International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/IJSRCSEIT

Video Summarization using Deep Learning

Mr. Basavaraj Muragod[1], Mr. Nagaraj Telkar[2], Ms. Pooja Bharamagoudra[2], Ms. Raksha Shet[2], Ms. Sheetal D Naik[2], Mr. Vinay L Patige[2]
Assistant Professor[1], UG Students[2]
Department of Computer Science and Engineering, SKSVMACET, Lakshmeshwar, Gadag, Karnataka, India

ARTICLEINFO

Article History:

Accepted: 01 May 2023
Published: 20 May 2023

Publication Issue
Volume 9, Issue 3
May-June-2023

Page Number
216-222

ABSTRACT

The advancements in digital video technology have empowered video surveillance to play a vital role in ensuring security and safety. Public and private enterprises use surveillance systems to monitor and analyze daily activities. Consequently, a massive volume of data is generated in videos that require further processing to achieve security protocol. Analyzing video content is tedious and a time-consuming task. Moreover, it also requires high- speed computing hardware. The video summarization concept has emerged to overcome these limitations. This paper presents a customized video summarization framework based on deep learning. The proposed framework enables a user to summarize the videos according to the Object of Interest (OoI), for example, person, airplane, mobile phone, bike, and car. Various experiments are conducted to evaluate the performance of the proposed framework on the video summarization (VSUMM) dataset, title-based video summarization (TVSum) dataset, and own dataset. The accuracy of VSUMM, TVSum, and own dataset is 99.6%, 99.9%, and 99.2%, respectively. A desktop application is also developed to help the user summarize the video based on the OoI. in the problem of video summarization, the goal is to select a subset of the input frames conveying the most important information of the input video. The collection of data proves to be a challenging task. The goal of video summarization is to shorten an input video to a summary video which conveys the most important information of the original video.

Keywords: Object of Interest (OoI); Video Summarization (VSUMM) dataset; Title-based Video Summarization (TVSum) dataset; Own dataset.

I. INTRODUCTION

In the world today, video is being recorded everywhere for many different reasons. Millions of hours of video are recorded everyday with online sites announcing that over 500 hours of video is uploaded every minute. These videos range from home videos like those of a birthday party or special event, to video

that is always being captured like a security camera or self- driving car cameras. Between the amount of video that we see on television, social media, or anywhere else, it is impossible to watch everything that is being presented to us on a daily basis. One solution that arises to solve this problem is to instead watch a summary of the important moments of videos. Video Summarization refers to the process of taking a video and creating a summary of it based on its important parts.

The goal of Video Summarization is that the resulting summary saves the user time by not having to watch the entire video to get a basic understanding of its content. According to YouTube, there were approximately 5.5 billion daily video views in the first quarter of 2022. We are experiencing an even stronger engagement of consumers with both online video platforms and devices (e.g., smart-phones, wearables etc.) that carry powerful video recording sensors and allow instant uploading of the captured video on the Web and YouTube is just one of the many video hosting platforms.

These cameras usually remain active round the clock and generate more than 2,500 petabytes of video data per day. The daily statistics of the real-world data produced by the video surveillance cameras. Considerable progress has already been made in developing video analytic tools that automatically perform content-based video interpretation, including motion detection facial recognition people counting and license plate recognition. The researchers have made several efforts to propose automatic VS. Most of the VS techniques generate a summary based on selecting keyframes representing the video through the skimming process. Feature-based approaches for VS produce a generalized video summary rather than focusing on a specific object. The shot boundary detection approaches are also well known for video summarization. These approaches show limitations in detecting the object precisely, hence failing to fulfill the user's requirements. Clustering and trajectory-based techniques summarize the video by focusing on similar activities, events, and objects. However, these approaches do not summarize any video containing information according to the user's interest. Consequently, these techniques limit the use of retrieval tasks and do not help enhance the users' observing experience. This study presents an effective VS framework based on the OoI to cope with the issues of video summarization.

