries at various dimensions are then analysed to produce a priority-based ranking of the numerous acts depicted in the video. To produce several summaries of an egocentric video in one shot, this technique employs a hierarchical (agglomerative) clustering algorithm. As a result, the proposed method is based on the

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0326
e ISSN: 2584-2854
Volume: 03
Issue:05 May 2025
Page No: 2080 - 2086

"analyse once, generate many" idea. The clustering process is used only once, and summaries of various lengths (multiple scales) are generated by splitting the resulting dendrogram at numerous levels. These many divides might be based on the needs of the user. It also shows how summaries of a specific Multiple scales of egocentric video can be studied to assign relative priority to the important acts in a first-person video.

2.5 Distinct frame patch index and appearance based linear clustering

Siva Priya et al. suggested a method [6] to select candidate frames using a patch-based method. The primary reason stems from two factors: Patches are less susceptible to image noise than pixels and hence provide a more reliable representation of how video material evolves spatially, patches are used to estimate the frame index with confidence, then to discover discrepancies across frames, and finally to obtain the distinct ones. Initially, each resized frame of a given video is divided into various patches, and these patches are then classed into distinct classes to estimate the frame's uniqueness, distinct frame patch (DFP) index, as the entropy. Whether these patches are similar in the YCbCr colour space is determined by the classification of these patches. This technique chose a group of good candidate frames based on this unique metric by presenting a unique appearance-based linear clustering (ALC) method to further refine these frames for the identification of different ones.

2.6 Domain independent redundancy elimination technique

The proposed approach by Jesna Mohan and Madhu S. Nair [7] is a redundancy elimination strategy based on uniform presampling followed by flow vector-based filtering out of confusing frames. For the calculation of flow vectors between consecutive frames, the approach employs the SIFT Flow algorithm. The magnitude of the flow vectors is then utilised to determine the abrupt shift in motion and to trace the frames when the scene changes. To reduce superfluous frames of input video while keeping keyframes, the extent of displacement is thresholded locally. This strategy works well on videos of all types. The resulting collection of frames can be used to summarise. The algorithm operates on a frame-by-frame basis.

2.7 Cluster validity index

The problem of video summarization has been presented as an automatic representative selection job by Ye Zhao et al. [8]. It begins with an examination of the video's syntactic structure. The video summary can properly capture the structure rules between shots and scenes, as well as the syntax semantics of the video. Because there are so many frames in the video, it is difficult for each frame to be a candidate frame. As a result, a more general solution would be to divide the movie into temporal parts dependent on frame rate. To keep things simple, the first frame of the segment is frequently chosen as the segment's representative frame. It is self-evident that the initial frame may not be the best representative frame of the segment, the candidate frames selection approach is inaccurate. To overcome the problem, they suggested the motion-based selection method, which evaluates candidate frames' forward and backward motion. The first step is to separate individual shots from a video clip. They then split shots into fixed length subshots and choose potential frames based on their forward and backward motion. Affinity Propagation is used in conjunction with the validity index to automatically select the best representatives from the candidate frame subset.

2.8 Online motion auto-encoder

Learning an internet dictionary unsupervised, this technique is a novel online motion-auto encoder (AE) model that, by continuously updating a designed recurrent auto-encoder network, may simulate online dictionary learning for memorising past states of object motions. The most recent Orange Ville benchmark, a fresh surveillance video dataset is gathered, allowing for objective assessment of our new field of video summarising algorithms. Yujia Zhang et al. [9] have provided spatial-temporal annotations for all important object motion clips to further video summarising research with granularities ranging from coarse to fine, outstanding performance for both object motion and frame-level summarization. They did extensive studies on several known video summarising benchmarks in addition to the key object motion-based summarization.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0326
e ISSN: 2584-2854
Volume: 03
Issue:05 May 2025
Page No: 2080 - 2086

2.9 Discriminative feature learning

Unsupervised video summarising approaches have been introduced by Yunjae Jung et al. [10]. These methods are based on variational autoencoders (VAE) and generative adversarial networks (GAN). It will begin with discriminative feature learning within a VAE-GAN architecture employing variance loss. Then, a chunk and stride network computer science network (CSNet) is developed to solve the difficulty of learning for long-length films, which is a drawback of most existing approaches. CSNet handles this challenge by looking at input features from both a local chunk and a global stride perspective. Finally, the difference in convolutional neural network (CNN) features between adjacent or wider spaced video to determine which region of the video is relevant are used.

