### Major Project

##### On

**3D OBJECT DETECTION FOR TRAFFIC MONITORING**

**For the partial fulfillment for the award of the degree of**

**Bachelor of Technology**

**In**

#### Applied Computational Science and Engineering

**Submitted By**

##### Ashwani Kumar Singh (2001921520018)

**Rishi Patel (2001921520047)**

##### Utkarsh Tripathi (2001921520064)

**Under the supervision of Mr. Dhirendra Siddharth**



##### G.L. BAJAJ INSTITUTE OF TECHNOLOGY & MANAGEMENT, GREATER NOIDA

**Affiliated to**

**DR. APJ ABDUL KALAM TECHNICAL UNIVERSITY, 2023-2024**

### Certificate

This is to certify that the Project report entitled **“3D Object Detection for Traffic Monitoring”** done by **Ashwani Kumar Singh(2001921520018), Rishi Patel (2001921520047) and Utkarsh Tripathi (2001921520064)** is an original work carried out by them in Department of Applied Computational Science & Engineering, G.L. Bajaj Institute of Technology & Management, Greater Noida under my guidance. The matter embodied in this project work has not been submitted earlier for the award of any degree or diploma to the best of my knowledge and belief.

Date:

**Mr. Dhirendra Siddharth Prof.(Dr.) Naresh Kumar**

**Signature of the Supervisor Dean of the department**

### Declaration

We hereby declare that the project work presented in this report entitled “**3D Object Oetection for Traffic Monitoring**”, in partial fulfillment of the requirement for the award of the degree

Of Bachelor of Technology in Applied Computational Science and Engineering, submitted to

A.P.J. Abdul Kalam Technical University, Lucknow, is based on my own work carried out at Department of Applied Computational Science & Engineering, G.L. Bajaj Institute of Technology & Management, Greater Noida. The work contained in the report is original and project work reported in this report has not been submitted by me/us for award of any other degree or diploma.

Signature: Signature:

Name: Ashwani Kumar Singh Name: Rishi Patel

Roll No: 2001921520018 Roll No: 2001921520047

Signature:

Name: Utkarsh Tripathi Roll No: 2001921520064

### ACKNOWLEDGEMENT

The merciful guidance bestowed to us by the almighty made us stick out this project to a successful end. We humbly pray with sincere heart for his guidance to continue forever.

We pay thanks to our project guide Mr. Dhirendra Siddharth who has given guidance and light to us during this project. His versatile knowledge has helped us in the critical times during the span of this project.

We pay special thanks to our Dean of Department Prof.(Dr.) Naresh Kumar who has been always present as a support and help us in all possible way during this project.

We also take this opportunity to express our gratitude to all those people who have been directly and indirectly with us during the completion of the project.

We want to thanks our friends who have always encouraged us during this project.

At the last but not least thanks to all the faculty of ACSE department who provided valuable suggestions during the period of project.

## Abstract

Machine Learning has played a major role in various applications including Autonomous Vehicles and Intelligent Transportation Systems. Utilizing a deep convolutional neural network, the article introduces a zero-calibration 3D Object recognition and tracking system for traffic monitoring. The model can accurately work on urban traffic cameras, regardless of their technical specification (i.e. resolution, lens, the field of view) and positioning (location, height, angle). For the first time, we introduce a novel satellite-ground inverse perspective mapping technique, which requires no camera calibrations and only needs the GPS position of the camera. This leads to an accurate environmental modeling solution that is capable of estimating road users’ 3D bonding boxes, speed, and trajectory using a monocular camera. We have also contributed to a hierarchical activity/traffic modeling solution using short- and long-term Spatio-temporal video analysis to understand the heatmap of the traffic flow, bottlenecks, and high-risk zones. The experiments are conducted on four datasets: MIO-TCD, UA-DETRAC, GRAM-RTM, and Leeds-Dataset including various use cases and traffic scenarios.

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
|  |  | Page No. |
| **Figure 1.1** |  | **11** |
| **Figure 5.1** |  | **23** |
| **Figure 5.2** |  | **24** |
| **Figure 5.3** |  | **25** |
| **Figure 6.1** |  | **29** |

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
|  |  | Page No. |
| **Table 1.1** | **N/A** | **N/A** |

##### TABLE OF CONTENT

Certificate (ii)

Declaration… (iii)

Acknowledgement (iv)

Abstract… (v)

List of Figures…………………………………………, (vi)

List of Tables (vii)

