2017

Automated Vehicle Recognition with Deep Convolutional Neural Networks

Yaw Okyere Adu-Gyamfi
University of Virginia

Noblis, Inc.

Anuj Sharma
Iowa State University, anujs@iastate.edu

See next page for additional authors

Automated Vehicle Recognition with Deep Convolutional Neural Networks

Abstract
In recent years there has been growing interest in the use of nonintrusive systems such as radar and infrared systems for vehicle recognition. State-of-the-art nonintrusive systems can report up to eight classes of vehicle types. Video-based systems, which arguably are the most popular nonintrusive detection systems, can report only very coarse classification levels (up to four classes), even with the best-performing vision systems. The present study developed a vision system that can report finer vehicle classifications according to FHWA's scheme and is also comparable to other nonintrusive recognition systems. The proposed system decoupled object recognition into two main tasks: localization and classification. It began with localization by generating class-independent region proposals for each video frame, then it used deep convolutional neural networks to extract feature descriptors for each proposed region, and, finally, the system scored and classified the proposed regions by using a linear support vector machines template on the feature descriptors. The precision of the system varied by vehicle class. Passenger cars and SUVs were detected at a precision rate of 95%. The precision rates for single-unit, single-trailer, and double-trailer trucks ranged between 92% and 94%. According to receiver operating characteristic curves, the best system performance can be achieved under free flow, daytime or nighttime, and with good video resolution.

Disciplines
Civil Engineering | Operational Research | Transportation Engineering

Comments
This article is published as Adu-Gyamfi, Yaw Okyere, Sampson Kwasi Asare, Anuj Sharma, and Tienaah Titus. "Automated Vehicle Recognition with Deep Convolutional Neural Networks." *Transportation Research Record: Journal of the Transportation Research Board* 2645 (2017): 113-122. DOI: 10.3141/2645-13. Posted with permission.

Authors
Yaw Okyere Adu-Gyamfi; Noblis, Inc.; Anuj Sharma; and Tienaah Titus

Automated Vehicle Recognition with Deep Convolutional Neural Networks

Yaw Okyere Adu-Gyamfi, Sampson Kwasi Asare, Anuj Sharma, and Tienaah Titus

In recent years there has been growing interest in the use of nonintrusive systems such as radar and infrared systems for vehicle recognition. State-of-the-art nonintrusive systems can report up to eight classes of vehicle types. Video-based systems, which arguably are the most popular nonintrusive detection systems, can report only very coarse classification levels (up to four classes), even with the best-performing vision systems. The present study developed a vision system that can report finer vehicle classifications according to FHWA's scheme and is also comparable to other nonintrusive recognition systems. The proposed system decoupled object recognition into two main tasks: localization and classification. It began with localization by generating class-independent region proposals for each video frame, then it used deep convolutional neural networks to extract feature descriptors for each proposed region, and, finally, the system scored and classified the proposed regions by using a linear support vector machines template on the feature descriptors. The precision of the system varied by vehicle class. Passenger cars and SUVs were detected at a precision rate of 95%. The precision rates for single-unit, single-trailer, and double-trailer trucks ranged between 92% and 94%. According to receiver operating characteristic curves, the best system performance can be achieved under free flow, daytime or nighttime, and with good video resolution.

Transportation agencies seeking to optimize traffic mobility and improve safety need accurate traffic data. Traditional technologies used in traffic data collection, such as piezoelectric sensors, magnetic loops, and pneumatic road tubes, have been popular among transportation agencies since the 1960s (*1*). However, these traditional methods are gradually giving way to emerging advanced collection technologies, such as active infrared or laser, radar, and video, for many reasons, including the damage caused by intrusive traditional methods, environmental conditions (e.g., snow) that inhibit their use, frequent equipment malfunctions, lack of consistent accurate data, and disruptions to traffic during installation (*2*). With the proliferation of advanced traffic data collection technologies, transportation professionals must identify the appropriate technology that suits an agency's data collection needs.

