UTM
UNIVERSITI TEKNOLOGI MALAYSIA
INTERNATIONAL JOURNAL OF
INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

Digital Video Summarization: A Survey

Sajjad H. Hendi*
Informatics Institute for Graduate Studies
Baghdad, Iraq
Email: phd202030562@iips.icci.edu.iq

Hazeem B. Taher
The University of Thi-Qar
College of Education for Pure Sciences
Thi-Qar, Iraq

Karim Q. Hussein
Mustansiryha University-Faculty of Science
Computer Science Dept.
Baghdad, Iraq

Submitted: 30/11/2022. Revised edition: 31/3/2023. Accepted: 31/3/2023. Published online: 13/9/2023
DOI: https://10.11113/ijic.v13n1-2.421

***Abstract*—Video summarization has arisen as a method that can help with efficient storage, rapid browsing, indexing, fast retrieval, and quick sharing of the material. The amount of video data created has grown exponentially over time. Huge amounts of video are produced continuously by a large number of cameras. Processing these massive amounts of video requires a lot of time, labor, and hardware storage. In this situation, a video summary is crucial. The architecture of video summarization demonstrates how a lengthy film may be broken down into shorter, story-like segments. Numerous sorts of studies have been conducted in the past and continue now. As a result, several approaches and methods—from traditional computer vision to more modern deep learning approaches—have been offered by academics. However, several issues make video summarization difficult, including computational hardware, complexity, and a lack of datasets. Many researchers have recently concentrated their research efforts on developing efficient methods for extracting relevant information from videos. Given that data is gathered constantly, seven days a week, this study area is crucial for the advancement of video surveillance systems that need a lot of storage capacity and intricate data processing. To make data analysis easier, make it easier to store information, and make it easier to access the video at any time, a summary of video data is necessary for these systems. In this paper, methods for creating static or dynamic summaries from videos are presented. The authors provide many approaches for each literary form. The authors have spoken about some features that are utilized to create video summaries.**

***Keywords*—Video summarization, static, dynamic, Surveillance video, Machine learning, Deep learning, Keyframe**

I. Introduction

Over the years, the Internet and social networking sites have experienced a massive surge of multimedia content, particularly video [1]. Now, video technologies are confronting a variety of issues and difficulties, most of which are attributable to the extraction of information in real time from a huge number of sources [2]. The information that was retrieved may be used to identify and detect several events that can aid in numerous studies, such as anomalous occurrences and human behavior, as well as to forecast events that typically take place in the scenarios [3]. Many scholars have recently concentrated their research efforts on developing efficient methods for extracting relevant information from films [4].

Given that data is gathered constantly, seven days a week, this study area is crucial for the advancement of video surveillance systems that need a lot of storage capacity and intricate data processing [5]. To make data analysis easier, make it easier to store information, and make it easier to view videos at any time, a summary of video data is necessary for these systems [6]. The process of summarizing may also be correlated with the type of scene (private or public), where the data analysis is dependent on the scene's dynamic or static nature, as well as its density (whether it is crowded or not) [7]. Since the summarizing method ought to take less time to complete and less storage space, pre-processing can be necessary to improve the process without erasing any data before the feature extraction work [8], [9].

Static, dynamic, picture and text summaries are the most often seen forms of summaries in literature. Each summary format has distinguishable properties. A group of frames from the input video that the summarization framework has determined to be keyframe are combined to create a static summary. Dynamic summaries, also known as video skims, are summaries created by grouping together playable video clips or images. Image summaries are made by putting together important things or objects from different frames. Instead of frames or shots, images are used as a summary. It can be a single picture or a group of pictures. Text summaries are created using Natural Language Processing (NLP) approaches using summarization frameworks [10].

II. Video Summarization

Static summaries, dynamic summaries, picture summaries, text summaries, hierarchical summaries, and multi-view summaries are the various types of summaries that may be created from a video. In this section of the article, we will provide an overview of the most significant research studies that have been published in the field of video summarization throughout the time period that was discussed in the section that served as the article's introduction. Only forms of static and dynamic video summaries are covered in this article.

