IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Enhanced Video Summarization With Real-Time Object Detection And Tracking Using Yolov3 And Deep Sort

[1]Sinchana C Poojari, [2]Navyashree J, [3]Siri M K, [4]Nithyasree M, [5]Kavita K Patil

[1]Student, [2]Student, [3]Student, [4]Student, [5]Associate Professor

[1]Information Science and Engineering,

[1]Global Academy of Technology, Bengaluru, India

Abstract: Video summarization is a fundamental task of computer vision and multimedia processing to condense lengthy videos into short representations without losing the valuable content and context. The study is founded on the use of object detection techniques in video summarization and relies on the capability of deep learning to automatically recognize and extract discriminative objects and events from video streams. Relying on the benefit of the latest object detection models and new summarization techniques, the study tries to enhance the efficiency and effectiveness of video summarization to allow users to quickly perceive the content and meaning of videos without the requirement of lengthy playback. The approach not only enhances video browsing and comprehension of content but also has its future areas of application in surveillance, video indexing, and content recommendation systems. Video summarization is a crucial element to successfully extract key moments from lengthy video recordings, reducing storage and processing costs, and enhancing the user experience. The work proposes a state-of-the-art video summarization technique founded on real-time object detection based on YOLOv3 and Deep SORT algorithms. Based on the fusion of the new approaches, the proposed method effectively extracts and tracks discriminative objects with improved efficiency, leading to informative and meaningful video summaries. Experimental results exhibit enhanced efficiency and accuracy compared to state-of-the-art summarization techniques, indicating the potentiality of the proposed methodology in its real-world applications like surveillance, sport analysis, and content generation.

Index Terms - Video Summarization, Object Detection, Object Tracking, YOLOv3, Deep SORT, Real-Time Processing

I. INTRODUCTION

With the video age, the sheer volume of videos shared across web portals, surveillance networks, and personal libraries placed in sharp relief the need for efficient means of summarizing and understanding the sheer volume of video content. Video summarization has thus emerged as the solution to the problem, providing a solution to generating informative and compact representations of video content. Most of the traditional video summarization methods utilize techniques such as key frame extraction, temporal clustering, and scene analysis. These are, however, prone to missing important visual features and events and, by proxy, generating suboptimal summarizations. The proliferation of video data in applications such as surveillance, entertainment, and sports analysis has rendered efficient summarization methodologies of critical importance. Traditional video summarization methods are based on key frame extraction, clustering, or motion analysis that do not necessarily capture the contextual significance of the events in a video.

Object detection, a subfield of computer vision, has witnessed remarkable advancements with the advent of deep learning. Convolutional Neural Networks (CNNs) have revolutionized object detection by enabling accurate identification and localization of objects within images and videos. Integrating object detection into the video summarization process presents a novel approach to capturing the most salient content within a video. By identifying key objects, actions, and interactions, the summarization process can provide a more comprehensive and contextually relevant summary.

This paper proposes an object-aware video summarization approach by leveraging state-of-the-art deep learning models—YOLOv3 for object detection and Deep SORT for object tracking. This methodology enables the creation of summaries that preserve crucial objects and interactions within a video sequence. The proposed approach is evaluated based on multiple performance metrics, demonstrating its effectiveness in improving video summarization quality. We have employed advanced object detection models, i.e., Faster R-CNN, YOLO (You Only Look Once), or SSD (Single Shot Multi Box Detector), to identify and track the objects of interest between video clips. The identified objects would then be employed as the foundation to build an interpretable video summary.

By extracting objects with higher semantic importance and contextual relevance, the resultant summary will be more representative of the original video content.

II. RELATED WORK

Video summarization methods have come a long way since the early years, with methods varying from handcrafted feature extraction to deep learning-based methods. Conventional methods, including keyframe selection and clustering, miss a lot of contextual information. Deep learning-based methods have included object detection and tracking features, improving summarization efficiency.

YOLOv3 achieved general applicability for object detection due to its fast detection and precision. Similarly, Deep SORT is a strong tracking algorithm that ensures object continuity across video frames. This current work extends earlier methodologies by integrating these two approaches for real-time object-aware summarization.

The aesthetic-driven method centers on picking attractive and meaningful shots. The ADUVS method introduces an aesthetics encoder that captures attributes like contrast, hue, saturation, and composition to improve the quality of video that is summarized. This provides efficient and engaging summaries by combining visual, audio, and textual features [1].

