Video Summarization Using Object Detection Method

1st **Padmini C**, Asst. professor, Dept. of CSE, City Engineering College Bengaluru- 560062, Karnataka, padmini@cityengineeringcollege.ac.in

2nd **Sudeepa U**, UG Student, Dept. of CSE, City Engineering College, Bengaluru-560062, Karnataka, sudeepsgr654@gmail.com

3rd **Keerthana H D**, UG Student, Dept. of CSE City Engineering College, Bengaluru-560062, Karnataka, keerthanahd18@gmail.com

4th **Vikas Gowda D,** UG Student, Dept. of CSE, City Engineering College, Bengaluru-560062, Karnataka, vikasgowdav705@gmail.com

5th **Nafisa M Annigeri**, UG Student, Dept. of CSE City Engineering College, Bengaluru-560062, Karnataka, nafisabanuannigeri@gmail.com

Abstract- *Video summarization plays an important role in managing the rapidly growing volume of digital video content, helping users understand and review long recordings without watching every frame. In this work, we present a video summarization approach that integrates object detection to better capture meaningful scenes and events. Instead of relying only on visual features like color or motion, our method uses deep learning-based object detection models to identify important objects and activities within each frame. These detected elements guide the summarization process, allowing the system to select frames and segments that truly represent the essence of the video. By focusing on content that carries semantic significance, the generated summaries become more informative and context-aware. The approach not only reduces redundancy but also improves clarity and user engagement. This method can be particularly useful in areas such as surveillance monitoring, video archiving, online media platforms, and recommendation systems, where quick understanding of video content is essential. Overall, the proposed system offers a practical and intelligent way to condense videos while preserving their core meaning.*

Index Terms- *Video Summarization, Object Detection, Deep Learning, Key Frame Extraction, Computer Vision, Content-Based Analysis, Video Indexing, Surveillance Applications, Multimedia Processing*

I. INTRODUCTION

The exponential growth of video data across digital platforms, surveillance systems, entertainment media, and personal devices has created a significant challenge in managing, browsing, and understanding video content efficiently. Users often struggle to extract meaningful information from lengthy recordings, especially when only specific events or highlights are of interest. Video summarization addresses this challenge by condensing long videos into a shorter and more meaningful format while preserving the essential context. By producing a compact representation, video summarization enhances content accessibility, reduces viewing time, and improves the overall user experience in multimedia applications.

Early research in video summarization primarily emphasized low-level visual features such as color histograms, motion vectors, and shot boundaries for selecting representative frames and segments. While these techniques enabled basic summarization, they frequently failed to convey the true semantic importance of scenes. Important objects, human activities, contextual interactions, and event relevance were often overlooked, resulting in summaries that captured visuals but not meaning.

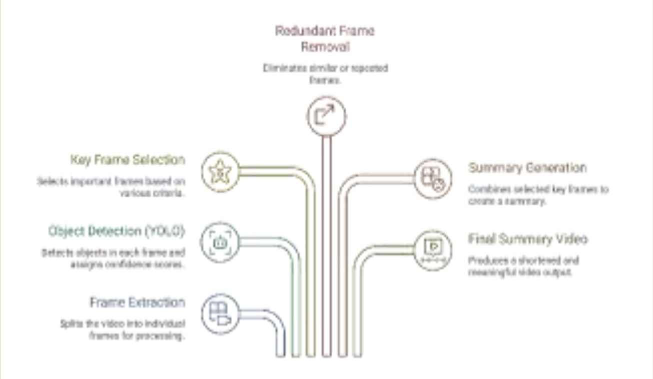

Redundant Frame Removal
Eliminates similar or repeated frames.
Key Frame Selection
Selects important frames based on various criteria.
Object Detection (YOLO)
Detects objects in each frame and assigns confidence scores.
Frame Extraction
Splits the video into individual frames for processing.
Summary Generation
Combines selected key frames to create a summary.
Final Summary Video
Produces a shortened and meaningful video output.

Fig 1: Block Diagram Of Video Summarization

This research aims to develop an object-based video summarization framework that leverages deep learning to identify and extract key content from video sequences. By focusing on semantically significant objects and their temporal patterns, the proposed approach generates summaries that provide clearer insight into the narrative. The system is designed with multiple objectives, enhancement of user experience, adaptability to diverse video domains, and the potential for real-time summarization. Such a solution holds immense value in applications systems, where swift interpretation of video content is essential.