In 2018 [11] proposed Adaptive mean shift-based keyframe extraction for video summarization (MSKVS), an effective approach for static video summaries based on keyframe extraction. Three main elements make up the suggested MSKVS first: a new feature representation is used to describe the visual content of the video; quick and easy. second: an algorithm is proposed to remove the majority of similar and redundant frames. third: an adaptive mean shift algorithm is utilized to choose the keyframes that best represent the scene. provide a novel verification method to assess the information maintained by the created summary and ensure that it merits showing the whole video stream regardless of the influence of the viewer's viewpoint. Using several evaluation measures, the experimental findings on six difficult datasets demonstrate that MSKVS provides state-of-the-art performance in a short computing time.

In 2020 [12] suggested an event summary method for monocular videos utilizing a deep learning architecture. To create a similarity matrix based on the visual attributes, a spatiotemporal similarity function is devised. According to an objective function, video frames are represented by the sparse matrix as graph vertices, and Highly Connected Subgraphs (HCS) are built as clusters. Finally, events are derived from such clusters under the presumption that the cluster centroid serves as a key frame for the event. As a result, the method avoids making an assumption about the number of clusters. Because of this, users can choose how many keyframes to use without having to wasted cost extra for processing. The suggested framework surpasses the state-of-the-art models on both Precision and F-measure, and it successfully recaptures the essential elements of the original video, as demonstrated by experiments conducted on two benchmark datasets.

In 2019 [13] A motion detection-based approach for summarizing videos has been presented. The main constraints of background removal techniques are sensor noise (noise of acquisition and digitization) and light variations in the scene. This study presents a method for handling these issues that combines background removal and the Structure-Texture-Noise Decomposition. The first step is to separate each grayscale picture in the series into three parts: Structure, Texture, and Noise. The background model is created by extracting the Structure and Texture elements from each image in the series. The background was subtracted from the absolute difference before computing the binary picture of moving objects. Additionally, the findings of the background removal were used to generate a summarization video. The change throughout the course of the series is computed using the produced background model. The testing findings show that the method is efficient and precise for detecting moving objects and produces a decent summary of the video stream.

In 2019 [14] The proposed summarizing technique in this work is based on choosing a threshold-based system that can choose keyframes that are best suited for storage and subsequent analysis. In order to determine the keyframes, a global threshold based on the Otsus approach is first applied to all of the frames of a surveillance video. Based on the threshold, a retrospective comparison using statistical methods is then performed on each frame. Based on the repeated comparison of frames using both global and local threshold comparison, a similarity index is produced. The local threshold is indexed using the following metrics: Tanimoto Similarity (TS), Tversky feature contrast model (TFCM), Pearson coefficient of mean square contingency, and Analysing Method Patterns to Locate Errors (AMPLE) (Pmsc). Every time a keyframe is chosen, the global threshold is revised based on a comparison between the local and global thresholds. The outcomes are compared with five surveillance videos and six keyframe identification techniques. The performance is measured using a statistic called Selection Rate. To decrease the number of video surveillance frames the suggested approach performs key-frame selection that is increasingly discrete. Six techniques were used for the study in order to compare the similarities between frames and extract keyframes. If the similarity index is smaller than the similarity index determined by the global threshold, the frames are selected as the keyframes. Based on the results of the approaches used, the global threshold changes itself and its value as shown in Fig. 1.

