

IIT
Bhubaneswar

**INDIAN INSTITUTE OF TECHNOLOGY BHUBANESWAR**

**SCHOOL OF ELECTRICAL SCIENCES**

**PROJECT REPORT ON**

“Object Detection & Tracking using Deep-Sort”

Submitted By:

Sarthak Siddhant Bharadwaj

Summer intern-2022

APPLICATION NO.: INT/ES/SUM/2022/0255

Under Guidance of :

Prof: Dr. Debi Prosad Dogra

**School of**Electrical Sciences

**E-mail ID:**[dpdogra@iitbbs.ac.in](mailto:dpdogra@iitbbs.ac.in" \t "_blank)

**CERTIFICATE**

This is to certify that Mr.Sarthak Sidhhant Bharadwaj has completed the project report on the topic “Object Detection & Tracking using Deep-Sort” at the School of Electrical Sciences, Indian Institute of Technology Bhubaneswar, under my supervision and guidance in the fulfillment of the requirements of the Summer internship program-2022.

**Prof: Dr. Debi Prosad Dogra Date:**

**School of Electrical Sciences Place:**

**ACKNOWLEDGEMENT**

I deem it a pleasure to acknowledge my sense of gratitude to my project guide Prof. Dr. Debi Prosad Dogra, under whom I carried out the project work. His incisive and objective guidance and timely advice encouraged me to continue the work with a constant flow of energy.

I wish to reciprocate the kindness shown by Mr. Arabinda (Ph.D. scholar) at the School of Electrical Sciences, IIT Bhubaneswar, who guided me with his valuable suggestions in completing my work.

This acknowledgment will remain incomplete if I fail to express our deep sense of obligation to my parents and God for their consistent blessings and encouragement.

Date: Sarthak S Bharadwaj

Place: Bhubaneswar, Odisha. Reg no.: 1901106431

Email: mr.sarthakbharadwaj@gmail.com

**ABSTRACT**

Computer vision has evolved in the last decade as a critical technology for numerous applications replacing human supervision. Timely detection of traffic violations and abnormal behavior of pedestrians in public places through computer vision and visual surveillance can effectively maintain traffic order in cities. However, despite a few recent computer vision-based techniques proposed to understand traffic violations or other types of on-road anomalies, no methodological survey provides a detailed insight into the classification techniques, learning methods, datasets, and application contexts. Thus, this study aims to investigate the recent visual surveillance–related research on anomaly detection in public places, particularly on the road. The study analyses various vision-guided anomaly detection techniques using a generic framework such that the critical technical components can be easily understood. Our survey includes definitions of related terminologies and concepts, judicious classifications of the vision-guided anomaly detection approaches, detailed analysis of anomaly detection methods including deep learning–based practices, descriptions of the relevant datasets with environmental conditions, and types of anomalies. The study also reveals vital gaps in the available datasets and anomaly detection capability in various contexts and thus gives future directions to computer vision–guided anomaly detection research. As anomaly detection is an essential step in automatic road traffic surveillance, this survey can be a valuable resource for interested researchers working on solving various issues of Intelligent Transportation Systems (ITS).

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| No**.** | TOPIC | PAGE NO. |
| **1** | **Objective of Project** | **06** |
| **2.** | **Introduction** | **07** |
| **3.** | **YOLO-Object detection** | **08** |
| **4.** | **Object tracking using deep sort** | **12** |
| **5.** | **Dataset Preparation** | **13** |
| **6.** | **Conclusion** | **15** |
| **7.** | **References** | **16** |

**OBJECTIVE**

This proposal aims to develop AI and computer vision-guided traffic monitoring systems with the following scientific objectives:

Development of AI and CV-guided system to detect mid-block congestion on city roads. Deliver real-time congestion information to the commuters through variable message signs.

Suggesting alternate routes through messages. Congestion will be measured by using Travel Time Index (TTI) and Area Occupancy.

Study the effect of congestion on connected signals by analyzing videos in real time. We aim to learn traffic patterns in an unsupervised way. We forecast the congestion level at the downstream intersection based on the observed (real-time) congestion at the immediate upstream intersection.

Implementation of deep learning guided method to automatically detect the available slots in designated roadside parking zones. We aim to deliver users' real-time information on available parking bays through mobile applications. It will also see illegal parking and warn the administrators.

