**PROJECT PHASE 1 REPORT**

ON

# Real-Time Segmentation and Labelling of Objects in Videos

Submitted by

**Ajay T Shaju (SJC20AD004)**

**Emil Saj Abraham (SJC20AD028)**

**Justin Thomas Jo (SJC20AD046)**

**Vishnuprasad K G (SJC20AD063)**

to

the APJ Abdul Kalam Technological University

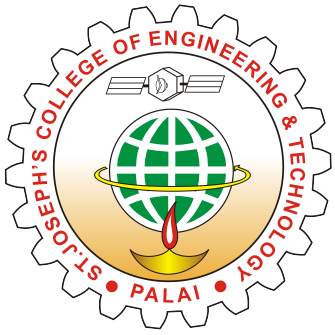
in partial fulfillment of the requirements for the award of the degree

of

Bachelor of Technology

in

**Artificial Intelligence and Data Science**




**Department of Artificial Intelligence and Data Science**
**St. Joseph's College of Engineering and Technology, Palai**

December : 2023

# Declaration

We undersigned hereby declare that the project phase 1 report on **"Real-Time Segmentation And Labelling Of Objects In Videos"**, submitted for partial fulfillment of the requirements for the award of the degree of Bachelor of Technology of the APJ Abdul Kalam Technological University, Kerala, is a bonafide work done by us under the supervision of **Dr. Deepa V**. This submission represents our ideas in our own words and where ideas or words of others have been included. We have adequately and accurately cited and referenced the original sources. We also declare that we have adhered to the ethics of academic honesty and integrity and have not misrepresented or fabricated any data, idea, fact, or source in our submission. We understand that any violation of the above will be a cause for disciplinary action by the institute and/or the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. This report has not been previously formed the basis for the award of any degree, diploma, or similar title of any other University.

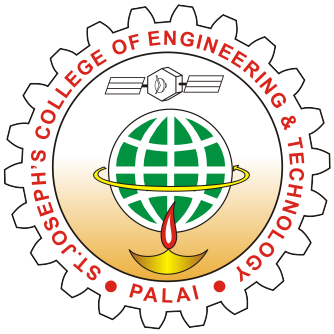
**Name and Signature of Students**

**Ajay T Shaju (SJC20AD004)**
**Emil Saj Abraham (SJC20AD028)**
**Justin Thomas Jo (SJC20AD046)**
**Vishnuprasad K G (SJC20AD063)**

Place: Choondacherry
Date: 07-12-2023

ST. JOSEPH'S COLLEGE OF ENGINEERING AND TECHNOLOGY, PALAI

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE




# CERTIFICATE

This is to certify that the report entitled "**Real-Time Segmentation And Labelling Of Objects In Videos**" submitted by **Ajay T Shaju (SJC20AD004), Emil Saj Abraham (SJC20AD028), Justin Thomas Jo (SJC20AD046), and Vishnuprasad KG (SJC20AD063)** to the APJ Abdul Kalam Technological University in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Artificial Intelligence and Data Science is a bonafide record of the project work carried out by them under my guidance and supervision.

**Project Guide**
Dr.Deepa V
Head of the Department
Department of AD

**Project Coordinator**
Mr.Jacob Thomas
Assistant Professor
Department of AD

**Head of Department**
Dr. Deepa V
Associate Professor
Department of AD

Place: Choondacherry
Date: 07-12-2023

## Acknowledgement

The success and final outcome of this project phase 1 required a lot of guidance and assistance from many people, and we are extremely privileged to have received their support throughout the completion of this project. All that we have accomplished is only possible due to their supervision and assistance, and we are sincerely grateful to them.

We would like to express our respect and gratitude to the management of St. Joseph's College of Engineering and Technology for providing us with the opportunity and platform to work on this project.

A special word of thanks goes to our beloved Principal, **Dr. V. P. Devassia**, for providing invaluable support and necessary facilities to carry out this project.

We are extremely indebted to **Dr. Deepa V**, Professor & Head of the Department of Artificial Intelligence and Data Science, for his valuable suggestions and encouragement throughout the course of this project work.

We would also like to express our gratitude to our project coordinator, **Prof. Jacob Thomas**, Assistant Professor in the Department of Artificial Intelligence and Data Science, for his valuable suggestions and guidelines during the entire duration of this project. We truly appreciate his contributions and technical support in preparing this report.

Our heartfelt thanks go to our project guide, **Dr. Deepa V**, Associate Professor in the Department of Artificial Intelligence and Data Science, displayed a keen interest in this project and provided guidance and all the necessary information for developing a robust system.

We are thankful and fortunate enough to have received constant encouragement, support, and guidance from all the staff members of the Department of Artificial Intelligence and Data Science. Their assistance played a crucial role in the successful completion of our project phase 1 work.

