EAI Endorsed Transactions
on Context-aware Systems and Applications Research Article EAI.EU

Efficient Key Frame Extraction from Videos Using Convolutional Neural Networks and Clustering Techniques

Anjali H Kugate[1], Bhimambika Y Balannavar[1], R H Goudar[1], Vijayalaxmi Rathod[1,*], Dhananjaya G.M[1] Anjanabhargavi Kulkarni[1], Geeta Hukkeri[2], Rohit B Kaliwal[1],

[1]Department of Computer Science and Engineering, Visvesvaraya Technological University, Belagavi, India.
[2]Department of Computer Science and Engineering, Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, India.

Abstract

One of the most reliable information sources is video, and in recent years, the amount of online and offline video consumption has increased to an unprecedented degree. One of the main difficulties in extracting information from videos is that, unlike images, where information can be gleaned from a single frame, a viewer must watch the entire video in order to comprehend the context. In this work, we try to use various algorithmic techniques, such as deep neural networks and local features, in conjunction with a variety of clustering techniques, to find an efficient method of extracting interesting key frames from videos in order to summarize them. Video summarization plays a major role in video indexing, browsing, compression, analysis, and many other domains. One of the fundamental elements of video structure analysis is key frame extraction, which pulls significant frames out of the movie. An important frame from a video that may be used to summarize videos is called a key frame. We provide a technique that leverages convolutional neural networks in our suggested model, static video summarization, and key frame extraction from movies.

Keywords: Video summarization, Key Extraction, Edge Detection, Motion Analysis, Key frames.

Received on 17 02 2024, accepted on 12 06 2024, published on 17 07 2024

doi: 10.4108/eetcasa.5131

*Corresponding author. Email: vijaylaxmirathod@gmail.com

1. Introduction

The digital camera is an indispensable tool in many domains these days, such as news, sports, education, entertainment, event planning, security monitoring, and advertising. A variety of security surveillance camera types, both stationary and mobile, are placed in public spaces, residences, businesses, airports, banks, and other locations that produce data around the clock. In today's social media-heavy culture and society, videos are arguably the primary source of information due to the advancements in efficient data storage and streaming technologies. It takes a large amount of time and effort to analyze video content to extract relevant or interesting information, as a viewer would need to watch the full film in order to acquire any or all pertinent information. ust as managing and manipulating large amounts of data differs from sending images via text, handling videos presents unique challenges due to their temporal structure. Identifying unusual, suspicious, or abnormal behavior from high-dimensional video data is particularly difficult and time-consuming. This task demands significant processing power, storage capacity, and human effort, requiring both focus and interaction. To address this issue, the concept of video summarization is introduced.

Video summarization involves creating a synopsis of lengthy videos by identifying and highlighting the most

interesting and informative content for prospective viewers. This condensed representation aids in efficiently navigating through a large number of videos and retrieving the particular ones desired. First and foremost, the condensed version of the video needs to include all of the important characters and scenes from the original, and it also needs to be devoid of repetition and unnecessary details. For the summarized video to accurately convey the entire narrative of the original, it must include every significant element.

1.1. Introduction to Key Frames

A simple and effective key frame extraction method can be utilized to produce static video summaries. Key frames, often referred to as representative frames, are the most evocative frames that successfully communicate a video's primary concepts and themes. Key frames can create summaries of the videos so that viewers can peruse them. The rapid advancement of digital video capturing and editing technologies has resulted in a significant increase of video data, necessitating the need for effective techniques for video retrieval and analysis.

A video summary is a condensed, time-limited version of the video, often created to meet viewing time constraints. This speeds up the decision-making process by allowing quick evaluation of the importance or relevance of information. Video summaries are especially valuable in scenarios where power, storage, and communication bandwidth are limited. They are primarily used in entertainment, military, and security contexts.

An MPEG (Moving Picture Expert Group) video is divided into numerous still pictures, called frames. Video summarization techniques have gained significant attention as effective methods for managing multimedia data. Video summaries must adhere to quality measurement standards. The initial phase in key frame extraction is shot segmentation, which involves recognizing transitions between consecutive shots. Video summarization has become a crucial process for making the viewing of large videos faster.

