Automated Vehicle Recognition with Deep Convolutional Neural Networks

Yaw Okyere Adu-Gyamfi, Ph.D. (Corresponding Author)
Department of Civil and Environmental Engineering
University of Virginia, Charlottesville, VA.
Tel: Email: yoa4q@virginia.edu

Sampson Kwasi Asare, Ph.D.
Intelligent Transportation System Analyst
Noblis Inc.
600 Maryland Avenue, S.W. Suite 700E
Washington, D.C. 20024
Tel: 202-863-2981 Email: Sampson.Asare@noblis.org

Anuj Sharma, Ph.D.
Civil, Construction and Environmental Engineering
Iowa State University, Ames, IA
Tel: Email: anujs@iastate.edu

Tienaah Titus
Geodesy and Geomatic Engineering
University of New Brunswick, Fredericton, NB, Canada.
Tel: Email: tttienaah@unb.ca

Submitted for Presentation and Publication at the 2017 TRB Annual Meeting

July 30

Word Count: 5143 + 9 Figures + 0 Tables = 7385 words

ABSTRACT

In recent years there has been a growing interest in the use of non-intrusive systems such as radar and infrared systems for vehicle recognition. State of the art non-intrusive systems can report up to 8 classes of vehicle types. Video based systems, which are arguably the most popular non-intrusive detection systems can only report very coarse classification levels (up to 4 classes) even with the best performing vision systems. The goal of this study is to develop a vision system that can report finer vehicle classifications according to FHWA's scheme as well as comparable to other non-intrusive recognition systems. The proposed system decouples object recognition into two main tasks: localization and classification. It begins with localization by generating class-independent region proposals for each video frame, then it uses Deep Convolutional Neural Networks (DCNN) to extracts feature descriptors for each proposed regions and finally, the system scores and classifies the proposed regions using a linear Support Vector Machines (SVM) template on the feature descriptors. The precision of the system varies for different vehicle classes. Passenger cars and SUVs can be detected at a precision rate of 95%. The precision rates for single-unit, single-trailer and double trailer trucks ranges between 92% and 94%. Based on Receiver Operating Characteristic (ROC) curves, the best system performance can be achieved under free flow, day or night time and good video resolution.

INTRODUCTION

As transportation agencies seek to optimize traffic mobility and improve safety, the need for accurate traffic data cannot be over-emphasized. Traditional traffic data collection technologies such as piezoelectric sensors, magnetic loops, and pneumatic road tubes have been very popular among many transportation agencies since the 1960's [1]. However, these traditional methods are gradually giving way to emerging advanced traffic data collection technologies such as active infrared/laser, radar, video, and others due to many reasons. These include the damage traditional methods cause as a result of their intrusive nature, environmental conditions (e.g. snow) that inhibit their use, frequent malfunctioning of equipment, lack of consistent accurate data, disruptions to traffic during installation, and many others [2]. With the proliferation in the number of advanced traffic data collection technologies, transportation professionals are often faced with the dilemma of identifying the appropriate technology that suits an agency's data collection needs.

Among the many data needs of transportation agencies is vehicle type classification [3]. Accurate vehicle type classification data is of fundamental importance to traffic operation, pavement design, and transportation planning [4]. For example, knowing the total number of trucks in a section of a roadway helps in computing corresponding passenger car equivalents needed for estimating the capacity of that roadway section [5]. Additionally, the geometric design characteristics of roadways (e.g. horizontal alignment, curb heights) are dictated by the types of vehicles that would be using such roadways [6].Under federal requirements for the Highway Performance Monitoring System, states are required to perform classified vehicle counts on freeways and highways and provide this information to the Federal Highway Administration (FHWA) every year [7]. Vehicle classification data is therefore very critical to the effective management and operation of transportation systems.

Many different techniques for acquiring vehicle type classification have been discussed in the literature, and prominent among them is the application of image processing techniques such as automated video-based classification system. In most instances, the classification is done based on the dimensions of vehicles. Lai et al. [8] demonstrated the estimation of accurate vehicle dimensions through the use of a set of coordinate mapping functions. Although they were able to estimate vehicle lengths to within 10% in every instance, their method requires camera calibration in order to map image angles and pixels into real-world dimensions. Commercially available video image processors such as the VideoTrack system developed by Peek Traffic Inc. are expensive and often require calibration to specific road surface information (e.g. distance between recognizable road surface marks) as well as camera information (such as elevation and tilt angle) which may not be easy to obtain [9]. Gupte, Masoud, Martin, and Papanikolopoulos [10], performed similar work by instead tracking regions and using the fact that all motion occurs in the ground plane to detect, track, and classify vehicles. Before vehicles may be counted and classified, their program must determine the relationship between the tracked regions and vehicles (e.g. a vehicle may have several regions or a region may have several vehicles). Unfortunately, their work does not address problems associated with shadows, so application of the algorithm is limited at the current stage.

Vehicle type classification using advanced techniques such as artificial intelligence has also been proposed in the literature. Zhou and Cheung proposed the use of Deep Neural Networks (DNN) to classify vehicles [11]. Since their test dataset was small compared to number of parameters inside DNN architecture, direct application of DNN was not possible. Therefore, they extracted features from a specific layer inside a properly-trained DNN, and transferred them to their specific classification task. This approach was used to classify cars, sedans, and vans. Hence, its performance on datasets that include trucks is unknown. Support Vector Machine (SVM) technique has also been used to conduct multi-class and intra-class vehicle type classifications [12]. In this study, two vehicle classification approaches were presented, using the SVM algorithm: 1) geometric-based approach and 2) appearance-based approach. While combining both geometry and appearance in classifying vehicles sounds encouraging, the proposed system only classifies vehicles into small, medium, and large categories; making its application limited.

In this paper, we propose a video-based vehicle detection and classification system for classifying vehicles according to FHWA's 13-vehicle type categories. The proposed approach takes advantage of recent advances in Deep Convolutional Neural Networks (DCNN), a machine learning technique that quickly and accurately learns unique vehicular features that can be used to report finer vehicle classes comparable to the state of the art non-intrusive recognition systems such as radar and microwave systems. The key algorithms of DCNN can be traced back to the late 1980s [13]. DCNNs saw heavy use in the 1990s. It however fell out of fashion with the rise of support vector machines. Interest in DCNNs was rekindled again in 2012 by [14] when they showed that substantially higher image classification accuracy could be achieved in the ImageNet dataset with DCNNs. Since its rebirth, profound improvements in the accuracy of object detection in complex scenes have been achieved.