The OoI refers to the objects such as person, car, mobile, and bike that a user selects to summarize the video by collecting all frames where the selected object appears. The proposed VS framework works in three steps: (i) the combination of the OoI selection phase, (ii) the object localization or detection phase, and (iii) the video summarization phase. Initially, the OoI selection is performed from the dictionary (a database of objects) to ignore the unnecessary noisy objects (other than the OoI) that are imperative for the segmentation of objects. After that, You Look Only Once (YOLOv3) detector is applied to localize the OoI. After localizing the OoI, the proposed VS algorithm summarizes the video based on the OoI. Based on the above discussion, the contributions of the proposed work can be summarized as follows: (i) In the OoI selection step, the proposed algorithm selects the object from the dictionary and ignores all the unnecessary objects automatically. After that, the YOLOv3 is used to detect the desired object. (ii) The proposed VS framework can detect single and multiple objects present in the video. (iii) The proposed VS algorithm effectively summarizes the video and overwhelms all the challenges shown in the VSUMM, TVSum, and own dataset. (iv) The experimental finding highlights that the proposed VS framework performs tremendously as compared to state-of-the-art methods in the area of video summarization.

Existing System

The generalized overview of the steps involved in an existing system for video summarization:

1. Video Segmentation: The video is segmented into shots or scenes based on changes in visual and/or

audio content. This can be achieved using various techniques such as threshold-based methods, clustering, or deep learning-based methods.

2. Feature Extraction: Key features are extracted from each shot, such as color histograms, motion vectors, and audio features. Different types of features can be used depending on the requirements of the summarization task.
3. Shot Ranking: The shots are ranked based on their importance or relevance to the overall video content. This can be done based on various criteria such as visual novelty, audio content, and motion saliency.
4. Keyframe Selection: The most representative or informative frames from each shot are selected as keyframes. This can be done using various techniques such as clustering, graph-based methods, or deep learning-based methods.
5. Summary Generation: The keyframes are combined to generate a summary of the video. This can be achieved using different methods such as greedy selection, optimization-based methods, or learning-based methods.
6. Post-Processing: The summary is post- processed to improve its coherence, readability, and overall quality. This can involve techniques such as clustering, smoothing, or transition generation.

Proposed System

Video summarization is the process of creating a shorter version of a longer video, which captures the most important and relevant parts of the original video. There are several approaches that can be used for video summarization, including keyframe extraction, object tracking, and motion analysis. Here's a proposed system for video summarization:

1. Video Preprocessing: The first step in video summarization is to preprocess the video. This can include tasks such as noise removal, stabilization, and color correction.
2. Shot Boundary Detection: Once the video has been preprocessed, the next step is to detect shot boundaries. Shot boundary detection involves identifying the points in the video where one shot ends and another begins. This can be done using techniques such as histogram analysis, edge detection, and motion analysis.
3. Keyframe Extraction: Once the shot boundaries have been detected, the next step is to extract keyframes from each shot. Keyframes are frames that capture the most important and relevant parts of a shot. Keyframe extraction can be done using techniques such as entropy-based analysis, clustering, and saliency detection.
4. Summary Generation: The final step in video summarization is to generate a summary of the video using the keyframes extracted from each shot. This can be done using techniques such as clustering, graph-based optimization, and deep learning-based models.
5. Evaluation: It's important to evaluate the quality of the video summary generated by the system. This can be done using metrics such as F-measure, precision, recall, and subjective user feedback.

Overall, the proposed system for video summarization involves several steps, including video preprocessing, shot boundary detection, keyframe extraction, summary generation, and evaluation. Different techniques can be used for each of these steps, depending on the specific requirements of the application.

II. METHODOLOGY

Here is a proposed methodology for video summarization:

1. Problem Definition: Define the problem statement, objectives, and scope of the video summarization project. Identify the type of video data to be summarized and the intended use case(s) of the summary.
2. Data Collection: Collect the video data to be summarized and any relevant metadata or annotations. Organize the data into a suitable

format for processing. Also, the tag, i.e.. Object of Interest (OoI) is given as the second input where that particular object is detected in the whole video as object detection.