There are two primary techniques for video summarization: dynamic video summarization (also known as video skimming) and static video summarization (also known as video summary). Static video summarization is a rapid and efficient method for distilling lengthy, stored videos. Different tactics are employed for various purposes, such as sports or environmental content. Video summarization techniques have recently garnered researchers' attention, and this field is continually evolving. Most approaches leverage low-level characteristics, with visual features being the most commonly utilized. These low-level features capture the global aspects of frames and include fuzzy color histograms, texture, mutual information, motion information, and color features.

At the high level, CNNs are used to extract feature vectors, which are then used to summarize the video. High-level features include edge detection, SIFT, and interest points from images. Video summarization is crucial for browsing large video collections more quickly and for indexing and accessing content more effectively. The main focus of this field of study is the automatic production of succinct descriptions for both dynamic and static videos. Static video summaries consist of a series of key frames extracted from the original video, while dynamic video summaries are composed of a collection of shots, constructed by considering the domain-specific relationships or similarities between all the video images.

2. Literature Overview

Frames make up a video, and while processing every single frame is often unnecessary, it consumes significant resources. Summarization facilitates adjusting the processing rate for such programs by focusing on key frames. Key frames are those that show a major distinction from earlier frames, making their extraction crucial for video content summarization. This process is essential for applications requiring content summaries, such as data storage, retrieval, and surveillance. One content-based video technique involved in color feature extraction Recovery (CBVR) is Block Truncation Coding (BTC) and its expanded version, Thepade's Sorted Ternary BTC (TSTBTC). [1]

The proliferation of video footage in recent years has posed challenges due to increased memory storage requirements and longer content analysis times. Consequently, efficient content summarization becomes essential. This work proposes a key frame method for perceptual video summarization (PVS) using co-occurrence matrix and permutation computation, integrating it into the Human Visual System (HVS) to recognize perceptually meaningful content. The method's effectiveness is evaluated using various video formats and subjective assessment ratings. [2]

Modern information recording often utilizes multimedia methods, with videos and pictures being the most common. Processing a video involves handling large amounts of information, and redundant frames can cause delays in retrieving essential data. Video summarization, which includes various techniques, accelerates this process. In this study, key frames are extracted using discrete wavelet transformations. Tests on high-definition films with frame rates of 356 and 7293 frames per second showed CPU runtimes of 17 seconds and GPU runtimes of 98 seconds. [3]

With the increasing use of mobile devices for video capture, with thousands of videos posted and downloaded from the Internet every second, managing complex video objects in terms of processing and storage has become challenging. Cutting original films into abstract summaries is an effective solution which allows users to choose content based on their interests. The proposed

video summarization method identifies key frames using shot segments and clusters to create static storyboards. These key frames, extracted from shot segments, help create abstract and summarized video content with minimal repetition. This study introduces a novel approach to creating static video storyboards and suggests a clustering method to identify key frames. Experiments using videos from the Open Video Project (OVP) show that the proposed method's performance is superior to Delaunay Triangulation video summaries and closely matches the OVP ground truth (video summary is assessed by percentage accuracy). [4]

One primary issue in video classification is shared human actions within videos. For example, running actions are common to both long jump and running sports videos. This paper presents a system combining a convolutional neural network (CNN) classifier with a key frame extractor using visual attention modeling. By selecting the top k frames with the highest saliency value, it may be feasible for the system to reduce shared action content and improve classification based on spatial attributes. [5] This approach aims to create a shortened version of the original video by extracting key frames using CNNs and Random Forest classifiers, analyzing videos frame by frame, and removing unnecessary frames using displacement vectors between successive frames. CNN extracts feature vectors at the high level.