To monitor the pedestrian-vehicle interaction (PVI) at traffic intersections and assess the risk pedestrians take while crossing the roads. This data can also be used to estimate the level of safety (PVI) at the intersection. This system can be used to identify vulnerable corners from a PVI perspective.

**INTRODUCTION**

With the growing population in India, the number of vehicles running on the roads is ghastly increasing. People everyday use transportation to travel to various places for work. There has been an increase in traffic jams in different metro cities. People often violate traffic rules, which are very hard to detect even by officers via old manual detection through CCTV. So to minimize the effort that is spent on detecting the anomalies, this project has been developed using AI and a computer vision-aided system.

We live in the big data era, where all areas of science and industry generate massive amounts of data. This confronts us with unprecedented challenges regarding their analysis and interpretation.

Humans glance at an image and instantly know what objects are in the picture. The human visual system is fast and accurate. Object detection is an advanced form of image classification where a neural network predicts objects in an image and points them out in the form of bounding boxes.

Object detection thus refers to the detection and localization of objects in an image that belongs to a predefined set of classes. The object detection algorithms would allow computers to drive cars without specialized sensors and enable assistive devices to convey real-time scene information to human users; Current detection systems repurpose classifiers to perform detection. To detect an object, these systems take a classifier for that object, evaluate it at various locations, and scale it in a test image. For this reason, there is an urgent need for novel machine learning and artificial intelligence methods to help utilize these data.

More recent approaches like R-CNN use region proposal methods[5] first to generate potential bounding boxes in an image and then run a classifier on these proposed boxes. After classification, post-processing is used to refine the bounding boxes, eliminate duplicate detections, and rescore the packages based on other objects in the scene. These complex pipelines are slow and hard to optimize because each component must be trained separately. But YOLO is swift. YOLO trains on full images and directly optimizes detection performance[4]. This unified model has several benefits over traditional methods of object detection.

A neural network runs on a new image at test time to predict detections. Furthermore, YOLO achieves more than twice the average precision of other real-time systems. YOLO is a much faster algorithm than its counterparts, running at as high as 45 FPS.

**YOLO-Object Detection**

YOLO is an acronym for “You Only Look Once” In machine learning terms, we can say that all objects are detected via a single algorithm run. It’s done by dividing an image into a grid and predicting bounding boxes and class probabilities for each cell in a grid.

YOLO’s raw output contains many bounding boxes for the same object. These boxes differ in shape and size. Some boxes are better at capturing the target object, whereas others offered by an algorithm perform poorly. Unlike sliding window and region proposal-based techniques, YOLO[7] sees the entire image during training and test time. Hence, it implicitly encodes contextual information about classes and their appearance. Fast R-CNN, a top detection method, mistakes background patches in an image for objects because it can’t see the larger context. YOLO makes less than half the background errors compared to Fast R-CNN.

YOLO learns generalizable representations of objects. When trained on natural images and tested on the artwork, YOLO outperforms top detection methods like DPM and R-CNN by a wide margin. Since YOLO is highly generalizable, it is less likely to break down when applied to new domains or unexpected inputs.

Yolo algorithm employs CNN to detect objects in real-time. The algorithm requires a single forward propagation through NN to detect objects. This means that prediction in the entire algorithm is made in a single algorithm run.

Yolo Algorithmworks using the following three techniques i) Residual blocks ii) Bounding box regression iii) Intersection over union (IOU).

First, the image is divided into various grids.

Each grid cell predicts B bounding boxes and confidence scores for those boxes. These confidence scores reflect how confident the model is that the box contains an object and how accurate it thinks it is that it predicts. Formally it is defined confidence as Pr(Object) ∗ IOU. The confidence score should be zero if no object exists in that cell. Otherwise, the confidence score should equal the intersection over union (IOU) between the predicted box and the ground truth.

Each bounding box consists of 5 predictions: x, y, w, h, and confidence. The (x, y) coordinates represent the box's center relative to the grid cell's bounds. The width and height are predicted to be close to the whole image.

