Portable Multi-Sensor System for Intersection Safety Performance Assessment

Final Report
July 2018

ctre
Center for Transportation Research and Education

Iowa State University
Institute for Transportation

Sponsored by
Iowa Department of Transportation
(InTrans Project 17-625)
National Highway Traffic Safety Administration

About InTrans and CTRE

The mission of the Institute for Transportation (InTrans) and Center for Transportation Research and Education (CTRE) at Iowa State University is to develop and implement innovative methods, materials, and technologies for improving transportation efficiency, safety, reliability, and sustainability while improving the learning environment of students, faculty, and staff in transportation-related fields.

Disclaimer Notice

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. The opinions, findings and conclusions expressed in this publication are those of the authors and not necessarily those of the sponsors.

The sponsors assume no liability for the contents or use of the information contained in this document. This report does not constitute a standard, specification, or regulation.

The sponsors do not endorse products or manufacturers. Trademarks or manufacturers' names appear in this report only because they are considered essential to the objective of the document.

ISU Non-Discrimination Statement

Iowa State University does not discriminate on the basis of race, color, age, ethnicity, religion, national origin, pregnancy, sexual orientation, gender identity, genetic information, sex, marital status, disability, or status as a U.S. veteran. Inquiries regarding non-discrimination policies may be directed to Office of Equal Opportunity, 3410 Beardshear Hall, 515 Morrill Road, Ames, Iowa 50011, Tel. 515 294-7612, Hotline: 515-294-1222, email eooffice@iastate.edu.

Iowa DOT Statements

Federal and state laws prohibit employment and/or public accommodation discrimination on the basis of age, color, creed, disability, gender identity, national origin, pregnancy, race, religion, sex, sexual orientation or veteran's status. If you believe you have been discriminated against, please contact the Iowa Civil Rights Commission at 800-457-4416 or the Iowa Department of Transportation affirmative action officer. If you need accommodations because of a disability to access the Iowa Department of Transportation's services, contact the agency's affirmative action officer at 800-262-0003.

The preparation of this report was financed in part through funds provided by the Iowa Department of Transportation through its "Second Revised Agreement for the Management of Research Conducted by Iowa State University for the Iowa Department of Transportation" and its amendments.

The opinions, findings, and conclusions expressed in this publication are those of the authors and not necessarily those of the Iowa Department of Transportation.

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
InTrans Project 17-625		

4. Title and Subtitle	5. Report Date
Portable Multi-Sensor System for Intersection Safety Performance Assessment	July 2018
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Anuj Sharma (orcid.org/0000-0001-5929-5120), Pranamesh Chakraborty (orcid.org/0000-0003-2624-5543), Shuo Wang (orcid.org/0000-0003-1016-7229), and Tongge Huang (orcid.org/0000-0002-5534-4623)	InTrans Project 17-625

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Center for Transportation Research and Education Iowa State University 2711 South Loop Drive, Suite 4700 Ames, IA 50010-8664	11. Contract or Grant No.

12. Sponsoring Organization Name and Address		13. Type of Report and Period Covered
Iowa Department of Transportation 800 Lincoln Way Ames, IA 50010	National Highway Traffic Safety Administration 1200 New Jersey Avenue, SE Washington, DC 20590	Final Report **14. Sponsoring Agency Code** Section 405C National Priority Safety Program - State Traffic Safety Information System Improvements

15. Supplementary Notes

Visit www.intrans.iastate.edu for color pdfs of this and other research reports.

16. Abstract

State departments of transportation (DOTs) and city municipal agencies install a large number of roadside cameras on freeways and arterials for surveillance tasks. It is estimated that there will be approximately a billion cameras worldwide by 2020. However, most of these cameras are used for manual surveillance purposes only. The main objective of this study was to investigate the use of these cameras as a sensor for traffic state estimation.

The scope of this project involved detecting vehicles, tracking them, and estimating their speeds. The research team adopted a tracking-by-detection framework for this study. The object detection task was performed using you only look once version 3 (YOLOv3) model architecture and the tracking was performed using the simple online and realtime tracking (SORT) algorithm. The team tested the framework on videos collected from three intersections in Ames, Iowa. The combined detection and tracking was performed at approximately 40 frames per second (fps) using GeForce GTX 1080 GPU, enabling it to be implemented online easily.

Camera calibration was performed by finding the edges of moving vehicles to automatically detect the vanishing points, and the scale factor was determined manually from a known fixed distance in the image and the real world. Although this methodology performed vanishing point determination automatically without any manual intervention, the speed estimation error came out to be quite high (~13 mph). The error can be reduced significantly by performing calibration and scale factor determination fully manually. However, since it requires full manual intervention, it is difficult to scale the algorithm across multiple cameras.

In the future, the detection task can be improved by training the model on a larger dataset, and further work can be done to improve speed estimation by extending automatic camera calibration to automatic scale estimation, which would improve accuracy simultaneously.

17. Key Words	18. Distribution Statement
deep learning—intersection safety—object detection—roadside camera surveillance—vehicle speed estimation—vehicle tracking	No restrictions.

19. Security Classification (of this report)	20. Security Classification (of this page)	21. No. of Pages	22. Price
Unclassified.	Unclassified.	44	NA

PORTABLE MULTI-SENSOR SYSTEM FOR INTERSECTION SAFETY PERFORMANCE ASSESSMENT

Final Report
July 2018

Principal Investigator
Anuj Sharma, Research Scientist
Center for Transportation Research and Education, Iowa State University

Co-Principal Investigators
Shauna Hallmark, Director
Institute for Transportation, Iowa State University
Peter Savolainen, Safety Engineer
Center for Transportation Research and Education, Iowa State University

Research Assistant
Pranamesh Chakraborty

Authors
Anuj Sharma, Pranamesh Chakraborty, Shuo Wang, and Tongge Huang

Sponsored by
Iowa Department of Transportation

Preparation of this report was financed in part
through funds provided by the Iowa Department of Transportation
through its Research Management Agreement with the
Institute for Transportation
(InTrans Project 17-625)

A report from
Institute for Transportation
Iowa State University
2711 South Loop Drive, Suite 4700
Ames, IA 50010-8664
Phone: 515-294-8103 / Fax: 515-294-0467
www.intrans.iastate.edu

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The research team would like to acknowledge the Iowa Department of Transportation for sponsoring this work using funding from the National Highway Traffic Safety Administration's Section 405C National Priority Safety Program for State Traffic Safety Information System Improvements.

EXECUTIVE SUMMARY

The number of Internet-connected cameras are increasing at a rapid pace. State departments of transportation (DOTs) and city municipal agencies install a large number of roadside cameras on freeways and arterials for surveillance tasks. The main objective of this study was to investigate the use of these cameras as a sensor for traffic state estimation.

The scope of this project involved detecting vehicles, tracking them, and estimating their speeds. The research team adopted a tracking-by-detection framework for this study. The object detection task was performed using you only look once version 3 (YOLOv3) model architecture and the tracking was performed using the simple online and realtime tracking (SORT) algorithm.

The team tested the framework on videos collected from three intersections in Ames, Iowa. The combined detection and tracking was performed at approximately 40 frames per second (fps) using GeForce GTX 1080 GPU, enabling it to be implemented online easily.

Camera calibration was performed by finding the edges of moving vehicles to automatically detect the vanishing points, and the scale was determined manually from a known fixed distance in the image and the real world. Although this methodology performed vanishing point determination automatically without any manual intervention, the speed estimation error came out to be quite high (~13 mph). The error can be reduced significantly by performing calibration and scale factor determination fully manually. However, since it requires full manual intervention, it is difficult to scale the algorithm across multiple cameras.

In the future, the detection task can be improved by training the model on a larger dataset. Specifically, the University at Albany's detection and tracking (UA-DETRAC) dataset can be used in the future to improve detection results. Tracking performance can be improved in the future by using Deep SORT or similar tracking algorithms that use appearance descriptions for tracking purposes. This can help in reducing the number of identity switches. Speed estimation can be improved in the future by extending automatic camera calibration to automatic scale estimation, which would also improve accuracy simultaneously.

CHAPTER 1. INTRODUCTION

The number of Internet-connected cameras is increasing at a rapid pace. It is expected that there will be a billion cameras worldwide by 2020. This breakthrough has great potential for making cities smarter and safer (Naphade et al. 2011). However, monitoring capability hasn't improved at the same pace. Most of the cameras installed on freeways and arterials are still used for manual surveillance purposes only.

State departments of transportation (DOTs) and city municipal agencies install a large number of roadside cameras for surveillance tasks like incident detection. These cameras are used by traffic incident managers (TIMs) who can zoom, tilt, and pan the cameras according to their needs. The objective of this project was to study the scope of using these cameras as a sensor system and use it for traffic state estimation or automatic surveillance purposes.

Traditionally, traffic state estimation is done using point-based sensors, which include inductive loops, piezoelectric sensors, and magnetic loops (Kotzenmacher et al. 2004). With the recent advances in active infrared/laser, radar sensors, these devices are gradually replacing the traditional point-based sensors (Zhong and Liu 2007). Also, with the increasing usage of navigation-based global positioning system (GPS) devices, probe-based data are emerging as a cost-effective way to collect network-wide traffic data (Feng et al. 2014).

Video monitoring and surveillance systems can also be used for calculating real-time traffic data (Ozkurt and Camci 2009). Recent advances in image processing techniques have improved vision-based detection accuracy. Deep learning methods like convolutional neural networks (CNNs) have been able to achieve human-level accuracy in image classification tasks (He et al. 2016).

The basic advantage of these methods are they don't require picking up hand-crafted features and, hence, can do away with the painstaking calibration tasks in using camera images for traffic state estimation (Bauza et al. 2010). The cameras can be used for various estimation or detection of various transportation-based problems such as the following:

- Traffic state determination (e.g., speed, volume, and occupancy)
- Queue length determination
- Origin-destination matrix estimation using vehicle re-identification
- Traffic conflict determination (vehicle-vehicle, vehicle-person)

Figure 1 gives a brief overview of the key steps in using cameras for traffic monitoring.