|  |  |  |
| --- | --- | --- |
| **Chapter 1.** | **Introduction ..........................................................................................** | Pg. No. |
|  | Introduction…………………………………...….................................. | 10-12 |
| **Chapter 2.** | **Objective……………………….............................................................** | Pg.No |
| 2.1 | Objective……………………………………………………………….. | 13-14 |
| **Chapter 3.** | **Existing System……….……………………………………………….** | Pg.No |
| 3.1 | Road user detection Research………………………………………..... | 15-16 |
| 3.2 | Camera Calibrations Research………………………………………… | 16-17 |
| 3.3 | Traffic Modelling Research……………………………………………. | 17-17 |
| **Chapter 4.** | **Motivation/Problem Statement............................................................** | Pg.No |
| 4.1 | Motivation…………………….....…………………………………….. | 18-19 |
| 5.1 | Road user detection and tracking..…………………………………….. | 20-21 |
| 5.2 | Camera Auto-Calibration……………………………………………… | 22-24 |
| 5.3 | Environment Modelling and Traffic Analysis………………………… | 24-25 |
| **Chapter 6.** | **Plan of work………....………………………..........................................** | Pg.No |
| 6.1. | Road user Detection Model…………………………………………… | 26-27 |
| 6.2. | Multiclass Object Tracking…………………………………………… | 27-27 |
| 6.3 | Camera Auto-calibration……………………………………………… | 27-27 |
| 6.4 | Environment Training…………………………………………………. | 27-29 |
| **Chapter 7.** | **Tools and Technology………………………………………………….** | Pg.No. |
| 7.1 | Hardware Infrastucture………………………………………………… | 30-30 |
| 7.2 | Software Infrastructure………………………………………………… | 30-31 |
| 7.3 | Integrated Development Environment………………………………… | 31-31 |

**REFEFERNCES 32**



**Chapter 1**

## Introduction



Smart video surveillance systems play a crucial role in various applications, ranging from enhancing security in crowded areas to managing traffic and congestion in urban environments. As technological advancements continue, the complexity of traffic scenes for automated surveillance has increased significantly, driven by factors such as urban development, the coexistence of classic and autonomous vehicles, population growth, and the rise in cyclists, pedestrians, and other vulnerable road users.

The proliferation of surveillance cameras, with over 20 million CCD cameras in the USA and UK alone and 10 lakhs cameras in the top 15 cities of India, emphasizes the importance of video surveillance for city councils, authorities, and governments. This extensive network of interconnected cameras provides a valuable platform for in-depth studies on traffic management and urban planning.

Automated video surveillance, as opposed to traditional traffic monitoring systems, has been a subject of extensive research aimed at improving traffic flow, safety, and the overall sustainability of transportation networks. Among the technologies investigated, Computer Vision stands out as a prominent tool for extracting diverse information from traffic scenes within Automated Traffic Monitoring Systems (ATMS) and Intelligent Transportation Systems (ITS).

In the realm of ATMS and ITS, computer vision techniques enable the extraction of valuable insights from traffic scenes. These include but are not limited to vehicle type recognition, vehicle counting, speed estimation, tracking, and trajectory estimation. These capabilities empower authorities to make informed decisions for better traffic management.

Fig. 1, as an illustrative example of the research output, demonstrates the intricacies of road environments. It showcases interactions between various road users, such as pedestrians, vehicles, and cyclists, detailing their moving trajectories, speeds, and the density of users in different zones of the road. The top row of Fig. 1 highlights a 3D road-user detection and classification system, while the bottom row provides a real-time digital twin and replication of the same scene from a bird's eye view. In

dynamic environments like these, the ability to process information in real-time and accurately detect simultaneous events becomes crucial for effective decision- making.



Fig 1.1

In addition to the broader contributions of advanced video surveillance and Computer Vision to security enhancement, traffic management optimization, and sustainable urban development, the efficiency of a traffic monitoring system is further emphasized by its ability to seamlessly integrate with a grid of interconnected cameras in diverse urban locations. This grid may encompass cameras with varying resolutions, different viewing angles, heights, and focal lengths. However, the integration of such a heterogeneous camera network presents challenges that need to be addressed for the system to operate effectively.

One significant challenge is the calibration of each individual camera within the network, considering their intrinsic parameters and mounting specifications. This calibration process is not only costly but also non-trivial, requiring meticulous attention to detail. Addressing this challenge is a key objective of the study, aiming to contribute in four distinct areas:

Custom Deep Convolutional Neural Network (DCNN) for 3D Road User

Recognition:

The study involves the adaptation of a specialized DCNN designed for the recognition of 3D road users, including vehicles, pedestrians, and bikes. This is achieved through the implementation of a single-stage and multi-head object detection architecture. This customized approach enhances the system's ability to accurately identify and categorize various road users in dynamic traffic scenarios.

Multi-Class Object Tracker for Continuous Trajectory Estimation:

To overcome occasional visual occlusions that may occur in dynamic traffic scenes, the study develops a multi-class object tracker. This tracker utilizes fused localization information to provide continuous trajectory estimation for road users. This innovation ensures that the surveillance system maintains accurate tracking even in situations where objects may be temporarily obscured from view.

Novel Satellite/Ground-Based Auto-Calibration Method:

Recognizing the challenges associated with the calibration of cameras within a diverse network, the study introduces a novel auto-calibration method. This method leverages both satellite and ground-based information to achieve accurate localization and distance estimation of road users captured by surveillance cameras. By automating the calibration process, the study aims to streamline this typically resource-intensive task.

Automated Short- and Long-Term Traffic Analysis:

Beyond real-time tracking, the study focuses on automated traffic analysis for both short- and long-term perspectives. This involves the identification and understanding of traffic bottlenecks, safety risks, and hazards for road users. By providing insights into the broader traffic dynamics, the system contributes to proactive measures for enhancing road safety and optimizing traffic flow.