3. Video Segmentation: Segment the video into shots or scenes based on visual and/or audio features. This can be achieved using techniques such as threshold-based methods, clustering, or deep learning-based methods.
4. Feature Extraction: Extract relevant features from each shot, such as color histograms, motion vectors, etc. Different types of features can be used depending on the requirements of the summarization task.
5. Shot Importance Ranking: All the shots or the frames which contains the object which was given in the form of tag is marked as the important frame or shot. All those frames which does not contain that object is marked as the not important frame and is simply ignored.
6. Keyframe Selection: Select the most representative or informative frames from each shot as keyframes. This can be done using techniques such as clustering, graph-based methods, or deep learning-based methods. Prioritize diversity in the selection process to improve the coverage of the summary.
7. Summary Generation: Generate a summary of the video by combining the selected keyframes. This can be achieved using different methods such as greedy selection, optimization-based methods, or learning-based methods. Use post-processing techniques such as clustering or smoothing to improve the coherence and readability of the summary. Here, all those frames which contains the necessary object is arranged in the form of chronological manner, i.e., in the sequential manner.
8. Evaluation: Evaluate the quality and effectiveness of the summary using metrics such as summary length, coverage, and relevance. Validate the summary against the intended use case(s) and solicit feedback from users.
9. Iteration: Iterate on the methodology and fine-tune the feature extraction, ranking, and summarization techniques based on the evaluation and feedback. Refine the methodology to improve the quality and efficiency of the video summarization process.

Overall, this proposed methodology aims to generate accurate, diverse, and informative summaries that can meet the requirements of various video summarization applications.

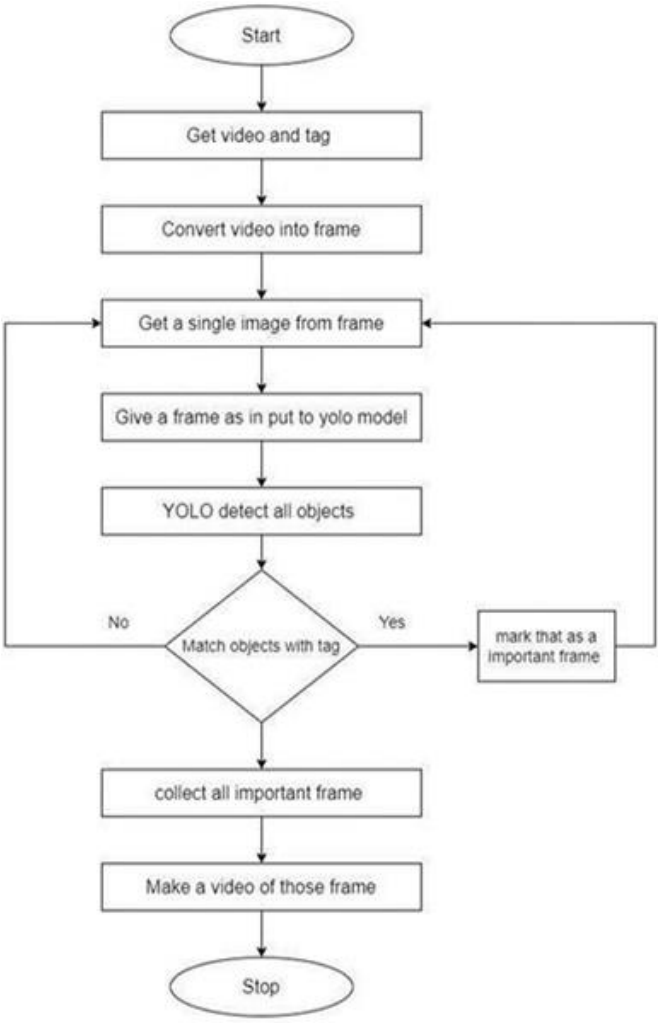

Figure 1: Flowchart of the working process

III. WORKING PRINCIPLE

The working principle of video summarization involves identifying the most important and relevant parts of a longer video and creating a shorter version

that captures the essence of the original video. Here's how the process typically works:

