REVIEW ARTICLE

SAMRIDDHI Volume 15, Issue 2, 2023 Print ISSN: 2229-7111 Online ISSN: 2454-5767

A Comparative Analysis of Video Summarization Techniques for Different Domains

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

[1]Gujarat Technological University, Ahmedabad, Gujarat, India.

[2]Electronics and commnication Department, GEC Gandhinagar, Gujarat, India.

Abstract

As technology progresses, a gigantic amount of video data is generated day by day. Pro-cessing such a huge video needs time and requires increased storage and computational power. Sometimes it is convenient for the user to watch a summary or highlight rather than a complete video, which is time-consuming. So, a fully automated solution is required to extract important segments from a 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 paper. Primarily, these methods are classified into different categories based on their methods or techniques used. Then an over-view of some the latest literature is presented with the dataset and evaluation approaches used. The review is also made related to the domain direction and concluded by presenting benefits and difficulties associated with current video summarization techniques.

Keywords: Video summarization, Video skimming, Static summarization, Keyframe.

SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology (2023); DOI: 10.18090/samriddhi.v15i02.11

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 that will reduce the cost of cataloging, indexing, and video archiving while also increasing the effectiveness, usability, and accessibility of stored content[1]. There is a huge need for videos. One crucial subject among all potential study subjects is enabling easy browsing of a sizable video data collection and explaining how to create effective content both 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 that requires expertise and is highly expensive.

In order to provide a preview of the video content, researchers started studies on almost 30 years ago with a target of generating keyframes, and then generated short clips which cover important video segments, also referred to as video highlights.[1] But the sequence of images was found insufficient for users to understand the video, particularly in lengthy videos.[3] These Keyframe-based summaries served the needs of thumbnail representation of the film as well as video browsing and indexing. Such a keyframe 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 improves the information conveyed by the summarization, which generates shorter videos, referred to as video skims, consisting of important segments with corresponding audio information.[3] The Video Skimming approach, which calls for dynamic video summarization, is used to overcome these issues by generating a temporally shortened and summarized version of a given video.[1-3] Because of its dynamic nature, video skimming represents a better understanding of the video from its summary. The following are some major advantages of video skimming and dynamic video summarization:

How will rapid browsing be enabled?

How can content access and representation be made more effective?

How can one present the plot in less time?

This study gives a complete survey on video summarization, concentrating on the large corpus of literature from the past.

Corresponding Author: Bijal U. Gadhia, Gujarat Technological University, Ahmedabad, Gujarat, India, e-mail: bij.1988@gmail.com

How to cite this article: Gadhia, B.U., Modasiya, S.S. (2023). A Comparative Analysis of Video Summarization Techniques for Different Domains. *SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology,* 15(2), 253-257.

Source of support: Nil

Conflict of interest: None

It presents a generic flow diagram to summarize and classify approaches/techniques used.

Generic flow of video summary

The block diagram in Figure 1 shows essential blocks of video summarization approaches that different researchers are developing.

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 these units are processed independently.[1-3] This small part denotes a set of a minimum number of frames, including activities to convey meaning. In[7] frames are first pre-processed by reducing their sizes to speed up the computation and then converted from the RGB color space to grayscale intensity.

Feature Extraction Technique

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

User Preferences

The user preferences block accepts user requirements for the summary to be performed. It typically includes no. of Keyframes, skim length, an input like which skimming to perform (summary sequence or highlights), and any other parameter customized for an application scenario. Highlights means a representation of an important events of the video, usually relevant for movies and surveillance video skimming, and summary sequence means a representation of the entire video content by the skim, usually relevant for sports.[3]

Unit Selection

On the basis of unit relevance, summary length, and other user factors, the unit selection and redundancy reduction block determines which units should be included in the video summary.[3] In order to generate a non-redundant video summary 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 will be generated which is either a Keyframe (static summary) or summary sequence (skim) or highlight (skim).