## Abstract

Automatic segmentation and labeling of objects in video is the task of identifying and assigning labels to individual objects in a video sequence. This is a challenging task due to the variations in object appearance, motion, and occlusion. However, it is an important task for many applications, such as video surveillance, self-driving cars, and medical imaging. This project aims to develop a system for the automatic segmentation and labeling of objects in video using deep learning. Deep learning models have been shown to achieve state-of-the-art results on a variety of computer vision tasks, including object segmentation and labeling. This kind of system comprises two core modules: a segmentation model for precise object identification and a labeling model dedicated to attributing accurate labels to these identified objects. Training these deep learning models necessitates a substantial dataset featuring annotated video sequences encompassing object segmentation and labeling. These systems have a wide-reaching utility across domains such as video surveillance, autonomous vehicles, and medical imaging, promising significant advancements in real-world applications.

# List of Abbreviations

A2D Actor and Action

ABME Asymmetric Bilateral Motion Estimation

BDD Berkeley DeepDrive

CamVid Cambridge-driving Labeled Video Database

CVPR Computer Vision and Pattern Recognition

DAVIS Densely Annotated VIdeo Segmentation

DDPG Deep Deterministic Policy Gradient

FBMS Freiburg-Berkeley Motion Segmentation

FCN Fully Convolutional Networks

LSTM Long Short-Term Memory

MOTS Multi-Object Tracking and Segmentation

R-CNN Region-Based Convolutional Neural Networks

SAM Segment Anything Model

SSD Single Shot MultiBox Detector

VIS Video Instant Segmentation

VOS Video Object Segmentation

VSPW Video Scene Parsing in the Wild

YOLO You Only Look Once

# List of Figures

# List of Tables

# Contents

# Chapter 1

# Introduction

In the dynamic landscape of autonomous systems, the paramount challenge lies in enabling machines to interpret and navigate through their environments with a level of sophistication close to human perception. At the core of this challenge is the processing of visual data, a fundamental aspect that governs the decision-making capabilities of autonomous entities such as self-driving cars and wheeled robots. This project report delves into the pivotal role played by automatic segmentation and labeling in advancing the perceptual capacities of these autonomous systems, addressing challenges in scalability, efficiency, and safety.

The main focus of this work is on the pivotal task of automatically segmenting and labeling objects within videos. This process involves dissecting visual data into distinct segments, each representing specific objects or areas of interest, and assigning categorical tags to these segments. The traditional method relies heavily on manual annotation, a time-consuming and resource-intensive approach that hampers the scalability and efficiency of training autonomous systems. The aim is to revolutionize this process by exploring innovative techniques that streamline object segmentation and labeling, paving the way for more efficient and scalable video analysis in complex environments.

| Approach | Definition | Key Techniques/Methods |
|---|---|---|
| **Unsupervised VOS** | Segments objects in a video without using annotated data. | Motion-based segmentation, optical flow, appearance modeling |
| **Semi-Supervised VOS** | Uses a combination of annotated and unannotated data for training. | Training on a small set of labeled frames, propagation methods |
| **Interactive VOS** | Involves human interaction to improve segmentation accuracy. | User inputs, corrections, annotations during segmentation |
| **Language-guided VOS** | Uses natural language instructions to guide the segmentation process. | Incorporates information from textual descriptions |

Table 1.1: Overview of Video Object Segmentation Approaches

| Annotation Method | Description | Use Cases |
|---|---|---|
| **Bounding Boxes** | Rectangles drawn around objects | Object detection, localization |
| **Polygons** | Outlining object shapes with vertices | Fine-grained object localization, non-rectangular objects |
| **Keypoint Skeletons** | Annotating specific points of interest | Pose estimation, object tracking |
| **Auto Annotation** | Automated generation of annotations | Speeding up large-scale annotation, initial model training |

Table 1.2: Overview of Annotation Methods

## 1.1 Background

The conventional method of manually annotating datasets, a foundational practice in the training of autonomous systems, has encountered formidable challenges in recent times. As the complexity and scope of these systems expand, there is a growing need for more extensive and diverse datasets. However, the manual annotation of such datasets has become a bottleneck, both in terms of time and resources. The labor-intensive nature of this process not only slows down the overall development and training of autonomous systems but also introduces the risk of human error, potentially undermining the accuracy and efficiency of environmental perception.

In response to the limitations imposed by manual annotation, the concept of automating segmentation and labeling has emerged as a pivotal advancement. Drawing upon the strides made in computer vision and machine learning, automated segmentation and labeling present an alternative that is not only more efficient but also highly scalable. By harnessing algorithms capable of discerning patterns and features in visual data, this approach liberates human annotators from the burdensome task of individually labeling each object in an image. Instead, machines take on the responsibility of segmenting and labeling, allowing human annotators to focus on more nuanced aspects of the training process. The challenges faced by autonomous systems extend beyond the realm of dataset annotation, particularly when these systems are deployed in real-time applications. Autonomous vehicles navigating through unpredictable traffic scenarios and real-time surveillance systems monitoring dynamic environments demand rapid and precise object identification. Traditional manual annotation processes struggle to meet the real-time demands of such applications, where decisions must be made swiftly. The need for instantaneous environmental perception necessitates the integration of automated segmentation and labeling, ensuring that these systems can process visual data on the fly, make split-second decisions, and adapt to dynamic changes in their surroundings. This intersection of real-time requirements and the limitations of manual processes underscores the urgency for innovative solutions that automated segmentation and labeling can provide.

