Object-based Video Summarization using YOLOv8 and SAM2

Sahil Sonawane, Ankit Kumar, Disha Soni, Himanshu Raj, Tavva Srinivas, Ashutosh Sahu, Aakash Dhoriyani, Rohit Patil
Indian Institute of Information Technology, Allahabad

***Abstract*—An Object-based Video Summarization system is leveraging the YOLOv8 and SAM2 models. The system aims to enhance the video summarization process by integrating real-time object detection and segmentation capabilities, allowing for the creation of concise and informative video summaries. YOLOv8 is used for efficient object detection, while SAM2 (Segment Anything Model 2) improves object segmentation, producing precise object boundaries. This combination ensures that the video summaries are contextually accurate and representative of the most important content, improving usability in scenarios like media libraries, surveillance, and video analytics.**

***Index Terms*—Object-based Video Summarization, YOLOv8, SAM2, Deep Learning, Object Detection, Segmentation, Video Processing.**

I. Introduction

A. Purpose

The purpose is to describe the functionalities and requirements of a video summarization system based on YOLOv8 and SAM2 integration. This system focuses on enhancing object-based video summarization by utilizing state-of-the-art deep learning techniques for object detection and segmentation. The system allows users to define objects of interest (OOIs), and the summarization will focus on frames containing these objects, leading to concise yet comprehensive summaries of videos.

B. Intended Audience and Reading Suggestions

This document is intended for:

- **Developers**: To understand the system architecture and technical requirements for implementation.
- **Project Managers**: To manage the development timeline, resources, and deliverables.
- **Testers**: To design test cases and ensure functionality meets the described requirements.
- **Documentation Writers**: To prepare user manuals, installation guides, and supplementary documentation.

Reading Suggestions:

- Begin with **Section 1** for an overview of the system and its goals.
- Refer to **Section 2** for details on the system architecture and methodologies.
- Explore **Section 4** for specific functionalities and system features.
- Review **Sections 7 and 8** for detailed diagrams, flowcharts, and non-functional requirements.

C. Project Scope

The scope of this project is to integrate the YOLOv8 object detection model with SAM2 (Segment Anything Model 2) to create a robust video summarization system. The system will:

- **Improve object detection** in real-time, focusing on specific user-defined objects of interest.
- **Enhance segmentation accuracy** for those objects, ensuring the output summary contains only the most relevant frames.
- **Provide temporal coherence**, maintaining the logical sequence and contextual relationships between frames.
- **Improve user experience** by simplifying video navigation in large datasets such as YouTube libraries, surveillance footage, and TikTok videos.

II. Comparative Analysis of Video Summarization Methods

A. Introduction

In recent years, several video summarization methods have been developed to enhance object detection and segmentation within video content. This section provides a comparative analysis of several state-of-the-art methods and highlights the strengths of integrating YOLOv8 and SAM2 for more effective summarization.

B. Existing Methods Overview

- **PGL-SUM**: Utilizes a combination of global and local attention mechanisms to estimate frame importance, addressing limitations in Recurrent Neural Networks (RNNs). This method focuses on temporal coherence.
- **STVT (Spatiotemporal Vision Transformer)**: Incorporates attention mechanisms to capture dependencies across frames. This model excels in handling both intra-frame and inter-frame correlations.
- **C2F (Coarse-to-Fine Network)**: This network refines predictions using hierarchical, multi-scale representations, which improves object detection accuracy and efficiency.

C. YOLOv8 and SAM2 Integration

The integration of YOLOv8 with SAM2 brings together real-time object detection and detailed segmentation capabilities. YOLOv8 detects objects efficiently in each frame, while SAM2 precisely segments them. This combination ensures that

only the most relevant portions of the video are included in the summary.

III. Methodology

A. YOLOv8 and SAM2 Integration

YOLOv8 is designed for real-time object detection, making it ideal for applications where speed and accuracy are critical. **SAM2** complements YOLOv8 by providing highly precise segmentation of objects, ensuring that irrelevant portions of the video are excluded from the summary. The integration involves:

- **Step 1**: Detect objects using YOLOv8 across all frames of the video.
- **Step 2**: Apply SAM2 to segment the detected objects for higher precision.
- **Step 3**: Generate the video summary by combining the relevant frames and maintaining temporal coherence.