2.10 Keyframe extraction and video skimming

The VSUMM method influences Shruti Jadon and Mahmood Jasim's [11] approach to video summarising. To begin, keyframes containing crucial information are retrieved. A fraction of the frames was used to reduce calculation time for video segmentation. Given that the sequence of frames is highly correlated, the variation between frames is expected to be very small when sampled at high frequencies, such as 30 frames per second. Instead, a low frequency rate of 5 frames per second had little effect on the results, but it significantly boosted the computing speed. It used a sampling rate of 5 frames per second and eliminated the unnecessary frames. After extracting all of the keyframes, cluster the frames to categorise them as intriguing or uninteresting. The cluster of important frames was used to construct the video summary. The video summary was decided to be around 15% of the duration of the original video. However, this summary was discontinuous and thus different from how a human observer would evaluate the summary, resulting in unsatisfactory ratings because the evaluation approach corresponds with how a human being scores the summary. This issue was solved by a 1.8 second skims from the extracted fascinating frame. This keeps the summary flowing and easy to understand. Frames are sampled at low frequency.

3. Discussion

From the literature study those techniques are very much efficient for the summarization of videos. In those videos the below mentioned Table.1. datasets have been used for the processing.

Merits

- To speed up browsing.
- To achieve efficient access.
- To save memory.

Demerits

- Inefficient process of finding corresponding timestamps.

The below table states the above illustrated techniques belong to which type of machine learning model and also which datasets have been used to process the video summarization. General Machine Learning approaches in video summarization.

- Supervised Learning Approaches: Requires labeled datasets where human-annotated summaries guide the model.
- Deep Learning-based Approaches: Recurrent Neural Networks (RNNs) and Long short-term memory (LSTM) networks capture temporal dependencies. Convolutional Neural Networks (CNNs) extract spatial features from frames. Transformers (e.g., VideoBERT) for context-aware summarization.
- Unsupervised Learning Approaches: Does not require labeled data; instead, it identifies patterns and clusters similar frames. Clustering Algorithms (e.g., K-means, DBSCAN) segment similar frames into clusters. Autoencoders learn compressed representations for efficient keyframe selection.
- Reinforcement Learning (RL) Approaches: Models are trained using reward-based learning. Encourages selection of the most informative and diverse frames. Example: Using Deep Q-Networks (DQN) for optimal keyframe extraction.
- Hybrid Approaches: Combines multiple techniques to improve summarization. Example: CNNs for feature extraction + RNNs for temporal analysis.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0326
e ISSN: 2584-2854
Volume: 03
Issue:05 May 2025
Page No: 2080 - 2086

Table 1 List of Datasets Used in The Above Discussed Techniques

S. No.	Name of the technique	Type of Machine Learning Model	Datasets
1	Graph-based structural Technique [1]	Supervised Learning	YOUTUBE [1]
2	Recurrent generative adversarial network [2]	Unsupervised Learning	SumMe, TVSum [2]
3	Self-attention binary neural tree [3]	Supervised Learning	SumMe, TVSum [3]
4	Global diverse attention [4]	Supervised Learning, Unsupervised Learning, Semi supervised Learning	SumMe, TVSum [4]
5	Egocentric based action ranking [5]	Supervised Learning	SumMe, TVSum [5]
6	Distinct frame patch index and appearance based linear clustering [6]	Supervised Learning, Unsupervised Learning	Own Dataset [6]
7	Domain independent redundancy elimination technique [7]	Supervised learning, Unsupervised Learning	VSUMM [7]
8	Cluster validity index [8]	Supervised Learning	SumMe, TVSum [8]
9	Online motion auto-encoder [9]	Unsupervised Learning	Base jumping dataset, SumMe, TVSum [9]
10	Discriminative Feature Learning [10]	Unsupervised Learning	SumMe, TVSum [10]
11	Keyframe extraction and video skimming [11]	Supervised Learning, Unsupervised Learning	VSUMM, sumMe [11]