The article suggests FIAS3, a new WCE video summary model with frame importance and a sparse subset selection approach towards maximum GI lesion coverage and redundancy minimization. The experimental work shows that FIAS3 performs better than state-of-the-art algorithms with 92% coverage at a 90% compression ratio. The model's flexibility to different datasets and potential improvement through domain adaptation and deeper architectures are also discussed [2].

This paper proposes a novel topological space and set theory-based moving target tracking method to enhance retrieval efficiency in large-scale multi-camera video scenes. The experimental results show improved retrieval performance, stability, and robustness compared to other existing algorithms, with less impact from video data volume and improved adaptability to complex road network topologies [3].

Keyshot-based summarization, unlike static frames, uses a key segment selection method for creating shorter versions of videos. In order to reduce computational complexity, the LTC-SUM framework offered a lightweight 2D CNN model to draw out features from thumbnails. Older summarization methods were dependent on centralized servers, whereas here a client- driven method like LTC-SUM is introduced for results in personalized summaries directly on their devices and decreases privacy concerns with quality summarization [7].

Table 2.1. Literature Survey Summary

Reference	Methodology	Observation
[4]	Real-time Event-driven CCTV Video Analytics	Developed a real-time system for monitoring road traffic events using CCTV.
[5]	Global Diversity and Local Context analysis.	Investigated the impact of global diversity and local context in summarization.
[6]	Convolutional Neural Network (CNN), HEVC Coding Features.	Improved video summarization using CNN and HEVC features.
[7]	2D Convolutional Neural Network (2D CNN).	Lightweight client-driven approach for personalized video summarization.
[8]	Machine Learning Algorithms.	Identified challenges and opportunities in video summarization.
[9]	Improved Clustering, Silhouette Coefficient.	Proposed keyframe generation method for effective video summarization.
[10]	Survey	Comprehensive review of video question-answering techniques and benchmarks.
[11]	Fully Convolutional Network (FCN)	Successful extraction of summaries from lecture videos using FCN.
[12]	Multi-Sensor Integration	Improved key-frame extraction from first-person videos using sensor integration.
[13]	Deep Feature Matching, Motion Analysis	Enhanced wireless capsule endoscopy video summarization with deep features.

III. RESEARCH METHODOLOGY

A. YOLOv3 for Object Detection

YOLOv3, or You Only Look Once Version 3, is an advanced deep learning model for fast and accurate object detection. It implements a multi-scale feature extraction method based on the Darknet-53 architecture, making detection robust irrespective of object size.

B. Deep SORT for Object Tracking

Deep SORT (Simple Online and Realtime Tracker) builds upon the original SORT algorithm with the addition of deep learning-based appearance features to enable more accurate tracking even under poor conditions with occlusions and overlapping objects.

C. Summarization Strategy

The proposed summarization method identifies video shots with tracked objects of interest, removing duplicate frames and keeping significant interactions and events. The selection criteria are as follows:

(i) Persistence of objects across frames
(ii) Motion intensity and movement patterns
(iii) Interaction frequency and event importance

Through an emphasis on these elements, the approach can sustain significant contextual content in the shortened video while diminishing relatively little from its total duration.

The Video Summarization Architecture includes several interconnected modules (Figure 1). The process of summarization starts with the user interaction, where original long videos are uploaded through a Flask-based web interface. As the video is uploaded, processing of it begins, where frames are extracted, and object detection is performed. The objects that have been detected are then tracked using a tracking algorithm, which generates movement data used for the summarization process. The summarization module processes this data for key frame identification based on object activity and relevance that results in the generation of a summarized video. The processed and summarized video will be stored in the database for further use. This proposed structural approach ensures efficient video summarization by integrating detection, tracking, and key frame selection.