The primary objective of this research is to develop an automated, video-based vehicle recognition system which:

1. Classifies vehicles according to the FHWA classification scheme.
2. Is robust to challenging real-world conditions such as high volume stop and go traffic, varying video resolution and lighting conditions.

The outline of the current study is as follows: First, an overview of the proposed approach to automated vehicle recognition and classification is provided. This section will highlight the machine vision algorithms selected for the purposes of this study. The training and fine-tuning of the algorithm will also be discussed. In the second section, a brief introduction of data used to train the deep learning model is given. Additionally, experiments conducted to test the efficiency of the vision system developed will be discussed in this section. The third section discusses results of experiments after using the developed vision system to process CCTV video data under varying conditions. Lastly, concluding remarks, recommendations, and additional research needs for future are presented in the fourth section.

PROPOSED APPROACH

The vision system developed in this study decouples object recognition into two main tasks: localization and classification. It begins with localization by generating class-independent region proposals with an algorithm called Selective Search [15]. Then it uses DCNN to extract unique feature descriptors on the proposed regions after warping them to a fixed square size (256 x 256). Finally, feature descriptors corresponding to each proposed region are classified through a linear SVM scoring system. Figure 1 below summarizes the proposed approach to automated vehicle recognition.

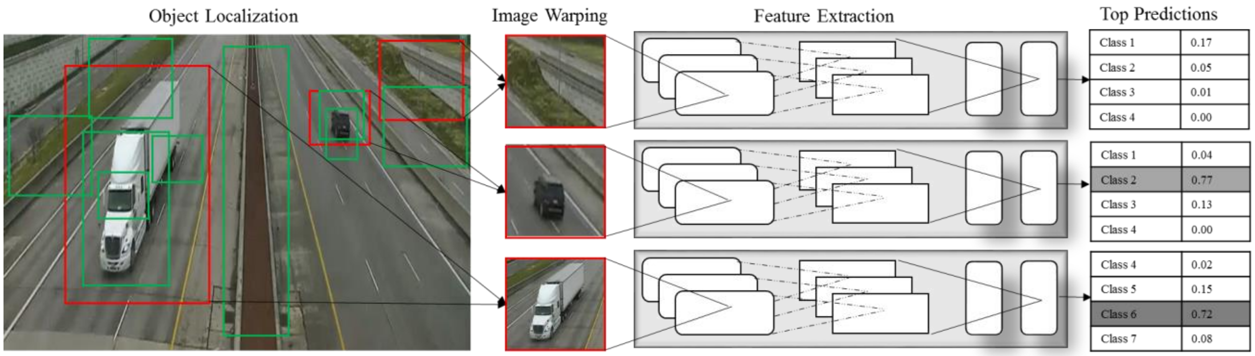


FIGURE 1 Approach to Vehicle Detection and Classification.

Object Localization with Selective Search

There are two main traditional approaches for object localization in images: Segmentation and Exhaustive Search. Segmentation tries to break a single partitioning of an image into its unique objects before any recognition [16]. This is sometimes extremely hard if there are disparate hierarchies of information in the image. A second approach is to localize objects by performing an exhaustive search within the image using various sliding window approaches [17]. The main challenge of using exhaustive search alone for object detection is that it fails to detect objects with low-level cues.

Uijlings et al. [15] developed Selective Search, an approach which combines the best of both worlds: segmentation and exhaustive search. It exploits the hierarchical structure of the image (segmentation) with the aim of generating all possible object locations (exhaustive search). The algorithm uses hierarchical grouping to deal with all possible object scales. Then, the color space of the image is used to deal with different invariance properties. Finally, region-based similarity functions are used to deal with the diverse nature of objects. We refer the reader to [15] for a detailed description of selective search. Figure 2 shows proposed object regions using selective search.

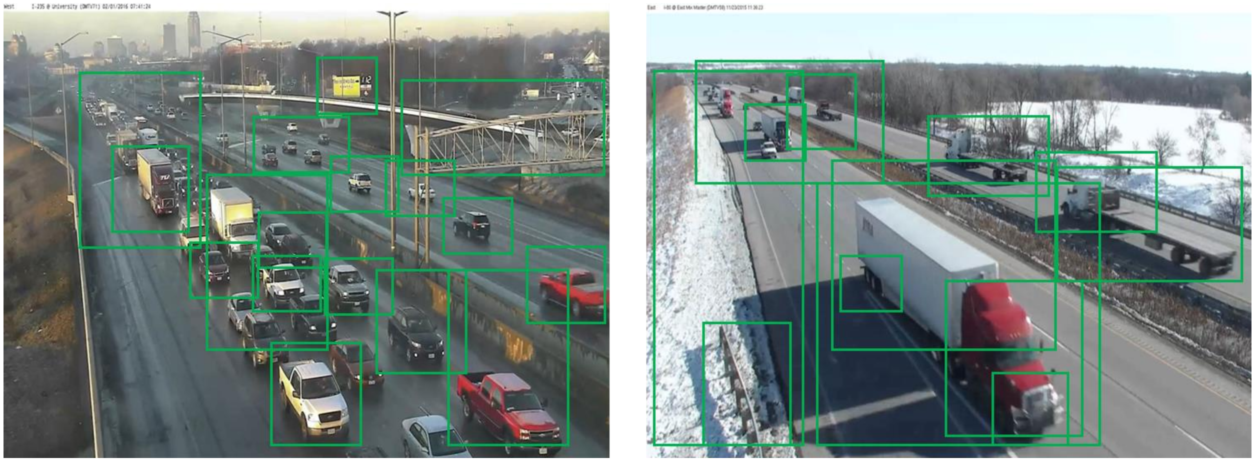


FIGURE 2 Candidate Region Proposals using Selective Search.

Object Classification with Deep Convolutional Neural Networks

After object localization with selective search, each detected object is fed through a Deep Convolutional Neural Network for classification. In the current study, a DCNN classifier is built to support key algorithms for classifying vehicles captured on CCTV cameras. The following section will explain the architecture used to build and test the DCNN classifier.