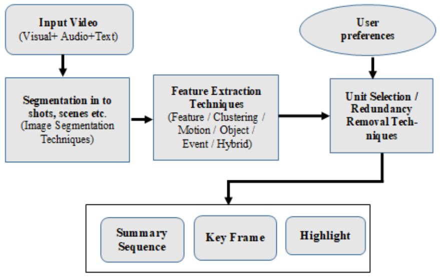
Input Video
(Visual+ Audio+Text)
User preferences
Segmentation in to shots, scenes etc. (Image Segmentation Techniques)
Feature Extraction Techniques (Feature / Clustering / Motion / Object / Event / Hybrid)
Unit Selection / Redundancy Removal Tech-niques
Summary Sequence
Key Frame
Highlight

Figure 1: Generic Block Diagram of video summary

Classification of video summary techniques

Feature-based summary

In order to generate feature based summaries low level features such as colour, motion, texture, edges are considered. Then histogram of mentioned features are obtained using certain frame difference measure. This frame difference measures represents a sudden changes happened in visual contents which is very sensible and works well to generate a summary.[1]

Suet Peng Yong *et al.*[15] presented the Keyframe extraction method using LUV color histogram and Texture features. Ying Li *et al.*[20] have presented the technique, where Keyframes selection is performed based color and low level descriptor properties. Initially, color histogram and SIFT (Scale Invariant Feature Transform) is used to extract Keyframe and later on clustering technique is applied to generate a summary. Similarly, Naveed Ejaz *et al.*[32] attempt to produce summaries that also uses three visual features like color histogram, wavelet statistics and edge direction.

Event-based summary

Event-based summaries are considered when events are identified during analysis and subsequently used for creating the video summary. These 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. Whereas for a popular movie, a fight or a love scene may be a more common event. Text or graphics may also be considered to explicitly represent events to generate highlights or to support the selection of suitable video segments based on the detection of specific events indicated.

Yanwei Fu *et al.* presented in[9] is 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 is proposed where each play scene (Event) is chosen according to the significance of play scene for which has three components: [i] Play ranks [ii] Number of Replays and making summary[iii] Play Occurrence Time.

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 Keyframe selection using motion analysis where motion metrics are calculated from two optical flow algorithms, using a different set of Keyframe selection criteria. P. D. Byrnes *et al.*[12] defined a method consisting of three main steps: 1) shot segmentation, to perform a preliminary grouping of informative frames into shots; 2) motion analysis, to ascertain the local motion between frames constituting individual shots; and 3) keyframe selection.

Clustering-based summary

This approach treats video frames as points in the feature space (color or edge histogram) and works on the assumption that the center of clusters can be used as keyframes for the entire video sequence. Clustering is the process of grouping a set of objects with other, but are dissimilar to objects in other clusters into classes or clusters so that objects within a cluster have similarities in comparison to one another but are dissimilar to objects in another cluster.

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

Deep Learning-based summary

Recently Deep learning-based model become very popular for producing human-like video summaries. Researchers are creating a model to perform statistical analysis on the given training examples. Then the models that were generated afterwards effectively link the underlying patterns depicted in the video with inferences (training) that can be used to generate highlights for new (testing) videos. A learned model to mimic humans is fundamentally different from the other techniques because it does not explicitly consider well-known human acts of human video understanding but also statistically estimates them.

In[10] a video frame is first partitioned into a number of patches, and each patch is represented with a global deep feature extracted from a pre-trained convolutional neural network (CNN). In [16] down sampled Videos are trained on Image Net. Then, Bidirectional LSTM (BiLSTM) is selected as the encoder to encode the temporal relation information in a sequence which then formulated as final Output. Mengjuan Fei *et al.*[21] predict the memorability score by using the trained deep network and calculating the images' entropy value. The image with the maximum entropy value and memorability score and in each shot was selected to produce the video summary. The study shows 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.

