An Evaluation-based Analysis of Video Summarising Methods for Diverse Domains

Bijal U. Gadhia[1], Dr. Shahid S. Modasiya[2]

[1]Research Scholar, G.E.C., Gandhinagar, Gujarat Technological University, Ahmedabad, Gujarat, India
[2]Assistant Professor, Electronics & Communication department, G.E.C., Gandhinagar, Gujarat, India

E-mail: [1]bijal.gadhia@gecg28.ac.in, [2]shahid@gecg28.ac.in

Abstract

As technology progresses, a gigantic amount of video data is generated day-by-day. Processing of such a huge video requires time, increased storage, and a computational power. Sometimes it is convenient for the user to watch a summary or highlight rather than watching a complete video, which is a time-consuming task. So, a fully automated solution is required to extract important segments from video. Researchers have proposed multiple approaches / techniques for summarizing the videos which resolve the problem of long videos and summarize them according to the video type. This survey and comparative evaluation of video summarizing techniques based on several domains are presented in this study. Primarily, these methods are classified into different categories based on their methods or techniques used. Furthermore, an overview of some of the latest literature is presented with the dataset and the evaluation approaches used. The review is also made related to the domain direction, and is concluded by presenting the benefits and difficulties associated with the current video summarization techniques.

Keywords: Video summarization, video skimming, static summarization, key frame, Summary sequence

1. Introduction

Every day, with the progression of technology, a tremendous amount of audio/video data is produced. This rapid development of digital video has led to a variety of new applications and, as a result, research and development of new technologies will reduce the cost of cataloguing, indexing, and video archiving while also increasing the effectiveness,

DOI: https://doi.org/10.36548/jiip.2023.2.005
Received: 18.04.2023, received in revised form: 18.05.2023, accepted: 02.06.2023, published: 12.06.2023

usability, and accessibility of stored content [1]. There is a huge need for videos. One crucial subject among all potential study subjects is how to enable easy browsing of a sizable video data collection and explain how to create effective content both as representation and access [1,2]. Similarly, generating a preview of the video takes a lot of time because one has to see the complete video and then has to perform a video editing task which requires expertise, and it is highly expensive.

To present a preview on video content, research has been carried out almost 30 years ago with a target of generating key frames, and then generated short clips which cover important segments of video, also referred to as highlights of video [1]. But the sequence of images was not enough to comprehend, particularly in lengthy videos [3]. These key frame-based summaries served the needs of thumbnail representation of the film as well as video browsing and indexing. Such a key frame-based summarization is referred to as static video summarization.

But in recent literature, there is a demand to generate short summaries of videos by processing aural and visual content, which is called dynamic summarization and also referred to as video skimming. Video Skimming enhances the information through summarization, which generates videos that are short referred to as video skims, consisting of important segments with corresponding audio information [3]. The Video Skimming approach, overcomes these issues by generating a temporally shortened and summarized version of a given video [1, 2, 3]. Because of its dynamic nature, video skimming enables to comprehend easily from summary. The following are some major advantages of video skimming and dynamic video summarization:

- Content summarization: Video skimming condenses the content of videos into concise summaries that capture the core essence of the information being presented.

- Effective information organization: Skimming methods help in organizing video content by generating structured metadata or tags associated with the video segments.

- Decision-making support: By extracting key insights video skimming enables decision-makers to quickly evaluate video content to make informed decisions.

Many prior surveys have predominantly focused on static summarization or have specifically concentrated on a particular domain within video summarization. This study gives a complete survey on video summarization (static + dynamic), concentrating on the large corpus of literature from the past. The presentation includes a standardized flow diagram for generating a comprehensive summary that incorporates both static and dynamic elements, along with the classification of the approaches and techniques employed in the process.

2. Standard Flow of Video Summary

The block diagram is represented in Figure 1 showing essential blocks of video summarization approaches that were developed by different researchers.