Model Training

DCNN models are computationally expensive, and this makes them very unattractive for practical applications. The recent interest in DCNNs could be attributed to the rise of efficient Graphical Processing Unit (GPU) implementations such as cuda-convnet [14], Torch [18] and Caffe [18]. In this study, a GeForce GTX Titan X GPU was used for model training and processing of videos. Model training involved two main steps: Supervised pre-training and Domain-specific fine-tuning.

Supervised Pre-Training: A DCNN model usually consists of thousands of parameters and millions of learned weights. This means that a very large training dataset (more than a million records) will be required to avoid over-fitting the model. Girshick et al. [20] however demonstrated that when labeled data is scarce, supervised pre-training for an auxiliary task with large training data followed by domain-specific fine-tuning (on a smaller dataset) could yield significant boost in performance. We adopted a similar approach by pre-training our DCNN model on a large auxiliary dataset (ILSVRC2012) [21] with image-level annotations. The resulting output is a rich feature detector which will later be fine-tuned to suit our purposes. Open source Caffe DCNN library was used for the pre-training model on 100 classes at a learning rate of 0.01.

Domain-Specific Fine-Tuning: To adapt the pre-trained model to the proposed task (vehicle recognition), the CNN model parameters are fine-tuned. First, the 100-way classification layer of the pre-trained model is replaced with 7 classes. We start Stochastic Gradient Descent (SGD) at a learning rate of 0.001, which allows fine-tuning to make progress while not clobbering the initialization. In each SGD iteration, we uniformly sample 20 positive windows for all classes and 70 background windows to construct a mini-batch of size 90. DCNN is used to extract a 4096 dimensional feature vector using caffe's implementation of CNN by Krizhesky [19]. Each mean subtracted candidate region proposal is forward-propagated through a network with five convolutional layers and two fully connected layers. The resulting feature vectors are then scored using linear SVMs trained for that specific class. The modeling architecture is shown in Figure 3 and summarized as follows:

1. Each class-independent region proposal from the previous step is warped to a 256 x 256 image.
2. The input warped image is filtered with 96 kernels of size 11X11, with a stride of 4 pixels. This is followed by max pooling in 3x3 grid.
3. Two subsequent convolutions with 384 kernels are carried out without pooling
4. Convolve output of fourth layer with 256 kernel, then apply spatial max pooling in 3x3 pixel grid.
5. Last 2 Layers: Fully connected layer of 4096 dimensions from the last layer.

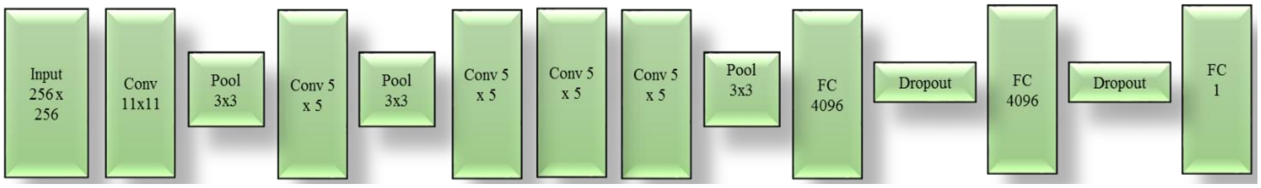
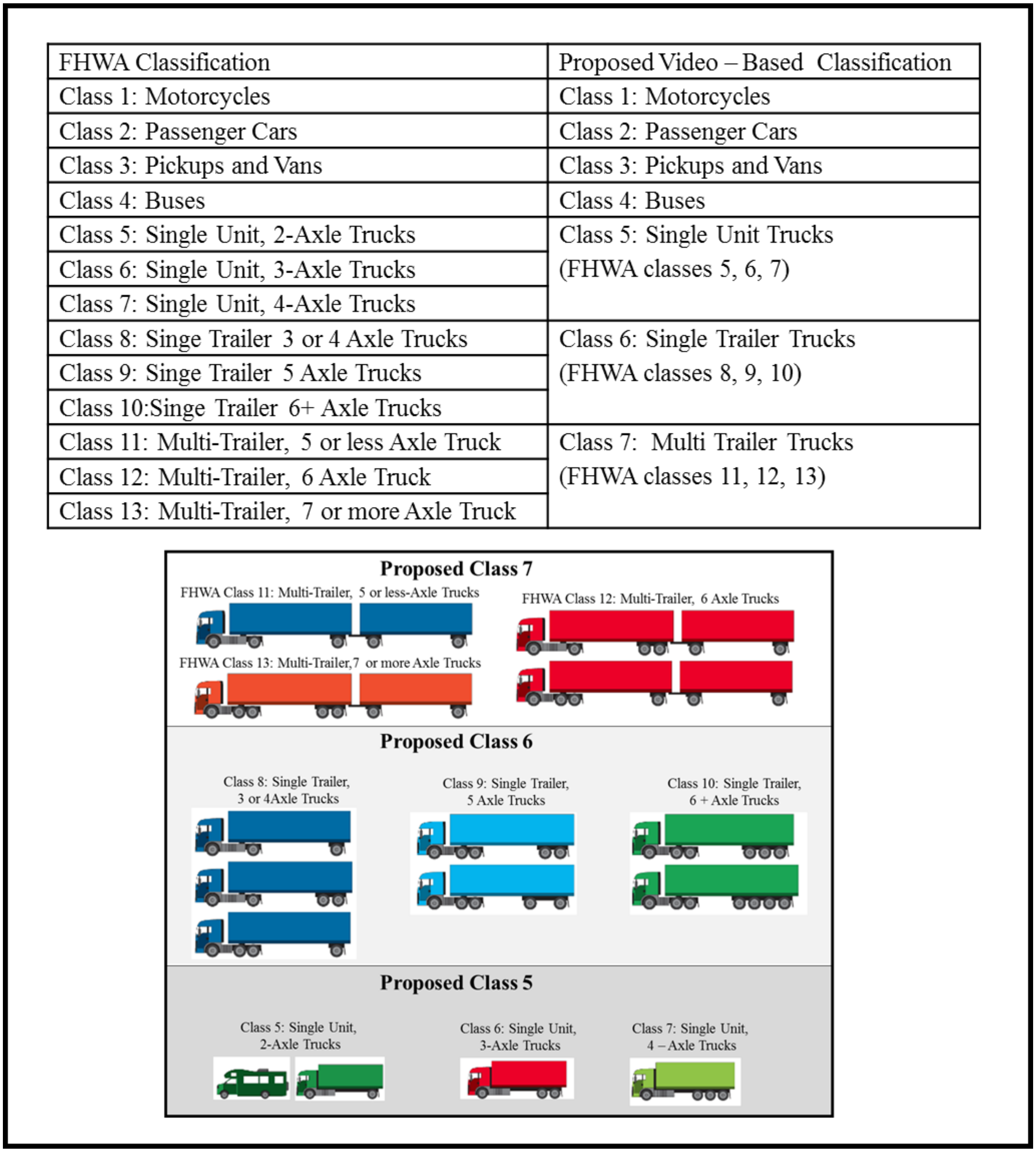