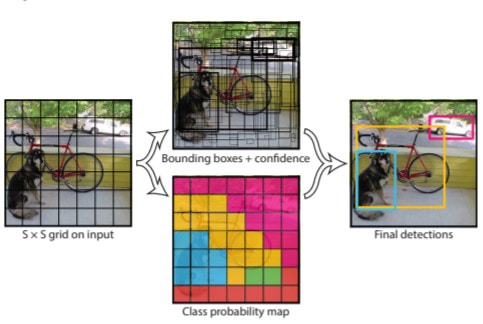
Finally, the confidence prediction represents the IOU between the predicted box and any ground truth box. Each grid cell also predicts C conditional class probabilities, Pr(Class-I |Object). These probabilities are conditioned on the grid cell containing an object. Only one set of class probabilities is predicted per grid cell, regardless of the number of boxes B. At test time, the conditional class probabilities are multiplied with the individual box confidence predictions

**Pr (Class-I|Object) ∗Pr (Object)∗IOU = Pr (Class-I) ∗IOU**

YOLO suppresses all bounding boxes with lower probability scores by first looking at the probability scores associated with each decision and taking the largest one. Following this, it suppresses the bounding boxes having the most prominent Intersection over Union with the current high probability bounding box.

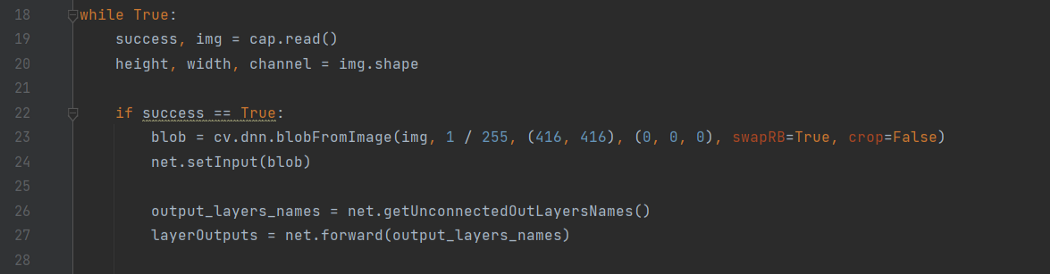
YOLO uses fully connected layers to predict bounding boxes instead of predicting coordinates directly from the convolution network, like in Fast R-CNN and Faster R-CNN. The initial convolutional layers of the network extract feature from the image, while the fully connected layers predict the output probabilities and coordinates.  
In the current version, instead of the fully connected layer and the anchor boxes are used to predict the bounding boxes.

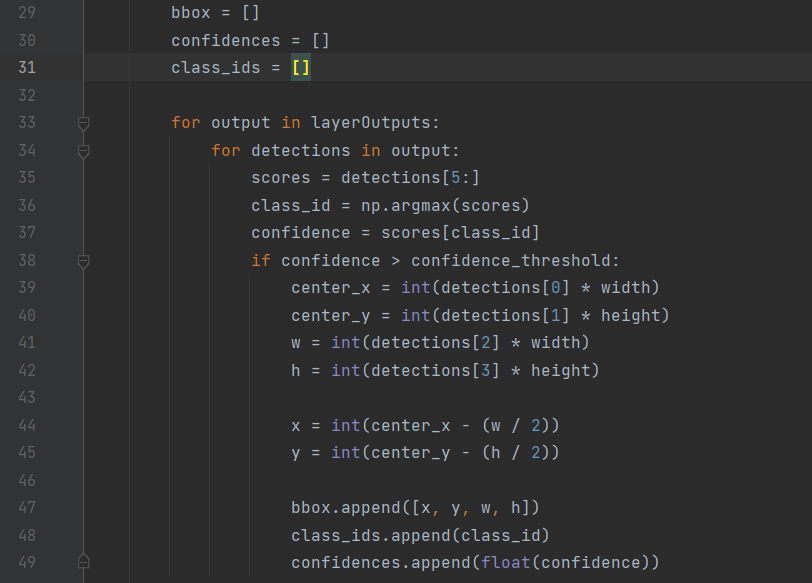
YOLOv2 gives state-of-the-art detection accuracy on the PASCAL VOC and COCO. It can run on varying sizes offering a trade-off between speed and accuracy. At 67 FPS, YOLOv2 can give an mAP of 76.8, while at 40 FPS, the detector provides an accuracy of 78.6 mAP, better than the state-of-the-model such as Faster R-CNN and SSD while running significantly faster than those models.