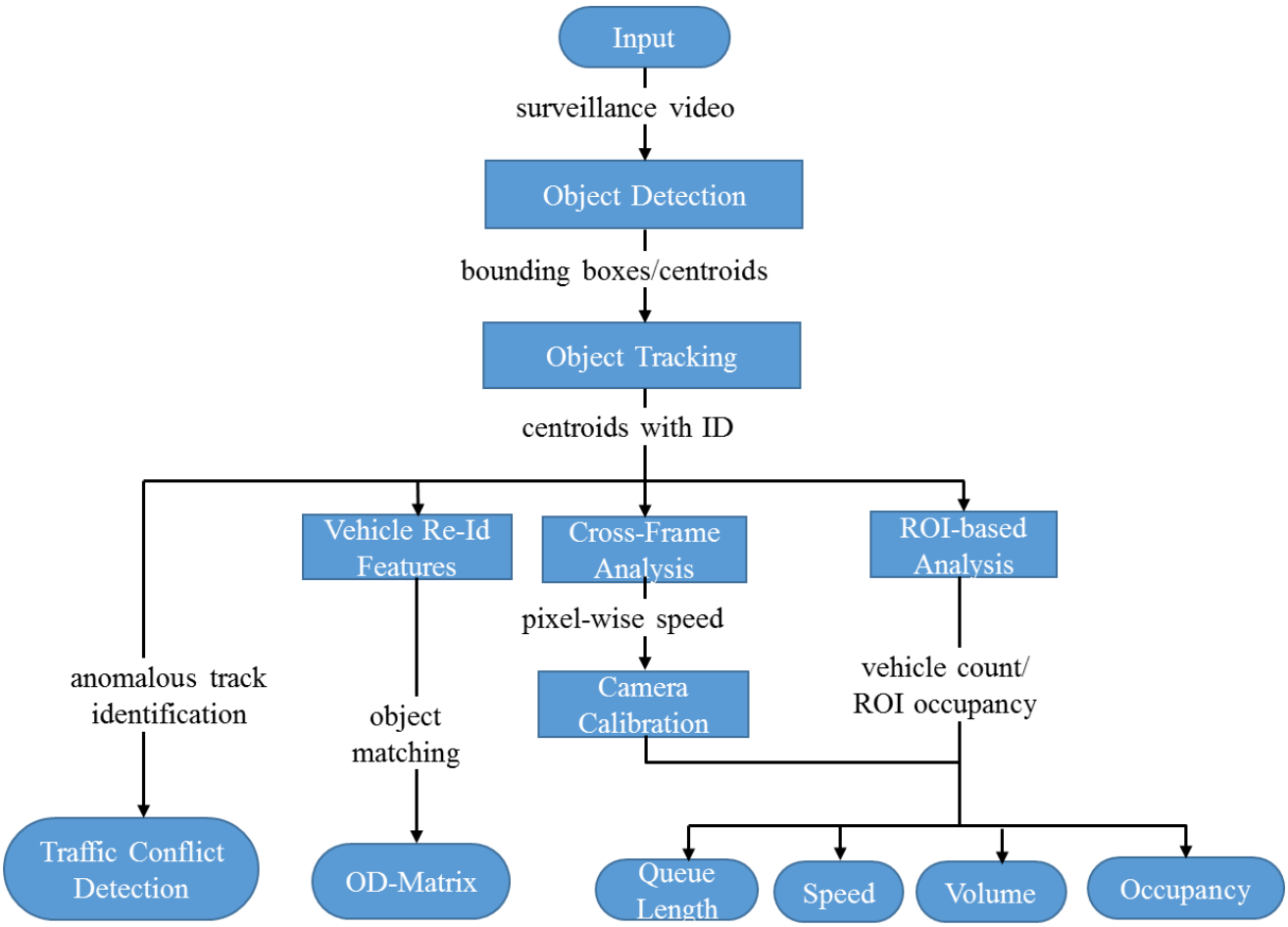


Figure 1. Overview of the key steps in using cameras as a sensor

The two main components involve object detection (vehicles and persons) and object tracking. The tracked objects can be directly used to find traffic conflicts using the trajectory information of the vehicles and people. The objects detected across multiple cameras can be used to determine the re-identification features and match the vehicles to find the origin-destination matrix.

Traffic state determination generally requires setting up the region of interest (ROI) first and then determining the camera calibration parameters. This information can be processed to determine the speed, volume, occupancy, and queue length, if present.

In this study, the research team performed proof-of-concept object detection, tracking, and camera calibration to determine the speed of all vehicles and, hence, the traffic volume within the camera field. Some preliminary analysis on anomalous trajectory identification was also performed. State-of-the-art object detection and tracking algorithms were adopted keeping in mind the necessity of real-time processing capabilities of the videos.

The next chapter provides an overview of the research done in object detection, tracking, and camera calibration. Chapter 3 provides the details of the data collected and used in this study, followed by the results obtained in Chapter 4. The final chapter provides the conclusions from this study and the future work that needs to be done.

CHAPTER 2. LITERATURE REVIEW

Significant research has been performed in the fields of object detection, tracking, and camera calibration. This chapter gives a brief overview of the research performed in each of these fields.

2.1. Object Detection

In recent years, the evolution of CNNs have resulted in significant improvements in object detection and classification performance. Results of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) point to dramatic improvements in object detection, localization, and classification (Russakovsky et al. 2015).

Regions with convolutional neural networks (R-CNNs) were among the first modern developments of CNN-based detection (Girshick et al. 2014). These developments involved cropping externally computed box proposals from an input image and running a neural-net classifier on these crops. However, overlapping crops led to significant duplicate computations, which, in turn, led to low processing speed. Fast R-CNNs involved pushing the entire input image only once through a feature extractor and cropping from an intermediate layer (Girshick 2015). This led to the crops sharing the feature extraction computation load and thereby increased processing speed.

Recent works have focused toward generating box proposals using neural networks instead of the external box proposals used in R-CNN and Fast R-CNN (Szegedy et al. 2013, Erhan et al. 2014, Ren et al. 2017, Redmon et al. 2016). These approaches involve overlaying a collection of boxes on the image at different locations, aspect ratios, and scales. These boxes are called anchors or priors.

Training is then performed to predict the discrete class of each anchor and the offset by which the anchor needs to be shifted to fit the ground truth bounding box. The accuracy and computation time of the object detection algorithm depends significantly on the choice of these anchors.

The following sections discuss the four such recent architectures: Faster R-CNN (Ren et al. 2017), single shot detector (SSD) (W. Liu et al. 2016), region-based fully convolutional networks (R-FCNs) (Dai et al. 2016), and you only look once (YOLO) (Redmon et al. 2016).

2.1.1 Faster R-CNN

Faster R-CNN performs detection in two stages. Stage 1, called the region proposal network (RPN), involves processing images by a feature extractor (VGG-16), and the class agnostic box proposals are predicted from the features obtained at some selected intermediate level (conv5).

In Stage 2, features from the same intermediate feature map are extracted using the box proposals and fed to the remainder of the feature extractor to predict the class and the class-

specific box refinement for each proposal. Faster R-CNN forms the basis on which most of the future object detection algorithms, including SSD and R-FCN, are developed.

2.1.2. SSD

SSD architecture is built on VGG-16 architecture. It uses a single feed-forward convolutional network to predict classes and anchor offsets, thereby evading the requirement for a second stage per-proposal classification operation.

In this approach, the output space of bounding boxes is discretized into a set of default boxes with different object scales and aspect ratios. During prediction, scores for presence of an object in each default box is generated by the network and, finally, adjustments are made to the box to match the object shape more accurately.

2.1.3. R-FCN

R-FCN is fundamentally derived from Faster R-CNN, but it is designed to work much faster than Faster-RCNN. In R-FCN, crops are extracted from the last layer of features prior to prediction instead of cropping features from the layer where region proposals are predicted. This minimizes the per-region computation and has shown to achieve comparable accuracy to Faster R-CNN with less computation time.

Previous research studies have proposed a position-sensitive cropping mechanism in place of the standard ROI pooling operation (Ren et al. 2017). A detailed comparison of these three algorithms (Faster R-CNN, SSD, and R-FCN), along with the speed-accuracy tradeoffs, can be found in a study by Huang et al. (2017).

2.1.4. YOLO

YOLO frames object detection as a regression problem (Redmon et al. 2016). A single neural network is used to predict the bounding boxes and associated class probabilities in a single evaluation over the entire image. Thus, the entire pipeline can be optimized end-to-end based on detection performance. This makes the algorithm very fast and images can be processed in real-time (45 frames per second [fps]). A detailed description of the YOLO model is provided in the Methodology chapter.

2.2. Multi-Object Tracking

Multi-object tracking (MOT) aims to estimate the states of multiple objects conserving their identification across time under motion and appearance variations. This involves determining the locations, velocities, and sizes of the objects across time. With the recent advancements in object detection, tracking-by-detection has emerged to be one of the predominant approaches for multi-object tracking. This generally involves associating the objects detected across multiple frames

in a video sequence. The two broad categories in a tracking-by-detection framework are batch and online tracking.

Batch methods usually involve determining object trajectories in a global optimization problem and processing the entire video at once. Short tracklets are generated; first, they start linking individual detections, and then, the tracklets are associated globally to form the object trajectory. Flow network formulations (Zhang et al. 2008, Berclaz et al. 2011) and probabilistic graphical models (Yang and Nevatia 2012, Andriyenko et al. 2012) are the two broad classes of algorithms in a batch MOT problem. However, the intensive iterative computation for generating globally associated tracks and the need for detection of the entire sequence beforehand limits the usage of these batch MOT approaches in real-time applications.

Online methods build trajectories sequentially by using information provided up to the present frame and associating the frame-by-frame objects detected. Thus, this approach can be easily implemented for real-time tracking. However, it makes these methods prone to fragmented trajectory generation under occlusion and object detection errors.

Traditional online MOT methods are multiple hypothesis tracking (MHT) (Reid 1979, Kim et al. 2015) and joint probabilistic data association filter (JPDAF) (Fortmann et al. 1983, Rezatofighi et al. 2015). The JPDAF method involves generating a single state hypothesis by weighting individual measurements with the association likelihoods. MHT, on the other hand, involves tracking all possible hypotheses, and then, applying pruning schemes for computational tractability. Both of these approaches require increased computational and implementation complexity, thereby limiting their real-time implementation.

Recently, Bewley et al. (2016) proposed simple online realtime tracking (SORT), which performs Kalman filtering in the image space and the Hungarian algorithm for frame-by-frame data associations. SORT ranks higher than MHT in the MOT Challenge dataset (Leal-Taixe et al. 2015) with a state-of-the-art object detection framework (Ren et al. 2017). However, SORT is known to perform poorly when state estimation uncertainty is high and is known to return substantially high identity switches.

To overcome this shortcoming, Wojke et al. (2017) proposed the Deep-SORT algorithm, which uses both motion and appearance information into the association metric, which increases robustness against occlusions or detection errors. Even more recently, Bae and Yoon (2018) proposed a robust online MOT method that uses confidence-based data association for handling track fragmentation and deep appearance learning for handling similar object appearance in tracklet association.

In this study, the research team used the SORT method for tracking purposes, primarily because of its simple framework and fast computation, which make it extremely suitable for real-time purposes. In the future, Deep SORT or other advanced algorithms can be implemented and relative improvements in tracking can be determined.

2.3. Camera Calibration

Camera calibration is the process of estimating camera parameters to map the pixel points in camera coordinates into real-world coordinates. This includes dealing with perspective projection, camera rotations, scene scale (distance of camera from the ground plane), and possible tangential and radial distortion. Mathematically, the camera calibration model can be represented as $P = K\,[RT]$, where K denotes intrinsic camera parameters, R denotes camera rotation, and T denotes camera translation. Rotation and translation, the extrinsic parameters, are relative to the world coordinate system defined.

Another alternative method of camera calibration is based on vanishing points of the road plane, which can be easily converted to the standard model. An attribute of camera calibration is the degree of automation, meaning whether the algorithm requires any manual input for each camera. This is particularly important for the scalability issue when the number of cameras installed grows significantly. Another important attribute is whether the algorithm can work from any arbitrary viewpoint or if it requires specific placement with respect to the road. Significant research has been done in this camera calibration domain keeping in mind the above challenges and requirements.

He and Yung (2007) proposed camera calibration based on the calibration pattern formed by road lane markings. Vehicle location on the ground plane is extracted using the shadows cast by rear vehicle bumpers. Various studies have been done using the detected line markings to determine the vanishing points (Cathey and Dailey 2005, Grammatikopoulos et al. 2005, You and Zheng 2016). Scene scale was determined from the average line marking stripe length and real-world stripe length or any other known dimensions in the world.

Camera calibration also has been performed using vehicle movement information. Schoepflin and Dailey (2002) and Dubská et al. (2015) obtained lane boundaries by detecting the vehicles and using the vehicle edges to determine the first vanishing point. The second vanishing point was determined by the intersection of lines formed by the bottom edge of the vehicles. Filipiak et al. (2016) used the license plates on vehicles to determine camera calibration parameters. The scene scale factor can be determined manually using the fixed length on roads (Schoepflin and Dailey 2002) or by fully automatically using the detected vehicle dimension distribution (Dubská et al. 2014). A detailed comparison of different camera calibration algorithms can be found in Sochor et al. 2017.

Thus, significant research has been done in using cameras to detect objects, track them, and use them for speed estimation. Also, various other traffic analyses (traffic conflict detection, origin-destination matrix generation, etc.) can be performed once vehicles are detected and tracked accurately.

In this study, the objective was to study the scope of using cameras as a sensor for traffic speed estimation. The next chapter provides a description of the methodology adopted for each step of speed estimation (object detection, tracking, calibration, and speed estimation).

CHAPTER 3. METHODOLOGY

The primary task in using cameras as a sensor was to detect objects and track them accurately. In this study, the research team adopted a tracking-by-detection framework to perform object tracking. In this framework, objects were detected in each frame. Then, detections from the current frame and previous frames were presented to the tracker.

The main advantage of this method is it can utilize the recent progress in object detection tasks using deep-learning techniques to detect each object correctly and, thereby, also perform tracking. Also, since this framework uses only the current and previous frames for tracking purposes, it can be easily implemented online.