FIGURE 3 Architecture of Deep Convolutional Neural Networks for Vehicle Classification.

DATA PROCESSING AND EXPERIMENTS

The primary sources of data for evaluating the proposed approach to vehicle recognition and classification included the Iowa DOT CCTV camera database and 511 Virginia. The CCTV cameras covered both freeways and non-freeway road types. They acquired videos of traffic scenes at sampling rates of 12, 25, and 30 frames per second. The conditions under which videos were acquired varied: day, night and dawn; snow, rain and sunshine; congested and non-congested traffic conditions. The cameras also had different views of traffic (top-down, side or front views) and were installed at varying heights above ground. In other words, the data being used to develop and evaluate the vision system captures the key challenges of conventional automated vehicle recognition and classification systems.

The vision system is trained to detect and classify vehicle according to the FHWA scheme. However, some of the classes in the FHWA scheme had to be merged because of the subtle differences between these classes which could not be visually differentiated in a video. The table in Figure 4 illustrates the differences between the FHWA classification scheme and the proposed video-based classification approach. In addition, a visual example of classes that were merged is also shown in Figure 4. Clearly, the key differences between merged classes is the number of axles. The view angle and height of CCTV cameras makes it challenging to distinguish different vehicle types based solely on axle configurations. Reducing the height and using a side camera view instead of a top-down view could be useful in this case. However, this configuration will increase occlusions especially during congested conditions.



FHWA Classification	Proposed Video – Based Classification
Class 1: Motorcycles	Class 1: Motorcycles
Class 2: Passenger Cars	Class 2: Passenger Cars
Class 3: Pickups and Vans	Class 3: Pickups and Vans
Class 4: Buses	Class 4: Buses
Class 5: Single Unit, 2-Axle Trucks	Class 5: Single Unit Trucks (FHWA classes 5, 6, 7)
Class 6: Single Unit, 3-Axle Trucks	
Class 7: Single Unit, 4-Axle Trucks	
Class 8: Singe Trailer 3 or 4 Axle Trucks	Class 6: Single Trailer Trucks (FHWA classes 8, 9, 10)
Class 9: Singe Trailer 5 Axle Trucks	
Class 10:Singe Trailer 6+ Axle Trucks	
Class 11: Multi-Trailer, 5 or less Axle Truck	Class 7: Multi Trailer Trucks (FHWA classes 11, 12, 13)
Class 12: Multi-Trailer, 6 Axle Truck	
Class 13: Multi-Trailer, 7 or more Axle Truck	

Proposed Class 7
FHWA Class 11: Multi-Trailer, 5 or less-Axle Trucks
FHWA Class 12: Multi-Trailer, 6 Axle Trucks
FHWA Class 13: Multi-Trailer,7 or more Axle Trucks
Proposed Class 6
Class 8: Single Trailer, 3 or 4Axle Trucks
Class 9: Single Trailer, 5 Axle Trucks
Class 10: Single Trailer, 6 + Axle Trucks
Proposed Class 5
Class 5: Single Unit, 2-Axle Trucks
Class 6: Single Unit, 3-Axle Trucks
Class 7: Single Unit, 4 – Axle Trucks

FIGURE 4 Merged Classes [22]

Training and Test Set

The CCTV camera data acquired from all the different sources were divided into a training and test sets. The training set is used to assist the DCNN model to learn unique features of the different types of vehicles. Whereas test set is used for evaluating how accurately the model has learned from the training data.

Training Database: The training database contains a set of positive and negative image samples. A positive image sample denotes images which contains either one or more of the 7 proposed vehicle classes. A negative sample on the other hand does not contain the target object to be identified. These images have associated bounding box annotation labels which indicates the specific location (top-left and bottom right corner) of the target object. The total number of positive training samples generated for each category class is shown in figure 5. The background objects of positive image samples were used as negative samples. For each positive image, we randomly sample 3 background objects as negatives.

The total time required for training the DCNN network on a Titan X GPU is approximately 3 hours.