The Random Forest Classifier is used to further classify the feature descriptors associated with frames into key-frames and non-key frames. VSUMM and OVP are the two benchmark datasets used to test the method. When compared to other cutting-edge video summarization methods, this suggested method produces superior results. Outcomes demonstrate that the technique can reliably produce excellent summaries for videos in every category. [6] Due to the internet's massive growth in video data, users need efficient ways to browse video content quickly. A popular topic in recent research is distinguishing important information from redundant information in video data. Video summarization can be categorized into two types: static and dynamic. Static video summarization involves creating a collection of key frames, which provides users with the core concept of the video through a sequence of these frames. Key frames are also useful for video retrieval services.

This paper presents a method that achieves shot segmentation by detecting transition frames and then extracting key frames from video shots. Various processing methods are applied to short and long shots. Different processing methods are applied to short and long shots, ensuring that each segmented shot produces a key frame or the most informative frame. The key frame is then rendered non redundant through the application of visual features. This study tests the proposed strategy on the Open Video Project dataset and compares it with VSUMM and other key frame extraction techniques. [7] Since experts across a range of fields and the general public consume increasing amounts of video, it is imperative to develop efficient techniques for indexing and categorizing content. A crucial first step in building a system for indexing and searching videos is video summarization, which is the process of compiling a synopsis and an overview of the most significant scenes in a video. Similar to extracting keywords from a text document, summarization involves extracting key frames to represent video content. Typically, clustering algorithms are applied to every video frame by summarization methods. Some computation methods are costly, however, due to the use of a dissimilarity matrix, which requires quadratic calculations.

To address this, within this framework we introduce a novel method for obtaining video summaries through modified shot segmentation. Hierarchical clustering is applied to extract the key frames from each shot, with the quantity of these key frames correlated with the movements and variations in the shot. Our proposal involves estimating local motion and accounting for motion in the shot by measuring the dissimilarity between shot frames using a co-occurrence matrix [8]. Digital video has gained appeal in several contexts due to its diverse applications and uses. Consequently, the quantity of video data that is accessible has grown exponentially. This is the reason why video summarization remains a focus point of a variety of research projects. In this study, we introduce a brand new method for video summarization that tracks local features between successive frames. This method processes the video stream directly and generate results instantly because it works in the uncompressed domain and only needs a small number of consecutive frames to function [9].

We evaluated our implementation using publicly available standard datasets and compared the outcomes with the most current research in the field. Over the past decade, object tracking in the Internet of Things (IoT) has has garnered significant interest, particularly for item-level activity recognition where video and radio-frequency identification (RFID) integration are crucial. There have been several approaches and uses proposed for optical object tracking, yet identifying the semantic properties of item-level objects in large volumes of video information is challenging, especially in supply chain management.

This paper presents a novel approach which uses IoT data to enable video summarization in order to mitigate such issues. Unlike popular methods of video summarization, important frames of the video content are selected during the background model establishment using information from the Internet of Things. We then align additional crucial frames with the background to extract important characteristics. In the end, a compact summary image for the queried items is produced through clustering analysis. Experiments have been conducted to confirm the effectiveness of the proposed method. [10]

The swift advancement of digital video has made room for numerous uses of multimedia. Every day, a massive

volume of video content is uploaded to the internet, requiring more storage space and internet bandwidth, which also necessitates the development of an efficient user-friendly browsing strategy for content summarization. Unlike previously published methods, we propose a video summarizing technique that utilizes color data collected from patches to determine shot boundaries. This approach accurately identifies video shots and enhances the method's robustness regarding various types of video transitions. Once video shots are identified, each video shot is then separated into sub shots based on the structural similarity between the frames; a key frame is then derived from the most representative sub shot of each unique video shot. Finally, the key frames recovered from each sub shot of the video shot are independently compared to remove any redundant frames. Experimental results, using videos from the Open Video Dataset, demonstrate the effectiveness of the proposed method. By comparing our results with state-of-the-art techniques, we validate that our method efficiently summarizes video content at the sub-shot level rather than analyzing the entire frame. [11]