One of the many data needs of transportation agencies is vehicle type classification (*3*). Accurate classification data are fundamental to traffic operation, pavement design, and transportation planning (*4*). For example, the total number of trucks in a section of a roadway is useful for computing the corresponding passenger car equivalents needed to estimate 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 will use such roadways (*6*). Under federal requirements for the Highway Performance Monitoring System, states must perform classified vehicle counts on freeways and highways and provide this information to FHWA every year (*7*). Vehicle classification data are therefore critical to the effective management and operation of transportation systems.

Many 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 systems. In most instances, classification is based on the dimensions of vehicles. Lai et al. demonstrated the estimation of accurate vehicle dimensions by using a set of coordinate mapping functions (*8*). Although they were able to estimate vehicle lengths to within 10% in every instance, their method requires camera calibration 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 et al. 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 (*10*). Before vehicles can 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.

Vehicle type classification with advanced techniques such as artificial intelligence has been proposed in the literature. Zhou and Cheung proposed the use of deep neural networks (DNN) to classify vehicles (*11*). Since their test data set was small compared with the 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 data sets that include

Y. O. Adu-Gyamfi, Department of Civil and Environmental Engineering, School of Engineering and Applied Science, University of Virginia, P.O. Box 400742, Charlottesville, VA 22904-4742. S. K. Asare, Noblis, Inc., Suite 700E, 600 Maryland Avenue, SW, Washington, DC 20024. A. Sharma, Civil, Construction, and Environmental Engineering, College of Engineering, Iowa State University, 352 Town Engineering, Ames, IA 50011. T. Titus, Geodesy and Geomatic Engineering, University of New Brunswick, P.O. Box 4400, Fredericton, New Brunswick E3B 5A3, Canada. Corresponding author: Y. O. Adu-Gyamfi, yoa4q@virginia.edu.

Transportation Research Record: Journal of the Transportation Research Board, No. 2645, 2017, pp. 113–122.
http://dx.doi.org/10.3141/2645-13

trucks is unknown. The support vector machine (SVM) technique has been used to conduct multiclass and intraclass vehicle type classifications (*12*). In that study, two vehicle classification approaches that use the SVM algorithm were presented: (*a*) a geometric-based approach and (*b*) an appearance-based approach. Although combining geometry and appearance to classify vehicles sounds encouraging, the proposed system classifies vehicles into only small, medium, and large categories, so its application is limited.

This paper proposes a video-based vehicle detection and classification system for classifying vehicles according to FHWA's 13 vehicle types. 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 state-of-the-art nonintrusive 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. However, they fell out of fashion with the rise of SVMs. Interest in DCNNs was rekindled in 2012 by Krizhevsky et al. (*14*), who showed that a substantially higher accuracy for image classification could be achieved in the ImageNet data set with DCNNs. Since its rebirth, profound improvements in the accuracy of object detection in complex scenes have been achieved.

The research presented here developed an automated, video-based vehicle recognition system that

1. Classifies vehicles according to the FHWA classification scheme and
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 this paper is as follows. First, an overview of the proposed approach to automated vehicle recognition and classification is provided. This section highlights the machine vision algorithms selected for this study, and the training and fine-tuning of the algorithm are 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 are discussed in this section. The third section discusses results of experiments using the developed vision system to process closed-circuit television (CCTV) video data under varying conditions. Concluding remarks, recommendations, and additional research needs 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 (256×256). Finally, feature descriptors corresponding to each proposed region are classified through a linear SVM scoring system. Figure 1 summarizes the proposed approach to automated vehicle recognition.

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 difficult 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 by using various sliding window approaches (*17*). The main challenge in use of exhaustive search alone for object detection is that it fails to detect objects with low-level cues.

Uijlings et al. developed Selective Search, an approach that combines the best of both worlds: segmentation and exhaustive search (*15*). It exploits the hierarchical structure of the image (segmentation) to generate 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 various invariance properties. Finally, region-based similarity functions are used to address the diversity of objects. Figure 2 shows proposed object regions resulting from use of selective search.