2.1 Segmentation

It is a pre-processing block that uses image segmentation techniques. In this block, the video is segmented into small pieces with a chronological segmentation and are independently processed [1, 2, 3]. This small part denotes a set of a minimum number of frames which includes activities to convey some meaning. In [7], frames were first pre-processed by minimizing the sizes in order to increase the computation speed and then converted to grayscale intensity from the RGB color space.

2.2 Feature Extraction Techniques

To make the summary content more accurate and pleasant, multiple approaches/techniques have been developed. A detailed review of the feature extraction technique is discussed in section 3.

2.3 User Preferences

The user preferences block allows users to specify their desired requirements for the summary generation process. This block typically includes parameters such as the number of key frames, the length of skims, the type of skimming (summary sequence or highlights), and other customized parameters based on specific application scenarios. In this context, highlights refer to significant events in the video, which are particularly relevant for movie and surveillance video skimming. On the other hand, summary sequence represents a condensed form complete content of the video, which is often relevant in sports-related skimming [3].

2.4 Unit Selection

On the basis of unit relevance, summary length, and other user factors, the selection of the unit and the reduction in redundancy determines the units that is included in the summary of the video [3]. To produce a non-redundant summary of video that fully covers the relevant information in the original video, this block also eliminates comparable summary units from the video skim. In the final output, three types of summary are generated which is either a key frame (static summary) or summary sequence (skim) or highlight (skim).

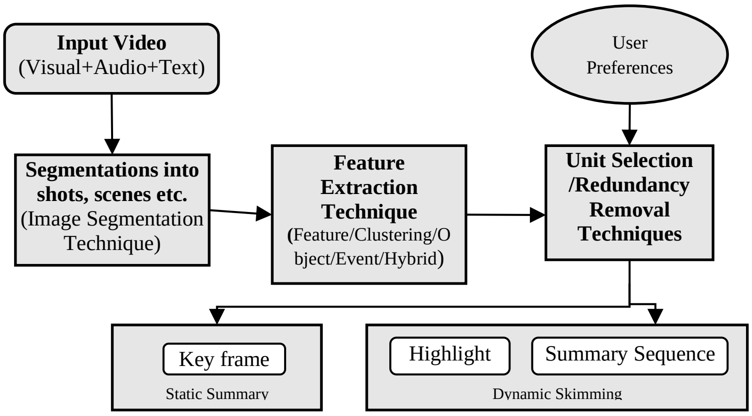
Input Video
(Visual+Audio+Text)
User Preferences
Segmentations into shots, scenes etc.
(Image Segmentation Technique)
Feature Extraction Technique
(Feature/Clustering/Object/Event/Hybrid)
Unit Selection /Redundancy Removal Techniques
Key frame
Static Summary
Highlight
Summary Sequence
Dynamic Skimming

Figure 1. Generic Block Diagram of Video Summary

3. Classification of Video summary Techniques

3.1 Feature-based Summary

To produce a feature-based summarization features of low level, colour, motion, texture, and edges are considered. Then, the histogram of mentioned features are obtained using certain frame difference measure. This frame difference measure represents a sudden change that occurred in visual contents which is very sensible and works well to generate a summary [1].

Suet Peng Yong et al. [15] presented the method of key frame extraction using LUV col-or histogram and texture features. Ying Li et al. [20] utilized the low-level properties in selecting the keyframe. Initially, color histogram and Scale Invariant Feature Transform were used to extract key frame, and later, the clustering technique was applied to generate a summary. Similarly, Naveed Ejaz et al. [32] attempted to produce summaries.

3.2 Event-based Summary

Event-based summaries are based on specific incidents from the actual video, like identifying unexpected changes such as a road accident, mobile snatching, violent activity, and for sports video, events like goals, penalty, and boundary- hit and so forth. In some cases, like commercial videos, either the text or graphics is considered.