## 1.2 Objective and Scope

Video segmentation and labeling is a critical domain within computer vision and machine learning, offering a comprehensive solution for understanding and interpreting visual data in the form of videos. The application of video segmentation and labeling is in fields such as autonomous systems, surveillance, sports analytics, and healthcare.

Video segmentation and labeling play a pivotal role in enabling machines to perceive and navigate through dynamic environments understanding the various objects and entities, enhancing safety and efficiency on the road or in the air.

Surveillance systems leverage video segmentation and labeling to detect and track objects or individuals in real time for public safety and security, where rapid identification of potential threats or anomalies is crucial. Additionally, surveillance systems can optimize processes and monitor equipment health.

Sports analytics benefits from video segmentation and labeling by enabling the tracking and analysis of player movements during games. Coaches and analysts use this data to gain insights into player performance, strategize, and enhance training regimens.

Video segmentation and labeling are used in medical imaging analysis. Videos from diagnostic tools, such as MRIs or CT scans, can be segmented to identify and track specific anatomical structures or abnormalities. This aids medical professionals in diagnosis and treatment planning.

Some of the objectives are as follows:

1. **Improved Object Recognition:**

   Video segmentation and labeling aim to improve object recognition in dynamic environments. By delineating objects or entities in a video and attaching relevant labels, the system can accurately recognize and track these elements, contributing to better decision-making in various applications.

2. **Real-time Processing for Dynamic Environments:**

   One of the key objectives is to enable real-time processing of video data in dynamic environments. This is particularly crucial in applications like autonomous systems and surveillance, where rapid decision-making is essential for safety and security.

3. **Automation for Reduced Manual Effort:**

   Automation is a key objective, aiming to reduce the manual effort required for annotating and labeling videos. By automating the segmentation and labeling processes, these systems enhance efficiency, accelerate data analysis, and make the technology more accessible.

4. **Scalabilty and Adaptability :**

   These systems should be capable of handling diverse video datasets and adapting to different scenarios, ensuring effectiveness and relevance across a wide range of applications.

5. **Semantic Understanding:**

   Beyond identifying objects, the system aims to comprehend the context and relationships between different objects in the video, contributing to more significant interpretations.

# Chapter 2

# Literature Review

The literature review provides a thorough overview of current techniques and challenges in automatically segmenting and labeling objects in videos. The review highlights trends like incorporating temporal information and attention mechanisms. It also covers advancements in deep learning and motion tracking while emphasizing the importance of diverse datasets.

1. **Video object segmentation and tracking: A survey** [1]

   The paper addresses challenges like fast motion and real-time processing in videos by combining segmentation and tracking. Challenges remain, such as issues with low-resolution videos and motion blur, affecting accuracy and flexibility in handling different object shapes.

2. **Deep Learning for Semantic Segmentation of Unmanned Aerial Vehicle Videos** [2]

   The paper proposes a model that combines Fully Convolutional Networks (FCN) and Long Short-Term Memory (LSTM) for segmentation. FCN handles each video frame on its own, while LSTM refines the results using temporal information from consecutive frames. However, accuracy is affected by noise in the frames, and performance drops with resized images.

3. **Video instance segmentation.** [3]

   The paper introduces YouTube-VIS, a big dataset for video instance segmentation, and suggested MaskTrack R-CNN for the job. However, it faces challenges when connecting objects, especially with things like overlapping and fast motion.

4. **Video Object Segmentation by Latent Outcome Regression** [4]

   The paper suggestes an unsupervised method where weights and outcomes are optimized together through iterations. The weight learning and segmentation inference work together to enhance quality, adjusting based on specific characteristics. However, a drawback is the extra time taken by the aggregation algorithm, making it slower, and less consistent.

5. **Semi-Supervised Video Object Segmentation Based on Local and Global Consistency Learning** [5]

   The paper utilizes more unlabeled frames to enhance robustness and generalization, considering both local and global video information. Achieved reduced complexity and memory usage, resulting in excellent segmentation and high prediction speed. However, the model's accuracy was somewhat insufficient (around 68%), primarily tested on a limited number of samples.

6. **Unsupervised Video Object Segmentation via Weak User Interaction and Temporal Modulation** [6]

   The paper incorporates a basic rectangle drawn around a person in the first frame to guide segmentation. Employs ETM and CTM modules for temporal information, boosting segmentation accuracy. However, struggles with tiny component contours in fast sequences and has limited learning ability for background areas.

7. **Deep Learning for Object Detection and Segmentation in Videos: Toward an Integration With Domain Knowledge** [7]

   The paper explores the differences between a two-stage and one-stage approach in CNN-based image object detectors, employing methods like optical flow, tracking, LSTM, GRU, self-attention mechanisms, and generative learning. Identified gaps in terms of data scarcity, generalizability, indistinguishable outputs, and a lack of reasoning in the results.