Object Classification with DCNN

After object localization with selective search, each detected object is fed through a DCNN for classification. In this study, a DCNN

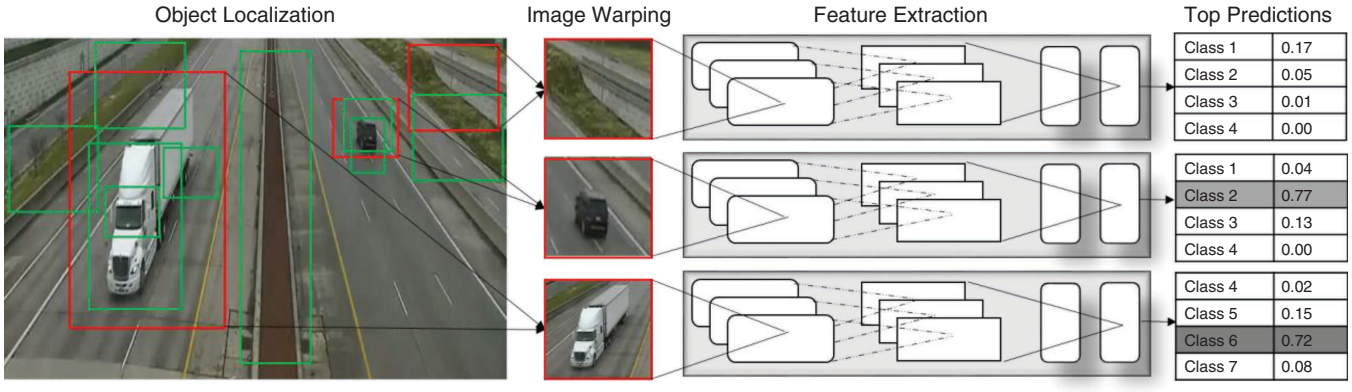
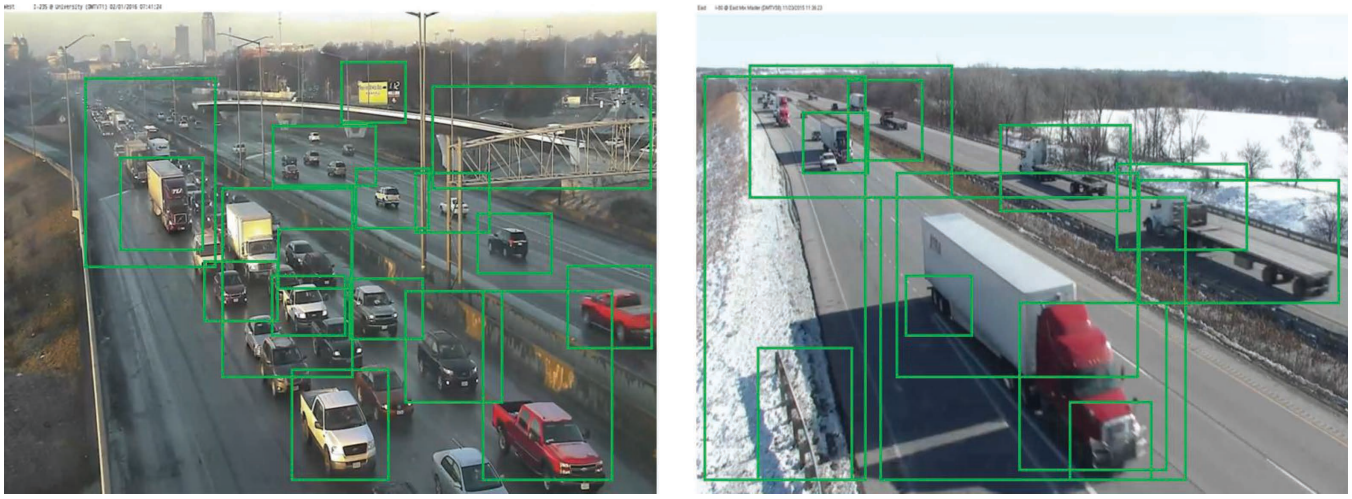


FIGURE 1 Approach to vehicle detection and classification.



