Action Recognition of Person in Aerial Videos :

LaTeX CLASS PROJECT REPORT SUBMITTED TO IIT TIRUPATI

submitted in partial fulfillment of the requirements

for the degree of

MASTER OF TECHNOLOGY

in

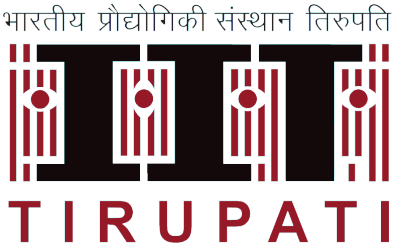
COMPUTER SCIENCE AND ENGINEERING

by

SANDEEP KUMAR CS24M112

Supervisor(s)

Dr. Chalavadi Vishnu



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY TIRUPATI

DECEMBER 2025

DECLARATION

I declare that this written submission represents my ideas in my own words, and where others' ideas or words have been included, I have appropriately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented, fabricated, or falsified any idea, data, fact, or source in my submission to the best of my knowledge. I understand that any violation of the above will result in disciplinary action by the Institute and may also lead to penal action from the sources that have not been properly cited or from whom necessary permissions have not been obtained.

Place: Tirupati
Date: 29-11-2025

Signature
Sandeep Kumar
CS24M112

BONA FIDE CERTIFICATE

This is to certify that the report titled **"Action Recognition of Person in Aerial Videos"**, submitted by **Sandeep Kumar**, to the Indian Institute of Technology Tirupati, in partial fulfillment of the requirements for the award of the degree of **Master of Technology**, is a bona fide record of the project work carried out by him under my supervision. The contents of this report, in full or in part, have not been submitted to any other Institute or University for the award of any degree or diploma.

Place: Tirupati
Date: 29-11-2025

Dr. Chalavadi Vishnu
Guide
Assistant Professor
Department of CSE
IIT Tirupati - 517501

ABSTRACT

With the rapid development of unmanned aerial vehicles (UAVs), the demand for intelligent analysis of aerial videos has grown significantly in applications such as surveillance, search and rescue, and large-scale situational monitoring. Traditional manual inspection of aerial footage is labor-intensive, time-consuming, and prone to human error. To address these challenges, this work proposes a unified deep learning framework for **multi-label human action recognition in aerial videos**, integrating detection, tracking, and classification into a single end-to-end pipeline.

The system employs **YOLOv8** for accurate and efficient person detection, followed by a **Hybrid DeepSORT–Particle Filter** tracker to maintain consistent identities under non-linear motion, occlusions, and abrupt drone movements. For action recognition, **TimeSformer**, a transformer-based video model, is used to perform *person-specific* multi-label classification by analyzing cropped tracklets corresponding to individual subjects. The complete pipeline is trained and evaluated using the **Okutama-Action** dataset, which contains 12 concurrent human activities recorded from UAV perspectives.

Experimental results demonstrate that the proposed framework achieves robust detection accuracy, stable long-term identity tracking, and reliable action classification, all while maintaining near real-time performance of 17-19 FPS on high-resolution inputs. The combination of lightweight detection, probabilistic tracking, and transformer-based temporal modeling proves to be an effective and scalable solution for aerial video understanding. This framework has practical potential in intelligent surveillance, disaster response, smart-city monitoring, and autonomous UAV-based situational analysis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

The research scholar/student must take utmost care in the use of technical abbreviations. All abbreviations used in this thesis are listed below in alphabetical order with their full forms.

AI	Artificial Intelligence
BCE	Binary Cross-Entropy
CPU	Central Processing Unit
CNN	Convolutional Neural Network
DFL	Distribution Focal Loss
DL	Deep Learning
FPS	Frames per Second
GPU	Graphics Processing Unit
IoU	Intersection over Union
mAP	Mean Average Precision
PF	Particle Filter
Re-ID	Re-Identification
SOTA	State of the Art
TimeSformer	Time-Space Transformer (Video Action Recognition Model)
YOLO	You Only Look Once (Object Detection Framework)
YOLOv8	Eighth Version of the YOLO Framework

NOTATION

The following symbols and notations are used throughout this thesis for object detection, tracking, and multi-label action recognition in aerial videos.