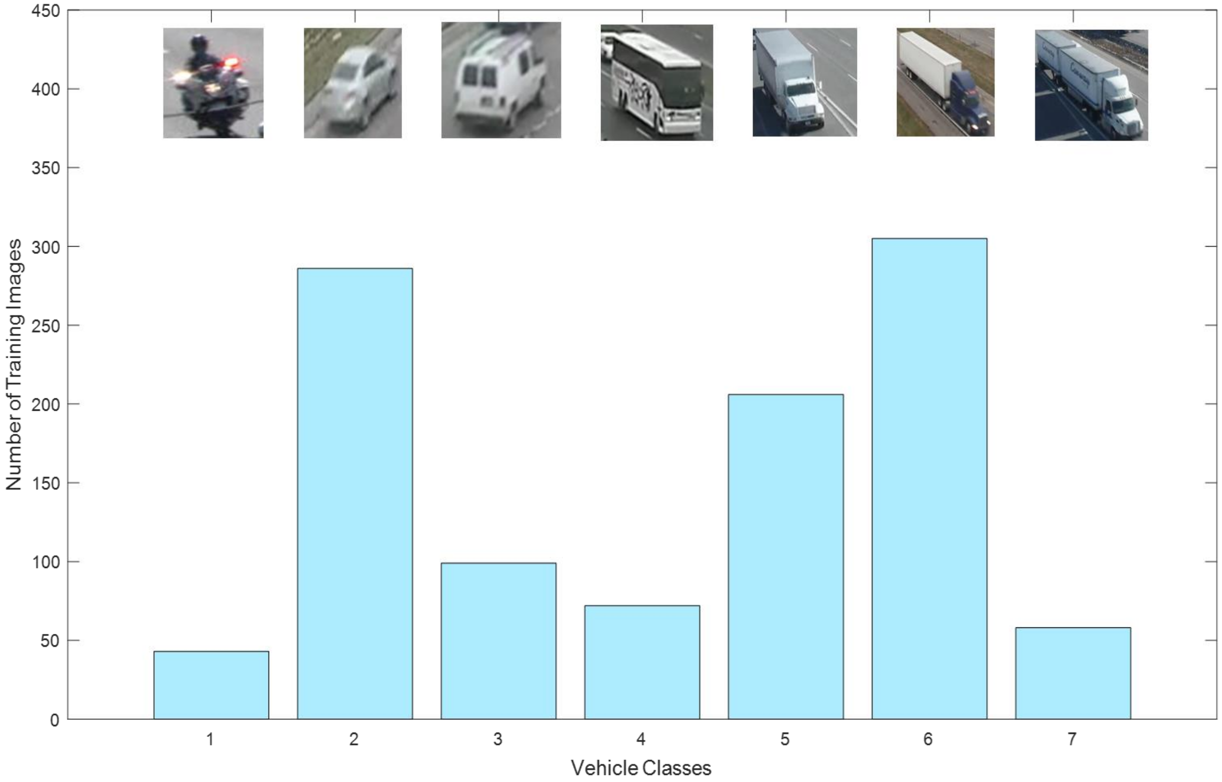


FIGURE 5 Histogram Showing Proportion of Training Image Set per Category Class.

Test Database: The test set consisted of 30 randomly selected videos, each with a total footage of approximately 5 minutes. The videos were manually tagged according to: vehicle location (top-left and bottom right corner), vehicle class (Class 1 through 7) and video frame number. All videos in our test set were analyzed using the DCNN model developed and OpenCV [23]. For each video frame, the selective search algorithm is used to identify candidate region proposals. Features are then computed for all region proposals and a linear SVM is used to classify each object proposal. Each frame of the video processed, returned an output indicating which of the 7 vehicle classes was detected.

EXPERIMENTAL RESULTS AND SYSTEM'S PERFORMANCE EVALUATION

To evaluate the performance of the system developed, outputs from the vision system were compared with results from the manually tagged videos in the test database. Precision and Recall rates defined in equations 1 and 2 are used as the measure of the systems' overall performance. A True Positive (TP) represents a detected and correctly classified vehicle which has a corresponding manually tagged object in the test database. A False Positive represents a detected and classified vehicle which has no corresponding manually tagged object in the test database. A detected but misclassified vehicle is also denoted as a False Positive even if it has a corresponding manually tagged object. A False Negative (FN) represents objects which were completely missed by the vision system.

$$precision = \frac{TP}{TP + FP} \qquad (1)$$

$$recall = \frac{TP}{TP + FN} \qquad (2)$$

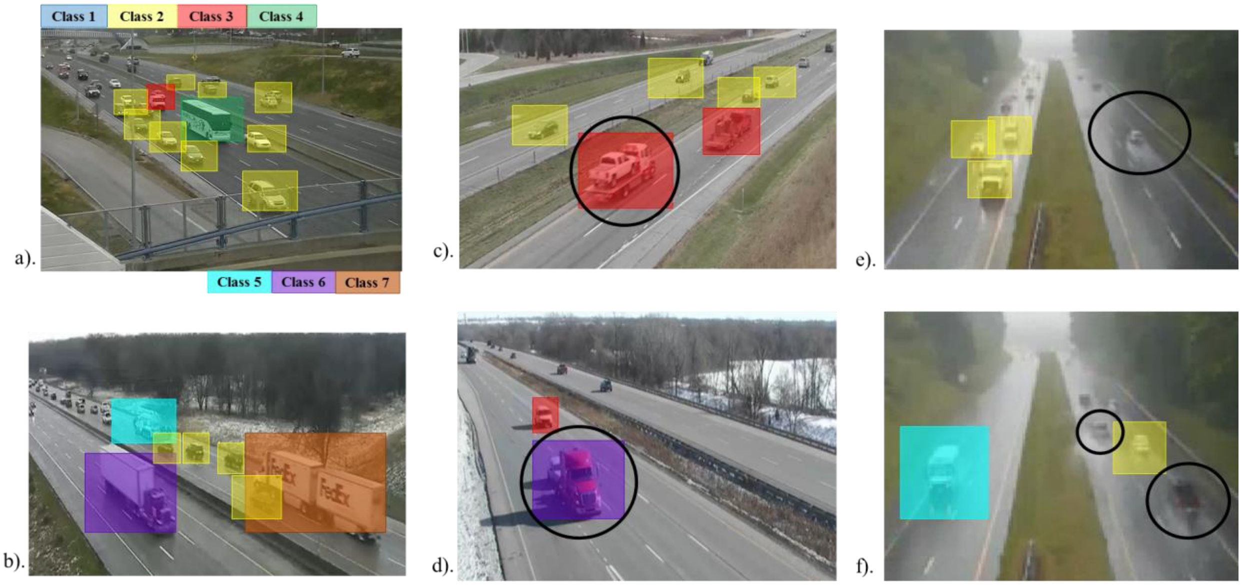


FIGURE 6 Vision System Output Examples: a – b): True Positives. c – d): False Positives. e – f): Misses.