B. Flowcharts and Diagrams

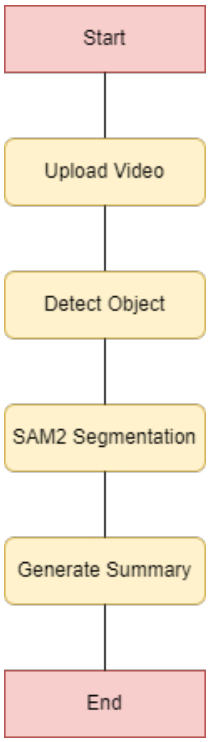

Fig. 1. YOLOv8 and SAM2 Integrated System Workflow

IV. Overall Description

A. Product Perspective

The product integrates state-of-the-art object detection and segmentation techniques to provide a real-time video summarization tool. The system improves over existing models by delivering more accurate and contextually coherent summaries, particularly in large video datasets.

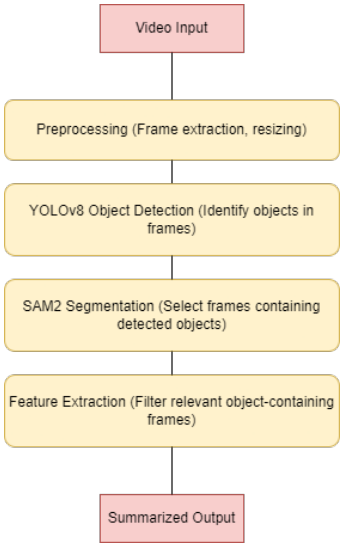

Fig. 2. Block Diagram of System Architecture

B. Product Functions

- **Real-Time Object Detection**: Utilizing YOLOv8 for fast object identification in video frames.
- **Precise Segmentation**: SAM2 enables accurate segmentation, ensuring that the summary includes only relevant objects.
- **User-Defined Focus**: The system allows users to select objects of interest, personalizing the summarization.

C. Operating Environment

The system is designed to operate in high-performance environments with access to GPU acceleration (e.g., Nvidia RTX), using frameworks such as **PyTorch** or **TensorFlow** for deep learning computations.

V. External Interface Requirements

A. User Interface

The system will provide a graphical user interface (GUI) where users can upload videos, select objects of interest, and configure settings for summarization. The summarized output will be displayed for further refinement.

B. Hardware and Software Interfaces

The system requires GPUs for both training and inference. The hardware must support multi-GPU setups for efficient summarization. The software interfaces include PyTorch for model implementation and OpenCV for video processing.

C. Communication Interfaces

The system offers API integration for uploading videos and retrieving summarized outputs, enabling seamless interaction with video management systems.

VI. System Features

A. YOLOv8 and SAM2-based Summarization

Components:

- **YOLOv8**: Detects objects in video frames in real-time.
- **SAM2**: Segments the detected objects for precise extraction.

Performance:

- **YOLOv8** achieves high object detection accuracy with minimal processing time.
- **SAM2** ensures that only relevant regions are included in the summary, improving precision.

VII. Non-Functional Requirements

A. Performance

- The system processes video at 30 frames per second using YOLOv8.
- SAM2 ensures that the segmentation is performed within 1-2 seconds per frame.

B. Safety and Security

- The system ensures data privacy by summarizing only user-defined objects and excluding unnecessary video frames.
- Videos are processed securely, with encryption during data transmission to protect sensitive content.

VIII. Conclusion

The integration of YOLOv8 and SAM2 in object-based video summarization offers significant advancements over existing methods. By focusing on user-defined objects and delivering precise segmentation, this system improves efficiency and accuracy in video summarization tasks. Future work may involve improving temporal coherence and optimizing the system for even larger datasets.

IX. Bibliography

References

[1] E. Apostolidis et al., "Combining Global and Local Attention with Positional Encoding for Video Summarization," *IEEE International Symposium on Multimedia (ISM)*, 2021.

[2] J. A. Ghauri, S. Hakimov, and R. Ewerth, "Supervised Video Summarization via Multiple Feature Sets with Parallel Attention," *TIB–Leibniz Information Centre for Science and Technology*, 2024.

[3] Y. Jin, X. Tian, Z. Zhang, P. Liu, and X. Tang, "C2F: An Effective Coarse-to-Fine Network for Video Summarization," *Harbin Institute of Technology*, 2024.