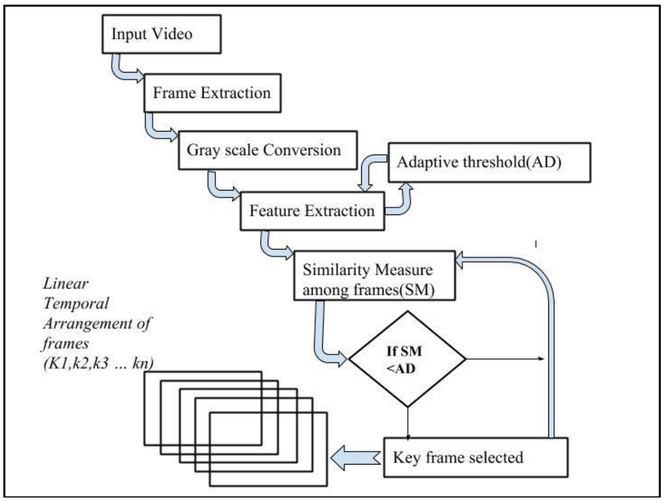

Fig. 1. Overall Sequence for Key-Frame Extraction

In 2019 [15] provided an online system for intelligent video recording, coarse and fine redundancy elimination, and summary creation. First, coarse redundancy elimination is applied to video data collected by devices with limited resources connected to an Industrial Internet of Things (IIoT) network and equipped with vision sensors. In order to pick potential keyframes, sequential characteristics are retrieved from the output frames and sent to the cloud for in-depth analysis, and this is the second step. Then candidate keyframes were then fine-tuned to identify those that would provide the most information in the summary. The coarse and fine refinement of video data implemented on devices with limited resources, as well as the presenting of significant data as a summary, are the main contributions of this study. When compared to state-of-the-art, experiments using publicly accessible datasets show a 0.3-unit gain in the F1 score with less time complexity. Additionally, delivered strong results on our recently produced dataset in an industrial setting, which is made accessible to the research community together with its labelled ground truth.

In 2020 [16] suggested a successful approach for video summaries. The system combines a type-specific summarization method with a cluster-based object tracking approach. An algorithm based on trajectories was developed to summarize surveillance video feeds. The suggested method collects foreground objects from input videos using a cluster-based tracking algorithm, then separates anomalous objects and categorizes them according to their movement patterns. A type-based summary technique is then put forth to effectively group moving objects with various pattern kinds within limited time endurance. Faster Region-based Convolutional Neural Networks (Faster-RCNN) was used to extract the foreground bounding box from each frame of the input video after being pre-trained on COCO as the human detector. Followed the cluster-based workflow for trajectory creation. Fifteen lengthy surveillance videos totaling ten minutes each from various street scenes were gathered as the training set for the summary module. as a consequence, this approach would be beneficial for precise and quick analysis of surveillance movies and produces positive results in the work of general surveillance video synopsis. The system also supports the user interaction feature, which enables users to locate an interesting foreground object from a summary video with ease.

In 2020 [17] suggested incorporating the concepts of both encoder-decoder attention and semantic preserving loss in a deep Seq2Seq framework for video summarization to pay close attention to discovering the inherent relationships between the original video and its summary while meanwhile minimizing the semantic information loss. To increase the robustness of the model against outliers, they additionally incorporate Huber loss to replace the well-known mean square error loss. On three publicly accessible standard datasets for extracting features, TVSum, and YouTube—they assess the proposed Deep Attentive and Semantic Preserving video summarization (DASP) technique. The suggested method regularly outperforms state-of-the-art ones, as shown by extensive tests on two benchmark videos summarizing datasets as shown in Fig. 2.

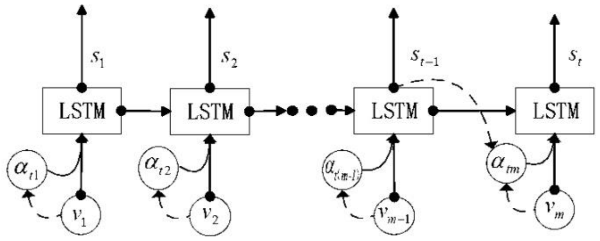

Fig. 2. The LSTM decoder's suggested attention process [11]

A key frame extraction method based on the fusion of four visual features—correlation of the Red-Green-blue (RGB) color channels, color histogram, mutual information, and moments of inertia—is suggested in 2020 [18]. To choose the most representative frames from the collection of frames following fusion, Kohonen Self Organizing Map is employed as a clustering method. Frames that have the greatest Euclidean distance inside a cluster are chosen as the final key frames after useless frames have been eliminated. In Fig. 3, the suggested framework is displayed. Pre-processing, feature fusion and cumulative value computation, Kohonen Self Organizing Map clustering, and representative frame selection make up its four stages. User-generated summaries, Video SUMMarization (VSUMM), and Video Key Frame Extraction via Dynamic Delaunay Clustering (VKEDDCSC), which demonstrate a significant increase in terms of fidelity and Shot Reconstruction Degree (SRD) score, are compared to the results of the suggested approach.