8. **A Reinforcement Learning Based Adaptive ROI Generation for Video Object Segmentation** [8]

   The paper explores ZVOS (Zero-Shot Video Object Segmentation), a unified RL framework using the Deep Deterministic Policy Gradient (DDPG) algorithm and a group co-attention mechanism. Identified a challenge in accurately distinguishing main objects from intricate backgrounds in the absence of prior object information.

9. **Spatio-Temporal Self-Attention Network for Fire Detection and Segmentation in Video Surveillance** [9]

   The paper develops a new method for fire detection in two stages, incorporating a spatial-temporal network. Applied self-attention to discriminative Spatio-Temporal features for improved segmentation masks. Created a video dataset with manually generated ground-truth segmentation masks. Challenges include the arbitrary shapes and sizes of fires, making learning more difficult, and the absence of large datasets with fire and ground-truth segmentation masks.

10. **Adaptive Template and Transition Map for Real-Time Video Object Segmentation** [10]

    This paper creates a lightweight semi-VOS model using two template matching methods: short-term for localization and long-term for fine mask generation. Introduced a transition map for an auxiliary loss to correct mis-estimated pixels from

previous frames, preventing error propagation. Identified a challenge when the target object disappears due to occlusion, causing performance degradation.

11. **The 2019 DAVIS Challenge on VOS: Unsupervised Multi-Object Segmentation** [11]

    The paper announces the third edition of the DAVIS Challenge series, introducing a new unsupervised multi-object track for video object segmentation. To support the unsupervised track, they've re-annotated existing sets and added new ones. The approach involves suggesting object proposals on each image without human supervision, prioritizing object semantics over motion patterns. Challenges include the demand for more accurate methods, improved evaluation metrics, and recognizing limitations in their annotation approach. The authors express the need for more diverse datasets to advance the field.

12. **Self-Supervised Deep TripleNet for Video Object Segmentation** [12]

    The paper introduces a self-supervised deep TripleNet model for video object segmentation, capable of learning from unlabeled video data. The model comprises two modules: the temporal motion module captures motion patterns between frames, and the appearance matching module generates segmentation masks based on the reference frame and its corresponding mask. This self-supervised learning approach eliminates the need for pixel-level annotations, making it more efficient than traditional methods. However, challenges include potential performance issues in complex backgrounds or varying lighting conditions and a limitation in generalizing to new datasets. The paper doesn't delve into the computational complexity, posing a consideration for real-time applications.

13. **Asymmetric Bilateral Motion Estimation for Video Frame Interpolation** [13]

    The paper introduces a new method called Asymmetric Bilateral Motion Estimation (ABME) to improve traditional video frame interpolation. ABME refines motion vectors for better accuracy, especially in dealing with challenges like occlusions and non-linear object motions. The paper aims to overcome limitations in traditional

methods related to symmetric bilateral motion vectors, which struggle with occlusions and non-linear motions. However, challenges with ABME might include the complexity of refining motion vectors and how well it handles complex motion patterns and real-world occlusions. The paper may not extensively discuss the practical limitations or failure cases of ABME, which could be crucial for understanding its usefulness.

14. **Video Frame Interpolation Based on Symmetric and Asymmetric Motions** [14]

    They propose a novel video frame interpolation network that incorporates both symmetric and asymmetric motion-based warping modules, addressing both linear and non-linear motions, as well as occlusions effectively. The symmetric warping module estimates symmetric motions to generate intermediate frames, while the asymmetric one predicts asymmetric motions to handle non-linear motions and occlusion problems. By combining the results from both modules, they achieve a more reliable reconstruction of intermediate frames. Additionally, they introduce a frame synthesis network to refine the combined warping results. Experimental results demonstrate that their proposed network outperforms state-of-the-art video interpolation algorithms, showcasing the effective complementary operation of the two types of warping modules across various benchmark datasets.

## 2.1 Survey Summary

These papers explore ways to understand and process videos. [1] combines segmentation and tracking but struggles with low-quality videos. [2] on UAV videos uses FCN and LSTM but has accuracy issues with noisy frames. [3] deals with video instance segmentation, facing challenges with connecting objects during fast motion.

[4] suggests an unsupervised method for video object segmentation but is slower and less consistent. [5] works with semi-supervised video object segmentation, achieving good results but not perfect accuracy.

[6] introduces unsupervised video object segmentation, making it more accurate but facing challenges in fast sequences. [7] compares two approaches in deep learning for object detection, pointing out gaps in data and generalization.

[8] explores ZVOS with reinforcement learning, dealing with challenges in distinguishing objects from complex backgrounds. [9] introduces a network for fire detection, handling challenges related to different fire shapes. [10] tackles real-time video object segmentation, considering issues in occlusion scenarios.