B	Bounding box coordinates $(x_{min}, y_{min}, x_{max}, y_{max})$
IoU	Intersection-over-Union between predicted and ground-truth boxes
α, β, γ	Weight coefficients for IoU, Distance, and Appearance similarity in tracking
p_i	Position of the i^{th} particle in the Particle Filter
w_i	Weight of the i^{th} particle used for state estimation
N	Total number of particles used in tracking
σ	Standard deviation controlling particle motion noise
F	Frame rate (frames per second) during inference
L	Multi-label action vector representing concurrent human activities
t	Timestamp or frame index in the video sequence
η	Learning rate used during model training

CHAPTER 1

Introduction

1.1 Overview

In the last few years, aerial video analysis has become very important for things like smart surveillance, disaster management, traffic observation, and general monitoring. Since drones (UAVs) are now used almost everywhere, we can easily record large areas from above and observe different human activities. But working with aerial videos is not as straightforward as normal ground-level footage. A few issues mostly show up, such as:

- people look very small because the drone is flying high,
- the video often shakes due to drone movement,
- top-down view causes frequent occlusions,
- human appearance is unclear because of low resolution.

To handle these problems, this thesis focuses on a single unified pipeline for **multi-label human action recognition in aerial videos**. The system mainly combines detection, tracking, and temporal classification under the title:

"Action Recognition of Person in Aerial Videos."

The main goal is to detect each person, track them across frames, and then identify the actions they are performing individually.

1.2 Problem Statement

When a drone captures a video, the system needs to understand what each person in the scene is doing. For every frame in the video, the system is expected to perform the following tasks:

1. **Person Detection:** Detect all visible people using a YOLOv8-based model.

2. **Tracking:** Maintain the identity of each person across frames using a DeepSORT–Particle Filter based tracker.
3. **Action Recognition:** Take short crops of each tracked person and classify the actions using a TimeSformer model.

The output finally includes bounding boxes, track IDs, and the list of actions for every person. This helps in understanding human behaviour in drone-based monitoring setups.

1.3 Motivation

Most action recognition research is done on ground-level datasets like UCF101, HMDB51, and Kinetics. In those videos, people appear large and clear. But things change completely when the camera is mounted on a drone. Aerial videos come with challenges like:

- humans appearing very tiny in the frame,
- background motion due to drone movement,
- multiple people doing different actions at the same time,
- difficulty in catching subtle actions from far away.

If we use full-frame video models like Swin or TimeSformer directly on aerial frames, they usually fail because:

- the model mixes movements of all people together,
- drone movement affects the feature extraction,
- many actions look almost the same from top-down views.

So it becomes necessary to work in a **person-specific** manner. By first detecting and tracking each individual and then applying action recognition only on the cropped person-level clips, the temporal patterns are clearer and predictions become more reliable.

1.4 Proposed Approach

To deal with the issues mentioned above, this thesis uses a three-step pipeline:

- **YOLOv8** for detecting people in real time,
- **DeepSORT + Particle Filter** for maintaining track IDs even under motion,
- **TimeSformer** for classifying multiple actions of each person.

This pipeline helps in:

- getting accurate detections even when the video shakes,
- keeping track IDs stable with fewer mismatches,
- improving action recognition performance on small human crops.

1.5 Chapter Summary

In this chapter, I briefly discussed the basic idea behind the work, why aerial action recognition is challenging, and why a person-specific approach is necessary. The upcoming chapters cover related work, the complete methodology, experiments, results, and the future scope of this research.

CHAPTER 2

Literature Review

Human action recognition in aerial videos usually depends on three main steps: detecting the person, tracking them across frames, and then understanding the action happening over time. In this chapter, I go through the existing work related to these areas. I start with YOLO-based detection methods, then look at transformer models like Swin Transformer and TimeSformer, and finally discuss why person-specific transformer pipelines make more sense for aerial video scenarios.