The detection task was performed using the YOLO version 3 (YOLOv3) algorithm, and the tracking was performed using SORT. Details of each algorithm are presented next.

3.1. Vehicle Detection: You Only Look Once (YOLO)

The research team adopted the YOLO model for vehicle detection and localization from videos (Redmon et al. 2016). Current object detection systems repurpose powerful CNN classifiers to perform detection. For example, to detect an object, these systems take a classifier for that object and evaluate it at various locations and scales in the test image. YOLO reframes object detection: instead of looking at a single image a thousand times to do detection, it only looks at an image once (but in a clever way) to perform the full detection pipeline (see Figure 2).

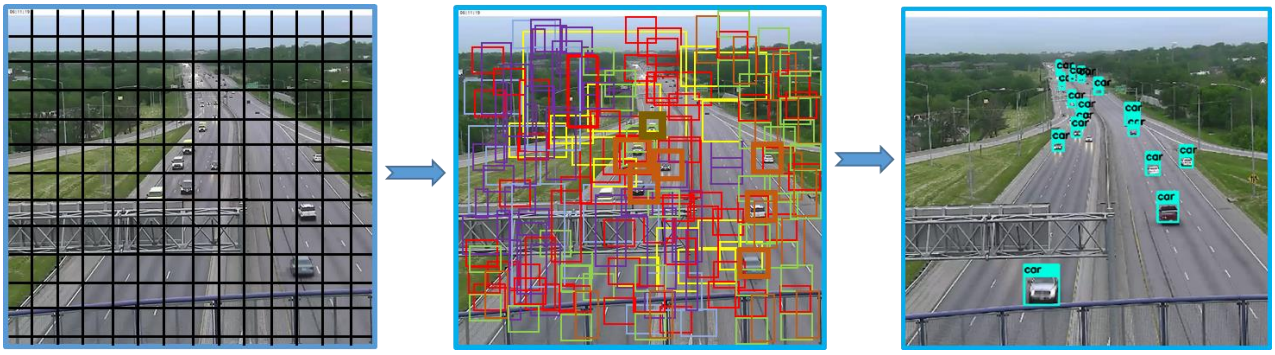


Figure 2. Confidence score prediction of bounding boxes by YOLO with colors and bounding box widths indicating confidence score probabilities

A single convolutional network simultaneously predicts multiple bounding boxes and class probabilities for those boxes. This makes YOLO extremely fast and easy to generalize to difference scenes.

In this study, the research team chose YOLOv3 for vehicle detection primarily because of its fast performance with reasonable accuracy, which makes it suitable for real-time performance (Redmon and Farhadi 2018). Specifically, the researchers used the YOLOv3-416 model trained on the Microsoft Common Objects in Context (COCO) dataset (Lin et al. 2014). The team chose

the classes of person, car, motorbike, bus, truck, and traffic light from the 80 classes in the COCO dataset for its vehicle detection module.

3.2. Multi-Object Tracking: Simple Online and Realtime Tracking (SORT)

The next task was to perform object tracking using the detection results obtained from Step 1 i.e., vehicle detection. To perform the multi-object tracking task, the team adopted the SORT algorithm formulated by Bewley at al. 2016.

This framework uses the traditional Kalman filter and Hungarian algorithm to perform tracking using the frame-by-frame object detection results. However, SORT only uses object location information to perform the tracking task and does not use any appearance descriptor. Therefore, it suffers from a significantly high number of identity switches due to occlusion issues or poor detection performance.

In the future, researchers can use the Deep-SORT algorithm, which solves the issue of identity switching using the appearance descriptor (Wojke et al. 2017). To perform the training task for the Deep-SORT algorithm, researchers can use the University at Albany's detection and tracking (UA-DETRAC) benchmark dataset (Wen et al. 2015) and vehicle re-identification (Ve-Ri) dataset (X. Liu et al. 2016). The Deep-SORT algorithm also involves determination of the appearance feature vector, which can be used for vehicle re-identification purposes and to determine the origin-destination matrix.

3.3. Camera Calibration and Scale Estimation

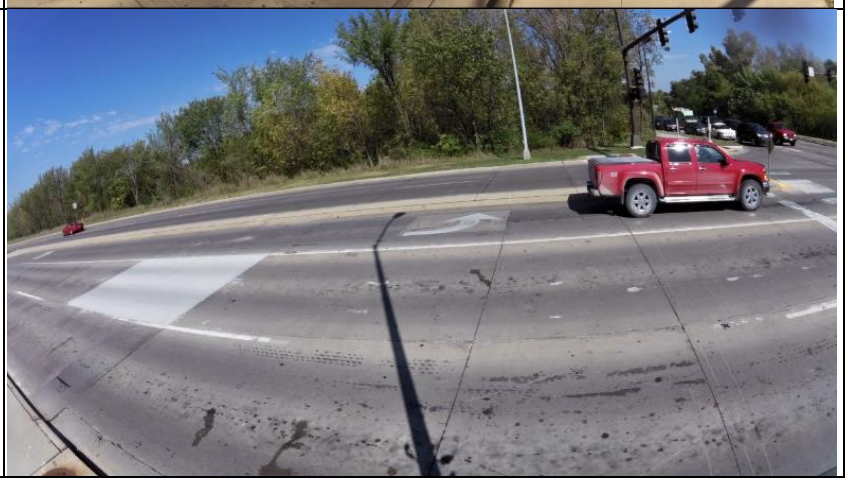
After obtaining the vehicle trajectories, the next step was to convert the displacement in pixels to actual displacement in the field. This was done by camera calibration and scaling. The research team adopted the methodology provided by Dubská et al. (2015). The ground plane position and vehicle movement direction relative to the camera were defined as described by three vanishing points (VPs). The first VP was parallel to the movement of vehicles. The second VP, perpendicular to the first VP, corresponded to the direction parallel to the road (or the ground plane). Vehicle edges and their movements across frames were used to determine the two VPs. The third VP, perpendicular to the ground plane, and the focal length were determined using the two VPs and assuming the principal point to be the middle of the image. The scaling was done by measuring a real-world fixed distance and its length in the image.

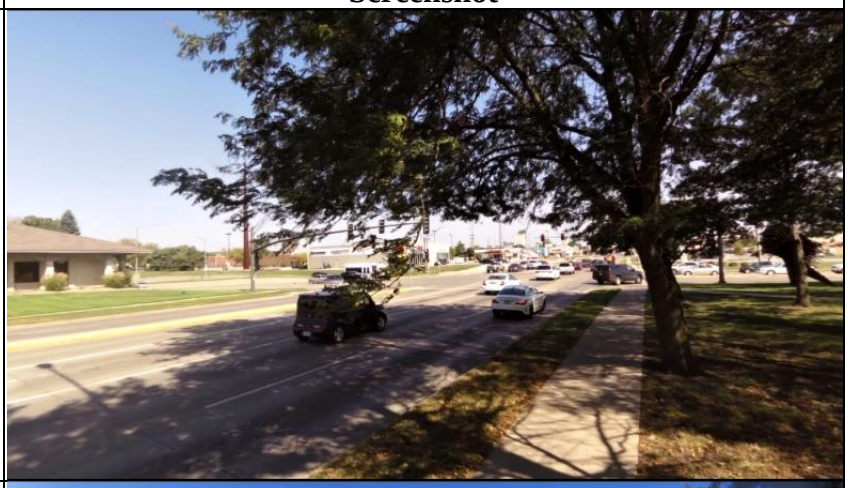
CHAPTER 4. DATA COLLECTION

Video data were collected in two phases in Ames, Iowa. In Phase 1, two cameras were used to record videos from two intersections in Ames for about 1.2 hours each. The two intersections were Lincoln Way and Grand Avenue and Stange Road and 13th Street. One camera was directed toward the traffic approaching the signal and the other camera was directed toward the signal with the traffic leaving the signal.

In Phase 2, four cameras were used to record videos at the Airport Road and South Duff Avenue intersection in Ames. Table 1 gives a summary of each video recording along with a sample screenshot. All videos were recorded in 30 fps with pixel dimensions of 1280×720. Figure 3 shows the locations where video recordings were done, with their corresponding IDs given in Table 1. Object detection and tracking were tested on this dataset.

Table 1. Ames, Iowa video recording location details

ID	Location	Direction	Dur-ation (mins)	Screenshot
1	13th St. – Stange Rd.	Signal	64	
		Approach	64	

ID	Location	Direction	Dur-ation (mins)	Screenshot
2	Lincoln Way – Grand Ave.	Signal	77	
		Approach	77	
3	Airport Rd. – South Duff Ave.	Signal	54	

ID	Location	Direction	Dur-ation (mins)	Screenshot
		Approach1	54	
		Approach2	54	
		Approach3	54	

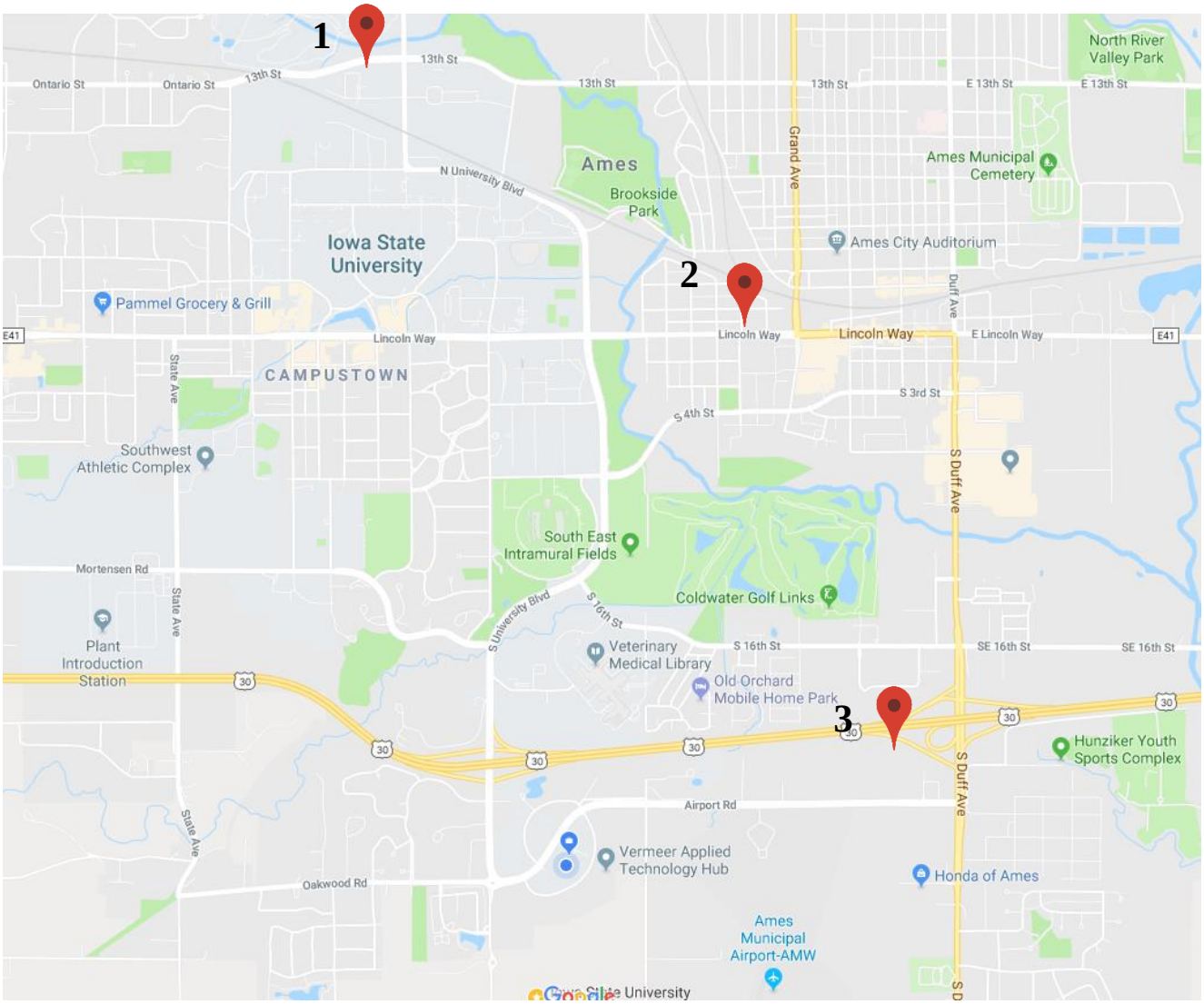
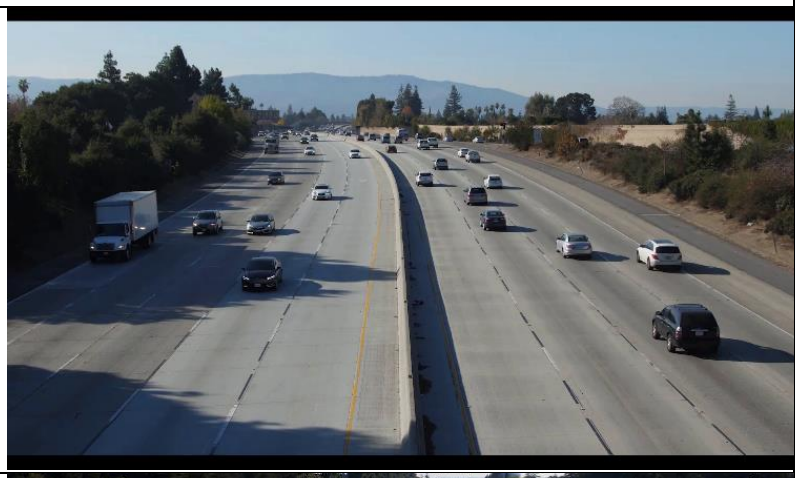
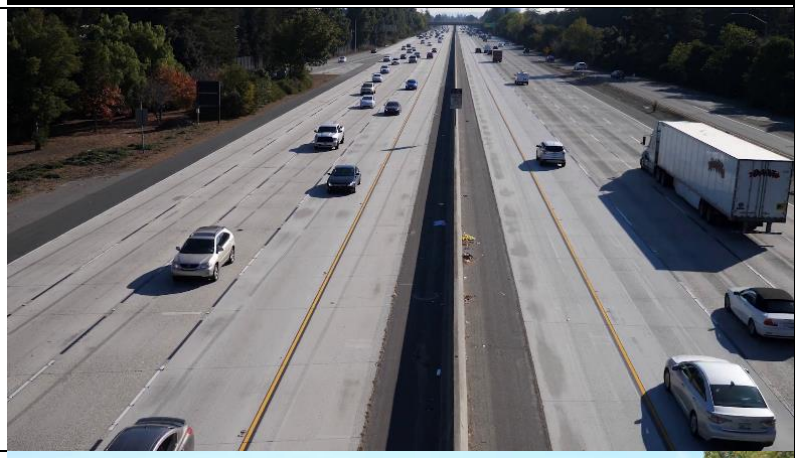