In Figure 6 above, we show examples of positive, false and missed calls from the vision system. Figures 6a and 6b show examples of vehicles that our vision system correctly recognized. False detections are shown in Figures 6c and 6d. In figure 6c, a class 5 vehicle towing a class 3 vehicle is falsely classified as class 3. In figure 6d the truck detected has no trailer and therefore could either belong to a class 6 or 7, however, the system classifies it as a class 6 type vehicle. Completely missed objects as shown in Figures 6e and 6f were mostly prominent in very poor resolution cameras. Also, distance between the camera and the object may influence the vehicle classification accuracy. For example, a double trailer truck may begin to look like a single trailer truck as the vehicle moves away from the camera. The table in figure 7 illustrates the average precision and recall rates of the vision system for detecting all 7 classes of objects from videos in our test database. On average, the vision system developed correctly detected and classified vehicles in the test database 95% (average precision) of the time. 93% (recall rate) of all vehicle types in the test database were detected and classified. The following observations can be deduced for each category class:

Class	Total Images	TP	FP	FN	Precision (%)	Recall (%)
Class 1	18	16	0	2	1.00	0.89
Class 2	8542	7250	410	882	0.95	0.89
Class 3	1348	1007	223	118	0.82	0.90
Class 4	109	106	0	3	1.00	0.97
Class 5	567	528	22	17	0.96	0.97
Class 6	3976	3600	323	53	0.92	0.99
Class 7	165	152	10	3	0.94	0.98

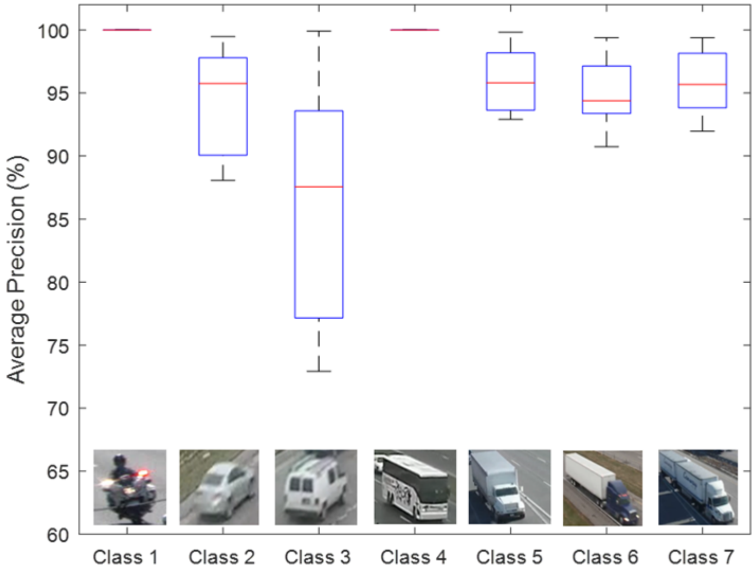


FIGURE 7 Performance Evaluations: Table shows average precision and recall rates of the system. Box plots shows range of precision rates for all 30 videos in the test database.

Classes 1 and 4: In spite of limited training data for Class 1 and 4 vehicle types, motorcycles and buses are the simplest objects to recognize using the system developed. Hence, a 100% precision rate. They are easily distinguishable from other classes as shown in the confusion matrix in Figure

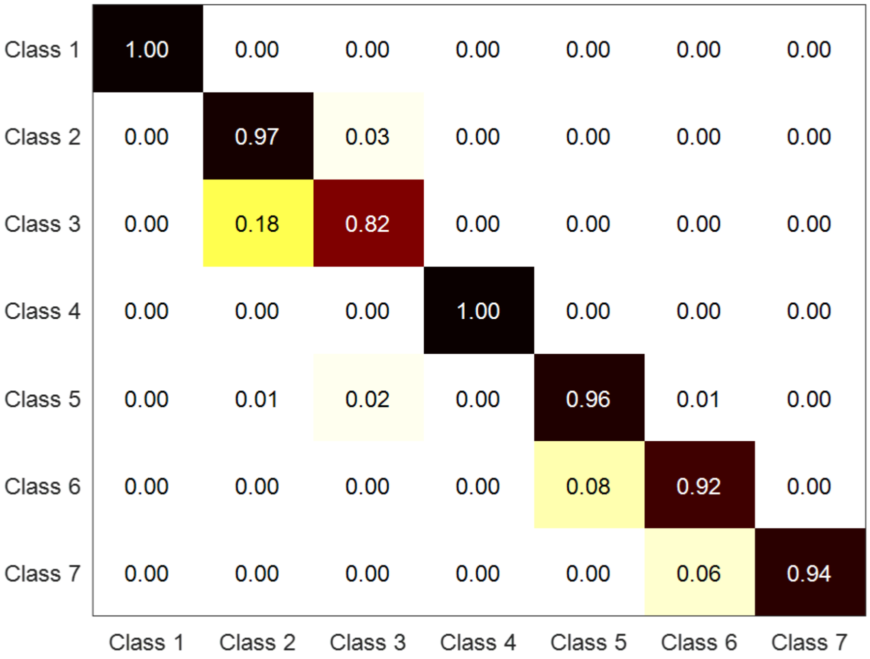
8. Due to the size of class 1 vehicle types, they are likely to be missed in poor resolution videos or occluded by larger trucks hence a relatively low recall rate.

Class 2 (Passenger Cars and SUVs): 95% (precision) of the time, the system was able to correctly recognize vehicles belonging to this class type. Class 2 type vehicles constitutes the largest proportion of vehicular traffic. Hence, this precision rate is appreciable. However, 89% (recall) of all class 2 type vehicles in the test database was recognized. The reason for a relatively lower recall rate is mainly due to occlusions from trucks.

Class 3 (Vans and Pickups): The system was least effective at recognizing class 3 vehicle types. It is able to correctly recognize vehicles belonging to this class type only 82% of the time although 90% (recall) of all class 3 type vehicles in the test database was recognized. The box plot of precision rates for class 3 shows the most variation, ranging between 73% and 98%. The drastic drop in precision rates is due to the inclusion of pickup trucks in this category class. Pickup trucks come in different forms: covered, uncovered, single or double cabin. The system confuses covered pickup trucks with SUVs which belongs to a different category class. Also, single cabin pickups are considered as class 2. The confusion matrix in Figure 8 confirms this observation. Distinguishing between single and double cabin pickups from a top-down view CCTV camera can be challenging even with the human eye.

Class 5 (Single Unit Trucks): The average precision and recall rates for class 5 shows that the vision system had few difficulties recognizing vehicles belonging to this category type. The majority of false positives (although few) in this category were related to single unit trucks towing a class 2 or class 3 type vehicle (see figure 6c). This confused the system and mostly led to misclassification. From the confusion matrix, this happens only 4% of the time. Some trucks were missed if the system could only see a distant rear-view and not the front-view of the truck. Truck to truck occlusions was also observed in some few cases.