Conclusion

This paper defines a quick summary of video summarization, approaches, and various techniques. The methodologies, applications, and datasets are the primary points of this research. As a result, certain image processing algorithms are inadequate for long recordings and require greater efficiency. Furthermore, video processing approaches such as key framing and skimming are better suited for summarization in dynamic environments such as surveillance footage where the camera is fixed. This study also emphasised the primary uses of video summarising in static and dynamic environments. The current study will help researchers to understand processing techniques, datasets, merits and demerits of many strategies utilised in the video summarization process.

References

[1]. C. Chai, L. Guoliang, W. Ruyun, L. Chen, L. Lei, Z. Peng and L. Hong, "Graph-based structural difference analysis for video summarization," Information Sciences, pp. 483-509, 2021.

[2]. L. Lan and Y. Chunxiao, "Recurrent generative adversarial networks for unsupervised WCE video summarization," Knowledge-Based Systems, 2021.

[3]. H. Fu and W. Hongxing, "Self-attention binary neural tree for video summarization. Pattern Recognition Letters," Pattern recognition letters, pp. 19-26, 2021.

[4]. P. Li, Y. Qinghao, Z. Luming, Y. Li, X. Xianghua and S. Ling, "Exploring global diverse attention via pairwise temporal

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0326
e ISSN: 2584-2854
Volume: 03
Issue:05 May 2025
Page No: 2080 - 2086

relation for video summarization," Pattern Recognition, 2021.

[5]. A. Sahu, A. S and Chowdhury, "Multiscale summarization and action ranking in egocentric videos," Pattern Recognition Letters, pp. 256-263, 2020.

[6]. S. Kannappan, L. Yonghuai and T. Bernard, "DFP-ALC: automatic video summarization using distinct frame patch index and appearance based linear clustering," Pattern Recognition Letters, pp. 8-16, 2019.

[7]. J. Mohan and N. Madhu S, "Domain independent redundancy elimination based on flow vectors for static video summarization," Heliyon, 2019.

[8]. Y. Zhao, G. Yanrong, S. Rui, L. Zhengqiong and G. Dan, "Unsupervised video summarization via clustering validity index," Multimedia Tools and Applications, 2020.

[9]. Y. Zhang, L. Xiaodan, Z. Dingwen, T. Min and X. Eric P, "Unsupervised object-level video summarization with online motion auto-encoder," Pattern Recognition Letters, pp. 376-385, 2020.

[10]. Y. Jung, C. Donghyeon, K. Dahun, W. Sanghyun and K. In So, "Discriminative feature learning for unsupervised video summarization," AAAI Conference on artificial intelligence, pp. 8537-8544, 2019.

[11]. S. Jadon and J. Mahmood, "Unsupervised video summarization framework using keyframe extraction and video skimming," EEE 5th International Conference on computing communication and automation (ICCCA), pp. 140-145, 2020.

[12]. E. Apostolidis, I. Alexandros, E. A. Metsai, M. Vasileios and P. Ioannis, "A stepwise, label-based approach for improving the adversarial training in unsupervised video summarization," AI for Smart TV Content Production, Access and Delivery, pp. 17-25, 2019.

[13]. C.-Y. Yang, Y. Heeseung, V. Srenavis and Y.-j. H. Jane, "A Mobile Robot Generating Video Summaries of Seniors Indoor Activities," Human-Computer Interaction with Mobile Devices and Services, pp. 1-6, 2019.

[14]. J. Park, L. Jiyoung, K. Ig-Jae and S. Kwanghoon, "Sumgraph: Video summarization via recursive graph modeling," Computer Vision, pp. 647-663, 2020.

[15]. J. Traver V and D. Dima, "Egocentric video summarisation via purpose-oriented frame scoring and selection," Expert Systems with Applications, 2022.

[16]. I. Puthige, H. Tanveer, G. Suneet and A. Mohit, "Attention over attention: An enhanced supervised video summarization approach," Procedia Computer Science, 2023.