Research [9] by Yanwei Fu et al. was based mostly on the recognition of important events while examining the spatiotemporal dynamics and visual characteristics of relevant objects. In [8], an automatic content - based video summarization method for large sports video archives was proposed, where each play scene (Event) was chosen according to the significance of play scene which had three components: [i] Play ranks [ii] Number of replays and making summary, and [iii] Play Occurrence Time.

3.3 Motion-based Summary

Motion of a video indicates the measure of the actions of the objects/background in a sequence of frames [13]. Motion is a more efficient feature to concentrate on the visual content of the shot, where more pans and zooms of camera are there. Mendi et al. [7] proposed key frame selection using motion analysis where motion metrics were calculated from two optical flow algorithms, using a different set of key frame selection criteria [12].

3.4 Clustering-based Summary

Clustering is the process of grouping a set of objects with another, the objects that are dissimilar to the objects in other clusters.

The research [9] presented that random walk are employed to cluster events centered on similar shots. In the random walk from each unsampled node, the most preferable sampled shot's cluster was determined. Finally, [17] applied clustering techniques on role community network in order to cluster particular movie segments into relevant groups to generate a final movie dynamic skim. In [13], a set of representative frames of the entire video was obtained using k-means clustering followed by motion detection.

3.5. Deep Learning -based Summary

Recently deep learning-based model has become very prevalent in producing human-like video summaries. A learned model is utilized to estimate and predict the video movements.

In [10], every frame in the video was divided into patches and then the patches were provided with the global deep features obtained from a CNN that is pretrained. In [16], down sampled videos were trained on Image Net. Mengjuan Fei et al. [21] utilized the maximum entropy and the memorability score of the images in a deep pretrained network to produce the summary of the video. The study showed that the majority of the applications of deep learning-based models outperform the conventional methods in a supervised task-based manner due to their ability to learn complex features.

3.6 Audio-based summary

Audio features are also considered as a sole source for the analysis of video. Many authors have also used combined audio, visual and textual attention for movie summarization just as presented in [17].

In [11], audio features were extracted through signal instantaneous amplitude and frequency to generate summary sequence. In [14], supervised audio classification was performed, which classified audio into four groups such as ball impact, cheer, silence or speech including both time-domain and frequency-domain.

A brief summary of the video summarization approaches is provided in Table 1 with the summary type and evaluation approach used.

Table 1. Brief Overview of Selected Summarization Techniques

Ref. No.	Approaches/ Methodologies	Dataset/Video Domain	Summary Type	Evaluation Approach
[10]	Deep Learning based	SumMe and TVSum Dataset	Key frame	Objective
[22]	Deep Learning based	SumMe and TVSum Dataset	Key frame	Subjective, Objective
[13]	Motion based, Clustering based	SumMe and TVSum Dataset	Key frame / Video Skim	Subjective

[09]	Clustering based	Office surveillance video		Objective
[11]	Color Based, Wavelet based, Text based, Audio based	Action, Horror, and Sci-fi movie	Video Skim	Subjective
[07]	Motion based	Rugby and Soccer Sports videos	Video Skim	Subjective
[16]	Deep Learning based	YoutTube videos, SumMe and TVSum Dataset	Key frame	Objective
[18]	Deep Learning based	SumMe and TVSum Dataset	Key frame	Subjective, Objective
[17]	Clustering based	Movie Videos	Video Skim	Subjective
[14]	Clustering based	Racquet Sports video	Video Skim	Subjective
[15]	Colour based	Wild Life videos, SumMe and TVSum Dataset	Video Skim	Subjective
[21]	Deep Learning based	Different Youtube Videos	Key frame	Subjective, Objective
[12]	Motion based	SumMe and TVSum Dataset	Key frame	Objective

4. Classification of summary based on domain

However, based on the domain, researchers have applied different techniques or approaches of video summarization. Therefore, a brief study and classification has been presented, that classify these domains into many groups classification of summarization techniques based on domain, which is shown in Table 2. For summarizing the videos, a variety of methods have been offered. For video summarizing, these techniques can be divided into numerous categories or domains.