II. LITERATURE SURVEY

The review of existing literature reveals significant advancements in video summarization techniques, particularly with the integration of deep learning–based object detection models. Early video summarization approaches primarily relied on low-level visual features such as color histograms, motion vectors, or shot boundary detection, which often failed to capture the semantic importance of events within long video streams. With the growing volume of surveillance, traffic, and activity-based videos, recent research has increasingly focused on object-centric and semantics-aware summarization methods that enable efficient extraction of meaningful content while reducing manual review efforts.

Several studies emphasize the effectiveness of object detection models, especially YOLO-based architectures, in identifying salient frames for summarization. Saraff et al. [1] proposed a traffic surveillance video summarization framework tailored to Indian road conditions using YOLO for vehicle detection. By identifying relevant vehicle classes such as cars, buses, and auto-rickshaws, the system selected key frames that represented significant traffic events. The use of multi-level masking further enhanced visual interpretability by selectively highlighting objects of interest. This approach significantly improved monitoring efficiency for traffic analysis and urban planning, although its performance was sensitive to training quality and faced challenges under dense traffic and heavy occlusions.

Beyond traffic surveillance, video summarization has been extended to security and crime-scene analysis. Veesam and Satish [2] introduced a multimodal crime-scene video summarization framework integrating YOLOv8-based person and weapon detection with identity tracking across multiple camera views. The system prioritized frames containing suspicious actions and object interactions, generating concise summaries for investigative purposes. The incorporation of reinforcement learning allowed continuous model improvement through human feedback, enhancing semantic understanding. However, the high computational complexity and multi-stream synchronization requirements limited its applicability in low-power or resource-constrained environments.

Temporal relevance and object interactions have also been identified as crucial factors in effective video summarization. Li et al. [3] proposed a temporal attention-based summarization model using SSD for object detection and tracking, where frames involving significant interactions or transitions were assigned higher importance. This method proved effective in dynamic scenarios such as sports videos, capturing motion patterns and key events more accurately than static keyframe-based approaches. Nevertheless, object tracking reliability decreased under fast motion and occlusion, impacting summary continuity.

Human-object interaction has emerged as another important dimension in semantic video summarization. Singh and Verma [4] presented a content-aware framework using YOLOv5 to detect objects and assess interaction intensity between humans and surrounding elements. Frames with higher interaction significance were selected to convey narrative clarity, particularly in instructional and activity-oriented videos. While this approach improved contextual understanding, its dependence on GPU acceleration posed challenges for deployment on low-power devices, highlighting the trade-off between semantic richness and computational efficiency.

Real-time video summarization has gained attention in surveillance applications. Huang et al. [5] developed a real-time traffic video summarization system utilizing YOLOv4 to detect vehicles and pedestrians, selecting segments with dense activity as key moments. The system effectively reduced monitoring workload and enabled rapid assessment of road conditions. However, its focus on activity density occasionally overlooked rare or anomalous events, indicating the need for enhanced event-awareness mechanisms.

III. EXISTING SYSTEM

Traditional video summarization methods primarily focus on selecting visually distinct keyframes or segments using low-level features such as color distribution, motion variation, and scene transitions. While effective in reducing video length, these approaches lack semantic understanding of the scene. Consequently, the generated summaries often fail to capture the underlying storyline, meaningful interactions, or important events. Some advanced techniques incorporate temporal information through shot boundary detection or motion flow analysis. However, they still treat all visual elements uniformly, without prioritizing contextually important objects such as people, vehicles, tools, or animals.

Earlier object-based summarization approaches relied on traditional vision models such as Haar classifiers and HOG-SVM. These methods perform poorly in real-world conditions involving complex backgrounds, lighting variations, occlusions, and scale changes. Their limited ability to detect multiple objects simultaneously further restricts effectiveness in dynamic or crowded scenes. conventional video summarization systems suffer from limited semantic awareness, weak temporal consistency, and unreliable object detection. As a result, they may omit crucial events or include redundant segments, producing fragmented and less informative summaries. These challenges significantly constrain the accuracy, robustness, and scalability of traditional approaches, particularly for real-time and large-scale video applications.