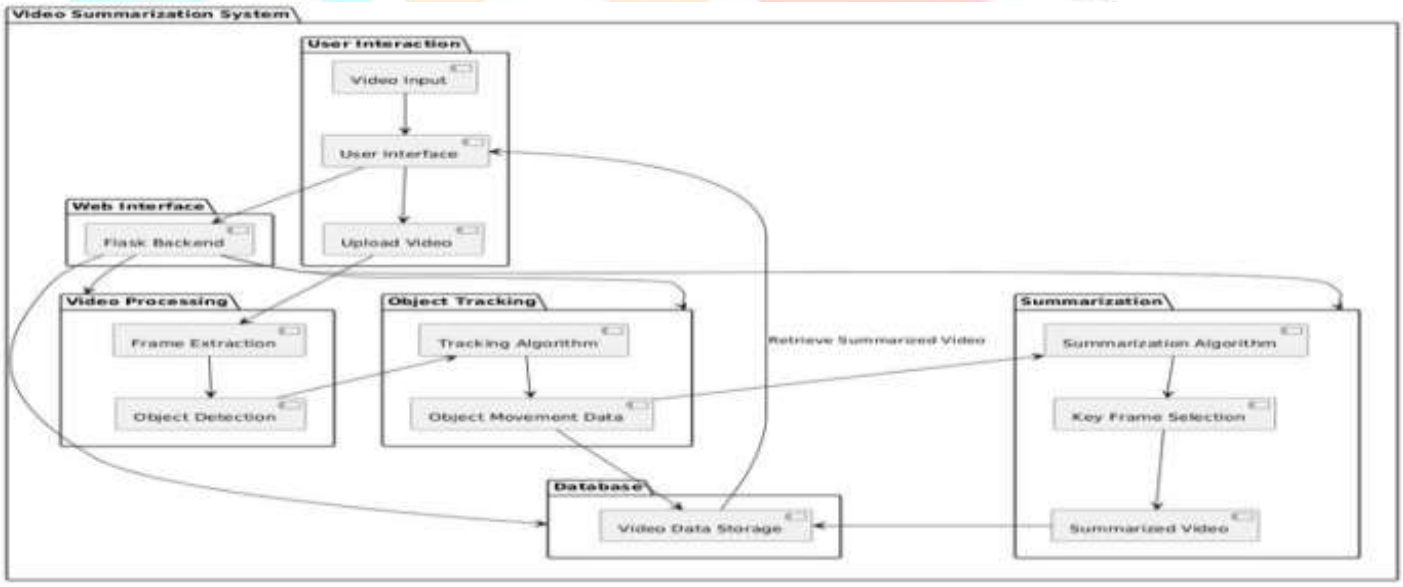

Figure 1. Video Summarization Architecture

IV. RESULTS AND DISCUSSION

The system was evaluated on benchmark data and real surveillance videos. Metrics such as precision, recall, F1-score, and processing time were compared to evaluate summarization performance.

(a)Experimental Setup: Coco dataset

(b)Evaluation Metrics: Precision, Recall, F1-score, Processing Speed.

(c)Implementation Tools: Python, OpenCV, TensorFlow/PyTorch.

Discussions: Quantitative results indicate the enhanced performance of the given approach compared to the conventional summarization with respect to precision and effectiveness. A sample qualitative outcome shows how the model performs well in retaining important moments at low redundancy cost.

The user login interface for the summarization platform provides summarized video content through authentication (Figure 2). The video upload interface enables users to select the files and upload to get a summarized video (Figure 3). The input long video as well as the short summarized video will be shown in the interface (Figure 5). Also, additional features like unique frames and common frames will be generated (Figure 6).

The live video recording interface allows users to capture the video content through CCTV or any other electronic device with a camera feature enabled. Also, it ensures a smooth transition providing keyframes and essential highlights for user review (Figure 7 – 9).