3. Method

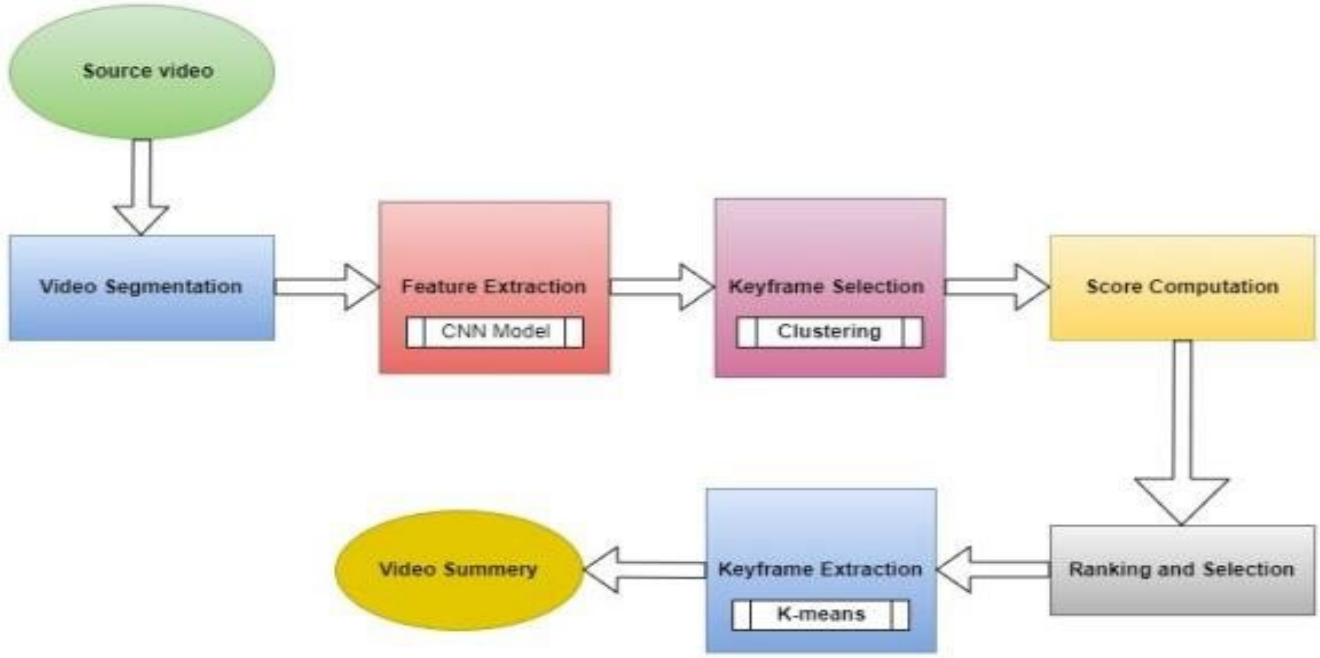

Figure 1: Proposed System Architecture

Source: The original video content intended for summarization is referred to as the "source video." This includes the entire, undisturbed sequence with all temporal and spatial details of the recorded events.

Video Segmentation: This entails segmenting the video into manageable chunks, such as scenes or clips. Often employed for video segmentation tasks, Convolutional Neural Networks (CNNs) divide each frame of a movie into distinct semantic areas or objects.

Feature Extraction: Representing the substance of each shot or scene is an essential first step. Color, texture, motion, and object recognition are among the visual features extracted. Here, we can extract different features using different algorithms.

1. Color Feature Extraction: The Color histogram algorithm is used to extract color features. It statistically represents thedistribution of colors in each frame, aiding segmentation by identifying dominant colors and their frequencies.

2.Motion Detection: Motion detection identifies activity in a scene by examining variations between a sequence of images,typically using frame referencing or pixel matching. Any alteration between frames is defined as a 'detection.'

Key frame Selection: Key frames that best capture the main ideas of each shot are chosen. These frames should be the mostrepresentative and informative, encapsulating the video's main ideas and aesthetic elements.

Clustering Algorithms: Grouping frames with comparable visual attributes can be accomplished by using clustering algorithms, such as K-Means. Key frames

from each cluster are chosen as representatives.