In summary, the study not only addresses the immediate challenges associated with traffic surveillance but also delves into the intricacies of camera calibration, trajectory estimation, and automated analysis. These advancements collectively contribute to the development of a sophisticated and robust traffic monitoring system capable of handling the complexities of modern urban environments.



**Chapter 2**

# Objective

The overarching objective of this project is to enhance the effectiveness of automated traffic surveillance systems through the integration of advanced technologies, particularly focusing on Computer Vision. The specific goals and objectives of the project can be summarized as follows:

Customized Deep Learning for 3D Road User Recognition:

* Adapt and implement a custom Deep Convolutional Neural Network (DCNN) designed specifically for 3D road user recognition.
* Target recognition of various road users, including vehicles, pedestrians, and bikes, using a single-stage and multi-head object detection architecture.

Multi-Class Object Tracking for Continuous Trajectory Estimation:

* Develop a multi-class object tracking system to ensure continuous trajectory estimation of road users.
* Address challenges posed by occasional visual occlusions by fusing localization information for robust and uninterrupted tracking.

Innovative Satellite/Ground-Based Auto-Calibration Method:

* Devise and implement a novel auto-calibration method that leverages both satellite and ground-based information.
* Automate the calibration process to account for varying intrinsic parameters and mounting specifications of each camera in the network.

Automated Short- and Long-Term Traffic Analysis:

* Implement algorithms for automated short-term and long-term traffic analysis.
* Identify and analyze traffic bottlenecks, safety risks, and hazards for road

users in real-time and over extended periods.

Real-Time Processing and Event Detection:

* Develop capabilities for real-time processing of video data from surveillance cameras.
* Implement algorithms for accurate and timely detection of simultaneous events in dynamic traffic environments.



Integration with Heterogeneous Camera Grid:

* Ensure seamless integration with a grid of interconnected cameras in different urban locations.
* Address challenges associated with varying resolutions, viewing angles, heights, and focal lengths within the camera network.

Contribute to Traffic Management and Urban Planning:

* Provide valuable insights and data for effective traffic management and urban planning.
* Support decision-making processes for city councils, authorities, and governments by offering comprehensive information on traffic flow, safety, and congestion.

System Robustness and Reliability:

* Ensure the developed system is robust and reliable in diverse environmental conditions.
* Conduct thorough testing and validation to verify the accuracy and performance of the implemented algorithms and methodologies.

By achieving these objectives, the project aims to contribute to the advancement of automated traffic surveillance systems, addressing the complexities of modern traffic scenarios and supporting the development of safer, more efficient, and sustainable urban environments.



**Chapter 3**

# Existing System

#### 3.1 Road-user detection research

In the realm of traffic surveillance, historical methods concentrated on detecting vehicles using motion-based solutions and background subtraction techniques. Nevertheless, these approaches faced challenges, such as difficulty in identifying stationary objects and struggles in assessing traffic on crowded urban roads.

Earlier studies employed techniques like support vector machines and dimensional reduction to improve background subtraction, attempting to address issues like partial occlusion and challenging illumination conditions. Despite these efforts, there were limitations, particularly in terms of adaptability to varying environmental conditions and complex traffic scenarios. For instance, approaches relying on corner detection algorithms or histogram of oriented gradients encountered difficulties in accurately pinpointing the location and size of vehicles, particularly in low-light conditions or when shadows were present.

Over the past decade, the advent of deep learning has brought about significant advancements in vehicle detection. Convolutional Neural Networks (CNNs) have been categorized into two types: two-stage detectors and single-stage detectors. Two-stage detectors, like the RCNN family, involve separate stages for region proposal and classification. On the other hand, single-stage detectors, such as Single-Shot Multi-Box Detector (SSD) and You Only Look Once (YOLO), treat detection as a regression problem, streamlining the process with a unified architecture for localization and classification.

Recent studies have showcased improved vehicle detection performance using deep learning methodologies such as Faster-RCNN, YOLO, and YOLOv3. Some have taken a hybrid approach, combining statistical models, random forest methods, and neural networks to predict traffic volume accurately. Others have explored innovative techniques, including 3D imaging of traffic scenes, clustering LiDAR point cloud data, and employing multi-modal systems with thermal sensors to enhance traffic monitoring capabilities.

While certain studies have incorporated complex methods involving multiple

cameras and sensors to obtain depth information, these approaches can be associated with considerable costs, particularly in large and crowded urban areas. As part of our research initiative, our aim is to mitigate such expenses by leveraging existing infrastructure. By doing so, we intend to enhance and optimize automated traffic surveillance systems, making them more practical and cost-effective for deployment in diverse urban settings. This approach aligns with the goal of developing efficient and scalable solutions for monitoring and managing traffic without imposing excessive financial burdens.

#### 3.2. Camera calibration research

In understanding how cameras capture images, it's important to recognize that they essentially convert the three-dimensional world into a two- dimensional image based on their unique settings and angles. This process is known as camera calibration, and it plays a pivotal role in various applications such as inverse perspective mapping, distance estimation, and vehicle speed determination.

Traditionally, calibration studies have utilized methods involving checkerboards, straight-line vanishing points, and manual procedures with multiple points. However, the focus on automating the camera calibration process has been relatively limited. Some studies, like the one conducted by Dubská and her colleagues, attempt to automatically calibrate cameras by leveraging the movement of detected cars and utilizing the Hough line transform algorithm. Nevertheless, this method faces challenges, especially in difficult lighting conditions and noisy environments.