[11] announces a challenge for video segmentation, highlighting the need for accurate methods. [12] introduces a self-supervised model for video object segmentation, removing the need for detailed annotations.

[13] and [14] focus on improving motion estimation and video frame interpolation, handling challenges like occlusions. However, more exploration is needed for practical limitations.

In summary, these papers contribute to making sense of videos, but each approach has its own challenges, showing that there's more work to do in this area. All these applications work in real-time, which is difficult to implement. Also, the model detection is not producing accurate results thus affecting the performance. Our work will be focused on resolving all these issues.

# Chapter 3

# Proposed Methodology

## 3.1 Introduction


Figure 3.1: High-Level Overview of the System

Figure 3.1 shows a high-level overview of the process of automatic segmentation and labeling of objects in videos. The steps involved are:

**Step 1: Video Input**

The first step is to input the video sequence into the system. This can be done by loading the video file from a disk or by streaming the video from a live feed.

**Step 2: Frame Identification**

Once the video input has been loaded, the system needs to identify the individual frames in the video sequence. This can be done by extracting each frame from the video file or

by using a more sophisticated motion detection algorithm to identify the frames where there is significant motion.

**Step 3: Object Segmentation and Labelling**

After the individual frames in the video have been identified, the system needs to segment the objects in each frame. This can be done using a variety of computer vision algorithms, such as Optical flow, Background subtraction, and Deep learning algorithms.

Then the system needs to label the objects. This can be done by using a variety of methods, such as Rule-based labeling, Machine learning, Human-in-the-loop labeling, etc.

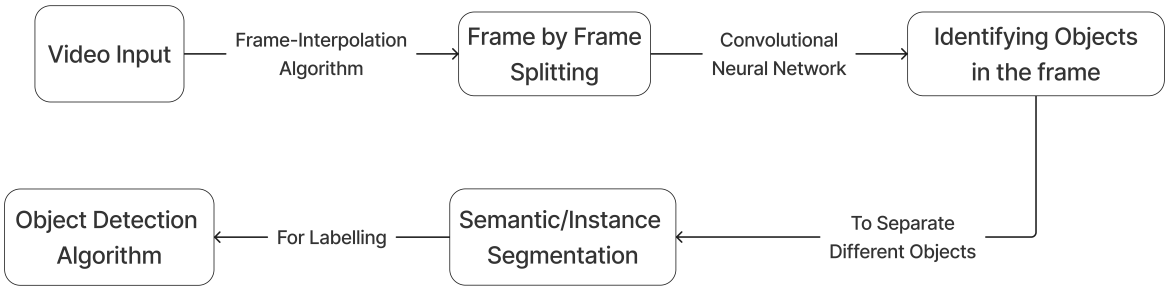



Figure 3.2: Block Diagram of the Proposed System

1. **Video Input:**

   The Video Input block serves as the system's entry point, accepting a continuous stream of video data. The video input is the raw material that undergoes subsequent processing stages to extract meaningful information and insights.

2. **Frame Interpolation Algorithm:**

   The Frame Interpolation Algorithm plays a crucial role in enhancing the temporal resolution of the video. By generating additional frames between existing ones, this algorithm smoothens motion transitions, resulting in a visually improved video. Techniques like optical flow analysis and interpolation methods are often employed to achieve this, contributing to a more seamless viewing experience.

3. **Frame by Frame Splitting**

   In the "Frame by Frame Splitting" stage, the continuous video stream is segmented into individual frames. Each frame is treated as an independent image, enabling detailed analysis and processing. This step sets the foundation for subsequent feature extraction and object recognition processes, as it allows the system to focus on individual frames for precise examination.

4. **CNN:**

   The CNN involves the application of Convolutional Neural Networks, a type of deep learning model designed for image-related tasks. In this context, the CNN extracts high-level features from each frame. These features may include edges, textures, and patterns that are crucial for understanding the content of the video. The CNN serves as a powerful tool for object recognition and scene understanding.

5. **Identifying Objects in Frame:**

   The "Identifying Objects in Frame" block employs an object detection algorithm to recognize and locate objects within each frame. This algorithm leverages the features extracted by the CNN to make predictions about the presence and position of objects. Common techniques for object detection include Region-Based CNNs (R-CNN), You Only Look Once (YOLO), and Single Shot MultiBox Detector (SSD).In this method, YOLO V8 was used.

6. **Separate Different Objects:**

   Following object identification, this stage involves a process to isolate and distinguish individual objects from one another within a frame. This separation is crucial for tracking and analyzing each object independently throughout the video sequence. Techniques like clustering or bounding box analysis may be employed to achieve this separation.

7. **Semantic/Instance Segmentation:**

   This block focuses on understanding the content of each frame at a pixel level. Semantic segmentation classifies each pixel into a specific class (e.g., person, car, background), while instance segmentation goes further by distinguishing between individual instances of the same class. This detailed segmentation provides a richer representation of the scene.

8. **Labeling:**

   In this stage, labels are assigned to segmented regions, providing a descriptive identification of the content. These labels convey information about the category or class of each object or region within the frame. The labeling step enhances the interpretability of the video data and facilitates subsequent analysis.