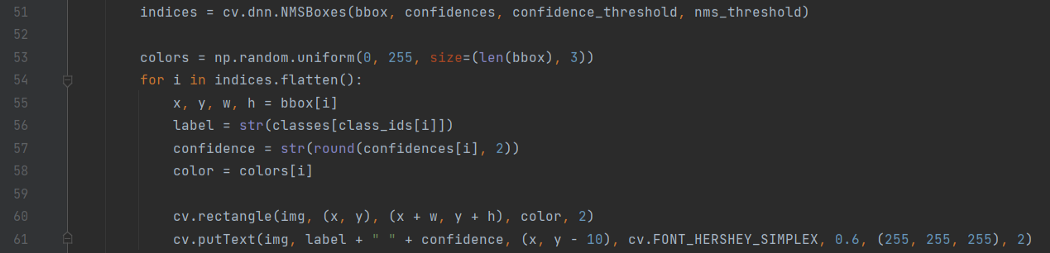


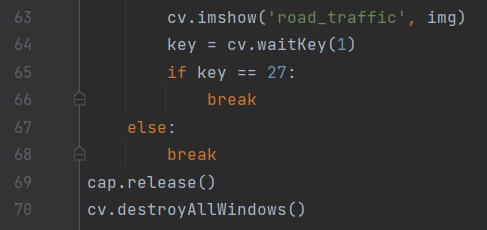
The codebase for implementation of the Yolo algorithm for object detection:











**Object Tracking using deep sort**

One of the most widely used objects tracking frameworks is Deep Sort, an extension to SORT (Simple Online Realtime Tracker). Deep Sort integrates appearance information to improve the performance of SORT. Due to this extension, able to track objects through more extended periods of occlusions, effectively reducing the number of identity switches.

Our tracking scenario is defined on the eight-dimensional state space (u, v, γ, h, x,˙ y,˙ γ, ˙ h˙) that contains the bounding box center position (u, v), aspect ratio γ, height h, and their respective velocities in image coordinates. Used a standard Kalman filter with constant velocity motion and linear observation model, where we take the bounding coordinates (u, v, γ, h) as direct observations of the object state.

We create a “Track” for each detection with all the necessary state information. It also has a parameter to track and delete tracks that had their last successful detection long back, as those objects would have left the scene. There is a minimum number of detections threshold for the first few frames to eliminate duplicate tracks.

Now the next problem is the assignment of new detection with new prediction because we have no idea how to associate track\_i with incoming detection\_k.

To solve this, two things were needed, A distance metric to quantify the association and an efficient algorithm to associate the data.

Deep SORT uses the squared Mahalanobis distance to incorporate the uncertainties from the Kalman filter. Thresholding this distance can give us an excellent idea of the actual associations. The standard Hungarian Algorithm is a very effective and straightforward combinatorial optimization algorithm that solves the assignment problem.

Deep sort also includes another distance metric based on the object’s appearance. The idea of obtaining a vector describing a given image’s features is quite simple. First, build a classifier over one dataset, train it till it achieves reasonably good accuracy, and then strip the final classification layer. Assuming a classical architecture, lastly, will be left with a dense layer producing a single feature vector, waiting to be classified.  
  
That feature vector becomes the “appearance descriptor” of the object.

The updated distance metric will be:

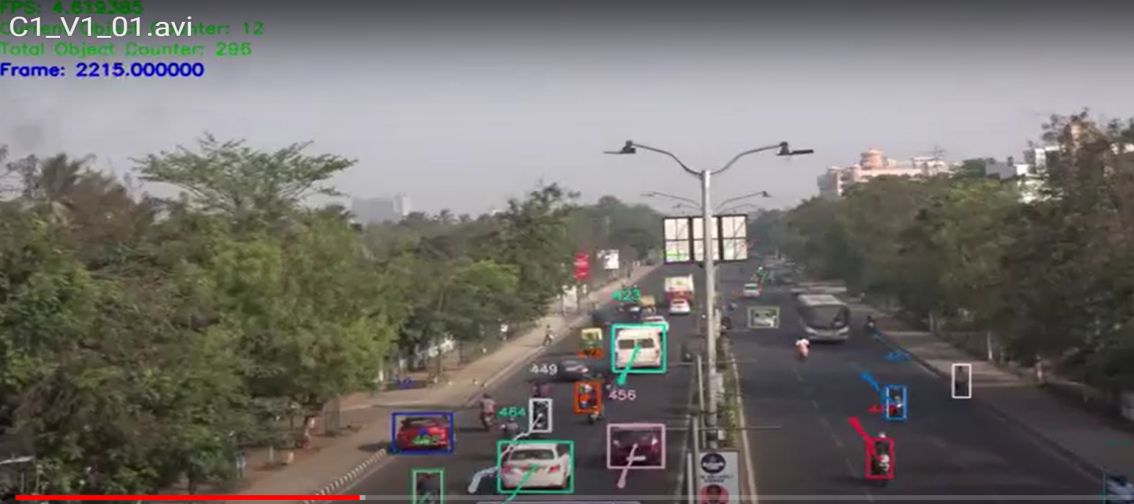
D = Lambda \* D\_k + (1 - Lambda) \* D\_a