Score Computation: A score is given to each key frame or segment to represent its significance. A number of factors,including object recognition and analysis of visual or audio content, are used to calculate saliency scores. The "SumMe"algorithm (Summarization of Media), for example, is a frequently used algorithm for video summarization score computationwhich automatically calculates frame scores based on based on features like motion, color, texture, and audio information to produce a representative and diverse video summary, calculating the overall significance of every frame in the video, and aiming to include only the highest-scoring frames in the summary.

Ranking and Selection: Segments or key frames are ranked according to their calculated scores and the key frame or segments with the highest ranking are chosen to be included in the summary. The "Graph-Based Visual Saliency" (GBVS) algorithm ranks and selects frames by generating a graph that illustrates the relationships between the different parts of an image or video frame. It uses a variety of low-level visual characteristics, including color, intensity, and orientation, to calculate saliency scores for each region. Through the process of identifying visually striking or significant areas within each frame, the chosen frames become more representative of the overall content of the video.

Key frame Extraction: This powerful approach involves identifying a set of summary key frames from a video sequence. The k-means clustering technique. The k-means clustering technique extracts key frames, dividing data points into k-clusters based on property similarities. In the context of keyframe extraction, each video frame is represented as a feature vector. These frames are then clustered into k clusters using the k-means method. Each video frame, represented as a feature vector, is clustered using the k-means method, and the centroid of each cluster serves as a representative key frame.

Video Summary: In video summarization, a summary video is a shortened version of the original, highlighting key or representative frames. This process condenses a video while retaining the most significant portions, making the content more digestible and manageable. Video summaries are particularly useful for large datasets, surveillance footage, and lengthy content where viewers need to quickly grasp key points or highlights, which is far more manageable than having to watch a video in its entirety.

3.1. Introduction to Key Frames

Step 1: Takes video as input.

Step 2: Segments the video into manageable chunks, such as scenes or shots.

Step 3: Convolution neural networks are used to extract the features from the frame.

Step 4: Key frame selection involves choosing frames from each shot that best capture the key elements. Every shot should contain the most representative and instructive frames possible. Among the requirements are diversity in content and image quality.

Step 5: A score is given to each key frame or segment to represent its significance. A number of factors, including object recognition and analysis of visual or audio content, are used to calculate scores.

Step 6: Key frames or segments are selected and ranked according to their calculated scores. Key frames or segments that receive the highest ranking are chosen to be included in the summary.

Step 7: Segments or key frames are chosen and arranged in order of their computed scores. The segments or key frames with the highest ranking are selected for inclusion in the summary.

Step 8: A video synopsis is then able to be created using a few chosen key frames.

4. Results and Discussion

Figure 2. The video segments of the provided source video

The technique of dividing a video into smaller, easier-to-manage segments, such scenes or snippets, is known as video segmentation. Convolutional neural networks, or CNNs, are used for this procedure.

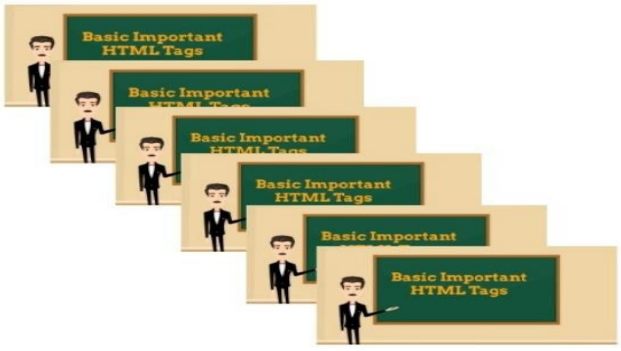

Figure 3. Feature extraction such as color and motion from video segments

Color feature extraction: The Color Histograms

algorithm is used to extract color features. Color histograms statistically depict the distribution of colors in each frame.

Motion detection: This technique can be used to spot movement in a scene. Frame referencing or pixel matching are typically employed. For this research, the Lucas-Kanade (LK) algorithm was used.