Figure 2. Login page

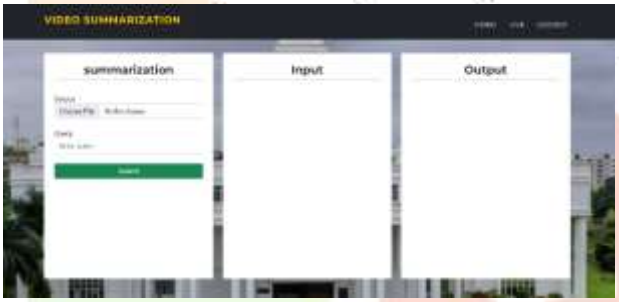

Figure 3. Video Upload

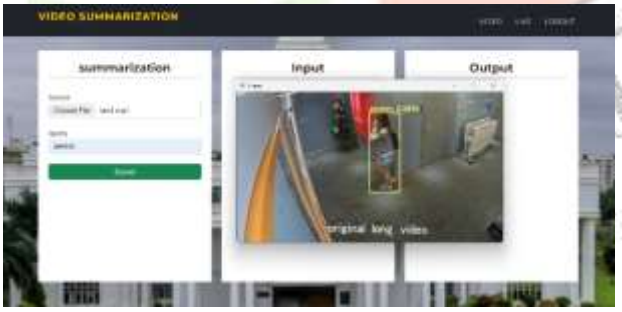

Figure 4. Processing video summarization

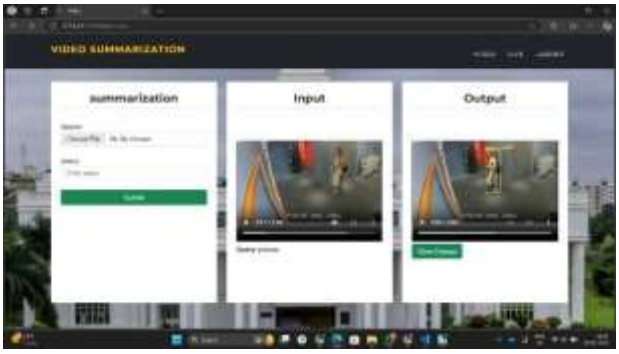

Figure 5. Input and Output

Figure 6. Unique Frames

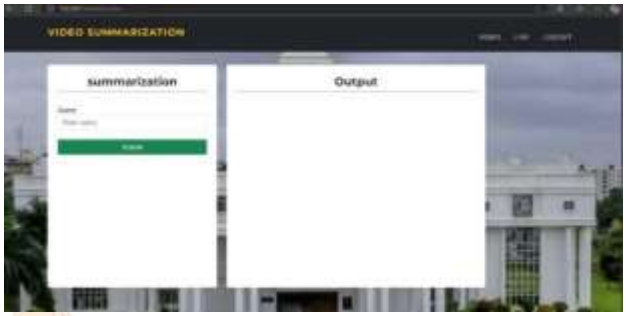

Figure 7. Live Video interface

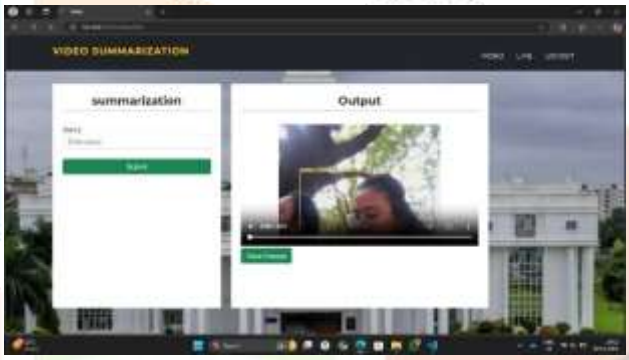

Figure 8. Live Video summarization output

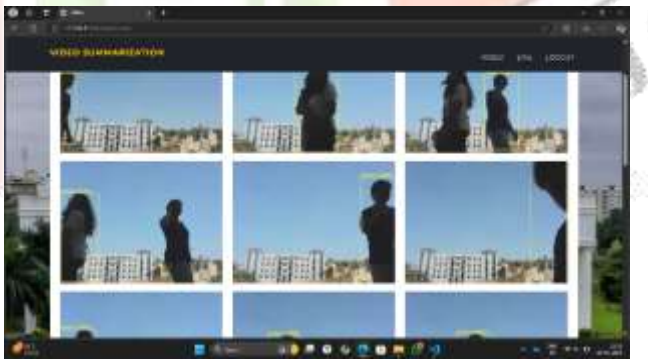

Figure 9. Unique Frames of Live Video

V. CONCLUSIONS

In this study, we proposed a new method for video summarization that uses object detection methods to improve the quality and usefulness of produced video summaries. By combining cutting-edge object detection models, we detected important objects and events in video frames. The "Object Aware Summizer" system we proposed ranks objects by importance scores and considers temporal relationships to build coherent summaries. Through exhaustive testing and comparison, we showed that our solution performs better than conventional video summarization techniques in generating summaries closer to users' expectations. The inclusion of object detection outputs and user customization functionalities makes our system able to serve various applications ranging from effective video browsing to supporting surveillance analysis and content recommendation. This research proves the efficiency of combining YOLOv3 and Deep SORT for improved video summarization. The suggested method effectively shortens video length without losing informative content, thus presenting a potential solution for real-time applications like surveillance

and automated content creation. Future work can be directed towards greater tracking robustness in cluttered scenes, efficiency of computation in order to provide speed of operation, and adaptive methods that have the ability to adaptively operate in varying video conditions.