Another approach involves using a Faster-RCNN model to detect car edges and extract points for calibration, as seen in the study by Sochor, Juránek, and Herout. Similarly, Song and his team employ an SSD object detector to track cars, extract points for calibration, and calculate vanishing points based on vehicle movements. However, a limitation in their approach is the assumption of fixed sizes for all cars, which may not accurately represent real-world scenarios.

It's worth noting that most of these calibration methods assume straight roads, making them less effective on curved roadways. In an alternative study by Kim, a different strategy is employed. This method considers multiple corresponding points between the roadside camera image and another view of the same scene. These points are then matched and used to estimate calibration parameters using a revised version of the RANSAC algorithm. While this approach addresses the curvature issue, it requires users

to manually set the corresponding points, making the calibration process less automated.

In summary, the quest for automating camera calibration involves various methods, each with its strengths and limitations. The challenge lies in developing approaches that are not only accurate and robust but also capable of handling real-world complexities, such as curved roads, without relying on extensive manual input.

#### 3.3 Traffic modeling research

Recent research has focused on estimating and predicting traffic flow using advanced techniques like LSTM modules and Restricted Boltzmann Machines (RBM) to extract spatial features from raw input progressively. A learning-based approach for predicting travel time has also been suggested. Some studies utilize 3D CNNs, graph-based CNNs, and Graph Convolutional Recurrent Neural Networks for dynamic traffic flow analysis, but they often rely on macro-scale traffic data, lacking fine-grained assessments for real- time visual monitoring.

In contrast, mathematical traffic modeling research aims to understand traffic flow based on microscopic urban and rural data, emphasizing factors like vehicle size and their lateral and longitudinal gaps. However, mathematical models can sometimes be too abstract, failing to reflect the naturalistic and region-specific conditions of roads and environments.

As part of our contribution, we propose developing an efficient digital twin of the environment through spatiotemporal analysis. This involves creating a live visual heatmap of the environment to identify highly congested areas, automatically recognize speeding violation zones, identify pedestrian-favored paths, and highlight potential crash zones for vehicles and pedestrians. This approach aims to provide a more realistic and region-specific representation of traffic conditions for improved monitoring and analysis.



**Chapter 4**

# Motivation/Problem Statement



The motivation behind the project is to enhance and advance the capabilities of traffic monitoring and management systems. The current state of traffic surveillance involves a variety of methodologies, including traditional mathematical models and recent advancements in deep learning and neural networks. However, there are certain limitations and gaps that the proposed project aims to address, thereby providing motivation for its implementation. Here are some key motivations:

Improving Accuracy and Real-Time Monitoring:

Many existing methods, including mathematical traffic models, can be abstract and may not accurately reflect real-world, region-specific conditions. The project seeks to enhance accuracy by leveraging advanced technologies like deep learning and neural networks, enabling real-time monitoring of traffic conditions.

Fine-Grained Assessments:

While some studies focus on macro-scale traffic data, the project recognizes the need for fine-grained assessments in real-time visual monitoring. By employing advanced techniques, the goal is to provide a more detailed and nuanced understanding of traffic flow, congestion, and potential safety issues.

Addressing Limitations of Current Approaches:

Current methodologies, such as relying on macro-scale data or abstract mathematical models, may have limitations in capturing the complexities of dynamic urban environments. The project aims to overcome these limitations by proposing a digital twin approach that integrates spatiotemporal analysis for a more comprehensive understanding.

Enhancing Traffic Safety:

Identifying speeding violation zones, pedestrian-favored paths, and potential crash zones are crucial for enhancing traffic safety. The

project is motivated by a desire to contribute to the reduction of accidents and incidents on the road by providing actionable insights to traffic management authorities.

Adaptability to Region-Specific Conditions:

The motivation stems from a recognition that traffic conditions can vary significantly based on the region. The proposed project aims to develop a model that is adaptable and reflective of the naturalistic and region-specific conditions of roads and environments.

Efficient Digital Twin Development:

The concept of a digital twin, represented by a live visual heatmap, serves as a motivating factor. Such a digital twin can offer a real-time, dynamic representation of the environment, allowing for the identification of congestion, speeding violations, and potential safety hazards.

Cost-Effective Implementation:

Leveraging existing infrastructure and minimizing costs in large and crowded cities is a practical motivation. The project aims to develop efficient solutions without imposing significant financial burdens.

In summary, the project is motivated by the desire to overcome the limitations of existing traffic monitoring approaches, provide a more accurate and real-time understanding of traffic conditions, and contribute to enhanced traffic safety through the development of an efficient digital twin of the environment.



**Chapter 5**

# Methodology



#### Road-users detection and tracking

Drawing inspiration from the advancements in YOLO (You Only Look Once) object detectors, particularly those trained on the COCO (Common Objects in Context) dataset, we have developed a sophisticated road user detection model named RUYOLO. This model is designed to efficiently perceive and analyze the dynamics of traffic scenes. In its architectural representation (refer to Fig. 3), key components are denoted by distinctive color-coded blocks.