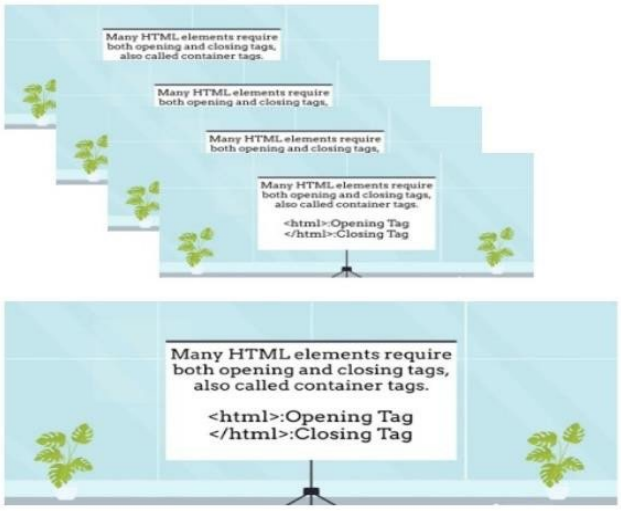

Figure 4. Keyframe selection using clustering

Key frame Selection: Among them, the keyframes that most effectively convey the primary concepts of each shot are selected. Keyframes are exemplary frames that have been chosen to best capture the central concepts and visual components of a video. For this, clustering algorithms are employed.

Figure 5. Score Computation to every frame based on features

Score Computation: Each key frame is assigned a score that indicates its importance. Scores are determined by a variety of factors, such as object color, motion, and examination of visual or auditory content. The "SumMe" algorithm is used to calculate the video summarization score. The video summary only contains frames with the highest scores.

Figure 6. Ranking selection based on score computation

Ranking and Selection: Key frames or segments are arranged in order of computed scores. The key frames or segments selected for inclusion in the summary are those with the highest ranking. The algorithm used for ranking and selecting frames in video summarization is Graph-Based Visual Saliency (GBVS).

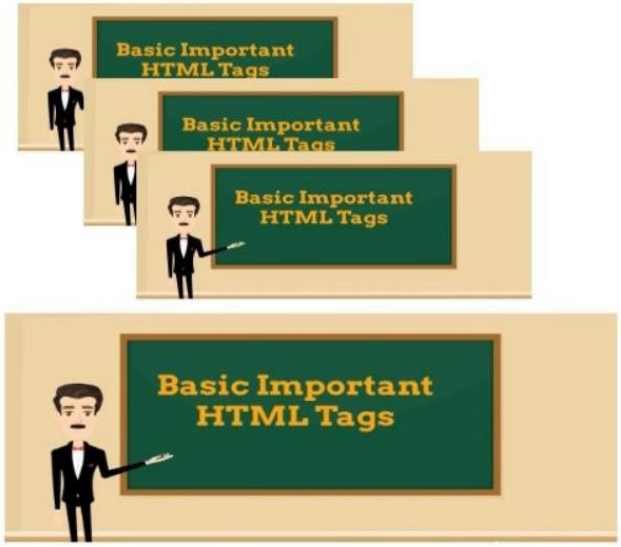

Figure 7. Key frame extraction using K-means algorithm

Key frame Extraction: Key frame extraction is a powerful tool for implementing video content, using a set of summary key frames to represent video sequences. The k-means algorithm is used in video processing to extract key frames. Each video frame is represented as a feature vector.

Figure 8. Final product of the video summary

Summary Video: The summary video condenses and presents the most engaging or educational content for viewers, typically found in a longer video document.

System Recovery: Recovering from a video summarization system failure requires a systematic approach, identifying where the failure occurred in the summarization process, then analyzing error logs to pinpoint issues by setting up logging mechanisms to record errors and diagnostic information during system operation. Log analysis tools are used to review error logs and identify patterns or trends that indicate common failure points, and logging and monitoring tools indicate troubleshooting failures within the video summarization system. By beginning by verifying data integrity, it is ensured that input data is correctly formatted and consistent. Algorithms and processes are debugged to identify and resolve issues causing failures, which could include errors during data preprocessing, feature extraction, key frame selection, or scoring computation. With systematic troubleshooting, data integrity is verified and algorithms debugged. Robust error handling mechanisms are implemented and thorough testing is conducted to ensure functionality. Continuous system monitoring is maintained with backups for swift recovery.