VI. Acknowledgment

The authors would like to thank the Global Academy of Technology for the facilities and computational resources that made this research possible. We also thank faculties and collaborators for their contributions and their feedback and suggestions during the preparation of this study.

References

[1] H. Huang, Z. Wu, G. Pang, and J. Xie, "An Aesthetic-Driven Approach to Unsupervised Video Summarization," IEEE Access, vol. 12, pp. 128768-128777, Jul. 2024, doi: 10.1109/ACCESS.2024.3434508.

[2] W. Xie, Z. Chen, Q. Li, Q. Ma, Y. Wang, T. Liu, Y. Fang, and Z. Zhao, "FIAS3: Frame Importance-Assisted Sparse Subset Selection to Summarize Wireless Capsule Endoscopy Videos," IEEE Access, vol. 11, pp. 10850-10863, Jan. 2023, doi: 10.1109/ACCESS.2023.3240999.

[3] W. Zeng, X. Min, Q. Deng, and X. Zhao, "A Moving Target Tracking Framework Based on a Set and Its Topological Space," IEEE Access, vol. 11, pp. 32882-32894, Mar. 2023, doi: 10.1109/ACCESS.2023.3262994.

[4] M. Tahir, Y. Qiao, N. Kanwal, B. Lee, and M. N. Asghar, "Real-time Event-driven Road Traffic Monitoring System using CCTV Video Analytics," IEEE Access, vol. 11, pp. 139097-139111, Dec. 2023, doi: 10.1109/ACCESS.2023.3340144.

[5] Y. Pan, O. Huang, Q. Ye, Z. Li, W. Wang, G. Li, and Y. Chen, "Exploring Global Diversity and Local Context for Video Summarization," IEEE Access, vol. 10, pp. 43611-43622, Mar. 2022, doi: 10.1109/ACCESS.2022.3163414.

[6] Issa, O. and Shanableh, T., "CNN and HEVC Video Coding Features for Static Video Summarization," IEEE Access, vol. 10, pp. 72080-72091, July 2022, doi: 10.1109/ACCESS.2022.3188638.

[7] G. Mujtaba, A. Malik, and E.-S. Ryu, "LTC-SUM: Lightweight Client-Driven Personalized Video Summarization Framework Using 2D CNN," IEEE Access, vol. 10, pp. 103041-103055, Sept. 2022, doi: 10.1109/ACCESS.2022.3209275.

[8] P. Kadam, D. Vora, S. Mishra, S. Patil, K. Kotecha, A. Abraham, and L. Abdelk, "Recent Challenges and Opportunities in Video Summarization With Machine Learning Algorithms," IEEE Access, vol. 10, pp. 122762-122785, Nov. 2022,doi: 10.1109/ACCESS.2022.3223379.

[9] F. Wang, J. Chen, and F. Liu, "Keyframe Generation Method via Improved Clustering and Silhouette Coefficient for Video Summarization," J. Web Eng., vol. 20, no. 1, pp. 147-170, Jan. 2021, doi: 10.13052/jwe1540-9589.2018.

[10] K. Khurana and U. Deshpande, "Video Question-Answering Techniques, Benchmark Datasets and Evaluation Metrics Leveraging Video Captioning: A Comprehensive Survey," IEEE Access, vol. 9, pp. 43799-43823, Feb. 2021, doi: 10.1109/ACCESS.2021.3058248.

[11] K. Davila, F. Xu, S. Setlur, and V. Govindaraju, "FCN-LectureNet: Extractive Summarization of Whiteboard and Chalkboard Lecture Videos," IEEE Access, vol. 9, pp. 104469-104484, July 2021, doi: 10.1109/ACCESS.2021.3099427.

[12] Y. Li, A. Kanemura, H. Asoh, T. Miyanishi, and M. Kawanabe, "Multi-Sensor Integration for Key-Frame Extraction From First-Person Videos," IEEE Access, vol. 8, pp. 122281-122291, Jul. 2020, doi: 10.1109/ACCESS.2020.3007150.

[13] B. Sushma and P. Aparna, "Summarization of Wireless Capsule Endoscopy Video Using Deep Feature Matching and Motion Analysis," IEEE Access, vol. 9, pp. 13691-13703, Dec. 2020, doi: 10.1109/ACCESS.2020.3044759.