9. **Object Detection Algorithm:**

   The final "Object Detection Algorithm" block signifies another instance of object detection, possibly applied after segmentation and labeling. This step may involve a more refined analysis to detect objects that were not initially identified or to improve the accuracy of the previous detection results. It contributes to a comprehensive understanding of the video content, ensuring that no relevant objects are overlooked.

In summary, this block diagram outlines a sophisticated video processing pipeline, highlighting the significance of each stage in extracting valuable information and insights from the input video stream. The combination of frame interpolation, deep learning-based feature extraction, object detection, segmentation, labeling, and refined object detection results in a holistic approach to video analysis.

## 3.2 Data Collection

The method of Data Collection by harnessing camera-captured images and videos for training, rather than relying solely on large pre-annotated datasets, introduces a paradigm shift in the efficiency of segmentation and labeling processes. By leveraging real-world visual data captured by cameras, we not only tap into the richness and diversity of dynamic environments but also minimize the need for extensive manual annotation efforts. Utilizing these unlabeled or minimally labeled datasets allows for the development of more robust and adaptable models. Self-supervised learning techniques, such as leveraging temporal coherence and spatial context within videos, enable the algorithm to learn intrinsic patterns and relationships autonomously. This approach not only streamlines the training pipeline but also ensures that the model gains a nuanced understanding of the complexities present in real-world scenarios, ultimately enhancing the generalization and real-world applicability of automated segmentation and labeling systems. Due to its lack of scalability, Large datasets can be used which are as follows:

- **YouTube-Objects Dataset:** It comprises videos sourced from YouTube by searching for the names of 10 object categories from the PASCAL VOC Challenge. It includes 9 to 24 videos per class, with each video lasting between 30 seconds and 3 minutes. The videos have weak annotations, ensuring that each one contains at least one object related to its respective class.

- **Freiburg-Berkeley motion segmentation (FBMS):** The dataset is a comprehensive benchmark consisting of 59 sequences. It provides precise ground truth annotations for moving objects at the pixel level.

- **Davis16:** A benchmark dataset for object tracking and segmentation in computer vision. It comprises 50 video sequences with challenging conditions like occlusions and appearance changes. Each frame has pixel-level annotations for the main object.

- **Davis17:** It expands on Davis16, offering 150 video sequences with varied challenges for object tracking and segmentation algorithms. It maintains high-quality annotations for evaluating algorithms in real-world complex scenarios.

  The community that holds DAVIS datasets runs an annual committee competition of Video Object Segmentation(VOS) and Video Instant Segmentation(VIS) in the DAVIS dataset.

- **YouTube-VOS:** A benchmark dataset and challenge designed for advancing the field of video segmentation in computer vision. Introduced by Google Research, YouTube-VOS comprises high-resolution video sequences with pixel-level annotations for object segmentation. This dataset is a vital resource for evaluating and benchmarking the performance of algorithms in the challenging task of segmenting objects across consecutive video frames.

- **A2D Sentence:** A2D, or "Actions in the Datasets," is a dataset specifically curated for action recognition in video sequences. This dataset provides a diverse collection of video clips capturing a wide range of human actions in various everyday scenarios. A2D is instrumental in training and evaluating algorithms designed for action recognition, enabling researchers and practitioners to enhance the performance of computer vision models. The dataset's rich content and labeled annotations contribute significantly to the development of robust systems capable of understanding and interpreting human actions in dynamic video environments.

- **CamVid:** Short for Cambridge-driving Labeled Video Database, is a widely used dataset in the field of computer vision and semantic segmentation. Created by the University of Cambridge, CamVid consists of high-resolution video sequences recorded from the perspective of a driving car. The dataset is annotated, providing pixel-level labels for various semantic classes such as road, building, pedestrian, and car. It is a valuable resource for training and evaluating algorithms in the challenging task of semantic segmentation, particularly in urban driving scenarios.

- **VSPW:** VSPW stands for Video Scene Parsing in the Wild. It is a large-scale dataset for video scene parsing that was introduced in the 2021 paper "VSPW: A Large-scale Dataset for Video Scene Parsing in the Wild" [15]. The VSPW dataset is annotated with pixel-level semantic labels for each frame. The labels are organized into 29 categories, including objects, such as cars, people, and trees; and scene categories, such as roads, sidewalks, and sky.

- **MOTSChallenge:** The Multiple Object Tracking and Segmentation (MOTS) Challenge is an annual competition that evaluates the performance of state-of-the-art algorithms for tracking and segmenting multiple objects in videos. The challenge is organized by the Multi-Object Tracking Benchmarking Taskforce (MOTChallenge) and is held in conjunction with the Conference on Computer Vision and Pattern Recognition (CVPR).

- **BDD100K:** BDD100K is a large-scale driving video dataset that was released in 2018. It is the largest and most diverse open-driving video dataset to date, with over 100,000 videos and 10 tasks. The dataset was collected by Nexar, a company that develops dashcams for cars. The videos were collected in a variety of cities around the world, including San Francisco, Los Angeles, and New York City.