Where D\_k is the Mahalanobis distance and D\_a is the cosine distance between the appearance feature vectors, and Lambda is the weighting factor.

**Dataset Preparation**

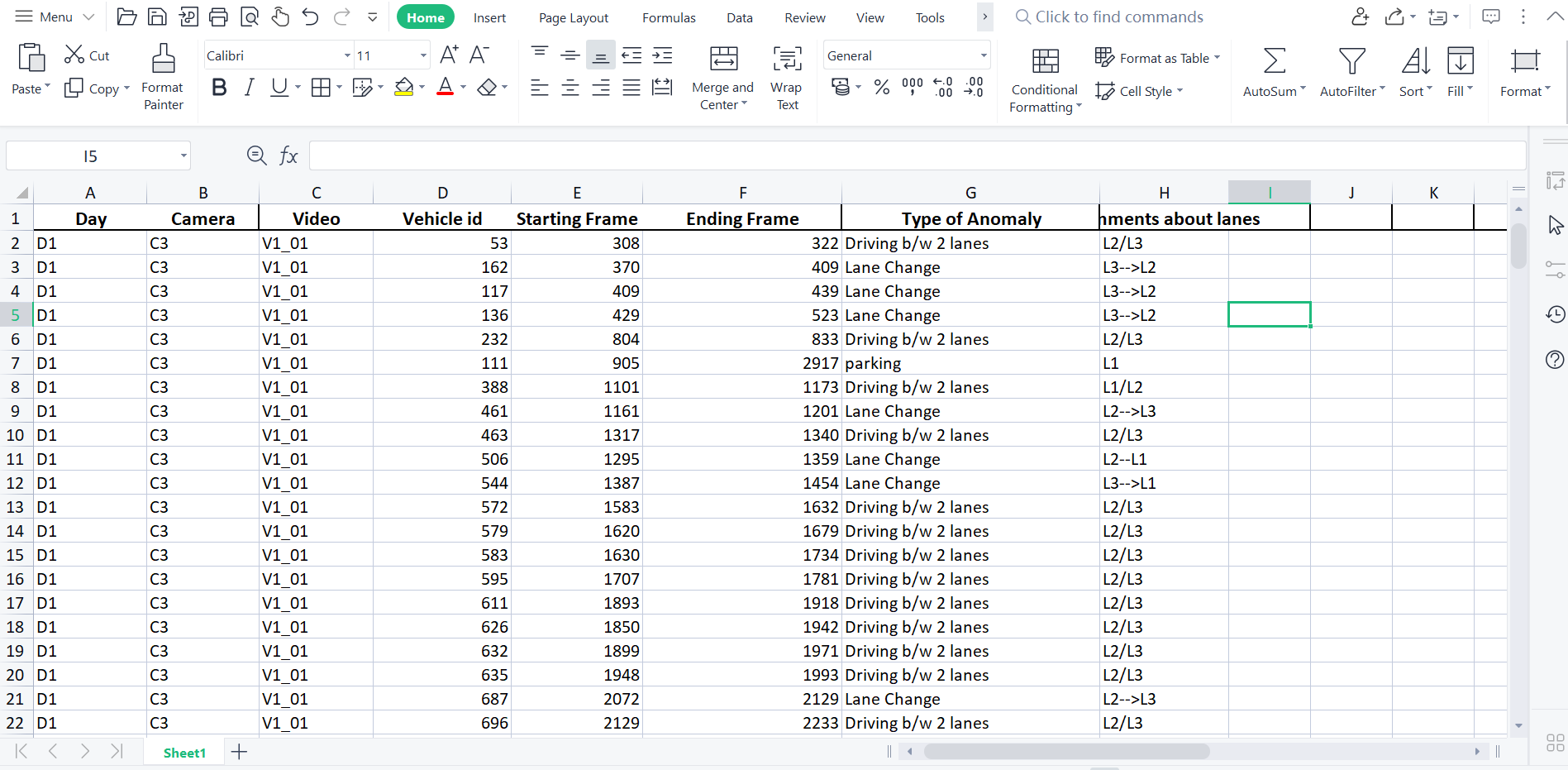
**Recordings-** First of all, for preparing the dataset, the live recordings of the most bustling roads were shot between the timeline of December 2021 to March 2022. Various CCTVs were set up at the front of Pal Heights, Bhubaneswar, and Kalinga Stadium, Bhubaneswar, Odisha. For recording the videos of traffic, Hikvision-PTZ Surveillance Camera was used. The camera was fixed on a 5 feet tripod placed on the footbridges with a height of approximately 20 feet above ground level. Thus, the recordings were taken at an elevation of 25 ft from the ground level under natural illumination conditions. The recordings were performed for different traffic phenomena. A total of 8-8.5 hours of video recordings was obtained at 25 frames per second for almost 3 to 4 days.