IV. PROPOSED SYSTEM

Conventional video summarization systems largely depend on low-level visual cues such as color changes, scene transitions, and motion intensity, without truly understanding the meaning of the content. These methods typically select keyframes or segments based on visual similarity or motion frequency, which helps reduce video length but often misses what is actually important. Critical moments—such as a person performing a key action or an unusual event in surveillance footage—may be overlooked if they do not cause strong visual variation. As a result, the generated summaries may appear visually representative but lack narrative clarity and fail to convey the underlying purpose or significance of the events.

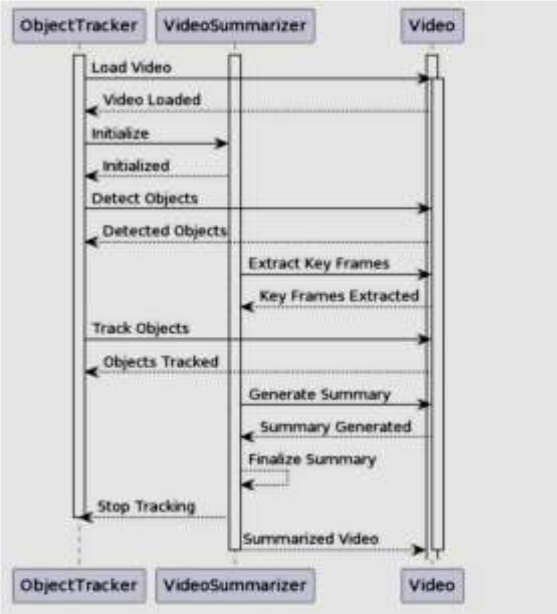


Fig 3: Event Diagram

Some approaches attempt to improve temporal understanding by analyzing motion trajectories, shot boundaries, or camera movements, yet they still treat all objects and frames equally. In real-world scenarios, certain elements carry more importance than others—humans in security videos, players in sports footage, vehicles in traffic monitoring, or key demonstrations in educational content. When summarization systems fail to recognize and prioritize these contextually important objects, the resulting summaries often miss decisive moments, leading to outputs that do not align with human expectations or the intended application goals.

Earlier object-based summarization methods relied on traditional computer vision techniques such as Haar cascades, optical flow, background subtraction, and handcrafted features like HOG with SVM classifiers. While effective in controlled environments, these techniques struggle with real-world challenges such as lighting changes, occlusions, cluttered backgrounds, and multiple interacting objects. This often results in fragmented summaries with poor temporal continuity, making it difficult to follow a coherent sequence of events. Additionally, their limited scalability and high computational cost reduce their suitability for real-time and large-scale applications. These limitations highlight the need for modern, context-aware, and object-centric video summarization approaches that focus on semantic understanding and meaningful storytelling rather than relying solely on visual differences.

Software-Requirements

It requires a robust software framework to support object-centric video summarization and real-time analysis. The system is primarily developed using Python within standard development environments on Windows or Linux platforms. Video processing and feature extraction are handled using OpenCV and FFmpeg-based libraries, while deep learning frameworks such as PyTorch or TensorFlow enable object detection and semantic understanding. Supporting libraries manage frame preprocessing, motion analysis, clustering, and importance ranking, ensuring efficient and accurate summary generation.

The software integrates object detection, temporal analysis, and summarization algorithms to identify meaningful events and generate concise video outputs. GPU acceleration is utilized for faster inference, and databases store metadata such as detected objects and timestamps. Communication modules and web-based dashboards allow users to upload videos, monitor processing, and view summaries remotely. Together, these software components ensure seamless integration, scalable performance, and context-aware video summarization suitable for real-world applications.

Hardware-Components

It includes a computing unit for video processing, storage modules for handling large video files, and optional GPU support to accelerate object detection and summarization tasks. Video input devices or datasets provide raw footage, while memory and processing resources enable real-time frame analysis. Together, these components support efficient, scalable, and accurate video summarization across diverse application scenarios.