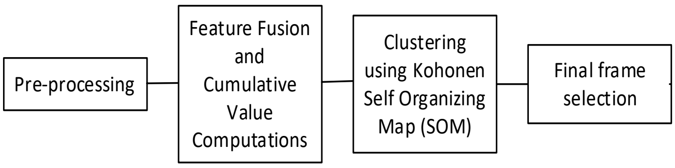
Pre-processing
Feature Fusion and Cumulative Value Computations
Clustering using Kohonen Self Organizing Map (SOM)
Final frame selection

Fig. 3. Framework for key frame extraction [12]

A novel approach known as a superintendence video summarization was suggested in 2020 [19] the goal of this solution is to choose keyframes from the video depending on two factors. The first factor, every object needs to be visible within the frame. The second factor, each thing must be visually visible and close to one another in order to only display related activities. In this research, the subject of interest is the surveillance system for Automated teller machines (ATMs), where the video summary from the ATM room camera should only show the portions of the video when a user is interacting with the device. The following processes are part of the suggested system for video summarization illustrated in Fig. 4: To minimize the number of inactive scenes in the video, the Frames that do not have any movement in them are removed. Then, by identifying several objects, verify the distance between them. Only representative frames and frames with valuable material should be chosen, and these frames will be used to recreate the condensed film. Thus, the final summary uses these important frames.

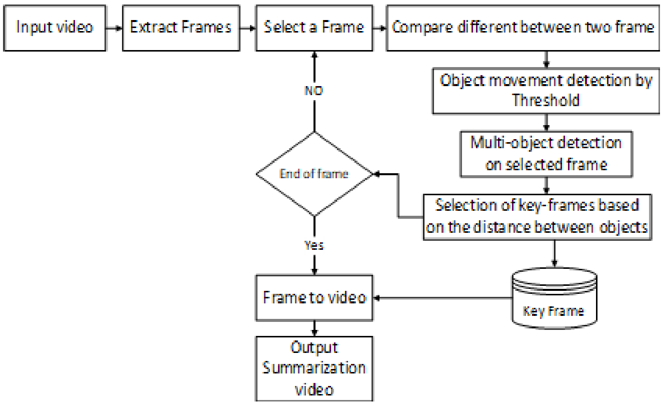
Input video
Extract Frames
Select a Frame
Compare different between two frame
Object movement detection by Threshold
Multi-object detection on selected frame
Selection of key-frames based on the distance between objects
NO
End of frame
Yes
Frame to video
Key Frame
Output Summarization video

Fig. 4. Block Diagram for superintendence video summarization

In 2021 [20] suggested employing normalized k-means and a quick-sort approach for video summarization in surveillance systems. The eight phases of the suggested method include pre-sampling, providing an ID number, feature extraction, feature selection, clustering, extracting frames, and video summary. With the suggested Three Step Cross Searching Algorithm approach, the video frames are first pre-sampled. Give each frame's ID number next. After that, features from the frames are extracted. Then, using Spider Monkey Algorithms based on Entropy, the required features are chosen. The features are aggregated using the Normalized K-Means method in the following step to choose the best candidate frames. The Key Frames selection is made by choosing the cluster set with the lowest Euclidean distance. Using the quick sort of approach, the video is finally organized and summarized. Finally, the suggested approach is contrasted with the standard practices in experimental assessment. The suggested video summarization produced superior results than the execution time-based Kmeans, Fuzzy C-Means Clustering (FCM), and K-medoids techniques.