**Pre-Processing of Videos -** Video data recorded were stabilized to reduce camera shaking. Then, the stabilized videos were converted to a suitable format and resized for further use.[13]



**Labeling and Annotation of the videos –** After the video are converted then; the videos are annotated for labeling or tagging video clips to train the model for the detection of objects. Video annotation involves annotating objects on a frame-by-frame basis to make them recognizable. Labeling was done for all vehicles and as well as for pedestrians, identifying things (cars, pedestrians) doing unusual activities on the road and classifying them under different categories [11] . The labeled data will be used to train the classification model during abnormality detection, traffic congestion detection, and pedestrian and vehicle interaction to validate the model—a sample picture of the stabilized formatted and annotated videos.

Creating dataset: The annotated videos were observed, and anomalies were tracked down and performed manually by various vehicles captured in the videos. The time frames between which the monster was detected were noted and documented. The type of anomaly that can be classified is also determined manually so that the model can be trained accordingly. The documentation was done in an excel sheet where each camera detects the details of each anomaly on various days are recorded with valid classification. The types of anomalies into which the data is classified are lane changing, parking on the road, driving in the wrong lane, etc.



**Conclusion**

India is a vast country with more than 4000 cities and towns. Out of these 4000 cities, 300 cities have a population of more than 100K. As per the government’s smart city mission, 20 cities have been upgraded to smart cities within a specified time frame. Intelligent transportation [1] is one of the primary features of a smart city. City dwellers must be provided with efficient, risk-free, affordable transportation systems. However, the primary hurdle to making transportation safer in Indian cities arises due to the high density of vehicles and lack of awareness. For example, city dwellers often violate the rules of lane driving, speeding, wrong driving, crossing roads without following the rules, on-road parking, etc. Despite initiating projects at various levels, city traffic administrators still depend on the manual observations of CCTV. Moreover, users do not get a real-time update on the traffic conditions at multiple junctions and information about the parking spaces. This proposal aims to develop AI, and computer vision-guided traffic [3] monitoring systems with the following scientific objectives:

a) Development of artificial intelligence and computer vision-guided systems to detect road congestion in real-time, infer the cause of congestion, and provide the road users with information through variable message signs about the state congestion and suggesting alternate routes.

b) Often, city roads are connected via multiple junctions. A lack of coordination in the signaling systems leads to substantial traffic delays. The problem occurs due to independent signal handling at hubs. The proposal aims to develop machine learning algorithms by analyzing the videos at subsequent junctions with the help of flow detection and segmentation.[6]

c) Delivering the current status of the available parking places to the users in a real-time and request-based fashion is another objective. Users often search for the best public parking place in terms of distance and cost. The proposed system will be able to guide the users with real-time updates and cost-effective paths to reach the nearest available parking zones.