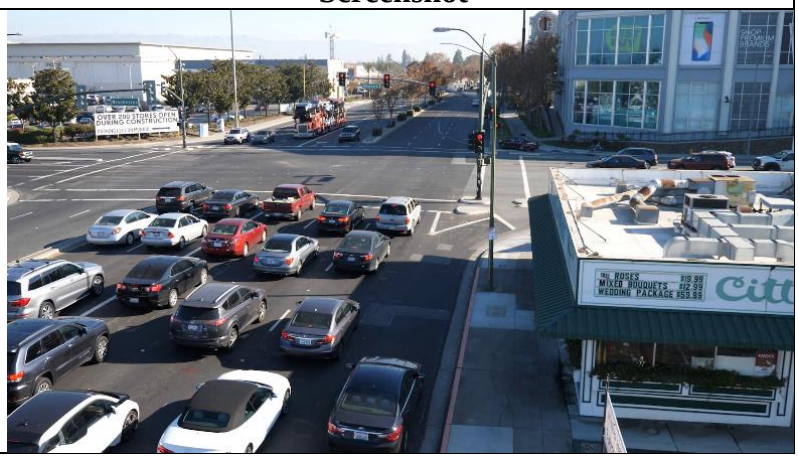
Figure 3. Study locations in Ames, Iowa where video recordings performed

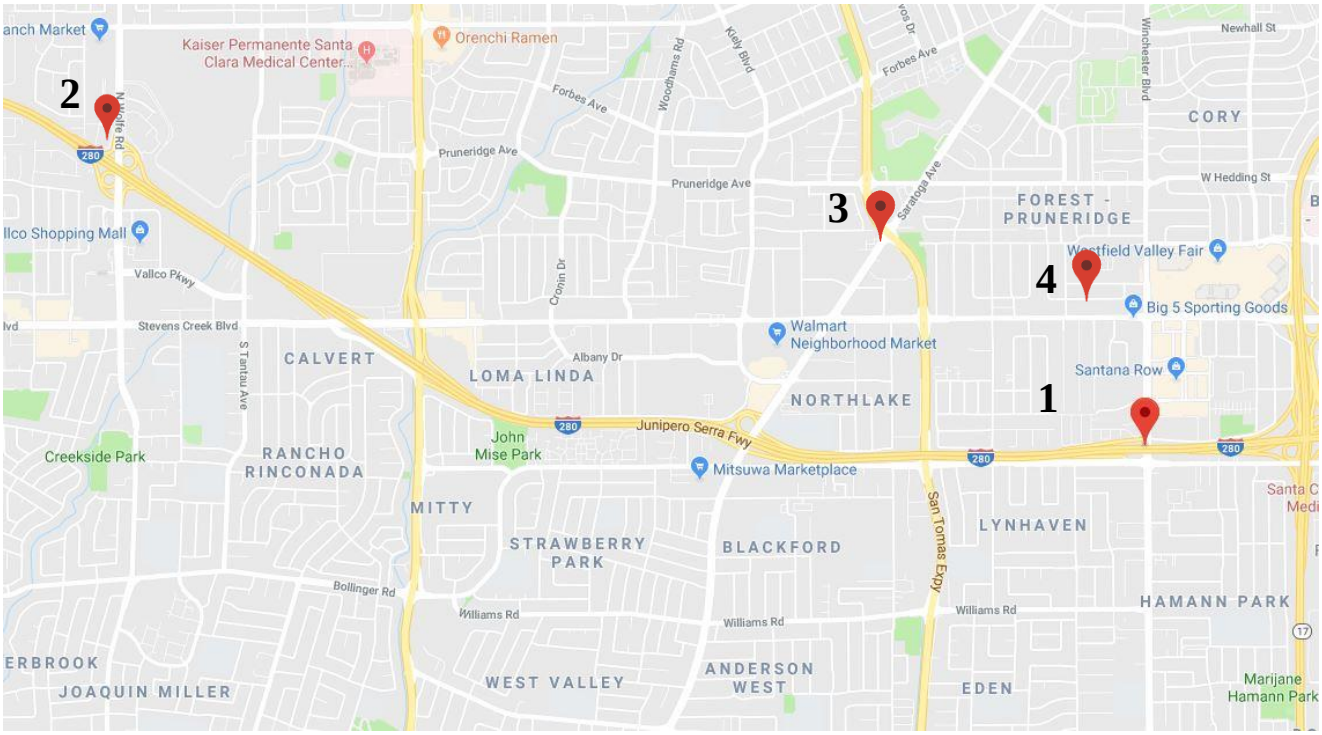
To test performance of the camera calibration and speed estimation methodology, the research team participated in Track 1 of the AI City Challenge 2018. In the challenge track, participating teams were asked to submit results for individual vehicle speeds in a test set containing 27 HD 1920×1080 videos, each 1-minute in length.

Table 2 gives a summary of each video recording location along with a sample screenshot. The videos were recorded at four locations in Silicon Valley, two on an area highway and two at city intersections, as shown in Figure 4. Performance was evaluated based on the ground truth generated by a fleet of control vehicles (with accurate GPS tracking) driven during the recording. Evaluation was based on the detection rate of the control vehicles and the root mean square error (RMSE) of the predicted control vehicle speeds.

Table 2. Video data available in AI City 2018 Challenge

ID	Location	Direction	Dur-ation (mins)	Screenshot
5	I-280 – Winchester	Approach Highway	8	
6	I-280 – Wolfe	Approach Highway	8	
7	San Tomas – Saratoga	Signal	6	

ID	Location	Direction	Dur-ation (mins)	Screenshot
8	Stevens Creek – Winchester	Signal	5	



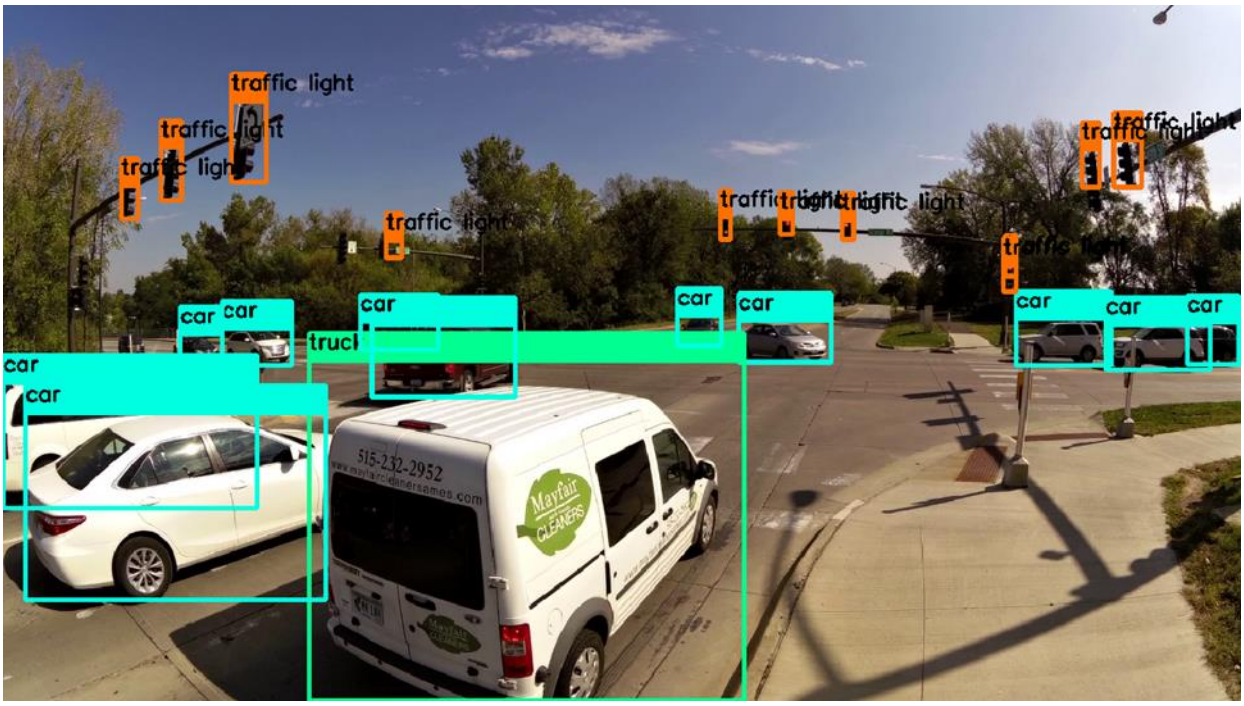
© Google 2017

Figure 4. Study locations where video recordings available from AI City 2018 challenge