In 2021 [21-26] developed a foreground extraction method for investigating the multi-frame, multi-scale, and generative adversarial network concepts (mFS-GANs). Additionally, a hybrid optimization technique known as HGWOSA that combines Simulated Annealing (SA) and Grey Wolf Optimizer (GWO) is suggested to guarantee a worldwide optimum outcome with minimal computational cost. Extensive simulations are used to assess and compare the performance of the proposed scheme to that of benchmark schemes. The tests are conducted at IIIT Bhubaneswar utilizing a self-generated surveillance video as well as several common surveillance video datasets (ChangeDetection.Net, MIT Surveillance Dataset, and UMN Dataset). The suggested approach outperforms the other competing methods in terms of both the quantitative and qualitative measures, according to the overall analysis and experimental assessments. The production of off-line video summaries, which may be applied to video surveillance applications for smart urban, can make significant use of the suggested approach.

III. Conclusions

By looking at the type of methods used by researchers in their researches that was mentioned in the previous section in order to obtain summarized videos that achieve the required accuracy and speed and not lose the semantic significance of the original videos, it was noted that most researchers followed the strategy of deep learning with the different types of methods they used and then With a smaller number of studies, researchers used machine learning, and lastly, A very small number of researchers, some traditional methods were used. From the above, it indicates that the current trend is to use deep learning methods in summarizing long videos, and this is due to the growing capacity of computing devices and because it saves effort on researchers, as well as gives results in the majority of them more accurate and better than other methods.

Acknowledgments

The authors would like to thank the Informatics Institute for Graduate Studies, Baghdad, Iraq for their help and encouragement in performing this work.

References

[1] Tiwari, V., & Bhatnagar, C. (2021). A survey of recent work on video summarization: Approaches and techniques. *Multimedia Tools and Applications, 80*(18), 27187-27221.
[2] Almeida, J., Torres, R. D. S., & Leite, N. J. (2010, December). Rapid video summarization on compressed video. *2010 IEEE International Symposium on Multimedia*, (pp. 113-120). IEEE.
[3] Almeida, J., Leite, N. J., & Torres, R. D. S. (2012). Vison: Video summarization for online applications. *Pattern Recognition Letters, 33*(4), 397-409.
[4] Almeida, J., Leite, N. J., & Torres, R. D. S. (2013). Online video summarization on compressed domain. *Journal of Visual Communication and Image Representation, 24*(6), 729-738.
[5] Salim, A. A., Ghoshal, S. K., & Bakhtiar, H. (2021). Tailored morphology, absorption and bactericidal traits of cinnamon nanocrystallites made via PLAL method: Role of altering laser fluence and solvent. *Optik, 226*, 165879.
[6] Mei, S., Guan, G., Wang, Z., Wan, S., He, M., & Feng, D. D. (2015). Video summarization via minimum sparse reconstruction. *Pattern Recognition, 48*(2), 522-533.
[7] Martins, G. B., Papa, J. P., & Almeida, J. (2016, October). Temporal-and spatial-driven video summarization using optimum-path forest. *2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)* (pp. 335-339). IEEE.
[8] Waheed, S. R., Suaib, N. M., Rahim, M. S. M., Adnan, M. M., & Salim, A. A. (2021, April). Deep Learning Algorithms-based Object Detection and Localization Revisited. *Journal of Physics: Conference Series, 1892*(1), 012001. IOP Publishing.
[9] Elharrouss, O., Almaadeed, N., Al-Maadeed, S., Bouridane, A., & Beghdadi, A. (2021). A combined multiple action recognition and summarization for surveillance video sequences. *Applied Intelligence, 51*(2), 690-712.
[10] Sreeja, M. U., & Kovoor, B. C. (2019). Towards genre-specific frameworks for video summarisation: A survey. *Journal of Visual Communication and Image Representation, 62*, 340-358.
[11] Adnan, M. M., Rahim, M. S. M., Al-Jawaheri, K., Ali, M. H., Waheed, S. R., & Radie, A. H. (2020, September). A survey and analysis on image annotation. *2020 3rd International Conference on Engineering Technology and its Applications (IICETA)*, (pp. 203-208). IEEE.
[12] Kumar, K. (2019). EVS-DK: Event video skimming using deep keyframe. *Journal of Visual Communication and Image Representation, 58*, 345-352.