1. Video Preprocessing: The first step in video summarization is to preprocess the video. This can include tasks such as noise removal, stabilization, and color correction.
2. Shot Boundary Detection: Once the video has been preprocessed, the next step is to detect shot boundaries. Shot boundary detection involves identifying the points in the video where one shot ends and another begins. This can be done using techniques such as histogram analysis, edge detection, and motion analysis.
3. Keyframe Extraction: Once the shot boundaries have been detected, the next step is to extract keyframes from each shot. Keyframes are frames that capture the most important and relevant parts of a shot. Keyframe extraction can be done using techniques such as entropy-based analysis, clustering, and saliency detection.
4. Summary Generation: The keyframes extracted from each shot are then used to generate a summary of the video. There are different techniques that can be used for summary generation, including clustering, graph-based optimization, and deep learning-based models. The goal is to create a summary that captures the essence of the video and conveys the most important and relevant information.
5. Evaluation: Finally, it's important to evaluate the quality of the video summary generated by the system. This can be done using metrics such as F-measure, precision, recall, and subjective user feedback.

Overall, the working principle of video summarization involves identifying the most important and relevant parts of a video and creating a summary that captures the essence of the original video. Different techniques can be used for each step of the process, depending on the specific requirements of the application.

IV.RESULTS AND DISCUSSION

The final output of a video summarization project is a shorter version of the original video that captures the most important and relevant parts of the video. This output can take different forms, depending on the requirements of the application. Here are some examples:

1. Keyframe-based summary: A keyframe-based summary consists of a sequence of keyframes extracted from the original video. These keyframes are selected based on their relevance and importance to the video content.

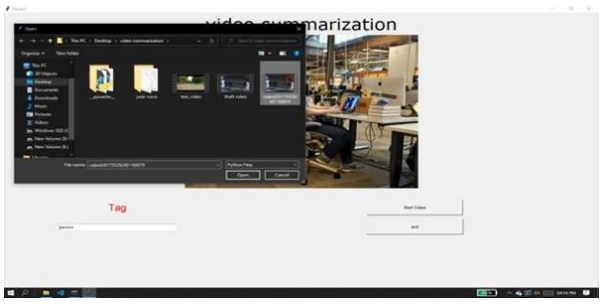

Figure 2: Input Video Selection

2. Summary video: A summary video is a shorter version of the original video that includes only the most important and relevant parts. The summary video can be generated by combining the keyframes extracted from each shot in the video.

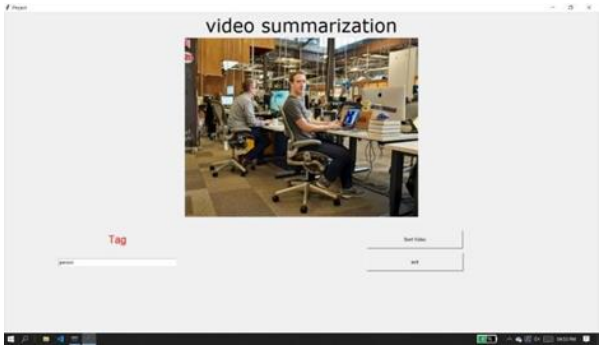

Figure 3: Input OoI or Tag Selection

3. Object detection: In the given video as the first input, only the particular important object which is necessary by the user was given as the second unput, i.e., Object of Interest (OoI) or Tag is considered and is marked as important frame. All the other frames which does not contain the object is marked and not important frame.

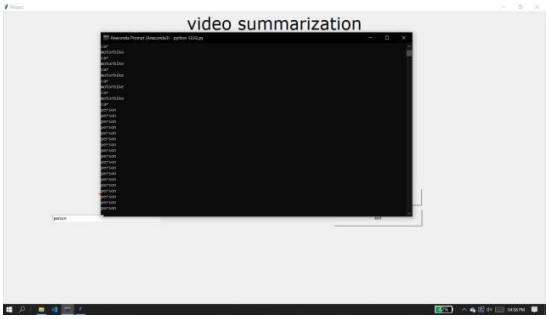

Figure 4: Object Detection Process in the video

4. Highlight reel: A highlight reel is a summary of the video content that includes only the most exciting or dramatic parts of the video. This can be useful for sports or action videos, where the most exciting parts of the video are the most important.

Figure 5: Final Summarized Output Video

FUTURE SCOPE

The future scope of our video summarization project is promising as the demand for efficient and effective video content consumption is increasing rapidly. Here are some potential areas for future developments and applications of video summarization:

1. Personalization: With the advancements in machine learning algorithms and data analytics, video summarization can be personalized to cater

V. to individual preferences. This can help in creating a personalized viewing experience for each user based on their interests and watching history.