A loss of human lives. The project aims to intelligently identify the vulnerable road segments of a city to understand the reasons for pedestrian behavior, thus can help the administrators with necessary feedback that can be used for alternate solutions. The project can be highly effective as Bhubaneswar smart city, and IIT Bhubaneswar are collocated. Therefore, the objectives above can be achieved with time-bound in Bhubaneswar. The project can be scaled up for other smart cities subsequently.

Sarthak Siddhant Bharadwaj

APPLICATION NO.: INT/ES/SUM/2022/0255

**References**

1. B. Tian, B. T. Morris, M. Tang, Y. Liu, Y. Yao, C. Gou, D. Shen, and S. Tang. 2017. Hierarchical and networked vehicle surveillance in ITS: A survey. IEEE Transactions on Intelligent Transportation Systems 18, 1 (2017), 25--48.[](http://scholar.google.com/scholar?hl=en%26q=B.+Tian,+B.+T.+Morris,+M.+Tang,+Y.+Liu,+Y.+Yao,+C.+Gou,+D.+Shen,+and+S.+Tang.+2017.+Hierarchical+and+networked+vehicle+surveillance+in+ITS:+A+survey.+IEEE+Transactions+on+Intelligent+Transportation+Systems+18,+1+(2017),+25--48.)[](https://dl.acm.org/doi/10.1109/TITS.2016.2552778)
2. B. T. Morris and M. M. Trivedi. 2008. A survey of vision-based trajectory learning and analysis for surveillance. IEEE Transactions on Circuits and Systems for Video Technology 18, 8 (2008), 1114--1127.[](http://scholar.google.com/scholar?hl=en%26q=B.+T.+Morris+and+M.+M.+Trivedi.+2008.+A+survey+of+vision-based+trajectory+learning+and+analysis+for+surveillance.+IEEE+Transactions+on+Circuits+and+Systems+for+Video+Technology+18,+8+(2008),+1114--1127.)[](https://dl.acm.org/doi/10.1109/TCSVT.2008.927109)
3. S. Vishwakarma and A. Agrawal. 2013. A survey on activity recognition and behavior understanding in video surveillance. The Visual Computer 29, 10 (2013), 983--1009.[](http://scholar.google.com/scholar?hl=en%26q=S.+Vishwakarma+and+A.+Agrawal.+2013.+A+survey+on+activity+recognition+and+behavior+understanding+in+video+surveillance.+The+Visual+Computer+29,+10+(2013),+983--1009.)[IMG_263](https://dl.acm.org/servlet/linkout?suffix=e_1_2_1_4_1%26dbid=16%26doi=10.1145/3417989%26key=10.1007/s00371-012-0752-6)
4. S. Sivaraman and M. M. Trivedi. 2013. Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis. IEEE Transactions on Intelligent Transportation Systems 14, 4 (2013), 1773--1795.[](http://scholar.google.com/scholar?hl=en%26q=S.+Sivaraman+and+M.+M.+Trivedi.+2013.+Looking+at+vehicles+on+the+road:+A+survey+of+vision-based+vehicle+detection,+tracking,+and+behavior+analysis.+IEEE+Transactions+on+Intelligent+Transportation+Systems+14,+4+(2013),+1773--1795.)[](https://dl.acm.org/doi/10.1109/TITS.2013.2266661)
5. M. S. Shirazi and B. T. Morris. 2017. Looking at intersections: A survey of intersection monitoring, behavior, and safety analysis of recent studies. IEEE Transactions on Intelligent Transportation Systems 18, 1 (2017), 4--24.[](http://scholar.google.com/scholar?hl=en%26q=M.+S.+Shirazi+and+B.+T.+Morris.+2017.+Looking+at+intersections:+A+survey+of+intersection+monitoring,+behavior+and+safety+analysis+of+recent+studies.+IEEE+Transactions+on+Intelligent+Transportation+Systems+18,+1+(2017),+4--24.)[](https://dl.acm.org/doi/10.1109/TITS.2016.2568920)
6. A. B. Mabrouk and E. Zagrouba. 2018. Abnormal behavior recognition for intelligent video surveillance systems: A review. Expert Systems with Applications 91 (2018), 480--491.[](http://scholar.google.com/scholar?hl=en%26q=A.+B.+Mabrouk+and+E.+Zagrouba.+2018.+Abnormal+behavior+recognition+for+intelligent+video+surveillance+systems:+A+review.+Expert+Systems+with+Applications+91+(2018),+480--491.)[](https://dl.acm.org/doi/10.1016/j.eswa.2017.09.029)