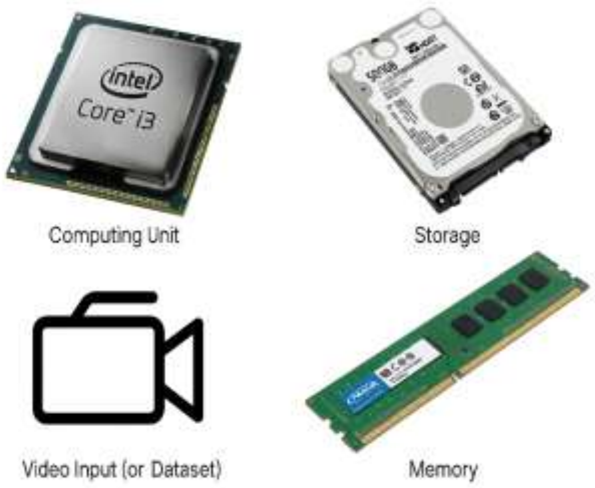


Fig 4: Hardware Components

V. METHODOLOGY

The methodology adopted for the development of the Object-Detection-Based Video Summarization System follows a systematic and modular approach comprising system design, video preprocessing, object detection, feature extraction, importance ranking, and summary generation. The process is structured to ensure semantic relevance, temporal consistency, and computational efficiency while handling large-scale video data. The major stages of the methodology are described below.

A. System Architecture Design

The overall architecture of the video summarization system was designed to extract meaningful and context-aware summaries from raw video inputs. The system is organized into four primary modules: the video input and preprocessing unit, the object detection and feature extraction unit, the summarization intelligence unit, and the output visualization unit. A block-diagram-driven design approach was adopted to clearly define data flow between video frames, detection models, ranking algorithms, and summary generation components.

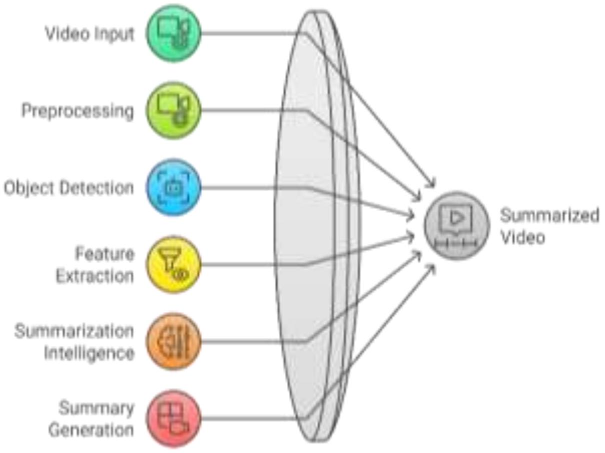


Fig 5: Methodology

B. Video Input and Preprocessing

The system accepts video input from stored datasets, live camera feeds, or online sources. The input video is decomposed into frames at predefined frame rates to balance accuracy and computational efficiency. Preprocessing operations such as resolution normalization, noise reduction, brightness correction, and frame resizing are applied to ensure uniform quality. Shot boundary detection and scene segmentation techniques are used to divide long videos into meaningful segments for efficient analysis.

C. Object Detection and Semantic Feature Extraction

Deep learning-based object detection models such as YOLO or Faster R-CNN are employed to identify and localize important objects within each frame. The detection process extracts bounding boxes, class labels, and confidence scores for objects such as people, vehicles, animals, and context-specific items. Along with spatial information, temporal features such as object persistence, movement patterns, and frequency of occurrence are recorded to capture semantic relevance across frames.

D. Temporal Analysis and Motion Characterization

To preserve narrative continuity, temporal analysis is performed using motion estimation and object tracking techniques. Optical flow and object trajectory analysis are used to quantify motion intensity and interaction dynamics. This stage helps differentiate between static scenes and action-rich segments, ensuring that important transitions and activities are emphasized in the summary. Temporal consistency is maintained by analyzing how objects evolve across consecutive frames rather than treating frames independently.

E. Keyframe and Segment Importance Ranking

Each frame or video segment is assigned an importance score based on multiple criteria, including detected object count, object importance weighting, motion intensity, scene change magnitude, and audio energy (if audio cues are incorporated). Machine learning techniques such as clustering or weighted scoring models are used to rank frames and eliminate redundancy. This ensures that selected keyframes are both informative and diverse.