(a) (b)

FIGURE 2 Candidate region proposals using selective search.

classifier was built to support key algorithms for classifying vehicles captured on CCTV cameras. The following section explains the architecture used to build and test the DCNN classifier.

Model Training

DCNN models are computationally expensive, which makes them 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 (*19*). 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 pretraining and domain-specific fine-tuning.

Supervised Pretraining

A DCNN model usually consists of thousands of parameters and millions of learned weights. Thus, a very large training data set (more than a million records) is needed to avoid overfitting the model. Girshick et al., however, demonstrated that when labeled data are scarce, supervised pretraining for an auxiliary task with large training data followed by domain-specific fine-tuning (on a smaller data set) could significantly boost performance (*20*). A similar approach was adopted here through pretraining the DCNN model on a large auxiliary data set (ILSVRC2012) (*21*) with image-level annotations. The resulting output is a rich feature detector that was fine-tuned to suit this study's purposes. The open source Caffe DCNN library was used for the pretraining model on 100 classes at a learning rate of 0.01.

Domain-Specific Fine-Tuning

To adapt the pretrained model to the proposed task (vehicle recognition), the CNN model parameters are fine-tuned. First, the 100-way classification layer of the pretrained model is replaced with seven classes. Stochastic gradient descent is started at a learning rate of 0.001, which allows fine-tuning to make progress while not clobbering the initialization. In each iteration of the stochastic gradient descent, 20 positive windows for all classes and 70 background windows are uniformly sampled to construct a minibatch of size 90. DCNN is used to extract a 4,096-dimensional feature vector with Caffe's implementation of CNN by Krizhevsky (*14*). 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 with linear SVMs trained for that specific class. The modeling architecture is shown in Figure 3 and is summarized as follows:

1. Each class-independent region proposal from the previous step is warped to a 256×256 image.
2. The input warped image is filtered with 96 kernels of size 11×11, with a stride of 4 pixels. This is followed by max pooling in a 3×3 grid.

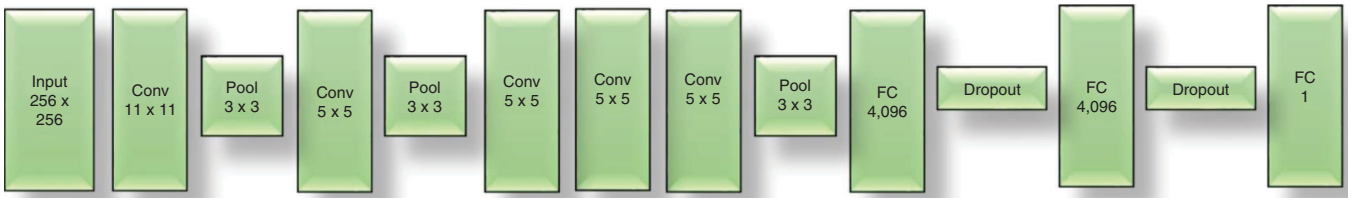


FIGURE 3 Architecture of DCNN for vehicle classification (conv = convolution; pool = pooling; FC = fully connected).

3. Two subsequent convolutions with 384 kernels are carried out without pooling.
4. The output of the fourth layer is convolved with 256 kernels, then spatial max pooling is applied in a 3- × 3-pixel grid.
5. For the last two layers, a fully connected layer of 4,096 dimensions is extracted from the last layer.