Audio-based summary

Audio features are also considered as sole source for analysis of video. Where the author look for content such as cheering, applause, music, speech excited speech to extract important event that occur in the video which used to general highlights. Similarly, many authors have also used combined audio, visual and textual attention for movie summarization, as presented in.[17]

In[11] audio features are extracted through signal instantaneous amplitude and frequency to generate summary sequence. In[14] supervised audio classification are performed which classifies audio into four groups like ball impact, cheer, silence or speech including both time-domain and frequency-domain. A brief summary of the video summarization approaches method is provided in Table 1 with the Summary Type and Evaluation approach used.

Classification of summary based on domain

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

Evaluation Approaches

To evaluate the performance of the summarization method, there are two evaluation approaches. It can be classified into 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] Here, the method presented in[7,8,13-15,17,19] uses an intrinsic evaluation approach, and methods proposed in[9,10,12,16,25] use an extrinsic evaluation approach. In[32] author evaluated the technique proposed by[6] in which they have determined two matrices called Accuracy Rate(CUSA) and Low Error Rate(CUSE) and compared their algorithm-generated summaries with user-generated summaries based on defines matrices. In the context of multi-view video summaries,[9] calculated length of summary and no. of events presented in the summary. Sometimes, in the extrinsic approach, users are invited to participate and asked to evaluate enjoyability and informativeness, represented using two parameters: recall and precision.

Other related work

Research related to video summarization has greatly improved recently. As a result, several different strategies have been created, other than the conventional approaches In Y. Takahashi *et al.*[8] proposed a method to generate keyframes 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.* has presented an effective approach,

Table 1: A brief overview of some selected Summarization Techniques

Ref. No.	*Approaches/ Methodologies*	*Data Set/Video Domain*	*Summary Type*	*Evaluation Approach*
[10]	Deep Learning based	SumMe and TVSum Dataset	Keyframe	Objective
[22]	Deep Learning based	SumMe and TVSum Dataset	Keyframe	Subjective, Objective
[13]	Motion based, Clustering based	SumMe and TVSum Dataset	Keyframe/Video Skim	Subjective
[09]	Clustering based	Office surveillance video	Video Skim	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	Keyframe	Subjective
[16]	Deep Learning Based	YouTube videos ,SumMe and TVSum Dataset	Keyframe	Objective
[18]	Deep Learning Based	SumMe and TVSum Dataset	Video Skim	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	Keyframe	Subjective
[21]	Deep Learning based	Different Youtube Videos	Keyframe	Subjective, Objective
[12]	Motion based	SumMe and TVSum Dataset	Keyframe	Objective
[32]	Colour, Motion based	Open video Project	Keyframe	Objective

Table 2: Classification of Domains with different summarization Techniques.

Domain	*Techniques/Approaches Used*	*Worked in Literature*
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/ Cartoon	Event, Color, Wavelet, Text, Audio, Clustering based approaches	[11,17,19,21,29,32]
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]

which measures physiological response measures like Heart Rate (HR), Blood Volume pulse (BVP). Based on the physiological responses of user an analytical framework has been generated to identify most entertaining segments. Finally, they have concluded that without requiring any conscious input from the users, external information in the form of physiological response considerably works well on movies like sci-fi action, horror and thriller.

Conclusion

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

In order to select the most appropriate technique, this comparative study will guide users. The first study shows that a clustering-based approach summarized videos more accurately and that most researchers have focused more on it than 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.

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.

This study also highlights that audio classification is more appropriate for the classification of domain-dependent videos such as sports and movies.

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". International 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 Transactions 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 Calcutta 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 keyframe 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 Attended Networks". AAAI, (2018).

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

[20] Yueting Zhuang, Ruogui Xiao and Fei Wu, "Key issues in video summarization and its application," 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 Sciences(2020)

[24] Amr Abozeid, Hesham Farouk, and Kamal ElDahshan. "Scalable Video Summarization: A Comparative 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 Detectors 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 Calcutta Conference (CALCON), pp. 407-411(2020).

[32] Naveed Ejaz, Tayyab Bin Tariq, and Sung Wook Baik. 2012. Adaptive Keyframe 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).