Table 2. Classification of Summarization Techniques based on Domain

Domain	**Techniques/Approaches Used**	**References**
TVSum and SumMe dataset	Color, Deep Learning, Motion, Clustering based approaches	[10][13][16][18][25]
Sports	Color, Motion, Clustering, Deep Learning, Graph, Audio, Event based approaches	[7][8][14][21][22][29][32]
Movies/	Event, Color, Wavelet, Text, Audio,	[11][17][19][21[29][32]

Cartoon	Clustering based approaches	
News Highlights	Color, Motion, Low level Descriptor, Clustering based approaches,	[21][28][29][32]
Lecture	Color , Motion and Clustering based approach	[23][30][29][32]
Surveillance	Graph and Event based approaches	[9]
Wild Life	Color and Clustering based approaches	[15]

5. Evaluation Approaches

The two evaluation approaches for video summarization are, intrinsic and extrinsic methods. When using extrinsic approaches, a video summary is typically evaluated with respect to how well it achieves a particular information retrieval task [33]. In intrinsic method, the quality of a generated video summary is judged directly based on summary analysis, where the criteria can be user judgment of fluency of the video summary [17]. The methods presented in [7], [8], [13], [14], [15], [17] and [19] use intrinsic evaluation approach and methods proposed in [9], [10], [12], [16] and [25] use extrinsic evaluation approach.

In [32], the author evaluated the technique proposed by [6] in which two matrices called Accuracy Rate (CUSA) and Low Error Rate (CUSE) were determined, and the algorithm generated summaries were compared with the user generated summaries based on defined matrices. In context of multi view video summaries, [9] calculated length of summary and number of events presented in summary. Sometimes, in extrinsic approach users are invited to participate and asked to evaluate enjoyability and informativeness which are represented using two parameters recall and precision.

6. Other Related Works

Research related to video summarization has greatly improved recently. As a result, several different strategies have been created, other than the conventional approaches. Y. Takahashi et al. [8] proposed a method to generate key frame and video posters by creating metadata that has a semantic description of video content. Then summaries are created according to the significance of each video content, which is normalized to handle large sports archives. In [19] Arthur G. Money et al. presented an effective approach, which measures physiological response measures like heart rate and blood volume pulse. Based on the physiological responses of users, an analytical framework has been generated to identify the most entertaining segments. Finally, it was concluded that without requiring any conscious input from the user’s external information in the form of physiological response considerably works well on movies like sci-fi action, horror, and thriller.

7. Conclusion and Findings

This study presents a brief comparative analysis of video summarization approaches based on domain, evaluation approach, and types of summary to be generated. Few observations and challenges from the above study are listed below:

- In order to select the most appropriate technique, this comparative study will guide users. At first, the study shows that a clustering-based approach summarized videos more accurately, and most of the researchers have focused more on it, compared to other techniques.
- It is critical to generalize one method, so feature-based summary techniques, especially color-based and low-level descriptors, are used with the aggregation of clustering-based approaches in order to provide relatively simple and effective solutions.
- This study also highlights that audio classification is more appropriate for the classification of domain-dependent videos such as sports and movies.
- In a recent development, deep learning techniques work well for the classification part. However, to generate a skimming video, deep learning requires a large amount of data for training and an efficient hardware specification, which is non-viable for most researchers.
- While recent developments have improved the accuracy of video summarization models, scaling these models to handle large-scale video datasets remains a challenge. Efficient algorithms and techniques are needed to summarize videos quickly and effectively.
- Video summarization is inherently subjective, as the definition of what constitutes important or representative content may vary depending on the user's preferences or the application domain. Developing universally satisfactory summaries remains a challenge.

Thus, the present work will help users choose specific strategies for the target domain.

References

[1] Ying Li, Tong Zhang, Daniel Tretter, “An overview of video abstraction techniques”. Proceedings of Tech. Rep., HP-2001-191, HP Laboratory (2001).