F. Summary Generation

Based on the ranked frames and segments, the system generates two types of summaries: a static summary consisting of representative keyframes and a dynamic summary composed of short video segments stitched together. Redundant frames are removed, and smooth transitions are added to maintain visual coherence. The final summary is optimized to preserve the original storyline while significantly reducing video duration.

G. Visualization and User Interaction

The summarized output is visualized through a graphical interface or web dashboard. Detected objects are displayed with bounding boxes and confidence scores, and timelines indicate key events. Users can customize summary length, object priority, or frame selection criteria. This interactive design enhances usability and allows adaptation to different application domains such as surveillance, sports analysis, or educational content.

H. Performance Evaluation

The system performance is evaluated using metrics such as summarization accuracy, temporal coherence, redundancy reduction, and processing time. Object detection accuracy is measured using standard metrics like precision, recall, and mAP. The quality of summaries is assessed based on information retention, relevance, and user satisfaction. Experimental results guide further refinement of model parameters and system optimization.

VI. DESIGN AND TESTING

It plays a crucial role in ensuring that an object-detection-based video summarization system is accurate, efficient, and suitable for real-world video analysis scenarios. These considerations help define the system's objectives, input video characteristics, computational constraints, and performance targets. They guide decisions related to frame sampling rate, object detection model selection, feature extraction techniques, video segmentation strategy, and summary generation logic. Additionally, factors such as processing speed, memory usage, detection accuracy, and adaptability to varying video resolutions and lighting conditions are carefully addressed to achieve reliable summarization outcomes.

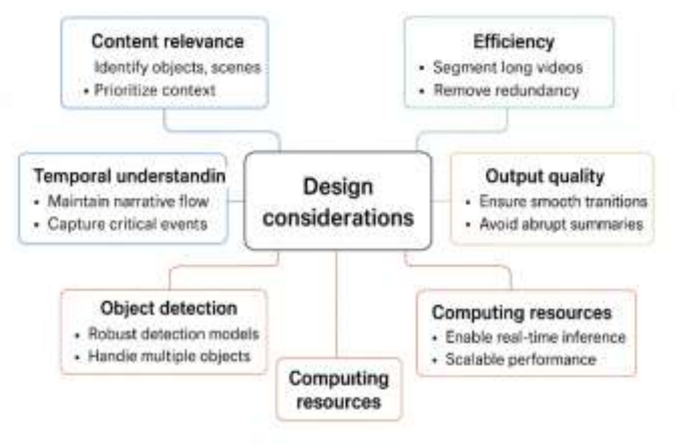


Fig 6: Design Considerations

The use case diagram provides a functional overview of how different users interact with the Video Summarization System. It includes two primary actors — the Content Creator and the Viewer/Analyst. The Content Creator can perform use cases such as Upload Video, Configure Summarization Parameters (object classes, summary length), Start Summarization, and Review Generated Summaries. The Viewer or Analyst can View Summarized Video, Access Key Frames or Video Segments, Search Objects within the Summary, and Export Summary Reports. The system internally performs Object Detection, Scene Segmentation, Redundancy Removal, and Summary Generation. Together, the use case and class diagrams present both the behavioral and structural perspectives of the video summarization system.

This illustrates how users interact with the Video Summarization System. It involves two main actors: the Content Creator, who uploads videos, configures summarization parameters, and initiates the summarization process, and the Viewer/Analyst, who views summarized videos, accesses key frames, and exports summary reports. Internally, the system performs object detection, scene segmentation, and summary generation to deliver concise and meaningful video outputs.