[13] Elharrouss, O., Al-Maadeed, N., & Al-Maadeed, S. (2019, June). Video summarization based on motion detection for surveillance systems. *2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC),* (pp. 366-371). IEEE.

[14] Moses, T. M., & Balachandran, K. (2019, March). A deterministic key-frame indexing and selection for surveillance video summarization. *2019 International Conference on Data Science and Communication (IconDSC),* (pp. 1-5). IEEE.

[15] Waheed, S. R., Rahim, M. S. M., Suaib, N. M., & Salim, A. A. (2023). CNN deep learning-based image to vector depiction. *Multimedia Tools and Applications,* 1-20.

[16] Li, Y., Lin, W., Wang, T., Guo, Q., Yang, R., & Xu, S. (2020, August). Video summarization via cluster-based object tracking and type-based synopsis. *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR),* (pp. 113-116). IEEE.

[17] Ji, Z., Jiao, F., Pang, Y., & Shao, L. (2020). Deep attentive and semantic preserving video summarization. *Neurocomputing, 405,* 200-207.

[18] Rani, S., & Kumar, M. (2020). Social media video summarization using multi-Visual features and Kohnen's Self Organizing Map. *Information Processing & Management, 57*(3), 102190.

[19] Chavan, T., Patil, V., Rokade, P., & Dholay, S. (2020, February). Superintendence Video Summarization. *2020 International Conference on Emerging Trends in Information Technology and Engineering (IC-ETITE),* (pp. 1-7). IEEE.

[20] Davids, D. M., & Christopher, C. S. (2021). An efficient video summarization for surveillance system using normalized k-means and quick sort method. *Microprocessors and Microsystems, 83,* 103960.

[21] Salim, A. A., Ghoshal, S. K., Danmallam, I. M., Sazali, E. S., Krishnan, G., Aziz, M. S., & Bakhtiar, H. (2021, April). Distinct optical response of colloidal gold-cinnamon nanocomposites: Role of pH sensitization. *Journal of Physics: Conference Series, 1892*(1), 012039. IOP Publishing.

[22] Salim, A. A., Bakhtiar, H., Shamsudin, M. S., Aziz, M. S., Johari, A. R., & Ghoshal, S. K. (2022). Performance evaluation of rose bengal dye-decorated plasmonic gold nanoparticles-coated fiber-optic humidity sensor: A mechanism for improved sensing. *Sensors and Actuators: A. Physical, 347,* 113943.

[23] Ghatak, S., Rup, S., Didwania, H., & Swamy, M. N. S. (2021). GAN based efficient foreground extraction and HGWOSA based optimization for video synopsis generation. *Digital Signal Processing, 111,* 102988.

[24] Waheed, S. R., Sakran, A. A., Rahim, M. S. M., Suaib, N. M., Najjar, F. H., Kadhim, K. A., ... & Adnan, M. M. (2023). Design a crime detection system-based fog computing and IoT. *Malaysian Journal of Fundamental and Applied Sciences, 19*(3), 345-354.

[25] Kadhim, K. A., Najjar, F. H., Waad, A. A., Al-Kharsan, I. H., Khudhair, Z. N., & Salim, A. A. (2023). Leukemia classification using a convolutional neural network of AML images. *Malaysian Journal of Fundamental and Applied Sciences, 19*(3), 306-312.

[26] Waheed, S. R., Saadi, S. M., Rahim, M. S. M., Suaib, N. M., Najjar, F. H., Adnan, M. M., & Salim, A. A. (2023). Melanoma skin cancer classification based on CNN deep learning algorithms. *Malaysian Journal of Fundamental and Applied Sciences, 19(*3), 299-305.