Conclusion

To sum up, video summarization is an essential procedure for handling the massive volume of video data produced in different fields. The literature review emphasizes the importance of key frames in producing condensed representations of video content and offers insights into various approaches and techniques used for video summarization. Extracting key frames from videos efficiently using convolutional neural networks (CNNs) and clustering techniques offers several advantages. Firstly, the automation enabled by CNNs streamlines the process, reducing the need for manual intervention and saving considerable time and effort. Secondly, CNNs can learn complex patterns in visual data, enhancing the accuracy of key frame identification compared to traditional methods. CNNs facilitate rapid extraction of key frames from extensive video datasets. The swift processing scalability makes the approach suitable for applications requiring the analysis of large amounts of video data. Additionally, the use of clustering algorithms, such as k-means, enables grouping frames with similar visual attributes, enhancing the efficiency of key frame selection and summary creation. Video summary is a vital technique that makes it easier for individuals to acquire and comprehend vast volumes of video content in the age of information overload.

To further enhance video summarization techniques, future research can explore the integration of additional modalities such as audio and text, allowing for a more comprehensive understanding of video content. Additionally, advancements in explainable AI and user-centric design principles can contribute to making video summarization systems more transparent, interpretable, and user-friendly. Empowering users with greater control over the summarization process and allowing for personalized preferences will enhance the usability and adoption of these technologies.

Conflicts of Interest

"The authors have no conflicts of interest to declare that are relevant to the content of this article".

Acknowledgment

This study which has been carried out was supported and guided by the Department of CSE, VTU Belagavi. Karnataka, India, hence we express our sincere gratitude for providing the required resources accessibility and knowledge centre facility to accomplish the work successfully.

References

1. S. D. Thepade and P. H. Patil, "Novel visual content summarization in videos using keyframe extraction with Thepade's Sorted Ternary Block truncation Coding and Assorted similarity measures," 2015 International Conference on Communication, Information & Computing Technology (ICCICT), Mumbai, India, 2015, pp. 1-5, doi: 10.1109/ICCICT.2015.7045726.
2. R. J. R, P. Nimmagadda, K. Sudhakar, B. C. J, P. Rajasekar and S. M. A, "Perceptual Video Summarization Using Keyframes Extraction Technique," 2023 3rd International Conference on Innovative Practices in Technology and Management (ICIPTM), Uttar Pradesh, India, 2023, pp. 1-4, doi: 10.1109/ICIPTM57143.2023.10118236.
3. C. Sharma and P. K. Sathish, "Parallelizing keyframe extraction for video summarization," 2015 International Conference on Signal Processing and Communication Engineering Sys- tems, Guntur, India, 2015, pp. 245-249, doi: 10.1109/SPACES.2015.7058258.
4. A. Tonge and S. D. Thepade, "A Novel Approach for Static Video Content Summarization using Shot Segmentation and k-means Clustering," 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon), Mysuru, India, 2022, pp.