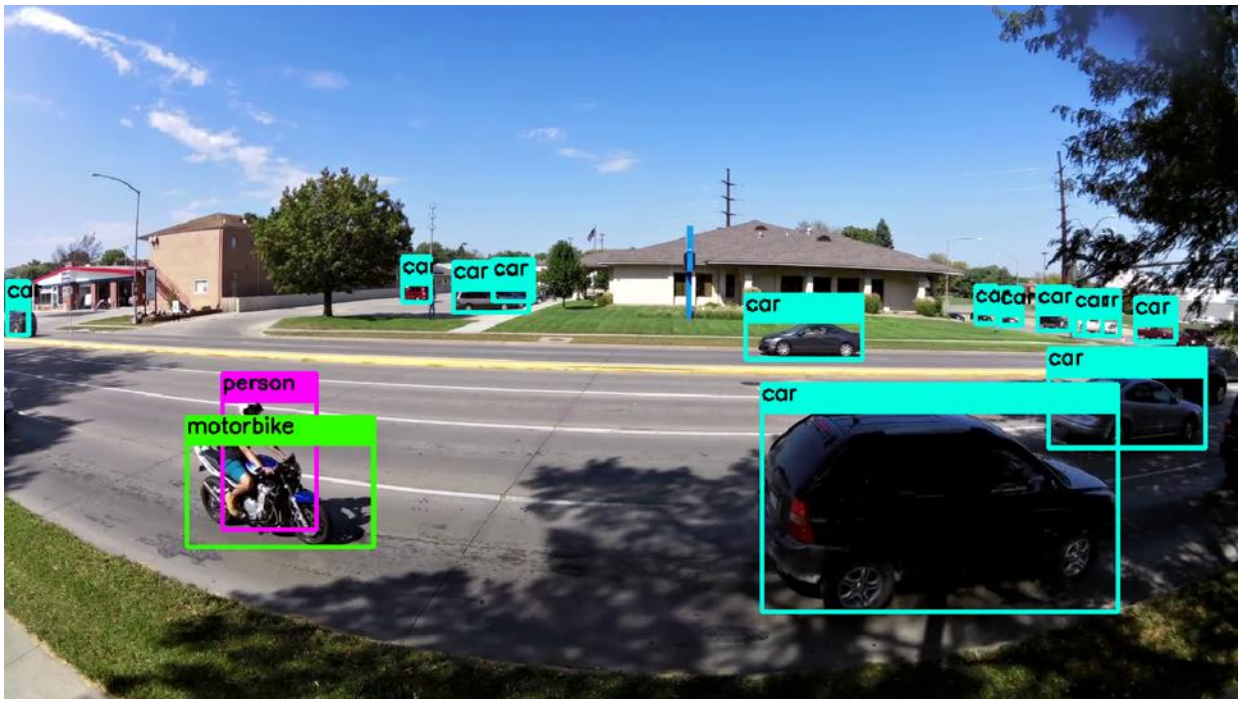
CHAPTER 5. RESULTS

5.1. Vehicle Detection

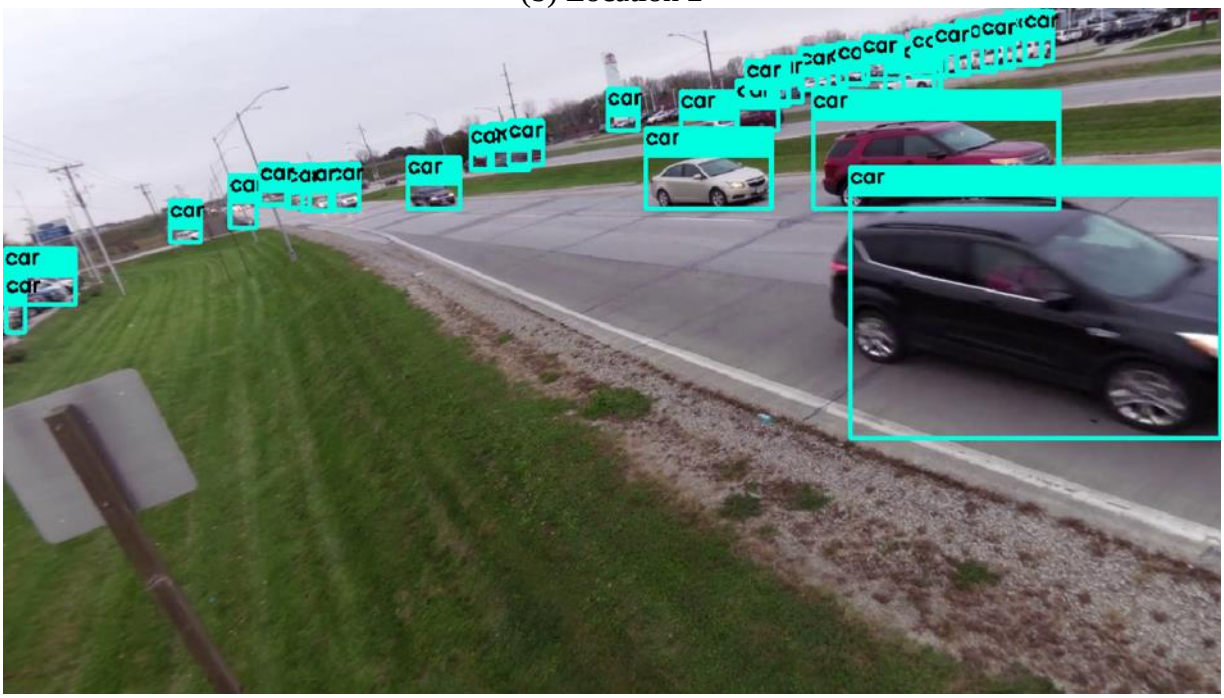
The object detection task using YOLOv3 architecture was tested to be performing at approximately 45 fps using NVIDIA's GeForce GTX 1080 Ti GPU, which implies that it can be implemented online easily. Figure 5 and Figure 6 give sample object detection results obtained from the videos recorded in the locations mentioned in Table 1 and Table 2, respectively.



(a) Location 1

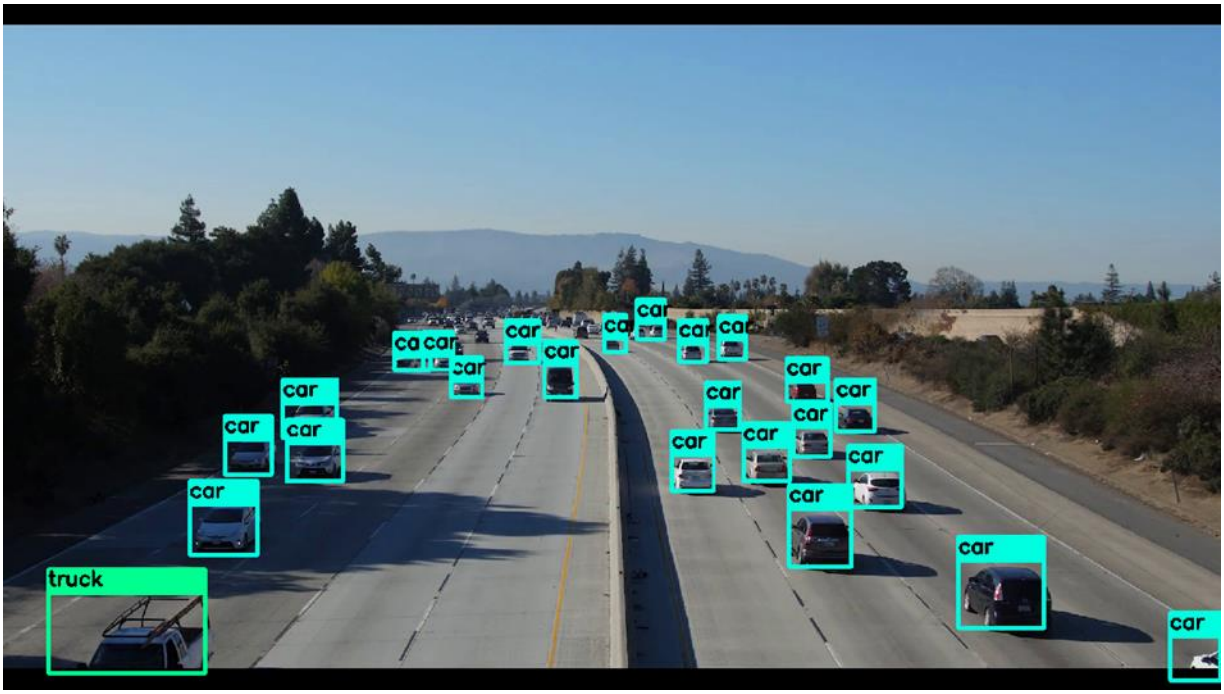


(b) Location 2

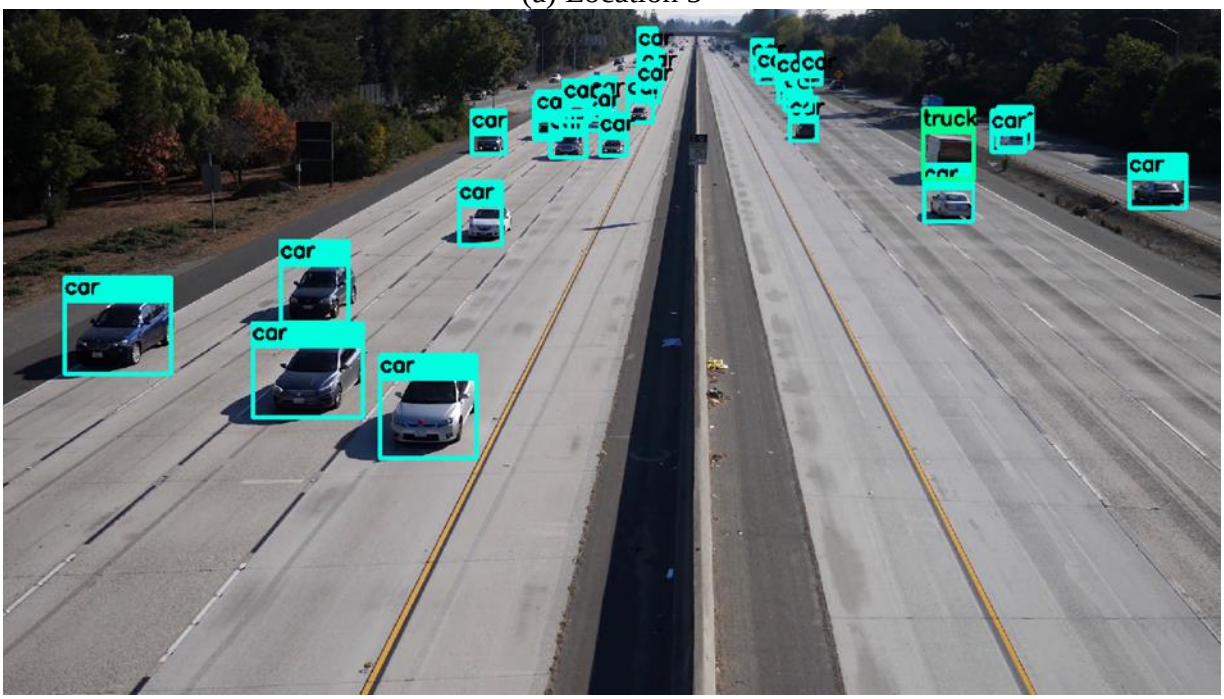


(c) Location 3

Figure 5. Sample vehicle detection results obtained from videos recorded in locations given in Table 1