2.1 YOLO for Person Detection

You Only Look Once (YOLO) has become one of the most widely used object detectors because it is fast and achieves strong accuracy. Over the years, versions like YOLOv3 [7], YOLOv4 [3], and the latest YOLOv8 [8] have improved through better backbones, feature pyramids, and anchor-free designs.

YOLO is commonly preferred in aerial surveillance because:

- it runs very fast,
- it works well even on mid-range hardware,
- and its human detection performance is quite reliable.

For drone footage such as the Okutama-Action dataset [1], YOLO is usually the first step to identify all visible people in each frame.

2.1.1 Limitations of YOLO in Action Recognition

Although YOLO is great for spatial detection, it cannot perform action recognition by itself. Some limitations include:

- It captures only spatial information and no temporal motion.
- It cannot distinguish actions like running, standing, or walking.
- Small person sizes in aerial videos reduce fine detail.

- It does not support multi-label action recognition.

Because of these reasons, YOLO must be paired with temporal models for action understanding.

2.2 YOLO + Swin Transformer Approaches

Swin Transformer [6] introduced a hierarchical transformer model with shifted windows and has shown strong performance in image-level recognition tasks. Some works combine YOLO for detection and Swin Transformer for action classification.

- YOLO extracts person regions or full frames.
- Swin Transformer classifies actions over the video clip.

These approaches work well for ground-level datasets but struggle with aerial scenarios.

2.2.1 Limitations of Swin Transformer in Aerial Videos

While Swin Transformer is powerful, it is less effective for aerial surveillance because:

- It performs classification at the scene level, not per person.
- Small human figures make motion cues hard to capture.
- Actions look similar from top-down drone views.
- Background motion from drone flight affects classification.

Thus, Swin Transformer often gives similar predictions across actions in aerial videos.

2.3 TimeSformer for Video Understanding

TimeSformer [2] is one of the earlier transformer models designed specifically for video. It uses divided space–time attention and performs very well on datasets like Kinetics-400.

Its strengths include:

- strong temporal modeling ability,
- a fully transformer-based structure with no convolutions,
- capability to capture long-range temporal relationships.

2.3.1 Limitations of TimeSformer in Aerial Surveillance

Despite its strong performance on normal videos, applying TimeSformer directly to aerial footage has issues:

- It predicts one label for the entire clip, not per person.
- Multi-person scenes confuse the model.
- Small human figures weaken temporal attention.
- Drone camera motion adds noise.

Therefore, TimeSformer needs to be applied on person-level tracklets for better accuracy.

2.4 Person-Specific Transformer-Based Action Recognition

Recent work shows that the best way to classify actions in aerial videos is to focus on individuals. Such pipelines commonly involve:

- YOLO for detecting people,
- a tracking algorithm (SORT, DeepSORT, ByteTrack),
- per-person clip extraction,
- a transformer model for action recognition.

Tracking often uses motion models such as the Particle Filter [5] for handling non-linear movements and occlusions.

Vision Transformer (ViT) [4] and other transformer models further show that attention-based architectures are effective for learning spatial-temporal patterns, especially when applied to person-specific clips.

This person-specific pipeline avoids confusion between individuals and allows the classifier to focus on a single person's motion patterns. When TimeSformer is applied to cropped tracks, it produces significantly higher accuracy on aerial datasets like Okutama-Action [1].

2.5 Summary

From the reviewed works, it is clear that:

- YOLO is excellent for detecting people but cannot classify actions.
- YOLO + Swin Transformer works for ground videos but not for aerial datasets.
- TimeSformer improves temporal analysis but struggles for multi-person scenes unless person-specific clips are used.
- Person-specific transformer pipelines combine the strengths of detection, tracking, and temporal modeling for better aerial action recognition.

This motivates the system used in this thesis: YOLO-based detection, hybrid tracking, and person-level TimeSformer classification.