1-7, doi: 10.1109/Mysu- ruCon55714.2022.9972379.
5. R. F. Rachmadi, K. Uchimura and G. Koutaki, "Video classification using compacted da- taset based on selected keyframe," 2016 IEEE Region 10 Conference (TENCON), Singa- pore, 2016, pp. 873-878, doi: 10.1109/TENCON.2016.7848130.
6. M. S. Nair and J. Mohan, "Video Summarization using Convolutional Neural Network and Random Forest Classifier," TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), Kochi, India, 2019, pp. 476-480, doi: 10.1109/TENCON.2019.8929724.
7. Y. Ding, D. Shen, L. Ye and W. Zhu, "A keyframe extraction method based on transition detection and image entropy," 2022 7th International Conference on Communication, Image and Signal Processing (CCISP), Chengdu, China, 2022, pp. 260-264, doi: 10.1109/CCISP55629.2022.9974364.
8. W. Sabbar, A. Chergui and A. Bekkhoucha, "Video summarization using shot segmentation and local motion estimation," Second International Conference on the Innovative Compu- ting Technology (INTECH 2012), Casablanca, Morocco, 2012, pp. 190-193, doi: 10.1109/INTECH.2012.6457809.
9. J. Iparraguirre and C. Delrieux, "Speeded-Up Video Summarization Based on Local Features," 2013 IEEE International Symposium on Multimedia, Anaheim, CA, USA, 2013, pp.370-373, doi: 10.1109/ISM.2013.70.
10. C. Luo, "Video Summarization for Object Tracking in the Internet of Things," 2014 Eighth International Conference on Next Generation Mobile Apps, Services and Technologies, Ox- ford, UK, 2014, pp. 288-293, doi: 10.1109/NGMAST.2014.20.
11. M. Asim, N. Almaadeed, S. Al-maadeed, A. Bouridane and A. Beghdadi, "A Key Frame Based Video Summarization using Color Features," 2018 Colour and Visual Computing Symposium (CVCS), Gjovik, Norway, 2018, pp. 1-6, doi: 10.1109/CVCS.2018.8496473.
12. A. S. Parihar, R. Mittal, P. Jain and Himanshu, "Survey and Comparison of Video Summarization Techniques," 2021 5th International Conference on Computer, Communication and Signal Processing (ICCCSP), Chennai, India, 2021, pp. 268-272, doi: 10.1109/ICCCSP52374.2021.9465347.
13. Thomas, Sinnu Susan, Sumana Gupta, and Venkatesh K. Subramanian. "Perceptual synoptic view of pixel, object and semantic based attributes of video." Journal of Visual Communication and Image Representation 38 (2016): 367-377.
14. You, Junyong, et al. "A multiple visual models based perceptive analysis framework for multilevel video summarization." IEEE Transactions on Circuits and Systems for Video Technology 17.3 (2007): 273-285.
15. Ajmal, Muhammad, et al. "Video summarization: techniques and classification." Computer Vision and Graphics: International Conference, ICCVG 2012, Warsaw, Poland, September 24-26, 2012. Proceedings. Springer Berlin Heidelberg, 2012.
16. Kapoor, Aditi, K. K. Biswas, and Madasu Hanmandlu. "Fuzzy video summarization using key frame extraction." 2013 Fourth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG). IEEE, 2013.
17. Yasmin, Ghazaala, et al. "Key moment extraction for designing an agglomerative clustering algorithm-based video summarization framework." Neural computing and applications 35.7 (2023): 4881-4902.
18. Sreeja, M. U., and Binsu C. Kovoor. "A multi-stage deep adversarial network for video summarization with knowledge distillation." Journal of Ambient Intelligence and Humanized Computing 14.8 (2023): 9823-9838.
19. Savran Kızıltepe, Rukiye, John Q. Gan, and Juan José Escobar. "A novel keyframe extraction method for video classification using deep neural networks." Neural Computing and Applications 35.34 (2023): 24513-24524.
20. Hsu, Tzu-Chun, Yi-Sheng Liao, and Chun-Rong Huang. "Video summarization with spatiotemporal vision transformer." IEEE Transactions on Image Processing (2023).
21. Issa, Obada, and Tamer Shanableh. "Static video summarization using video coding features with frame-level temporal subsampling and deep learning." Applied Sciences 13.10 (2023): 6065.
22. Khan, Habib, et al. "Deep multi-scale pyramidal features network for supervised video summarization." Expert Systems with Applications 237 (2024): 121288.
23. Sabha, Ambreen, and Arvind Selwal. "Data-driven enabled approaches for criteria-based video summarization: a comprehensive survey, taxonomy, and future directions." Multimedia Tools and Applications 82.21 (2023): 32635-32709.
24. Rahman, Mohammad Rajiur, et al. "Enhancing lecture video navigation with AI generated summaries." Education and Information Technologies (2023): 1-24.
25. Derdiyok, Seyma, and Fatma Patlar Akbulut. "Biosignal based emotion-oriented video summarization." Multimedia Systems 29.3 (2023): 1513-1526.