The orange blocks symbolize the backbone of the model, employing stacked convolutional layers to extract spatial features from input images. On the other hand, the blue blocks represent the neck slots of the model, enriched with multiple partial connections, including Contextual Spatial Pyramid (CSP) introduced by Wang et al. in 2020. These connections play a vital role in mitigating the vanishing gradient problem and enhancing feature propagation throughout the neural network.

Additionally, we have incorporated a path aggregation method to augment both semantic and spatial information within the model.

Conducting a comprehensive series of experimental studies, we focused on optimizing our model. This involved the exploration of various loss functions, with a particular emphasis on Focal-Loss (Lin et al., 2017) and Distance-IoU loss (Zheng et al., 2020) in the head part of the network. The objective was to enhance the classification results, specifically in the context of detecting vehicles and pedestrians, during the training process.

The training of RUYOLO was executed utilizing the MIO-TCD dataset (Luo et al., 2018). This dataset provides a diverse array of annotated traffic video samples, encompassing 11 classes. These classes include pedestrians, articulated trucks, bicycles, buses, cars, motorcycles, motorized and non-motorized vehicles, pickup trucks, single-unit trucks, and work vans. This meticulous selection aimed to ensure the development of a highly accurate and customized model tailored to our application requirements.

The final output of the RUYOLO model yields a set of parameters (𝑥𝑏, 𝑦𝑏, 𝑤𝑏, ℎ𝑏, s,

𝐜) for each object in the image stream. Here, 𝑥𝑏, 𝑦𝑏 denote the center points, 𝑤𝑏, ℎ𝑏 represent the width and height of the detected bounding boxes, s indicates the objectness confidence score, and 𝐜 is a vector of classification probabilities corresponding to the number of classes. The coordinates of the middle point at the bottom side of each bounding box are considered as the reference point for detected objects, representing the closest contact point of vehicles and pedestrians to the

ground (the road surface). This is mathematically expressed as (̂𝑥, ℎ𝑏/2).

) = (𝑥𝑏, 𝑦𝑏 +

In essence, RUYOLO functions as an advanced system for road user detection, leveraging cutting-edge techniques to ensure accurate classification and localization in dynamic traffic scenarios.

The SORT (Simple Online and Realtime Tracking) algorithm, as introduced by Bewley, Ge, Ott, Ramos, and Upcroft in 2016, has gained prominence for its effectiveness in generic object tracking applications. This algorithm assigns a distinct identification (ID) to each object by evaluating the Intersection over Union (IoU) between detected bounding boxes in consecutive frames of a given video. However, it is noteworthy that this process is designed for single-class tracking, necessitating the separate treatment of each class. A challenge arises when the object detector assigns a new class to an existing object, causing a misalignment with the object tracker's estimation. This common issue leads to the SORT tracker perceiving the same object as new, resulting in the assignment of a new ID and the loss of the previous tracking.

To address this challenge, we introduced a category vector denoted as 𝐜́ ∈ W1×11 to augment the baseline SORT tracker. This augmentation is followed by a Hungarian intra-frame association, culminating in our multi-class object tracking (MCOT) solution. Notably, this integration mitigates the problem of false multi-ID assignment to the same object. The category vector is a one-hot encoded representation of the detected class vector 𝐜, where the highest class probability is set to 1, and the remaining probabilities are suppressed to 0.

Taking advantage of the smoothing effect inherent in the tracker, we effectively filter out the bouncing of detected categories across frames. Additionally, this integration enables the tracker to calculate IoU between the bounding boxes of different categories. Consequently, this approach facilitates a multi-object and multi-category ID assignment, enhancing the overall robustness and accuracy of the tracking system.

#### Camera auto-calibration

The aim of this segment of the study is to create an automatic Inverse Perspective Mapping (IPM) camera calibration approach applicable to scenarios where intrinsic camera details and mounting specifications are unknown. Traditional IPM approaches usually require multi-point camera calibration within the scene. However, our goal is to develop a methodology universally applicable to all Closed-Circuit Television (CCTV) traffic surveillance cameras.

With only the geo-location of a camera (specifically, GPS coordinates) as input, we extract a top-view satellite image corresponding to the location of the CCTV camera. The focus is on creating an auto-calibrated Satellite-Ground-based Inverse Perspective Mapping (SG-IPM). This method estimates a planar transformation matrix , which transforms the camera perspective image into a bird's-eye view image.

Using pixel coordinates (𝑥, 𝑦) in a digital image container, where 𝑤 and ℎ are the width and height of the image, respectively, we distinguish between perspective space

(̂) and inverse perspective space. The surveillance camera image is denoted as 𝐈, the

satellite image as ̀𝐈, and the bird's-eye view image as , calculated through a linear

transformation 𝐆 ∶ → .

Since the coordinates of the bird's-eye view image approximately match those of the

satellite image, a transformation function (̌𝑥,

) = 𝛬((̂𝑥,

), 𝐆) facilitates the

transformation of pixel locations from 𝐈

to 𝐈

. Conversely, 𝐆⁻¹ inverts the mapping

process, implying that (̂𝑥,

) = 𝛬((̌𝑥,

), 𝐆⁻¹) transforms pixel locations from

to .