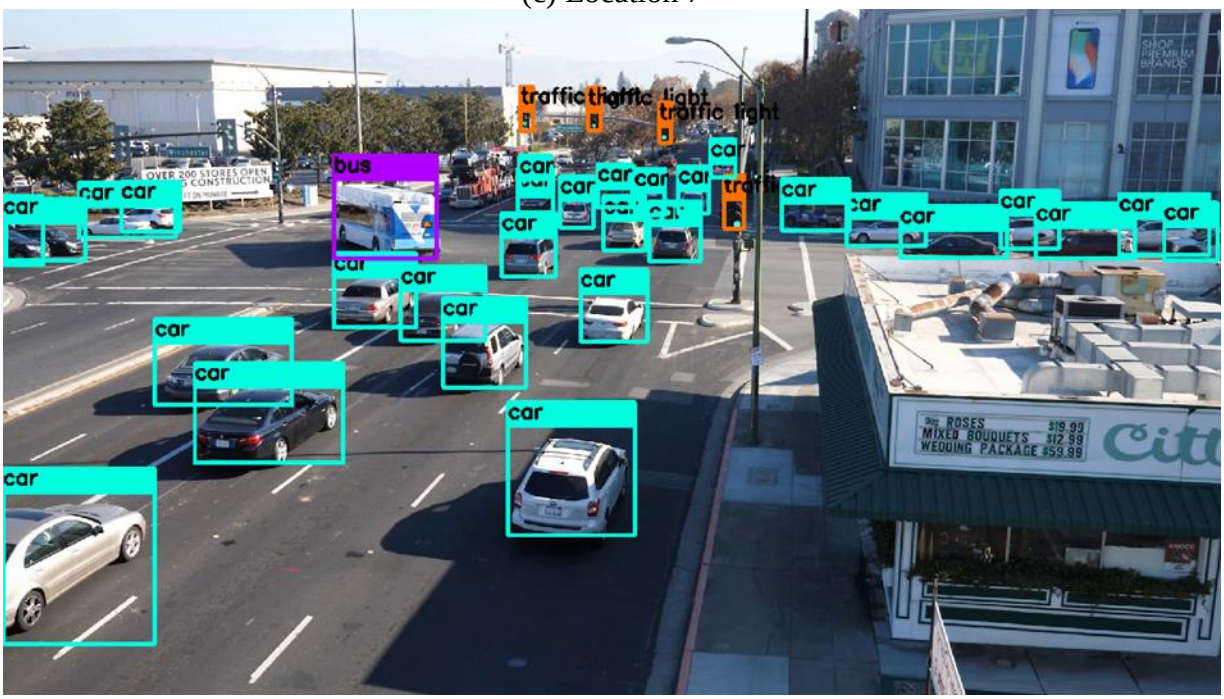


(a) Location 5



(b) Location 6

(c) Location 7



(d) Location 8

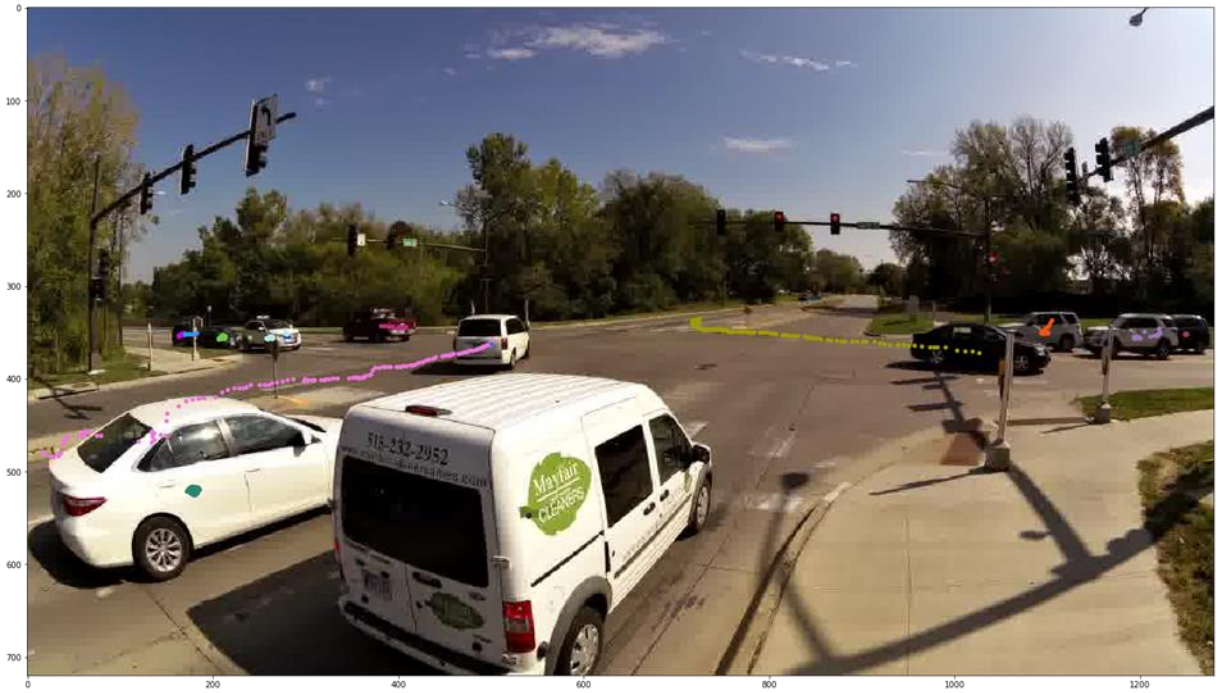
Figure 6. Sample vehicle detection results obtained from videos recorded in locations given in Table 2

It can be seen that, although the model can successfully detect nearby objects, it fails to detect some distant objects (Figure 6a). The results can be improved in the future by fine-tuning the trained model using a richer traffic dataset with diverse varieties, shapes, and sizes of vehicles. The publicly available UA-DETRAC benchmark dataset (Wen et al. 2015) and manual

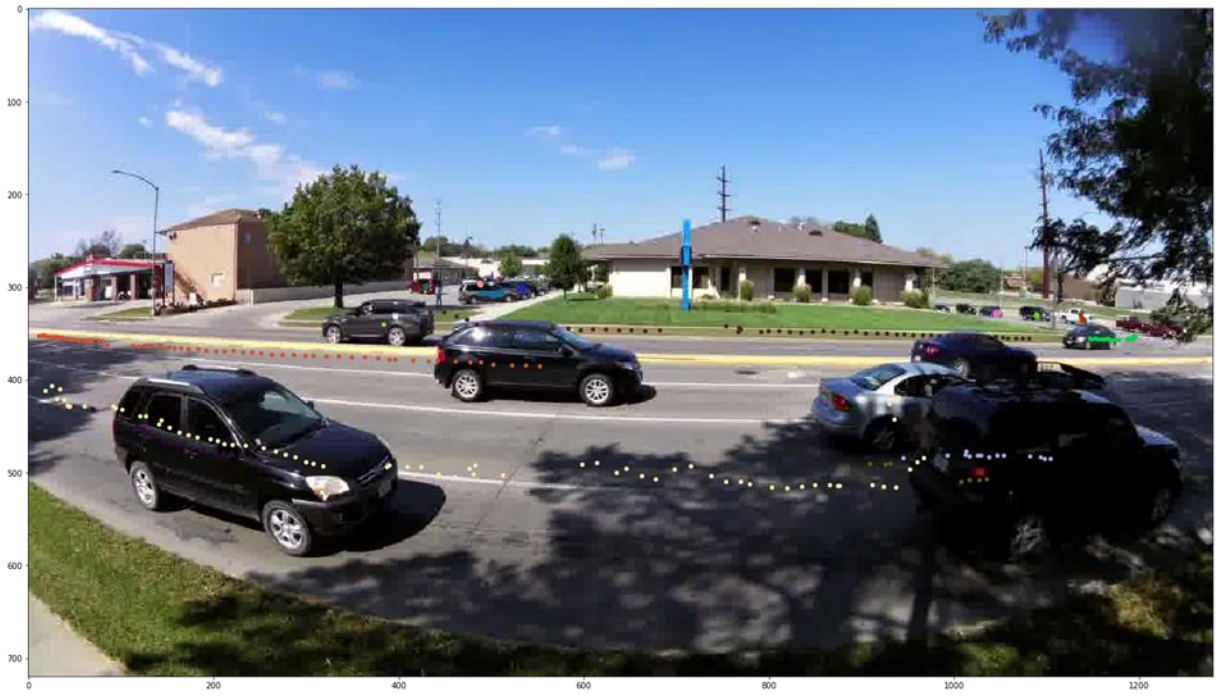
annotation of vehicles in some images also can be used to fine-tune the model and improve object detection results.

5.2. Vehicle Tracking

The bounding boxes obtained from the detection task were then fed to the tracking algorithm. Figures 7 and 8 show examples of tracking results obtained from the study locations in Ames and Silicon Valley (AI City Challenge dataset), respectively. Each color in the images represents a unique tracking ID provided to the vehicles. However, as already stated, the SORT tracking algorithm returned a high number of identity switches (Figures 7b and 8a).



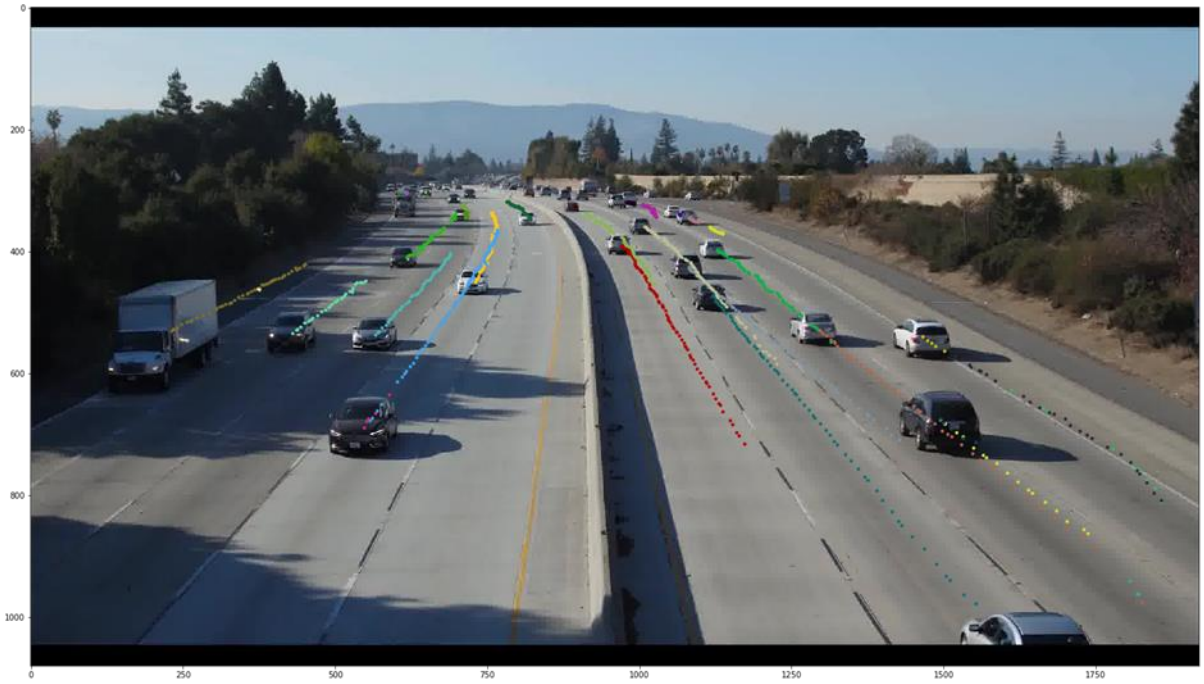
(a) Location 1



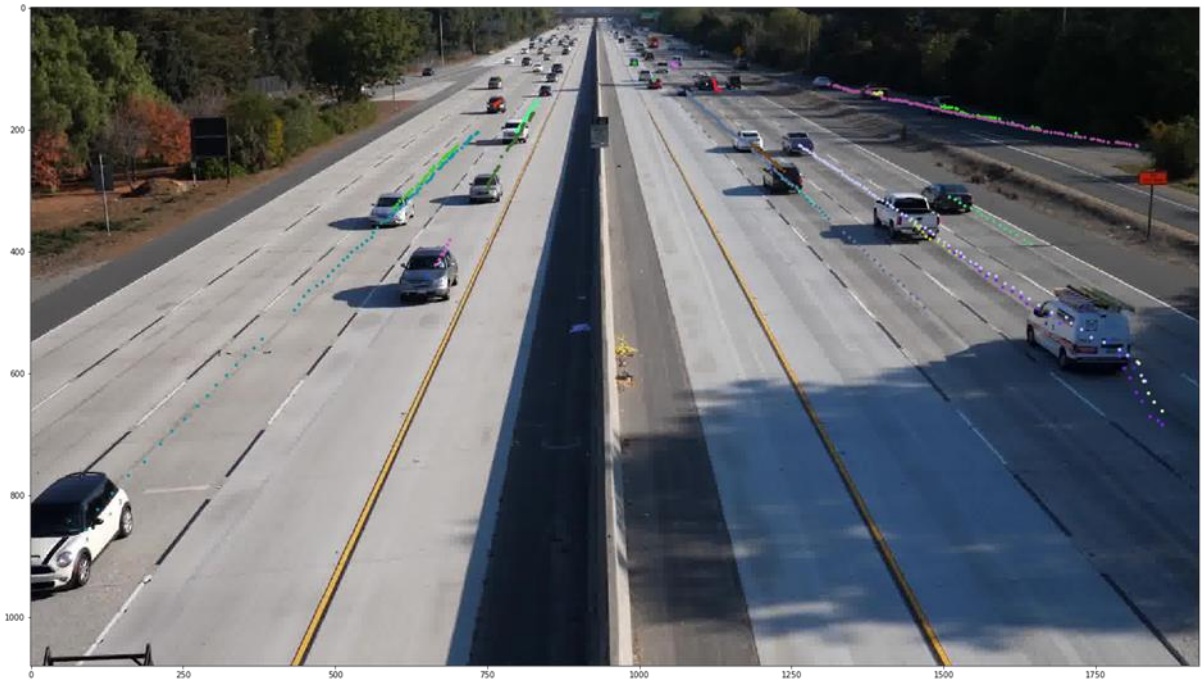
(b) Location 2

(c) Location 3

Figure 7. Sample vehicle detection results obtained from videos recorded in locations given in Table 1



(a) Location 5



(b) Location 6

(c) Location 7

(d) Location 8

Figure 8. Sample vehicle detection results obtained from videos recorded in locations given in Table 2

A sample tracking video file from the Location 3 video recording (Duff Avenue and Airport Road intersection in Ames) is available at the following link: https://iastate.app.box.com/v/vehicle-track-pc. The video also shows the identity switching issue.