See Table 3.1 for short summary of the datasets explained above.

Overall, the process of automatic segmentation and labeling of objects in video represents a transformative leap in the realm of computer vision, offering unparalleled efficiency and scalability. The collection of data for this purpose is a cornerstone in training robust machine learning models. By automating the segmentation and labeling tasks, we alleviate the burden of manual annotation, which is not only time-consuming but also prone to human errors. The datasets curated for automatic segmentation and labeling serve as invaluable resources, capturing the diversity and complexity of real-world scenarios.

| Dataset | Year | Description |
|---|---|---|
| CamVid | 2009 | 4 videos with pixel-level annotations for 32 object classes. |
| YouTube-Objects | 2012 | 1,407 videos with object-level annotations for 10 object classes. |
| FBMS | 2014 | 59 videos with object-level annotations for 10 object classes. |
| DAVIS16 | 2016 | 50 videos with object-level annotations for 30 object classes. |
| DAVIS17 | 2017 | 150 videos with instance-level annotations for 30 object classes. |
| YouTube-VOS | 2018 | 4,519 videos with pixel-level annotations for 65 object classes. |
| A2D Sentence | 2018 | 3,782 videos with annotations for referring to objects in natural language. |
| MOTSChallenge | 2019 | 4 sequences of images from surveillance cameras with annotations for 2D tracking of vehicles and pedestrians. |
| BDD100K | 2020 | 1 Lakh driving scenes images with pixel-level annotations for 8 object classes. |
| VSPW | 2021 | 3,536 videos with pixel-level annotations for 8 object classes. |

Table 3.1: Summary of Datasets with Annotations

The significance of data collection lies in shaping algorithms for autonomous systems, surveillance, and various applications. Labeled datasets enable accurate interpretation of visual information, aiding machines in navigating and interacting seamlessly. Moreover, datasets for segmentation and labeling drive advancements in real-time applications. Swift object identification for autonomous vehicles and anomaly detection in surveillance videos enhance system responsiveness. Collected data forms the foundation for intelligent, adaptive systems in complex environments.

## 3.3 Work Done So Far

This chapter deals with the results and inferences derived from the work done for phase 1 of this project. It includes the procedure and result of automatic image segmentation and labeling (via detection).

Before the main work of automatic video segmentation and labeling, work was done on automatic image segmentation and labeling using the Segment Anything (SAM) model for Image Segmentation and You Only Look Once for Object Detection. Figure 4.1 is the input image that is fed into the combination of Object Segmentation Detection Models. Figure 4.2 is the output after segmentation and detection.


Figure 3.3: Input for Automatic Image Segmentation and Labeling

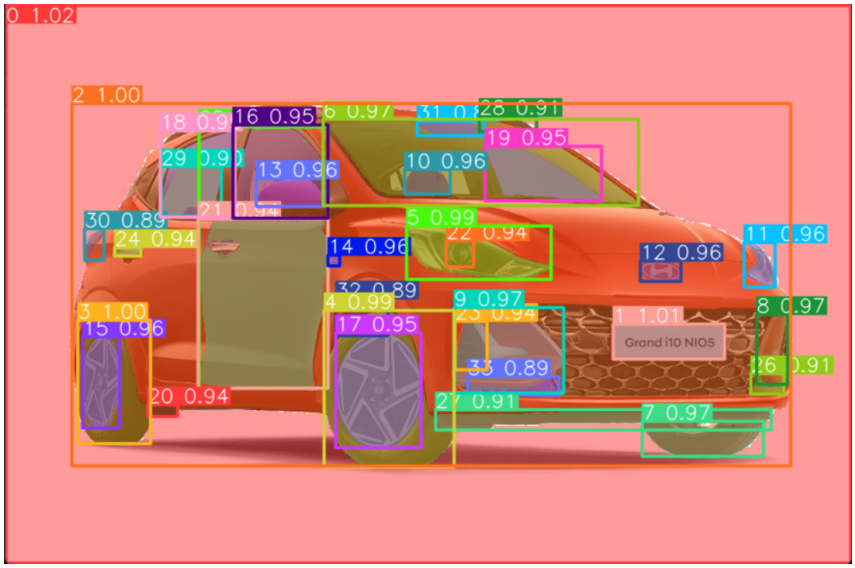



Figure 3.4: Output from Automatic Image Segmentation and Labeling

When an image is inputted to a segmentation and object detection model together, the segmentation model focuses on dividing the image into distinct segments or regions based on pixel-level classification. This means it assigns a label to each pixel, outlining the boundaries and identifying areas belonging to different objects or categories within the image. On the other hand, the object detection model detects and localizes specific objects within the image. It identifies bounding boxes around these objects and assigns them labels, indicating the class or category of each object detected.