To solve the linear system involved in this process, a minimum of four pairs of corresponding points in the ground-based image (̂𝐈) and satellite-based image (̀𝐈) is required. These feature points must demonstrate robustness and invariance to various transformations such as rotation, translation, scale, tilt, and partial occlusion, especially in scenarios with high affine variations.

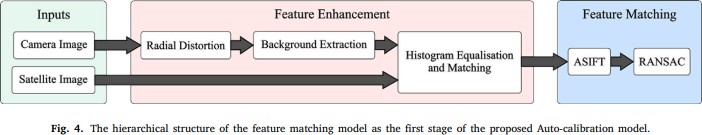


Fig 5.1

##### Feature Mapping

To enhance the visual similarity between the camera image and the satellite image for identifying corresponding points, we utilized a technique to correct distortions caused by imperfections in the camera lens and optics.

To distinguish background objects from foreground objects and road users in the traffic scene. In general it is seen that we need 70 frames to remove foreground images/data from a given image it may vary depending upon the quality of the data provided.

##### Feature Enhancement

In our approach, inspired by the Affine Scale Invariant Feature Transform (ASIFT), we apply random affine transformations to generate various perspectives of both the camera image (̂𝐈\_g) and the satellite image (̀𝐈\_g). This strategy allows us to create multiple sample images with different viewing angles, providing a broader range of data compared to the original camera and satellite images. The goal is to increase the

likelihood of finding similar feature pairs between \_g and ̀𝐈\_g. However, this process

may introduce some outliers, which are mismatched features that can lead to inaccurate estimates of the transformation matrix.

To address this issue, we employ the Random Sample Consensus (RANSAC) algorithm. RANSAC is an iterative learning algorithm used for parameter estimation. In each iteration, the algorithm randomly selects four corresponding pairs from all

matching points between \_g and ̀𝐈\_g. It then computes the transformation matrix

using these sampled pairs. Next, it conducts a voting process on all matching feature pairs to identify the best-matching samples. This iterative and voting-based approach helps eliminate outliers and enhances the accuracy of estimating the transformation matrix.

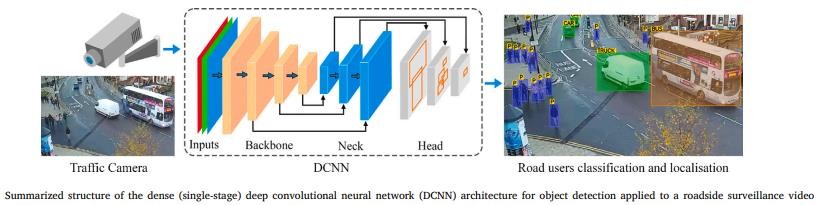


Fig 5.2

#### Environment Modelling and traffic Analysis

Automated analysis of traffic scene videos via surveillance cameras is a complex task. This is mainly due to the existence of various types of objects such as trees, buildings, road users, banners, etc in various sizes and distances. Occlusion and lighting conditions are additional parameters that make it non-trivial. In this section, we elaborate on our techniques for providing an abstract visual representation of the environment, objects of interest, traffic density, and traffic flow. In order to achieve a 3D modeling and representation of the road users (e.g. vehicles), we require to identify and recognize the following properties for the road users and the road scene:

* Vehicle’s speed (𝖯)
* Vehicle’s heading angle (𝜃)
* Road boundaries

Before proceeding any further, we need to ensure an accurate mapping from the image plane to the bird’s eye view image. We have observed bird’s eye view projected points normally suffer from some noise caused due transformation (𝐆). To alleviate this noise and recover the natural movement of objects (especially the vehicle’s dynamics, and moving trajectory, we applied a constant velocity Kalman filter. This significantly improves the tracking performance to provide a more accurate speed and heading angle estimation.

##### 2D to 3D Bounding Box Conversion

The process involves transitioning from a 2D bounding box representation in the surveillance camera image (̂𝐈) to a cubical 3D bounding box, thereby estimating eight

corners of the cube. The floor of this cube comprises four corner points, corresponding to a rectangle in the satellite image (̀𝐈). This rectangle denotes the

occupied ground plate area by the object and is characterized by the center (̌𝑥, ),

height (ℎ̌,𝑏), and width (𝑤,̌ 𝑏). The dimensions ℎ̌,𝑏 and 𝑤̌,𝑏 are determined based on ground truth information about the approximate 2D dimensions of the corresponding object's class in the real world.

The method encompasses 2D to 3D bounding box conversion for various vehicle classes, including bicycles, motorcycles, articulated trucks, pickup trucks, single-unit trucks, vans, minibusses, double-decker buses, mini cars, and standard sedans.

Utilizing the center point and class information, the width and length of the object are extracted from the object's class ground truth. For instance, the dimensions (2.55 m ×

4.95 m) for Double-Decker buses in the UK.

The height calculation involves fitting the 3D bounding box tangent to the upper edge of the 2D bounding box. To represent the object's movement direction, the rectangle is rotated and aligned with the estimated heading angle (𝜃̃

𝑣𝑖). The resulting rectangle's corners, when converted back to the camera image domain (̂𝐈) using the inverse perspective transformation matrix (𝐆−1), form the cube's floor. By adding a predetermined height (ℎ3𝐷) to the 𝑦-axis of the floor corners, the roof's points are established.