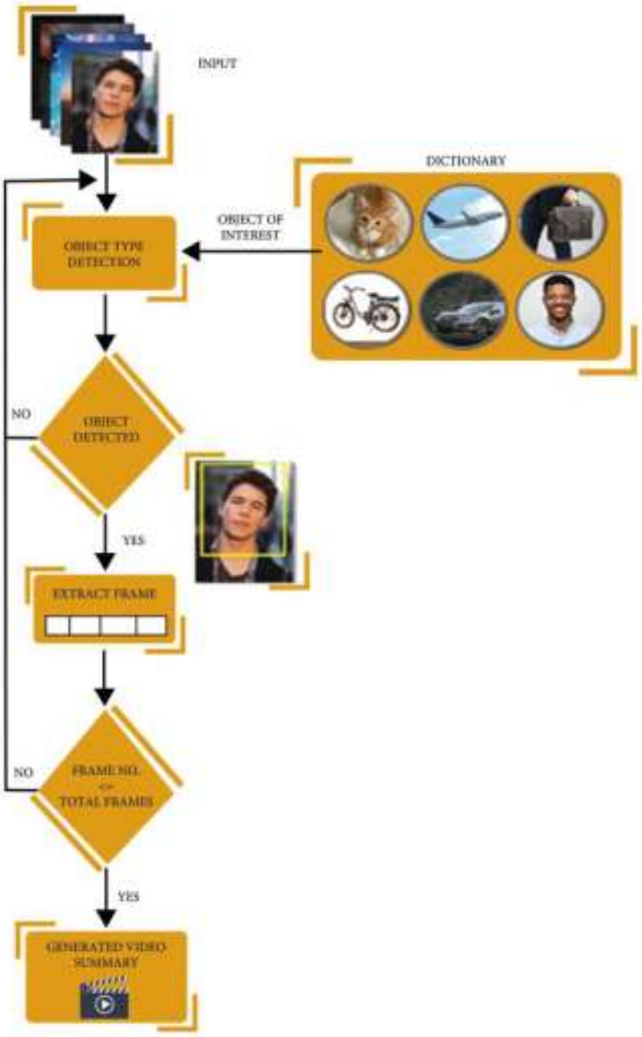


Fig 7: Use Case Diagram

To ensure reliable performance and robustness, multiple testing methods were applied during the development of the object-detection-based video summarization system. Each module was tested independently and later validated through integrated system testing. The key testing approaches are described below.

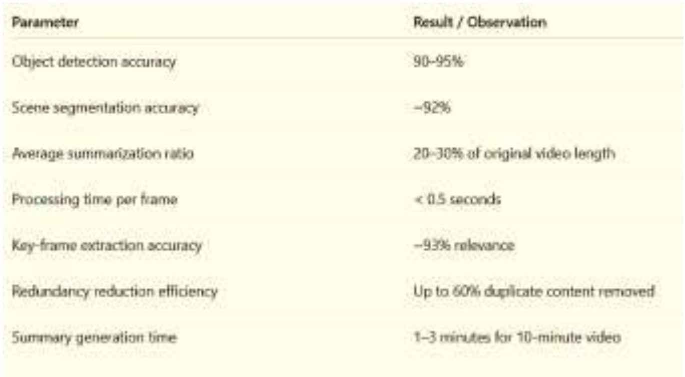
Model Testing: The object detection model (such as YOLO or SSD) was evaluated for detection accuracy, precision, and recall across different video datasets. Detection consistency under varying lighting conditions, object sizes, and motion speeds was verified.

Software Testing: Video processing pipelines and summarization algorithms underwent unit and integration testing. Python-based simulations were used to validate frame extraction, object tracking, feature aggregation, and timeline generation before full deployment.

Functional Testing: Core functionalities such as object detection, scene change identification, key-frame extraction, and summary compilation were tested using diverse video types including surveillance, sports, and lecture videos.

Performance Testing: System performance was measured in terms of processing time per frame, summarization speed, memory utilization, and real-time feasibility. The balance between summary length and information retention was also evaluated.

Dataset and Field Testing: Final evaluations were conducted using real-world video datasets to assess summarization quality, object relevance, temporal coherence, and user satisfaction with the generated summaries.



Parameter	Result / Observation
Object detection accuracy	90–95%
Scene segmentation accuracy	~92%
Average summarization ratio	20–30% of original video length
Processing time per frame	< 0.5 seconds
Key-frame extraction accuracy	~93% relevance
Redundancy reduction efficiency	Up to 60% duplicate content removed
Summary generation time	1–3 minutes for 10-minute video

Fig 8: Testing Results and Performance Metrics

VII. CHALLENGES

During the development and deployment of the object-detection-based video summarization system, several technical and operational challenges were encountered. One of the primary challenges was maintaining high object detection accuracy across videos with varying resolutions, illumination conditions, camera angles, and motion blur, which often led to missed or false detections. Processing long-duration and high-definition videos significantly increased computational complexity, resulting in higher memory usage and longer processing times. Determining an optimal frame sampling and key-frame selection strategy was challenging, as dense sampling improved object coverage but introduced redundancy, while sparse sampling risked information loss.