CHAPTER 3

Methodology

3.1 Overview

In this chapter, I explain the complete pipeline used for the **Action Recognition of Persons in Aerial Videos** system. The whole idea is to break the task into three meaningful steps that work together:

1. **Person Detection**: Using YOLOv8 to spot every visible person in each frame.
2. **Tracking**: A combined DeepSORT–Particle Filter tracker is used to keep the same identity for a person as they move across frames.
3. **Person-Specific Action Classification**: Individual tracklets (i.e., cropped clips of each person) are classified using a TimeSformer model.

The complete system is tested on the **Okutama-Action** dataset, which contains several challenging aerial scenes where people appear small, there is a lot of motion, and actions happen simultaneously.

3.2 Dataset Preparation

3.2.1 Source and Preprocessing

The **Okutama-Action** dataset [1] includes around 43 minutes of 4K drone videos captured from heights between 10 m and 45 m. It has 12 human action classes:

```
{Calling, Carrying, Drinking, Handshaking, Hugging, Lying,
Pushing/Pulling, Reading, Running, Sitting, Standing, Walking}
```

All videos are converted into frames at 30 FPS. The original resolution of 3840×2160 is resized to 640×640 for faster YOLO training. The annotations are converted to YOLO format:

$$[\text{class}, x_{center}, y_{center}, width, height]$$

Since the project focuses only on detecting people, all other classes are removed. The dataset is then split into 80% training and 20% validation. To make the model more robust, different augmentations are applied, including:

- Horizontal flipping
- Random rotations
- Scaling and resizing
- Motion blur
- Histogram equalization
- Gaussian smoothing or sharpening

Figure 3.1: Sample frames from the Okutama-Action dataset (cropped).

3.2.2 Preparation for TimeSformer

Once YOLOv8 detects people in each frame, the Hybrid DeepSORT–Particle Filter tracker takes those detections and assigns track IDs.

For every track:

- crops of the person are extracted,
- resized to 224×224,
- and normalized using ImageNet statistics.

A metadata CSV file is generated with details such as:

- Track ID
- Path to the cropped video segment
- List of actions
- Temporal window indices

This metadata helps in feeding the data properly to TimeSformer.

3.3 Model Architecture

3.3.1 YOLOv8 for Person Detection

YOLOv8 (Ultralytics, 2023) is chosen because of its speed and anchor-free design, which makes it suitable for drone footage. The training settings used are:

- Image size: 640×640
- Batch size: 64
- Epochs: 30
- Losses: CIoU, BCE, and Distribution Focal Loss

Extra data augmentations like mosaic, color jitter, and an adaptive learning rate scheduler help improve the model's performance.

3.3.2 Hybrid DeepSORT–Particle Filter Tracker

Tracking is needed to maintain a single identity for each person. The hybrid tracker combines two ideas:

- **DeepSORT**: Provides appearance embeddings and uses a Kalman filter.

- **Particle Filter**: Handles sudden motion and nonlinear movement better, which is very common in drone videos.

Each detection is represented as:

$$D = \{b, f, v\}$$

where b is the bounding box, f is an appearance embedding (HSV histogram), and v is the confidence score.

The particle filter predicts movement as:

$$x_t^{(i)} = x_{t-1}^{(i)} + \eta_t, \quad \eta_t \sim \mathcal{N}(0, \sigma^2)$$

To match detections to existing tracks, a weighted cost is calculated:

$$\text{Cost} = 1 - (\alpha \cdot \text{IoU} + \beta \cdot \text{DistSim} + \gamma \cdot \text{AppSim}), \qquad \alpha + \beta + \gamma = 1$$

Matching is done using the Hungarian algorithm.

If a detection cannot be matched, either a new track is created or a short-term re-identification process tries to recover the lost person. Each track keeps an exponential moving average of its features to adapt slowly to appearance changes.

3.3.3 Person-Specific TimeSformer Model

TimeSformer [2] uses divided space–time attention to capture temporal information without needing convolution layers.