[2] Arthur G. Money and Harry Agius. 2008. “Video summarisation: A conceptual framework and survey of the state of the art”. J. Vis. Comun. Image Represent. 19, 2,121–143 February (2008).

[3] Vivekraj V. K., Debashis Sen, and Balasubramanian Raman. 2019. “Video Skimming: Taxonomy and Comprehensive Survey”. ACM Comput. Surv. 52, 5, Article 106, October (2019).

[4] Haq, Hafiz Burhan & Asif, M & Bin, Maaz. “Video Summarization Techniques: A Review”. Inter-national Journal of Scientific & Technology Research. 9. 146-153 (2021).

[5] Song, Yale, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes, ”TVSum: Summarizing web videos using titles,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,5179- 5187 (2015).

[6] Gygli, Michael and Grabner, Helmut and Riemenschneider, Hayko and Van Gool, Luc, ”Creating Summaries from User Videos,” European conference on computer vision, Zurich,505-520 (2014).

[7] Mendi, Engin & Clemente, Hélio & Bayrak, Coskun. “Sports video summarization based on motion analysis. Computers & Electrical Engineering”. 39. 790–796 (2013).

[8] Y. Takahashi, N. Nitta and N. Babaguchi, "Video Summarization for Large Sports Video Archives," 2005 IEEE International Conference on Multimedia and Expo, pp. 1170-1173 (2005).

[9] Y. Fu, Y. Guo, Y. Zhu, F. Liu, C. Song and Z. Zhou, "Multi-View Video Summarization," in IEEE Transactions on Multimedia, vol. 12, no. 7, pp. 717-729 (2010).

[10] S. Mei, M. Ma, S. Wan, J. Hou, Z. Wang and D. D. Feng, "Patch Based Video Summarization With Block Sparse Representation," in IEEE Transactions on Multimedia, vol. 23,732-747 (2021).

[11] G. Evangelopoulos et al., "Multimodal Saliency and Fusion for Movie Summarization Based on Aural, Visual, and Textual Attention," in IEEE Transactions on Multimedia, vol. 15, no. 7, 1553-1568, (2013).

[12] P. D. Byrnes and W. E. Higgins, "Efficient Bronchoscopic Video Summarization," in IEEE Transac-tions on Biomedical Engineering, vol. 66, no. 3, pp. 848-863 (2019).

[13] I. Alam, D. Jalan, P. Shaw and P. P. Mohanta, "Motion Based Video Skimming," 2020 IEEE Cal-cutta Conference (CALCON), pp. 407-411(2020).

[14] Liu, Chunxi & Jiang, Shuqiang & Xing, Liyuan & Ye, Qixiang & Gao, Wen. “A framework for flexible summarization of racquet sports video using multiple modalities”. Computer Vision and Image Understanding. 113. 415-424 (2009).

[15] Yong, Suet & Deng, Jeremiah & Purvis, Martin. “Wildlife video key-frame extraction based on novelty detection in semantic context”. Multimedia Tools and Applications, (2013).

[16] Ji, Zhong & Jiao, Fang & Pang, Yanwei & Shao, Ling. “Deep Attentive and Semantic Preserving Video Summarization”. Neurocomputing. 405,(2020).

[17] C. Tsai, L. Kang, C. Lin and W. Lin, "Scene-Based Movie Summarization Via Role-Community Networks," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 23, no. 11,1927-1940 (2013).

[18] Wei, H., Ni, B., Yan, Y., Yu, H., Yang, X., & Yao, C. “Video Summarization via Semantic Attend-ed Networks”. AAAI, (2018).

[19] Money, Arthur & Agius, Harry. “Analysing user physiological responses for affective video summa-rization”. Displays. 30. 59-70 (2009).

[20] Yueting Zhuang, Ruogui Xiao and Fei Wu, "Key issues in video summarization and its applica-tion," Fourth International Conference on Information, Communications and

Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint, 448-452(2003).