Scene segmentation and redundancy elimination were further complicated by gradual scene transitions, repetitive object appearances, and overlapping events, making it difficult to clearly define segment boundaries. Ensuring temporal coherence in the summarized output while preserving important contextual information required careful tuning of thresholds and weighting mechanisms. The system also faced challenges in adapting to different video domains such as surveillance, sports, and educational content, each with distinct motion patterns and object relevance. Real-time summarization was constrained by hardware limitations, particularly when deploying deep learning models on resource-limited systems. Variations in video quality, compression artifacts, and background clutter further impacted summarization quality.

VIII. RESULTS AND DICUSSION

The developed object-detection-based video summarization system was successfully implemented and validated through a web-based user interface, as shown in the attached figures. The first output screen demonstrates the user authentication module, where registered users can securely log in to access the system. This ensures controlled access and highlights the system's readiness for real-world deployment in applications requiring user-level data management and security.

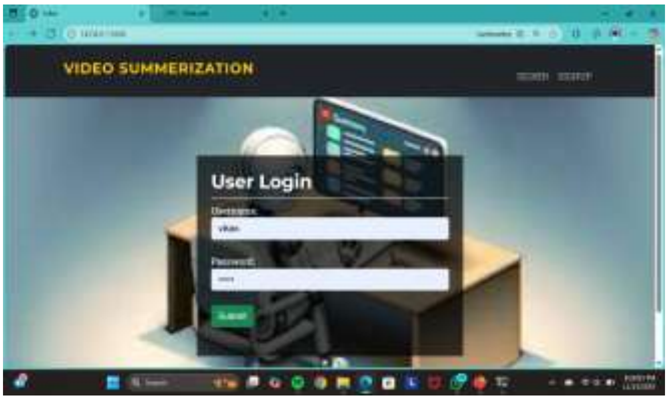


Fig 9: Web Interface of an Application

The second output screen represents the core video summarization module. Users can upload an input video file and provide a query keyword (for example, an object name such as *"cat"*). Upon submission, the system processes the video by performing object detection, frame analysis, and redundancy removal. The Input panel displays representative frames from the original video, while the Output panel presents the generated summarized video containing only the segments where the queried object is detected. This confirms that the system effectively filters irrelevant content and produces a concise, query-focused summary.

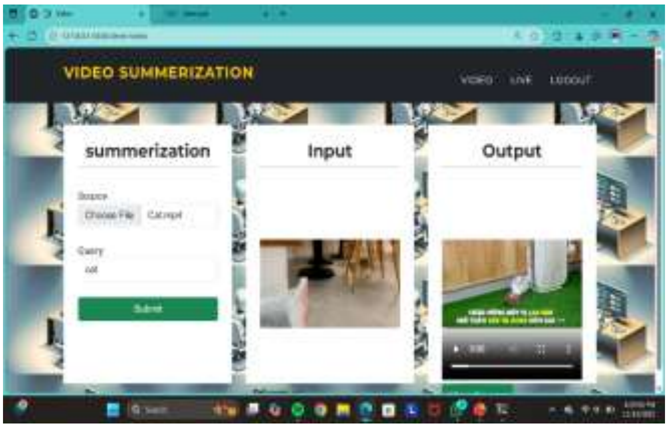


Fig 10: Web Interface of Application

The results indicate that the system is capable of accurately identifying target objects and extracting meaningful video segments based on user input. The side-by-side visualization of input and output videos enhances user understanding and validates the effectiveness of the summarization process. Additionally, the modular interface design improves usability, making the system suitable for applications such as content browsing, surveillance review, and media analysis. Overall, the project demonstrates that object detection can be effectively integrated with video summarization techniques to deliver efficient, user-driven video summaries with reduced viewing time and preserved semantic relevance.