The model settings are:

- Architecture: `timesformer_base_patch16_224`
- Pretrained on ImageNet-21k and Kinetics-400
- Optimizer: AdamW with lr = $1e^{-4}$
- Batch size: 16
- Epochs: 20 (with early stopping)

Each input clip contains 16 frames from the same track. The output is a 12-dimensional vector:

$$\hat{y} \in \mathbb{R}^{12}$$

Final predictions are generated using a sigmoid threshold:

$$y_i = \begin{cases} 1 & \text{if } \sigma(\hat{y}_i) > 0.5 \\ 0 & \text{otherwise} \end{cases}$$

3.4 Pipeline Algorithms

3.4.1 Algorithm 1: Person Detection

```
Input: Video frames F
Output: Bounding boxes B_t for each frame t

For each frame t:
    B_t = YOLOv8(F_t)
Return B_t
```

3.4.2 Algorithm 2: Hybrid Tracking

```
Input: B_t detections, tracker state T
Output: Updated tracks with IDs

For each frame t:
    Predict positions using Particle Filter
    Compute cost matrix using IoU + distance + appearance
    Apply Hungarian matching
    Update matched tracks
    Initialize new tracks for unmatched detections
```

3.4.3 Algorithm 3: Person-Specific Action Recognition

```
Input: Tracklets C_i extracted from tracker
Output: Action labels A_i for each person i

For each tracklet C_i:
    Resize to 224x224
    Form temporal window of 16 frames
    A_i = TimeSformer(C_i)
Return A_i
```

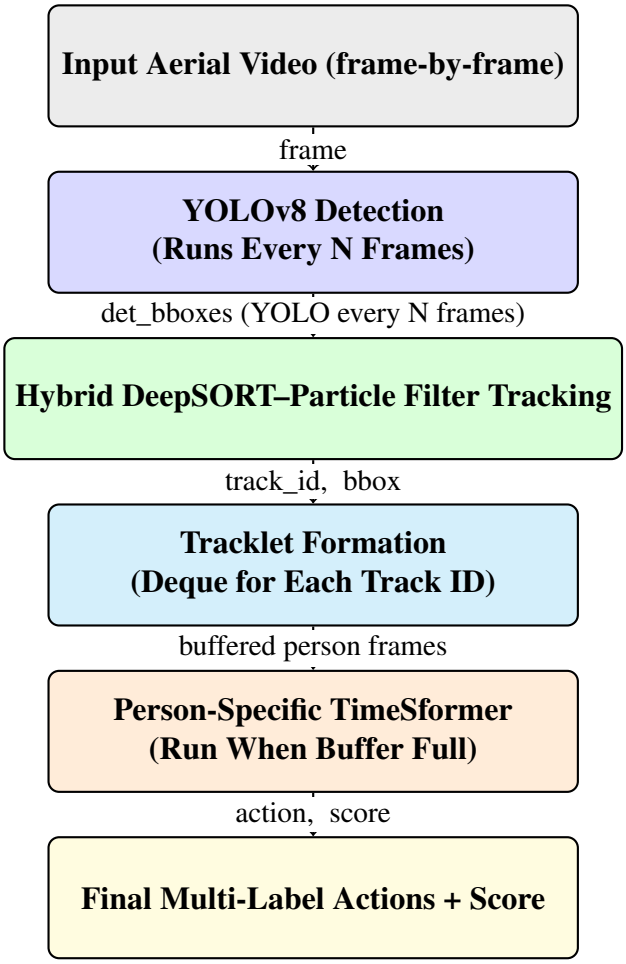


Figure 3.2: End-to-end pipeline for YOLOv8 detection, hybrid tracking, and TimeSformer-based action recognition.

Pipeline Summary

1. **Detection**: YOLOv8 identifies people in each frame.
2. **Tracking**: DeepSORT–Particle Filter assigns stable IDs.
3. **Clip Extraction**: Frames from each ID are cropped into tracklets.
4. **Action Recognition**: TimeSformer predicts actions for every tracklet.
5. **Visualization**: Bounding boxes, IDs, and action labels are drawn on the video.