7. S. Omar, A. Ngadi, and H. H. Jebur. 2013. Machine learning techniques for anomaly detection: An overview. International Journal of Computer Applications 79, 2 (2013).[](http://scholar.google.com/scholar?hl=en%26q=S.+Omar,+A.+Ngadi,+and+H.+H.+Jebur.+2013.+Machine+learning+techniques+for+anomaly+detection:+An+overview.+International+Journal+of+Computer+Applications+79,+2+(2013).)[IMG_271](https://dl.acm.org/servlet/linkout?suffix=e_1_2_1_8_1%26dbid=16%26doi=10.1145/3417989%26key=10.5120/13715-1478)
8. X. Li and Z.-M. Cai. 2016. Anomaly detection techniques in surveillance videos. In International Congress on Image and Signal Processing, BioMedical Engineering and Informatics. IEEE, 54--59.[](http://scholar.google.com/scholar?hl=en%26q=X.+Li+and+Z.-M.+Cai.+2016.+Anomaly+detection+techniques+in+surveillance+videos.+In+International+Congress+on+Image+and+Signal+Processing,+BioMedical+Engineering+and+Informatics.+IEEE,+54--59.)[IMG_273](https://dl.acm.org/servlet/linkout?suffix=e_1_2_1_9_1%26dbid=16%26doi=10.1145/3417989%26key=10.1109/CISP-BMEI.2016.7852681)
9. K. Yun, H. Jeong, K. M. Yi, S. W. Kim, and J. Y. Choi. 2014. Motion interaction field for accident detection in traffic surveillance video. In International Conference on Pattern Recognition. IEEE, 3062--3067.[](http://scholar.google.com/scholar?hl=en%26q=K.+Yun,+H.+Jeong,+K.+M.+Yi,+S.+W.+Kim,+and+J.+Y.+Choi.+2014.+Motion+interaction+field+for+accident+detection+in+traffic+surveillance+video.+In+International+Conference+on+Pattern+Recognition.+IEEE,+3062--3067.)
10. D. Pathak, A. Sharang, and A. Mukerjee. 2015. Anomaly localization in the topic-based analysis of surveillance videos. In Winter Conference on Applications of Computer Vision. IEEE, 389--395.[](http://scholar.google.com/scholar?hl=en%26q=D.+Pathak,+A.+Sharang,+and+A.+Mukerjee.+2015.+Anomaly+localization+in+topic-based+analysis+of+surveillance+videos.+In+Winter+Conference+on+Applications+of+Computer+Vision.+IEEE,+389--395.)
11. W. Sultani, C. Chen, and M. Shah. 2018. Real-world anomaly detection in surveillance videos. In Conference on Computer Vision and Pattern Recognition. IEEE, 6479--6488.[](http://scholar.google.com/scholar?hl=en%26q=W.+Sultani,+C.+Chen,+and+M.+Shah.+2018.+Real-world+anomaly+detection+in+surveillance+videos.+In+Conference+on+Computer+Vision+and+Pattern+Recognition.+IEEE,+6479--6488.)
12. S. Lee, H. G. Kim, and Y. M. Ro. 2018. STAN: Spatio-temporal adversarial networks for abnormal event detection. In International Conference on Acoustics, Speech and Signal Processing. IEEE, 1323--1327.[](http://scholar.google.com/scholar?hl=en%26q=S.+Lee,+H.+G.+Kim,+and+Y.+M.+Ro.+2018.+STAN:+Spatio-temporal+adversarial+networks+for+abnormal+event+detection.+In+International+Conference+on+Acoustics,+Speech+and+Signal+Processing.+IEEE,+1323--1327.)
13. B. Tian, Q. Yao, Y. Gu, K. Wang, and Y. Li. 2011. Video processing techniques for traffic flow monitoring: A survey. In Conference on Intelligent Transportation Systems. IEEE, 1103--1108.[](http://scholar.google.com/scholar?hl=en%26q=B.+Tian,+Q.+Yao,+Y.+Gu,+K.+Wang,+and+Y.+Li.+2011.+Video+processing+techniques+for+traffic+flow+monitoring:+A+survey.+In+Conference+on+Intelligent+Transportation+Systems.+IEEE,+1103--1108.)