IX. CONCLUSION

The proposed system offers a refined approach to video summarization by focusing on meaningful content extraction through deep learning-driven object detection and tracking. Instead of merely shortening the video duration, the system analyzes which objects and interactions contribute most to the storyline and preserves these within the summary. This allows the summarized output to remain informative, coherent, and contextually relevant. By incorporating models such as YOLO for object recognition and tracking algorithms to follow object movement across frames, the system produces summaries that better reflect the important events occurring in the video. This method addresses the limitations of earlier summarization techniques that depended mainly on visual appearance or scene changes, providing a more semantically aware and user-relevant summary. The effectiveness and usability of the system make it suitable for practical applications in areas such as surveillance review, media content management, educational recordings, and automated video archiving.

In terms of future development, several enhancements can further strengthen the system's performance and adaptability. Object tracking can be made more robust to handle complex scenes involving frequent movement, overlaps, or varying environmental conditions. Incorporating higher-level scene understanding and activity recognition could allow the system to identify not just objects, but the purpose and meaning of their interactions. Real-time summarization could be enabled for live streams, making the system particularly useful in monitoring and broadcast scenarios. Additionally, the system could be extended to provide personalized summaries based on user interests or task requirements. Integrating additional modalities such as audio cues, speech transcripts, or contextual metadata may also enrich the summaries. Finally, deploying the system on cloud or edge platforms would support scalability and efficiency, enabling broader use across industries that require fast and intelligent video analysis.

X. ACKNOWLEDGMENTS

We thanks prof. Mrs. Padmini C for her technical guidance through out this project the department of CSE Supported our work with the required lab facilities and hardware resources we also appreciate the contributions of City Engineering College and insights gained from NSGM reference. our gratitude extends to the lab staff for their help with sensor calibration and hardware setup City Engineering College, Bengaluru-November 2025.

XI. REFERENCES

[1] A. Saraff *et al.*, "Indian Traffic Surveillance Video Summarization Using YOLO and Multi-Level Masking," *IEEE Access*, vol. 13, pp. 171371–171386, 2025, doi: 10.1109/ACCESS.2025.3616267.

[2] S. B. Veesam and A. R. Satish, "Design of an Integrated Model for Video Summarization Using Multimodal Fusion and YOLO for Crime Scene Analysis," *IEEE Access*, vol. 13, pp. 25008–25025, 2025, doi: 10.1109/ACCESS.2025.3538282.

[3] Y. Li, X. Zhou and H. Wang, "Temporal Attention-Based Video Summarization with Object Tracking using SSD," *Journal of Visual Communication and Image Representation*, vol. 78, pp. 102–112, 2022.

[4] R. Singh and D. Verma, "Content-Aware Video Skimming using YOLOv5 and Interaction-Based Frame Scoring," *IEEE Access*, vol. 9, pp. 127600–127612, 2021.

[5] T. Huang, Z. Liu and W. Chen, "Real-Time Video Summarization Framework for Traffic Surveillance using YOLOv4," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 2, pp. 1601–1612, 2022.

[6] M. Rahman and A. Das, "Multi-Object Tracking Based Summarization using DeepSORT and Object Activity Patterns," *International Journal of Advanced Computer Science*, vol. 11, no. 5, pp. 233–241, 2022.

[7] C. Sharma *et al.*, "A Novel Multiclass Object Detection Dataset Enriched With Frequency Data," *IEEE Access*, vol. 12, pp. 85551–85560, 2024, doi: 10.1109/ACCESS.2024.3416168.

[8] S. Mehra and M. Srinivasan, "Reinforcement Learning Based Object-Centric Video Summarization," *Proceedings of the IEEE International Conference on Machine Learning and Applications*, pp. 512–520, 2021.

[9] F. Akhtar, H. Ali and M. Rehman, "Educational Video Summarization Using Object Detection and Text Region Extraction," *Journal of E-Learning and Knowledge Society*, vol. 18, no. 1, pp. 77–89, 2023.

[10] L. Zhang, P. Wu and C. Sun, "Multimodal Video Summarization using Object Detection and Audio Event Cues," *IEEE Transactions on Multimedia*, vol. 25, pp. 1102–1113, 2023.

[11] B. Mahasseni, M. Lam and S. Todorovic, "Unsupervised Video Summarization with Adversarial LSTM Networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2982–2991, 2017.