The results from Figure 4.2 can be divided into two parts, segmentation and object labeling. The segmentation has worked well as the SAM Model is trained on tens of thousands of wide variety data, but the detection using YOLOv8 is not that accurate as the model is trained on some general classes like person, bicycle, car, bus, etc... but not specific parts of a car( as per figure 4.1) like door, mirror, and light.

# Chapter 4

# Conclusion

In conclusion, Image and video segmentation and labeling hold incredible importance across numerous domains due to their multiple roles. These processes are instrumental in object recognition, enabling the identification and detection of specific elements within visual data, a vital function for applications such as autonomous vehicles and medical imaging. The semantic understanding derived from segmentation aids in comprehending context, supporting tasks like satellite imagery analysis and surveillance. Moreover, these techniques provide annotated data crucial for training machine learning models, particularly in computer vision, enhancing the accuracy of algorithms. Segmentation and labeling also underpin augmented and virtual reality applications by enabling the seamless integration of digital information into real-world environments. In video analysis, these processes facilitate object tracking, motion analysis, and event recognition, which are pivotal in fields like surveillance and sports analysis. The initial phase of this project focused on real-time image segmentation and labeling, employing the SAM model for segmentation and YOLO for object detection. Which, on the segmentation process worked effectively, but object detection needs more training in more classes. Future enhancements could involve refining the object detection model to encompass finer-grained categorizations for improved accuracy and specificity in labeling object components within images and videos.

# References

[1] R. Yao, G. Lin, S. Xia, J. Zhao, and Y. Zhou, "Video object segmentation and tracking: A survey," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 4, may 2020. [Online]. Available: https://doi.org/10.1145/3391743

[2] Y. Wang, Y. Lyu, Y. Cao, and M. Y. Yang, "Deep learning for semantic segmentation of uav videos," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019, pp. 2459–2462.

[3] L. Yang, Y. Fan, and N. Xu, "Video instance segmentation," 2019.

[4] L. Zhang and Y. Lu, "Video object segmentation by latent outcome regression," *IEEE Access*, vol. 8, pp. 30 355–30 367, 2020.

[5] H. Liang, L. Liu, Y. Bo, and C. Zuo, "Semi-supervised video object segmentation based on local and global consistency learning," *IEEE Access*, vol. 9, pp. 127 293–127 304, 2021.

[6] J. Fan, K. Zhang, Y. Zhao, and Q. Liu, "Unsupervised video object segmentation via weak user interaction and temporal modulation," *Chinese Journal of Electronics*, vol. 32, no. 3, pp. 507–518, 2023.

[7] A. Ilioudi, A. Dabiri, B. J. Wolf, and B. De Schutter, "Deep learning for object detection and segmentation in videos: Toward an integration with domain knowledge," *IEEE Access*, vol. 10, pp. 34 562–34 576, 2022.

[8] U. A. Usmani, J. Watada, J. Jaafar, I. A. Aziz, and A. Roy, "A reinforcement learning based adaptive roi generation for video object segmentation," *IEEE Access*, vol. 9, pp. 161 959–161 977, 2021.

[9] M. Shahid, J. J. Virtusio, Y.-H. Wu, Y.-Y. Chen, M. Tanveer, K. Muhammad, and K.-L. Hua, "Spatio-temporal self-attention network for fire detection and segmentation in video surveillance," *IEEE Access*, vol. 10, pp. 1259–1275, 2022.

[10] H. Park, J. Yoo, G. Venkatesh, and N. Kwak, "Adaptive template and transition map for real-time video object segmentation," *IEEE Access*, vol. 9, pp. 116 914–116 926, 2021.

[11] S. Caelles, J. Pont-Tuset, F. Perazzi, A. Montes, K.-K. Maninis, and L. V. Gool, "The 2019 davis challenge on vos: Unsupervised multi-object segmentation," 2019.

[12] K. Xu, L. Wen, G. Li, and Q. Huang, "Self-supervised deep triplenet for video object segmentation," *IEEE Transactions on Multimedia*, vol. 23, pp. 3530–3539, 2021.

[13] J. Park, C. Lee, and C.-S. Kim, "Asymmetric bilateral motion estimation for video frame interpolation," 2021.

[14] W. Choi, Y. J. Koh, and C.-S. Kim, "Video frame interpolation based on symmetric and asymmetric motions," *IEEE Access*, vol. 11, pp. 22 394–22 403, 2023.

[15] J. Miao, Y. Wei, Y. Wu, C. Liang, G. Li, and Y. Yang, "Vspw: A large-scale dataset for video scene parsing in the wild," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 4131–4141.

# VISION & MISSION OF THE DEPARTMENT

## Vision

To achieve excellence in Artificial Intelligence and Data Science to cater to the ever-changing industrial and socio-economic needs.

## Mission

- To provide high-quality and value-based technical education in the Artificial Intelligence and Data Science program.
- To establish an infrastructure fostering industry-institute interaction in order to meet global expectations and requirements.
- To empower students to become globally competent and effective problem-solvers to develop entrepreneurial skills and higher studies.