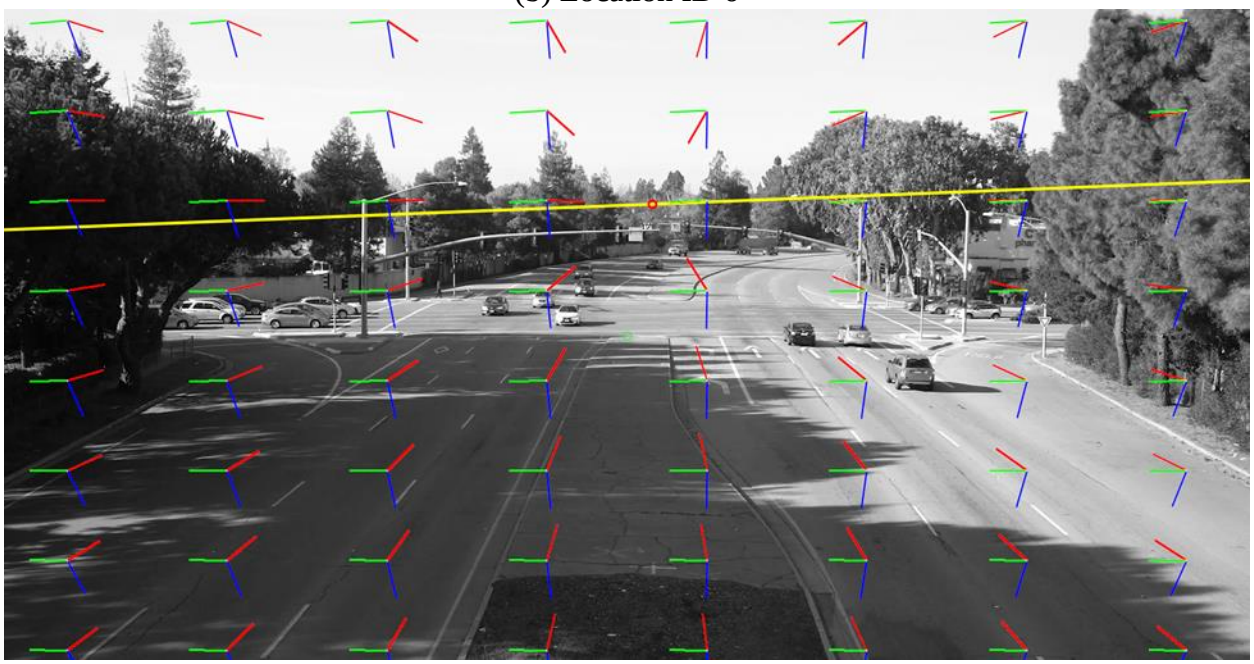
In the future, the Deep-SORT algorithm can be implemented, which can reduce this issue using the appearance description information (Wojke et al. 2017).

5.3. Camera Calibration

Camera calibration and speed estimation were tested on the AI City Challenge videos. Figure 9 shows the vanishing points determined in each of the four locations. The red, green, and blue lines in the images indicate the first, second, and third VP, respectively. It can be seen that the second VP estimation is not always perfect (ID 5, 7, and 8 in Figures 9 a, c, and d respectively).

(a) Location ID 5

(b) Location ID 6

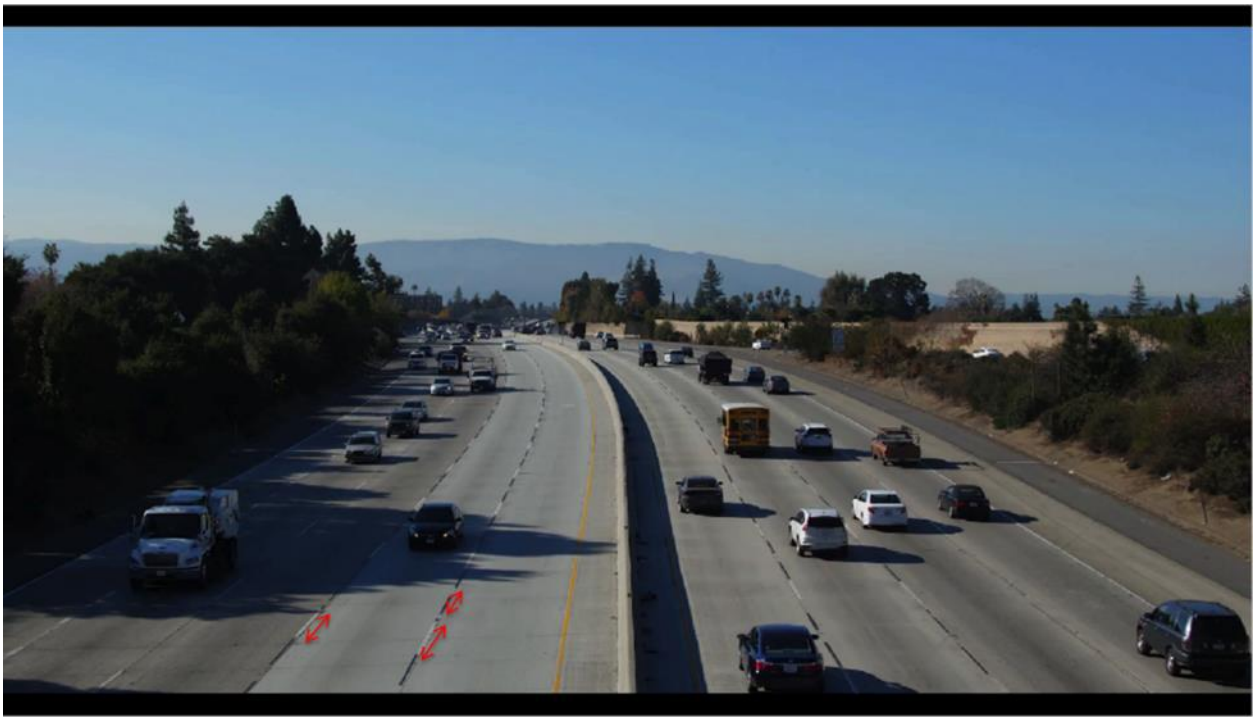


(c) Location ID 7

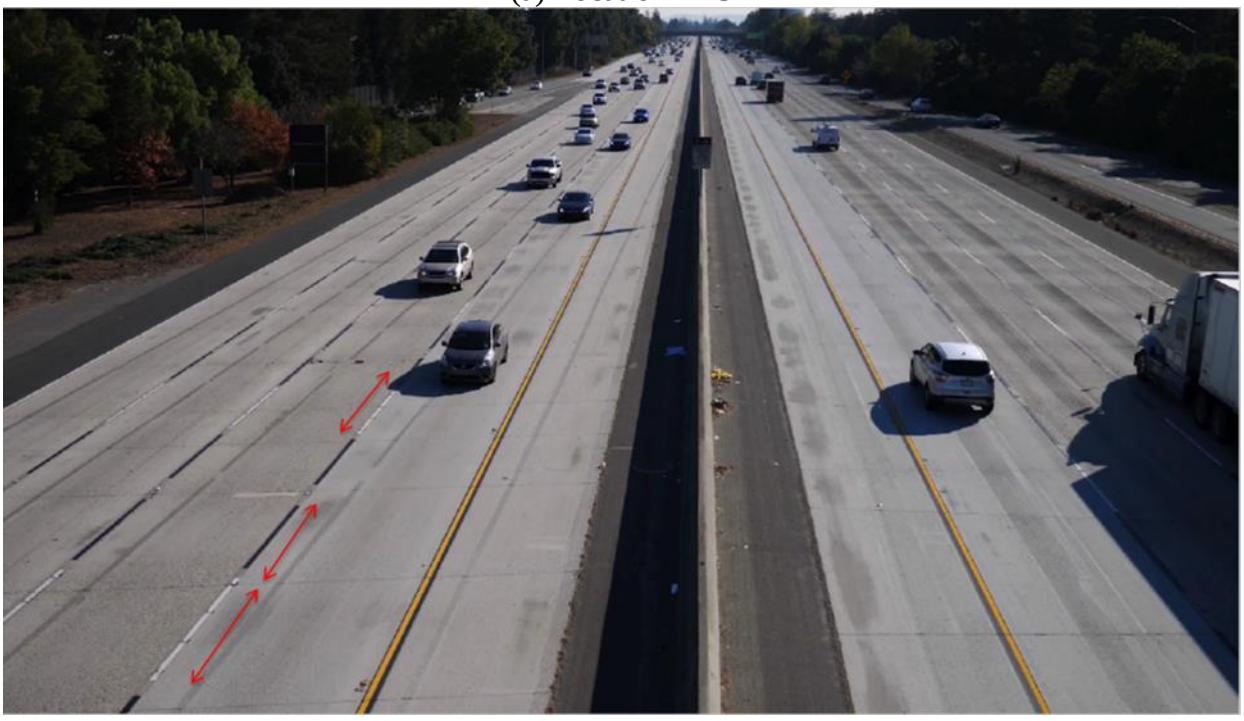
(d) Location ID 8

Figure 9. Vanishing points determined in location IDs provided by AI City Challenge

A similar observation was also noted by Dubská et al. 2015. Although the camera calibration approach used here was fully automatic and doesn't require any manual work, the scale estimation needed a manual annotation of a fixed distance from the real world in a sample image for each location. Multiple fixed length distances were annotated from a sample image at each location, and the average scale factor was used for speed estimation. Figure 10 shows the fixed length distances annotated for each location.



(a) Location ID 5



(b) Location ID 6

(c) Location ID 7

(d) Location ID 8

Figure 10. Fixed length distances, shown by red arrows, used for scale estimation in location IDs provided by AI City Challenge

The research team's AI City Challenge submission received a 96.3% detection ratio, which means 96.3% of the control fleet vehicles were detected for at least 30% of frames with intersection-over-union (IOU) of 50% or more. This shows that the object detection works fairly

well in test conditions. The RMSE in speed estimation is reported to be 12.9 mph on the challenge website. This shows a significant scope of improvement in speed estimation. A detailed analysis needs to be done in the future using speed data obtained from control vehicles in a test bed to find out the scope of improvement needed for the methodology adopted.

Calibration also can be done fully manually to determine the vanishing points along with the scale factor. The winning team submission in the AI City Challenge involved manually labeling two parallel line pairs orthogonal to each other on the three-dimensional (3D) ground plane (Tang et al. 2018). These line pairs were used to derive the vanishing points. A set of line segments on the ground plane, each defined by two endpoints, were manually selected, with ground-truth 3D lengths measured using Google Maps.

Using the calculated camera parameters, the two-dimensional (2D) endpoints of the line segments were back projected to 3D. Their Euclidean distances represented the estimated 3D lengths of the line segments. The absolute differences between estimations and ground truths were summed to describe the reprojection error. The objective was to minimize the reprojection error.