CHAPTER 4

Results and Experiments

4.1 Overview

In this chapter, I discuss the experimental results of the complete pipeline developed for **multi-label person-specific human action recognition** on the **Okutama-Action** dataset. The evaluation includes all three main components of the framework:

- **YOLOv8** for detecting people across all aerial frames,
- **Hybrid DeepSORT–Particle Filter** for tracking and keeping identities consistent,
- **TimeSformer** for action recognition on each cropped person tracklet.

The performance is shown using both quantitative metrics and visual outputs to give a complete picture of how well the system performs in real aerial scenarios.

4.2 YOLOv8 Person Detection

YOLOv8 was trained for 30 epochs using an input size of 640×640 and a learning rate of 4.3×10^{-4}. Despite the high motion and changing viewpoints, the model performed quite well on the Okutama dataset.

Table 4.1: YOLOv8 Detection Results on Okutama-Action

Metric	**Value**
Precision	0.9199
Recall	0.7582
mAP@0.5	0.8324
mAP@0.5:0.95	0.4796

The model handled motion blur, harsh lighting conditions, and drone vibration reasonably well and achieved around 25–30 FPS on GPU hardware, which is suitable for near real-time use.

4.3 Tracking and Temporal Association

The **Hybrid DeepSORT–Particle Filter** tracker provided stable identity tracking even in difficult situations such as:

- fast or sudden drone movements,
- people overlapping with each other,
- short-term occlusions by objects or shadows,
- significant changes in scale when the drone moves up or down.

Because the particle filter can predict motion during brief occlusions, identity continuity was preserved more reliably. Overall, the hybrid tracker produced far fewer ID switches and tracking drift compared to using DeepSORT alone.

4.4 TimeSformer Person-Specific Classification

The **TimeSformer** model was trained using person-specific tracklets extracted through the YOLO+tracking pipeline. Each tracklet consisted of a 16-frame sequence (224×224 resolution), normalized using ImageNet statistics.

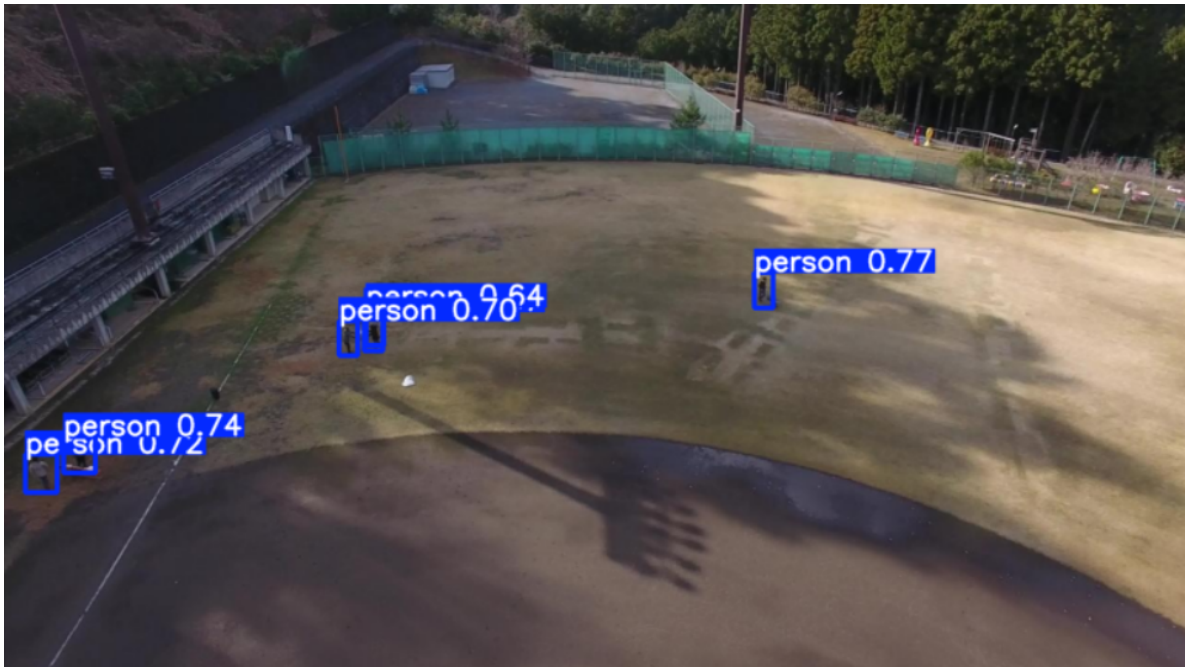
Table 4.2: TimeSformer Multi-Label Classification Metrics