[21] Mengjuan Fei, Wei Jiang, Weijie Mao,"Memorable and rich video summarization", Journal of Visual Communication and Image Representation,Volume 42, Pages 207-217,ISSN 1047-3203 (2017).

[22] Melissa Sanabria, Frédéric Precioso, Thomas Menguy. Hierarchical Multimodal Attention for Deep Video Summarization. 25th International Conference on Pattern Recognition, Milan, Italy, (2021)

[23] Salim, Fahim & Haider, Fasih & Luz, Saturnino & Conlan, Owen. Automatic Transformation of a Video Using Multimodal Information for an Engaging Exploration Experience. Applied Scienc-es(2020)

[24] Amr Abozeid, Hesham Farouk, and Kamal ElDahshan. "Scalable Video Summarization: A Compar-ative Study". In Proceedings of the International Conference on Compute and Data Analysis (ICCDA '17). Association for Computing Machinery, New York, NY, USA, 215–219 (2017).

[25] Naveed Ejaz, Irfan Mehmood, Sung Wook Baik, "Feature aggregation based visual attention model for video summarization", Computers & Electrical Engineering,Volume 40, Issue 3,Pages 993-1005,ISSN 0045-7906 (2014).

[26] Psallidas, T.; Koromilas, P.; Giannakopoulos, T.; Spyrou, E. "Multimodal Summarization of User-Generated Videos". Appl. Sci., 11, 5260 (2021).

[27] Avola D., Cinque L., Foresti G.L., Martinel N., Pannone D., Piciarelli C. "Low-Level Feature De-tectors and Descriptors for Smart Image and Video Analysis: A Comparative Study". In: Kwaśnicka H., Jain L. (eds) Bridging the Semantic Gap in Image and Video Analysis. Intelligent Systems Reference Library, vol 145. Springer, (2018).

[28] Liang B, Li N, He Z, Wang Z, Fu Y, Lu T. "News Video Summarization Combining SURF and Color Histogram Features". Entropy.; 23(8):982 (2021).

[29] Enabzadeh, Roya and Behrad, Alireza. 'Video Summarization Using Sparse Representation of Local Descriptors'. 1: 315 – 327 (2019).

[30] Badri Narayan Subudhi, Thangaraj Veerakumar, Sankaralingam Esakkirajan, Santanu Chaudhury, “Automatic lecture video skimming using shot categorization and contrast based features, Expert Systems with Applications”, Volume 149,(2020).

[31] I. Alam, D. Jalan, P. Shaw and P. P. Mohanta, "Motion Based Video Skimming," 2020 IEEE Cal-cutta Conference (CALCON), pp. 407-411(2020).

[32] Naveed Ejaz, Tayyab Bin Tariq, and Sung Wook Baik. 2012. Adaptive key frame extraction for video summarization using an aggregation mechanism. J. Vis. Comun. Image Represent. 23, 7 1031–1040, (2012).

[33] Taskiran, Cuneyt.”Evaluation of Automatic Video Summarization Systems”. Proceedings of SPIE - The International Society for Optical Engineering. (2006).

Author's Biography

Bijal U. Gadhia is pursuing Ph.D. in Computer Engineering from Gujarat Technological University (State University), Gujarat, India. Currently, She is a faculty member at Government Engineering College, Gandhinagar (Government Employee), Gujarat, India, and has served several governmental activities around the university and outside. Area of teaching includes Compiler Design, Python for DataScience, Object oriented Programming in Java, and Advanced Java. Her research interests are the application of Deep Learning, Machine Learning, Image Processing, and Data Science.

Dr. Shahid S. Modasiya is an Assistant Professor at the Department of Electronics and Communication Engineering at Government Engineering College, Gandhinagar under the affiliation of Gujarat Technological University. His research interest areas are RF & Microwave and antenna design. He has also published various papers in the field of his research interest.