For vehicles, the cube's height (except for pedestrians) is determined as ℎ3𝐷 = 𝛽 × ℎ̂

𝑏, where 𝛽 = 0.6 serves as a height coefficient. Notably, the cube's height for pedestrians equals the height of the detected bounding box in the perspective domain (ℎ3𝐷 = ℎ̂,𝑏). This comprehensive approach facilitates a robust representation of the occupied space by each road user in both the camera and satellite images.

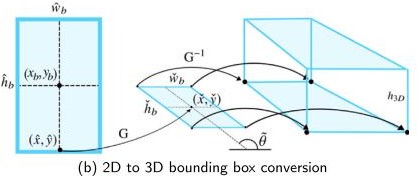


Fig 5.3



#### Chapter 6

**Plan of work**



#### Road User Detection Model: RUYOLO

##### Architectural Framework

Inspired by YOLO object detectors trained on the COCO dataset, our model, named RUYOLO, boasts an advanced architectural framework. Distinctive color-coded blocks, such as orange for the backbone and blue for neck slots, signify key components. The incorporation of Contextual Spatial Pyramid connections in the neck slots addresses challenges like vanishing gradients, augmenting feature propagation.

##### Optimization Strategies

Extensive experimental studies were conducted to optimize RUYOLO. Various loss functions, including Focal-Loss and Distance-IoU loss, were explored in the head part of the network. These optimizations specifically targeted improving the classification results pertaining to vehicles and pedestrians during the training process.

##### Training Dataset

RUYOLO was meticulously trained on the MIO-TCD dataset, providing a diverse set of annotated traffic video samples covering 11 classes. This approach ensures the development of a highly accurate and customized model tailored to our application requirements.

##### Output Parameters

The output of RUYOLO includes a set of parameters (𝑥𝑏, 𝑦𝑏, 𝑤𝑏, ℎ𝑏, s, 𝐜) for each object in the image stream. These parameters encompass center points, bounding box dimensions, objectness confidence score, and classification probabilities. The model provides a detailed representation of detected objects, offering valuable insights for traffic

analysis.

#### Multi-Class Object Tracking: MCOT

Building upon the Simple Online and Realtime Tracking (SORT) algorithm, we introduce a category vector 𝐜́ to augment the baseline SORT tracker. This enhancement, coupled with Hungarian intra-frame association, forms our multi-class object tracking (MCOT) solution. By addressing the issue of false multi-ID assignment, MCOT significantly improves the tracking accuracy across different object classes.

#### Camera Auto-Calibration: SG-IPM Approach

##### Goal and Methodology

This segment focuses on developing an automatic Inverse Perspective Mapping (IPM) camera calibration approach applicable to scenarios with unknown intrinsic camera details. The goal is to create a universally applicable method for all Closed-Circuit Television (CCTV) traffic surveillance cameras.

##### Geo-Location-Based Calibration

Utilizing only the geo-location (GPS coordinates) of a camera, we extract a top-view satellite image corresponding to the location of the CCTV camera. This forms the basis for developing an auto-calibrated Satellite-Ground-based Inverse Perspective Mapping (SG-IPM) technique.

##### Feature Mapping and Enhancement

To enhance feature matching between camera and satellite images, we apply a radial distortion correction technique. Furthermore, inspired by the Affine Scale Invariant Feature Transform (ASIFT), random affine transformations are employed to generate diverse perspectives for both camera and satellite images. The Random Sample Consensus (RANSAC) algorithm is then applied to eliminate outliers and enhance the accuracy of the transformation matrix.

#### Environment Modeling and Traffic Analysis

##### 2D to 3D Bounding Box Conversion

In the context of achieving a 3D modeling and representation of road users, we employ a technique for transitioning from a 2D bounding box representation in the surveillance camera image to a cubical 3D bounding box. This involves estimating eight corners of the cube and determining dimensions based on ground truth information about the object's class.

##### Speed and Heading Angle Estimation

Accurate mapping from the image plane to the bird’s eye view image is crucial. To address noise in projected points, we apply a constant velocity Kalman filter, improving tracking performance. Speed and heading angle estimations are derived from the transformed and filtered data.

##### Comprehensive Environment Modeling

In our approach, the modeling process includes automated analysis of traffic scenes, considering various objects and factors such as occlusion and lighting conditions. A detailed representation of occupied space by each road user in both camera and satellite images is achieved through 2D to 3D bounding box conversion.

#### Experimental Results and Analysis

##### RUYOLO Performance Metrics

The performance of RUYOLO is evaluated using various sizes and numbers of head architectures. Learning parameters, including sizes such as 'small,' 'medium,' 'large,' and 'xlarg,' are considered. The model is trained and optimized with different head modules, and the performance metrics are analyzed to ascertain the optimal configuration.

##### Training and Validation Processes

One-cycle learning rate strategy is employed to choose optimal learning rates and avoid prolonged training times. Analytic graphs illustrate the training and validation processes. Overfitting concerns are addressed by selecting optimal weights, ensuring the model learns generalized features.