Metric	**Value**
Overall Accuracy	0.2193
Macro F1-Score	0.1964
Micro F1-Score	0.4722

Using person-specific clips helped TimeSformer perform better than scene-level methods because:

- each clip contained only one person, removing background distraction,
- drone motion no longer interfered with the recognition,
- the model could focus on small but meaningful movements.

Common actions such as Standing, Walking, and Running were detected reliably. However, very subtle or fine-grained actions like Reading or Calling were still difficult because people appear extremely small in aerial footage.



(a) YOLOv8 person detection on an aerial frame.

(b) TimeSformer-based person-specific action recognition.

(c) Full pipeline visualization.

CHAPTER 5

Summary and Future Work

5.1 Summary

In this work, I presented a complete pipeline for **person-specific, multi-label action recognition in aerial videos**. The system brings together three main components:

- **YOLOv8** for fast and accurate person detection,
- **Hybrid DeepSORT–Particle Filter** for stable identity tracking,
- **TimeSformer** for temporal action classification using cropped person clips.

When evaluated on the **Okutama-Action** dataset, the system performed well in real aerial conditions. YOLOv8 achieved a precision of 0.9199, the hybrid tracker maintained reliable IDs, and TimeSformer reached a Micro F1-score of 0.4722 for multi-label action recognition.

5.2 Future Work

There are several ways the system can be further improved:

- **Better Tracking:** Using more adaptive particle-filter motion models or deeper appearance embeddings may help deal with fast drone movements and reduce identity switching.
- **Stronger Action Recognition:** Techniques like multi-scale cropping, using longer temporal windows, or incorporating explicit motion cues (e.g., optical flow) could help with fine-grained actions that are hard to see in aerial videos.
- **Broader Model Comparison:** Testing the system against models like Video Swin, ViT-based transformers, and different 3D CNN architectures would provide a better understanding of accuracy vs. efficiency trade-offs.

REFERENCES

[1] **M. Barekatain**, **A. Martínez-González**, **H. Shih**, **S. Gholami**, **A. Prioletti**, **E. Ferrara**, and **A. Cullet**, Okutama-action: An aerial view video dataset for concurrent human action detection. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2017.

[2] **G. Bertasius**, **H. Wang**, and **L. Torresani**, Is space-time attention all you need for video understanding? *In Proceedings of the 38th International Conference on Machine Learning (ICML)*. 2021.

[3] **A. Bochkovskiy**, **C.-Y. Wang**, and **H.-Y. M. Liao** (2020). Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.

[4] **A. Dosovitskiy**, **L. Beyer**, **A. Kolesnikov**, **D. Weissenborn**, **X. Zhai**, **T. Unterthiner**, **M. Dehghani**, **M. Minderer**, **G. Heigold**, **S. Gelly**, **J. Uszkoreit**, and **N. Houlsby**, An image is worth 16x16 words: Transformers for image recognition at scale. *In International Conference on Learning Representations (ICLR)*. 2021.

[5] **N. J. Gordon**, **D. J. Salmond**, and **A. F. M. Smith** (1993). Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEE Proceedings F (Radar and Signal Processing)*, **140**(2), 107–113.

[6] **Z. Liu**, **Y. Lin**, **Y. Cao**, **H. Hu**, **Y. Wei**, **Z. Zhang**, **S. Lin**, and **B. Guo**, Swin transformer: Hierarchical vision transformer using shifted windows. *In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021.

[7] **J. Redmon** and **A. Farhadi** (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

[8] **Ultralytics** (2023). Yolov8: Ultralytics real-time object detection. https://github.com/ultralytics/ultralytics. Accessed: 2025-02-01.