Class 6 (Single-Trailer) and 7 (Multi-Trailer): Single and multi-trailers also had appreciable precision and recall rates even in very poor resolution videos and congested conditions. From the confusion matrix in Figure 8, they are easily distinguishable from the other 5 classes. However, a multi-trailer begins to look like a single – trailer as the truck moves away from the camera. This is one reason for a relatively higher false positive rate for the class 6 category. Another source of false positive detections is when a truck has removed its trailer (see figure 6d). Such situations mostly confused the vision system. The system missed some trucks belonging to this category class if it could only see a distant rear-view and not the front-view of the truck. Truck to truck occlusions was observed in some few cases.



	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7
Class 1	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Class 2	0.00	0.97	0.03	0.00	0.00	0.00	0.00
Class 3	0.00	0.18	0.82	0.00	0.00	0.00	0.00
Class 4	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Class 5	0.00	0.01	0.02	0.00	0.96	0.01	0.00
Class 6	0.00	0.00	0.00	0.00	0.08	0.92	0.00
Class 7	0.00	0.00	0.00	0.00	0.00	0.06	0.94

FIGURE 8 Confusion Matrix for the 7 Vehicle Classes.

Sensitivity Analysis

Finally, we investigated conditions and configurations which could influence the performance of the vision system developed. The sensitivity of the proposed system to three key conditions was evaluated: first, we looked at the influence of time of day vehicle recognition is carried out (day and night time); secondly, the prevailing traffic conditions (free flow and congested, stop and go traffic); and lastly, the camera resolution (blurring, sampling rates, rain, and snow). To evaluate the sensitivity of the system, the test database is partitioned into 8 subgroups according to the combination of the factors influencing the effectiveness of the vision system developed.

The vision system is used to process videos from each of these subgroupings. Receiver Operating Characteristics (ROC) curves are then used to compare the performance of the system per each subgroup based on the true positive vs. false positive rates. True Positive and False Positive Rates are as defined in equations 3 and 4.

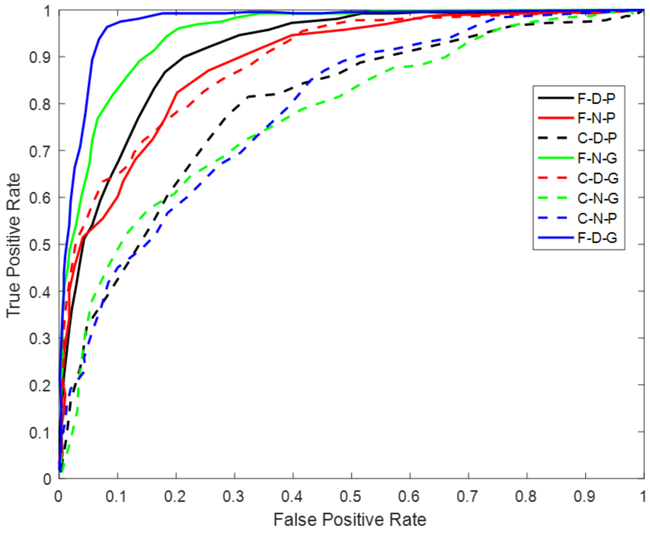
$$TPr = TP/(TP + FN) \tag{3}$$

$$FPr = FP/(FP + TN) \tag{4}$$

Where TPr is the True Positive rate and FPr is the False Positive Rate. TP, FP and FN are as defined in equations 1 and 2. TN represents the total number of non-vehicular objects that were not classified as vehicles. For a poorly constructed vision system, as its sensitivity (true positive rate) increases, it also loses the ability to discriminate between vehicular and non-vehicular objects such as shadows, buildings, trees, etc. As a result, the true positive and false positive rates are almost directly proportional. On the other hand, the mark of a good vision system is that its true positive rates are marginally higher than the corresponding false positive rates. Figure 9 shows the

ROC curves for each subgrouping and a table of the calculated area under each curve. The following deductions can be drawn.

In general, Figure 9 shows that the true positive rates are marginally higher than the corresponding false positive rate irrespective of traffic condition or camera configuration. It is however evident that the prevailing traffic conditions clearly has an impact on the performance of the vision system. Under congested conditions, the system can reach a high true positive rate (90% or more) only if it incurs a false positive rate between 25% and 55%. On the other hand, during free flow conditions, the system generally incurs between 5% and 30% false positive rate in order to reach a high true positive rate of 90% or more. The table in Figure 9 also shows an observable difference between the areas under the curve for free flow and that for congested conditions. The influence of the video quality on the systems performance is minimal during free flow conditions. Marginal effects of poorly resolved videos are however observed when traffic conditions are congested and the time of day is at night. Generally, the system is relatively more effective at processing day time videos as compared to night time. Under free flow conditions, the influence of time of day is insignificant. The influence of the time of day is critical when video resolution is low and traffic condition is congested.



Condition	Area Under Curve
Free Flow, Day Time, Good Resolution (F-D-G)	0.972
Free Flow, Day Time, Poor Resolution (F-D-P)	0.911
Free Flow, Night Time, Good Resolution (F-N-G)	0.948
Free Flow, Night Time, Poor Resolution (F-N-P)	0.886
Congested, Day Time, Good Resolution (C-D-G)	0.888
Congested, Day Time, Poor Resolution (C-D-P)	0.793
Congested, Night Time, Good Resolution (C-N-G)	0.785
Congested, Night Time, Poor Resolution (C-N-P)	0.782

FIGURE 9 Receiver Operating Characteristics (ROC) curves comparing the influence of a combination of factors that influence the performance of the vision system.

Figure 9 also suggests that although the system is robust to conditions such as time of day and video resolution, the combined effect of these factors could drastically degrade the performance of the system. For example, if traffic condition is congested and at the same time video resolution is poor and the time of day is night, the false positive rate reaches 45% for a true positive rate greater than 85%. In order to get the best results out of the system in such conditions, the use of a camera with frame rates greater than 30 frames per second is suggested. Also, areas where video cameras are mounted should be well illuminated. This will reduce the combined influence of camera resolution and time of day on the performance of the system.