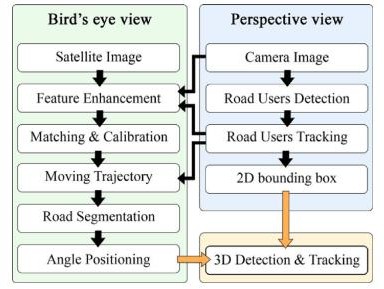


Fig 6.1 : Overall structure of hierarchical stages from 2D to 3D object detection



**Chapter 7**

# Tools and Technology



#### Hardware Infrastructure

##### Processing Units

The computational backbone of our research relies on high-performance processing units to efficiently handle complex algorithms and models. Central Processing Units (CPUs) with multiple cores and Graphics Processing Units (GPUs) are employed to ensure parallel processing capabilities, optimizing the execution of computationally intensive tasks.

##### Memory Architecture

Robust Random Access Memory (RAM) and Graphics Double Data Rate (GDDR) memory modules are integrated into the hardware infrastructure. Adequate memory facilitates seamless data handling, enabling the simultaneous processing of large datasets and complex neural network models.

##### Storage Solutions

To accommodate the storage requirements of extensive datasets, high-capacity Solid State Drives (SSDs) or Hard Disk Drives (HDDs) are employed. The choice between SSDs and HDDs depends on the specific demands of data access speed and storage capacity.

##### Network Connectivity

Efficient communication between hardware components is crucial for real-time data processing and model updates. High-speed network interfaces, such as Gigabit Ethernet or, in advanced setups, 10 Gigabit Ethernet, are integrated to ensure low-latency data exchange.

#### Software Stack

##### Operating Systems

The choice of operating systems (OS) is a critical aspect of our research framework. Depending on the specific requirements and compatibility with tools and libraries, we

consider operating systems such as Linux distributions (e.g., Ubuntu, CentOS) for their stability, security, and extensive support for open-source tools.

##### Development Frameworks

Our research heavily relies on powerful and versatile development frameworks for machine learning and computer vision. TensorFlow and PyTorch are prominently featured, providing a robust ecosystem for designing, training, and deploying deep learning models.

##### Computer Vision Libraries

OpenCV (Open Source Computer Vision Library) is instrumental in image and video processing tasks. Its comprehensive set of functionalities, including image manipulation, feature detection, and object tracking, greatly contributes to the success of our road user detection and tracking algorithms.

##### Version Control

To manage the collaborative nature of research and ensure version control of codebases, Git, and platforms such as GitHub or GitLab are employed. This facilitates seamless collaboration among researchers, version tracking, and the integration of new features or improvements.

#### Integrated Development Environments (IDEs)

##### Code Development

Sophisticated Integrated Development Environments enhance the efficiency of code development. IDEs such as Visual Studio Code, PyCharm, or Jupyter Notebooks provide features like code autocompletion, debugging tools, and visualization capabilities, streamlining the development process.

##### Experimentation and Prototyping

For rapid prototyping and experimentation, platforms like Google Colab or Jupyter Notebooks are utilized. These platforms provide an interactive and collaborative environment for testing algorithms, adjusting parameters, and analyzing results in real- time.

**References:**

Abdollahi, M., Khaleghi, T., & Yang, K. (2020). An integrated feature learning approach using deep learning for travel time prediction. Expert Systems with Applications, 139,

Article 112864. Expert Systems With Applications 227 (2023) 12025316

M. Rezaei et al.

Ahmad, A. (2021). Assessment of automated road features xtraction algorithm from UAV images (Ph.D. thesis), Universiti Teknologi Mara Perlis. Alldieck, T.,

Bahnsen, C. H., & Moeslund, T. B. (2016). Context-aware fusion of RGB and thermal imagery for traffic monitoring. Sensors, 16(11).

Arinaldi, A., Pradana, J. A., & Gurusinga, A. A. (2018). Detection and classification of vehicles for traffic video analytics. Procedia Computer Science, 144, 259–268, INNS

Conference on Big Data and Deep Learning. Arnold, E., Al-Jarrah, O. Y., Dianati, M., Fallah, S., Oxtoby, D., & Mouzakitis, A. (2019). A survey on 3d object detection methods for autonomous driving applications. IEEE Transactions on Intelligent Transportation Systems, 20(10), 3782–3795.

Bewley, A., Ge, Z., Ott, L., Ramos, F., & Upcroft, B. (2016). Simple online and realtime tracking. In 2016 IEEE international conference on image processing. IEEE.

Bhoi, A. (2019). Monocular depth estimation: A survey. arXiv:1901.09402. Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.

Brunetti, A., Buongiorno, D., Trotta, G. F., & Bevilacqua, V. (2018). Computer vision and deep learning techniques for pedestrian detection and tracking: A survey.

Neurocomputing, 300, 17–33. Bui, K.-H. N., Yi, H., & Cho, J. (2020). A multi- class multi-movement vehicle counting framework for traffic analysis in complex areas using CCTV systems. Energies, 13(8).

Chen, K., Chen, F., Lai, B., Jin, Z., Liu, Y., Li, K., et al. (2020). Dynamic spatio- temporal graph-based CNNs for traffic flow prediction. IEEE Access, 8, 185136–185145