2. Real-time video summarization: Video summarization can be extended to real-time video streams, such as live events and surveillance cameras, where important events need to be captured and summarized in real-time.
3. Multimodal video summarization: The current focus of video summarization is on visual content, but in the future, it can be extended to include other modalities such as audio and text. Multimodal video summarization can provide a more comprehensive summary of the content.
4. Semantic video summarization: Semantic video summarization aims to capture the meaning and context of the video content. This can help in creating more informative and relevant summaries.
5. Cross-modal video summarization: Cross-modal video summarization involves summarizing a video using different modalities, such as using text or audio to summarize a video. This can help in creating summaries that are accessible to people with different sensory abilities.
6. Multiple tags input: Multiple tags can be given as the input as Object of Interest (OoI) where it helps in identifying multiple objects in the single given input video.

In conclusion, video summarization has a bright future ahead, and with the development of new technologies and algorithms, we can expect to see more innovative applications of video summarization in various domains.

VI.CONCLUSION

In conclusion, video summarization is an important field that can help people save time and effort by providing concise and informative summaries of longer videos. It involves a variety of techniques such as keyframe extraction, object detection, and scene segmentation to identify important segments of a video and create a summary that captures the essence of the original content.

Video summarization has many practical applications, including in education, journalism, marketing, and entertainment. It can be used to create shorter versions

of lectures or presentations for students, to summarize news stories or events for journalists, to showcase product demos or advertisements in a shorter format for marketing purposes, and to create previews or highlights of longer videos for entertainment. It becomes very useful when it comes to investigation situations of any crime.

Developing a video summarization project requires a good understanding of the underlying algorithms and techniques, as well as the ability to work with large datasets and advanced machine learning models. It also requires careful consideration of user needs and preferences, as well as the development of user-friendly interfaces and tools for editing and sharing the generated summaries.

Overall, video summarization is a fascinating field that holds great potential for improving the way we consume and interact with video content. As technology continues to evolve, we can expect to see more advances in this area, and more innovative applications of video summarization in a wide range of fields and industries.

VII. REFERENCES

[1]. Security Info Watch, "Data generated by new surveillance cameras to increase exponentially in the coming years," 2016.

[2]. B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised Video Summarization with Adversarial LSTM Networks," in 2017 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2982–2991.

[3]. E. Cosgrove, "One billion surveillance cameras will be watching around the world in 2021, a new study says," 2019.

[4]. X. Yan, S. Z. Gilani, M. Feng, L. Zhang, H. Qin, and A. Mian, "Self-supervised learning to detect key frames in videos," Sensors, vol. 20, no. 23, p. 6941, 2020.

[5]. B. Korbar, D. Tran, and L. Torresani, "Scsampler: sampling salient clips from video for efficient action recognition," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6232–6242, IEEE, Seoul, Korea (South), November 2020.

[6]. J. Huo and T. L. van Zyl, "Unique faces recognition in videos," in Proceedings of the 2020 IEEE 23rd International Conference on Information Fusion (FUSION), pp. 1–7, IEEE, Rustenburg, South Africa, July 2020.

[7]. S. Manna, S. Ghildiyal, and K. Bhimani, "Face recognition from video using deep learning," in Proceedings of the 2020 5th International Conference on Communication and Electronics Systems (ICCES), pp. 1101–1106, IEEE, Coimbatore, India, June 2020.

[8]. J. Zhang, S. Chen, S. Tian, W. Gong, G. Cai, and Y. Wang, "A crowd counting framework combining with crowd location," Journal of Advanced Transportation, vol. 2021, Article ID 6664281, 12 pages, 2021.

[9]. W. Ullah, A. Ullah, T. Hussain et al., "Artificial Intelligence of Things-assisted two- stream neural network for anomaly detection in surveillance Big Video Data," Future Generation Computer Systems, vol. 129, pp. 286–297, 2022.

[10]. S.-H. Zhong, J. Lin, J. Lu, A. Fares, and T. Ren, "Deep semantic and attentive network for unsupervised video summarization," ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 18, no. 2, pp. 1–21, 2022.