SUMMARY AND CONCLUSIONS

The performance of video-based recognition systems for Intelligent Transportation System (ITS) purposes have stagnated in recent years. The best performing systems can only report up to 3-4 classes as compared to 13 classes required by FHWA's vehicle classification scheme. The current study takes advantage of recent advances in machine vision and high performance computing to accurately learn unique vehicular features that can be used to report finer classifications comparable to other non-intrusive recognition systems.

The vision system developed achieved average precision rates between 82% and 100% and average recall rates between 89% and 99% for 7 classes of vehicles. Motorbikes and buses are the most easily recognizable class categories by the system, followed by passenger cars, and single and double trailer trucks. Class 3 type vehicles which included vans and pickup trucks were the most challenging. ROC curves were used to evaluate the sensitivity of the system to different camera configurations, traffic, and lighting conditions. Overall, the best system performance can be achieved under free flow traffic, during day or night time with good video resolution. Under congested conditions, the user is likely to incur between 15 – 30 percent false positive rates in order to achieve a true positive rate greater than 90 percent. However, it is noteworthy that the performance of the proposed vision system under congested conditions during day time is significantly better than during the night.

This performance was achieved through two main tasks. First, the selective search algorithm was used to generate class-independent region proposals in order to localize and segment objects. Secondly, DCNN descriptors for each proposed region were extracted and classified through a linear SVM scoring system.

Future studies should look at model architectural designs which could be used to increase the number of classes that can be accurately distinguished by the vision system. Also, tracking algorithms could be built to aid in vehicle counting and other traffic management tasks such as congestion detection, stranded vehicle detection, etc. A comparison to other existing automated video-based vehicle recognition systems will also be expedient.

REFERENCES

1. Minge, E., K. Jerry, and S. Peterson. *Evaluation of Non-Intrusive Technologies for Traffic Detection*. Publication MN/RC 2010-36. Minnesota Department of Transportation, 2010.
2. Fekpe, E., D. Gopalakrishna, and D. Middleton. *Highway Performance Monitoring System Traffic Data for High-Volume Routes: Best Practices and Guidelines*. Office of Highway Policy Information. FHWA, U.S. Department of Transportation, 2004.
3. *Traffic Monitoring Handbook*. Transportation Statistics Office of Florida Department of Transportation. http://www.dot.state.fl.us/planning/statistics/tmh/tmh.pdf. Accessed July 2, 2016.
4. Zhang, G., R.P. Avery, and Y. Wang. A Video-based Vehicle Detection and Classification System for Real-time Traffic Data Collection Using Uncalibrated Video Cameras. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1993*, Transportation Research Board of the National Academies, Washington D.C., 2007, pp. 138–147.

5. Transportation Research Board (TRB). *Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 2010.
6. American Association of State Highway and Transportation Officials (AASHTO). *AASHTO Guide for Design of Pavement Structures*. AASHTO, Washington D.C. 1993.
7. Federal Highway Administration (FHWA). *Highway Performance Monitoring System*. http://www.fhwa.dot.gov/policyinformation/hpms/reviewguide.cfm. Accessed July 2, 2016.
8. Lai, A.H.S., G.S.K. Fung, and N.H.C. Yung. *Vehicle Type Classification from Visual-Based Dimension Estimation*. Proceedings of the IEEE Intelligent Transportation Systems Conference, Oakland, CA, 2001, pp. 201-206.
9. Avery, R. P., Y. Wang, and G. S. Rutherford. *Length-Based Vehicle Classification Using Images from Uncalibrated Video Cameras*. Proceedings of the 7th International IEEE Conference on Intelligent Transportation Systems, 2004, pp. 737-742.
10. Gupte, S., O. Masoud, R.F.K. Martin, and N.P. Papanikolopoulos. *Detection and Classification of Vehicles*. IEEE Transactions on Intelligent Transportation Systems, Vol. 3, No. 1, 2002, pp. 37-47.
11. Zhou, Y., and N. Cheung. *Vehicle Classification Using Transferrable Deep Neural Network Features*. http://arxiv.org/pdf/1601.01145.pdf. Accessed July 2, 2016.
12. Moussa, G.S. Vehicle Type Classification with Geometric and Appearance Attributes. *International Journal of Civil, Environmental, Structural, Construction and Architectural Engineering* Vol: 8, No. 3, 2014, pp. 277-282.
13. LeCun, Y., B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation* Vol. 1, 1989, pp. 541-551.
14. Krizhevsky, A., I. Sutskever, and G. Hinton. *ImageNet Classification with Deep Convolutional Neural Networks*. Proceedings from Advances in Neural Information Processing Systems Conference, 2012, pp. 1106-1114.
15. Uijlings, J.R.R., K.E. Van de Sande, T. Gevers, and A.W.M. Smeulders. Selective Search for Object Recognition. *International Journal of Computer Vision*, Vol.104 No. 2, 2013, pp.154-171.
16. Abramov,K.V., Skribtsov, P.V and Kazantsev, P.A. *Image Segmentation Method Selection for Vehicle Detection Using Unmanned Aerial Vehicle.* Modern Applied Science; Vol. 9, No. 5; 2015
17. Felzenszwalb, P. F., Girshick, R. B., McAllester, D. and Ramanan, D. *Object detection with discriminatively trained part based models*. TPAMI, 32:1627–1645, 2010.
18. Collobert, R., K. Kavukcuoglu, and C. Farabet. Torch7: A Matlab-like Environment for Machine Learning. In *BigLearn NIPS Workshop*, 2011.
19. Jia, Y. Caffe: *An Open Source Convolutional Architecture for Fast Feature Embedding*. http://ucb-icsi-vision-group.github.io/caffe-paper/caffe.pdf. Accessed July 2, 2016.
20. Girshick, R.B., J. Donahue, T. Darrell, and J. Malik. *Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation*. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2014, pp. 580-587.
21. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C. and Fei-Fei, L. *ImageNet Large Scale Visual Recognition Challenge*. *IJCV*, 2015

22. Randall, J.L. *Traffic Recorder Instruction Manual*, http://onlinemanuals.txdot.gov/txdotmanuals/tri/vehicle_classification_using_fhwa_13category_scheme.htm. Accessed July 2, 2016.

23. Bradski, G. The OpenCV Library. Dr. Dobb’s Journal of Software Tools, 2000.