The non-linear optimization problem was iteratively solved by the estimation of distribution algorithm (EDA). This methodology resulted in a RMSE of 4 mph. Thus, it can be seen that manual calibration to determine the vanishing points and scale factor can significantly improve the speed estimation results. However, it comes at the expense of manually calibrating cameras, which impacts scalability.

CHAPTER 6. CONCLUSION

In this study, the research team performed vehicle detection, tracking, and speed estimation from camera videos. The team adopted a tracking-by-detection framework. The object detection task was performed using the YOLOv3 model architecture, and the tracking was performed using the SORT algorithm. The team tested the framework on videos collected from three intersections in Ames, Iowa. The combined detection and tracking was performed at approximately 40 fps using the GeForce GTX 1080 GPU. Therefore, it can be implemented online easily.

Camera calibration was performed by finding the edges of moving vehicles to automatically detect the vanishing points, and the scale was determined manually from a known fixed distance in the image and the real world. Although this methodology performed vanishing point determination automatically without any manual intervention, the speed estimation error came out to be quite high (~13 mph). The error can be reduced significantly by performing calibration and scale factor determination fully manually. However, since it requires full manual intervention, it is difficult to scale the algorithm across multiple cameras.

The detection task can be improved in the future by training the model on a larger dataset. Specifically, the UA-DETRAC dataset can be used in the future to improve detection results. Tracking performance result can be improved in the future by using Deep SORT or similar tracking algorithms, which use appearance description for tracking purposes. This can help in reducing the number of identity switches. Speed estimation can be improved in the future by extending automatic camera calibration to automatic scale estimation, which would also improve accuracy simultaneously.

REFERENCES

Andriyenko, A., K. Schindler, and S. Roth. 2012. Discrete-Continuous Optimization for Multi-Target Tracking. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June, Providence, RI, pp. 1926–1933.

Bae, S. H. and K. J. Yoon. 2018. Confidence-Based Data Association and Discriminative Deep Appearance Learning for Robust Online Multi-Object Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 40, No. 3, pp. 595–610.

Bauza, R., J. Gozalvez, and J. Sanchez-Soriano. 2010. Road Traffic Congestion Detection through Cooperative Vehicle-to-Vehicle Communications. Paper presented at the 35th Annual IEEE Conference on Local Computer Networks, October 10-14, Denver, CO.

Berclaz, J., F. Fleuret, E. Türetken, and P. Fua. 2011. Multiple Object Tracking Using K-Shortest Paths Optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 33, No. 9, pp. 1806–1819.

Bewley, A., Z. Ge, L. Ott, F. Ramos, and B. Upcroft. 2016. Simple Online and Realtime Tracking. In Proceedings of the International Conference on Image Processing, September 25–28, Phoenix AZ, pp. 3464–3468.

Cathey, F. W. and D. J. Dailey. 2005. A Novel Technique to Dynamically Measure Vehicle Speed Using Uncalibrated Roadway Cameras. In IEEE Procedings – Intelligent Vehicles Symposium, June 6–8, Las Vegas, NV, pp. 777–782.

Dai, J., Y. Li, K. He, and J. Sun. 2016. R-FCN: Object Detection via Region-based Fully Convolutional Networks. 30th Conference on Neural Information Processing Systems, May, Barcelona, Spain. https://www.robots.ox.ac.uk/~vgg/rg/papers/dai16nips.pdf

Dubská, M., A. Herout, R. Juránek, and J. Sochor. 2015. Fully Automatic Roadside Camera Calibration for Traffic Surveillance. *IEEE Transactions on Intelligent Transportation Systems,* Vol. 16, No. 3, pp. 1162–1171.

Dubská, M., J. Sochor, and A. Herout. 2014. Automatic Camera Calibration for Traffic Understanding. Presented at the British Machine Vision Conference, September 1–5, Nottingham, England. http://www.bmva.org/bmvc/2014/files/paper013.pdf.

Erhan, D., C. Szegedy, A. Toshev, and D. Anguelov. 2014. Scalable Object Detection Using Deep Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 23–28, Washington, DC, pp. 2155–2162.

Feng, Y., J. Hourdos, and G. A. Davis. 2014. Probe Vehicle Based Real-Time Traffic Monitoring on Urban Roadways. *Transportation Research Part C: Emerging Technologies,* Vol. 40, pp. 160–178.

Filipiak, P., B. Golenko, and C. Dolega. 2016. NSGA-II Based Auto-Calibration of Automatic Number Plate Recognition Camera for Vehicle Speed Measurement. In Proceedings of Applications of Evolutionary Computation, 19th European Conference, Evo Applications, March 30–April 1, Porto, Portugal, pp. 803–818.

Fortmann, T.E., Y. Bar-Shalom, and M. Scheffe. 1983. Sonar Tracking of Multiple Targets Using Joint Probabilistic Data Association. *IEEE Journal of Oceanic Engineering,* Vol. 8, No. 3, pp. 173–184.

Girshick, R. 2015. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision. https://www.cvfoundation.org/openaccess/content_iccv_2015/papers/Girshick_Fast_R-CNN_ICCV_2015_paper.pdf

Girshick, R., J. Donahue, T. Darrell, and J. Malik. 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587. https://www.cv-foundation.org/openaccess/content_cvpr_2014/papers/Girshick_Rich_Feature_Hierarchies_2014_CVPR_paper.pdf

Grammatikopoulos, L., G. Karras, and E. Petsa. 2005. Automatic Estimation Of Vehicle Speed From Uncalibrated Video Sequences. Paper presented at International Symposium on Modern Technologies, Education and Professional Practice in Geodesy and Related Fields, November 3–4, Sofia, Bulgaria.

He, K., X. Zhang, S. Ren, and J. Sun. 2016. Deep Residual Learning for Image Recognition. Presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition, https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7780459. pp. 770–778.

He, X. C. and N. H.C. Yung. 2007. A Novel Algorithm for Estimating Vehicle Speed from Two Consecutive Images. In Proceedings of the IEEE Workshop on Applications of Computer Vision, February 21–22, Austin, TX.

Huang, J., V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy. 2017. Speed/accuracy Trade-Offs for Modern Convolutional Object Detectors. Presented at the IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, Honolulu, HI. http://openaccess.thecvf.com/content_cvpr_2017/papers/Huang_SpeedAccuracy_Trade-Offs_for_CVPR_2017_paper.pdf

Kim, C., F. Li, A. Ciptadi, and J. M. Rehg. 2015. Multiple Hypothesis Tracking Revisited. Presented at the IEEE International Conference on Computer Vision, December 7-13, Santiago, Chile, pp. 4696-4704,

Kotzenmacher, J., E. D Minge, and B. Hao. 2004. *Evaluation of Portable Non-Intrusive Traffic Detection System*. Minnesota Department of Transportation, Minneapolis, MN.

Leal-Taixe, L., A. Milan, I. Reid, S. Roth, and K. Schindler. 2015. MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking. https://arxiv.org/pdf/1504.01942.pdf.

Lin, T. Y., M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*, Springer, New York, NY, pp. 740-755.

Liu, W., D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C Berg. 2016. Ssd: Single Shot Multibox Detector. In *European Conference on Computer Vision*, Springer, New York, NY, pp. 21–37.

Liu, X., W. Liu, H. Ma, and H. Fu. 2016. Large-Scale Vehicle Re-Identification in Urban Surveillance Videos. Paper presented at the IEEE International Conference on Multimedia and Expo, July 11-15, Seattle, WA.

Naphade, M., G. Banavar, C. Harrison, J. Paraszczak, and R. Morris. 2011. Smarter Cities and Their Innovation Challenges. *Computer*, Vol. 44, No. 6, pp. 32–39.

Ozkurt, C. and F. Camci. 2009. Automatic Traffic Density Estimation and Vehicle Classification for Traffic Surveillance Systems Using Neural Networks. *Mathematical and Computational Applications*, Vol. 14, No. 3, pp. 187–196.

Redmon, J., S. Divvala, R. Girshick, and A. Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. Paper presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, Las Vegas, NV, pp. 779–788.

Redmon, J. and A. Farhadi. 2018. *YOLOv3: An Incremental Improvement*. https://arxiv.org/pdf/1804.02767.pdf.

Reid, D. B. 1979. An Algorithm for Tracking Multiple Targets. *IEEE Transactions on Automatic Control*, Vol. 24, No. 6, pp. 843–854.

Ren, S., K. He, R. Girshick, and J. Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, No. 6, pp. 1137–1149.

Rezatofighi, S. H., A. Milan, Z. Zhang, Q. Shi, A. Dick, and I. Reid. 2015. Joint Probabilistic Data Association Revisited. Presented at the IEEE International Conference on Computer Vision, December 7–13, Santiago, Chile, pp. 3047–3055.

Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, Vol. 115, No. 3, pp. 211–252.

Schoepflin, T. N., and D. J. Dailey. 2002. Dynamic Camera Calibration of Roadside Traffic Management Cameras. In Proceedings of the IEEE 5th International Conference on Intelligent Transportation Systems, September 3–6, Singapore, pp. 25–30.

Sochor, J., R. Juránek, J. Špaňhel, L. Maršík, A. Široký, A. Herout, and P. Zemčík. 2017. BrnoCompSpeed: Review of Traffic Camera Calibration and Comprehensive Dataset for Monocular Speed Measurement. https://www.researchgate.net/publication/313879450_BrnoCompSpeed_Review_of_Traffic_Camera_Calibration_and_Comprehensive_Dataset_for_Monocular_Speed_Measurement

Szegedy, C., A. Toshev, and D. Erhan. 2013. Deep Neural Networks for Object Detection. In Proceedings of the 26th International Conference on Neural Information Processing Systems, Volume 2, December 5–10, Lake Tahoe, NV, pp. 2553–2561.

Tang, Z., G. Wang, H. Xiao, A. Zheng, and J.-N. Hwang. 2018. Single-Camera and Inter-Camera Vehicle Tracking and 3D Speed Estimation Based on Fusion of Visual and Semantic Features. Presented at the IEEE Conference on Computer Vision and Pattern Recognition Workshops, June 18-22, Salt Lake City, UT. http://openaccess.thecvf.com/content_cvpr_2018_workshops/w3/html/Tang_Single-Camera_and_Inter-Camera_CVPR_2018_paper.html

Wen, L., D. Du, Z. Cai, Z. Lei, M.-C. Chang, H. Qi, J. Lim, M.-H. Yang, and S. Lyu. 2015. UA-DETRAC: A New Benchmark and Protocol for Multi-Object Tracking. https://arxiv.org/abs/1511.04136.

Wojke, N., A. Bewley, and D. Paulus. 2017. Simple Online and Realtime Tracking with a Deep Association Metric. Paper presented at IEEE International Conference on Image Processing, September 17-20, Beijing, China.

Yang, B. and R. Nevatia. 2012. Multi-Target Tracking by Online Learning of Non-Linear Motion Patterns and Robust Appearance Models. Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition, June 16–21, Los Angeles, CA, pp. 1918–1925.

You, X. and Y. Zheng. 2016. An Accurate and Practical Calibration Method for Roadside Camera Using Two Vanishing Points. *Neurocomputing*, Vol. 204, pp. 222–230.

Zhang, L., Y. Li, and R. Nevatia. 2008. Global Data Association for Multi-Object Tracking Using Network Flows. Paper presented at the 2008 IEEE Conference on Computer Vision and Pattern Recognition, June 23–28, Anchorage, AK, pp. 1–8.

Zhong, M. and G. Liu. 2007. Establishing and Managing Jurisdiction-Wide Traffic Monitoring Systems: North American Experiences. *Journal of Transportation Systems Engineering and Information Technology*, Vol. 7, No. 6, pp. 25–38.

THE INSTITUTE FOR TRANSPORTATION IS THE FOCAL POINT FOR TRANSPORTATION AT IOWA STATE UNIVERSITY.

InTrans centers and programs perform transportation research and provide technology transfer services for government agencies and private companies;

InTrans manages its own education program for transportation students and provides K-12 resources; and

InTrans conducts local, regional, and national transportation services and continuing education programs.

IOWA STATE UNIVERSITY

Visit www.InTrans.iastate.edu for color pdfs of this and other research reports.