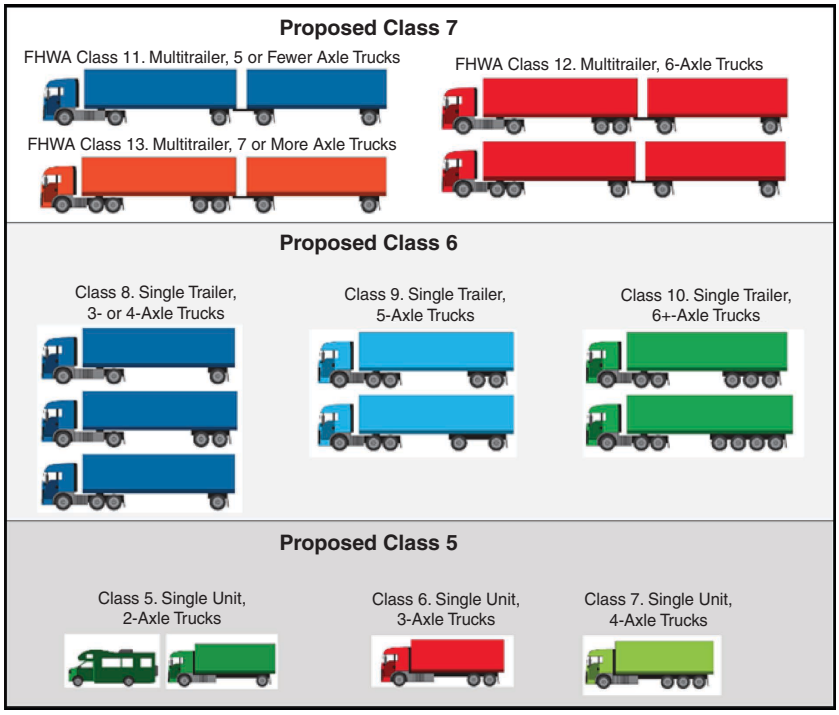
DATA PROCESSING AND EXPERIMENTS

The primary sources of data for evaluating the proposed approach to vehicle recognition and classification were the Iowa Department of Transportation CCTV camera database and 511 Virginia. The CCTV cameras covered both freeways and nonfreeway roads. 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; and congested and noncongested traffic conditions. The cameras had different views of traffic (top-down, side, or front views) and were installed at varying heights above ground. In other words, the data used to develop and evaluate the vision system captured the key challenges of conventional automated vehicle recognition and classification systems.

The vision system was trained to detect and classify vehicles according to the FHWA scheme. However, some of the classes in the FHWA scheme had to be merged because of the subtle differences between them that could not be visually differentiated in a video. Figure 4*a* illustrates the differences between the FHWA classifica-

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. Single-trailer 3- or 4-axle trucks	Class 6. Single-trailer trucks (FHWA Classes 8, 9, 10)
Class 9. Single-trailer 5-axle trucks	
Class 10. Single-trailer 6+-axle trucks	
Class 11. Multitrailer, 5 or fewer axle trucks	Class 7. Multitrailer trucks (FHWA Classes 11, 12, 13)
Class 12. Multitrailer, 6-axle trucks	
Class 13. Multitrailer, 7 or more axle trucks	

(a)



(b)

FIGURE 4 Merged classes (*22*).

tion scheme and the proposed video-based classification approach. An illustration of the classes that were merged is given in Figure 4*b*. The key difference between merged classes is the number of axles. The view angle and height of CCTV cameras make it challenging to distinguish different vehicle types solely on the basis of axle configurations. Reducing the height and using a side camera view instead of a top-down view could be useful in this case. However, such a configuration will increase occlusions, especially during congested conditions.

Training and Test Set

The CCTV camera data acquired from all the sources were divided into training and test sets. The training set is used to help the DCNN model learn unique features of the types of vehicles. The test set is used to evaluate how accurately the model learns from the training data.

Training Database

The training database contains a set of positive and negative image samples. A positive image sample denotes images that contain either one or more of the seven proposed vehicle classes. A negative sample does not contain the target object to be identified. These images have associated bounding box annotation labels that indicate the specific location (top-left and bottom-right corners) of the target object. The total positive training samples generated for each category class are shown in Figure 5. The background objects of positive image samples were used as negative samples. For each positive image, three background objects were randomly sampled as negatives.

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

Test Database

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

EXPERIMENTAL RESULTS AND SYSTEM PERFORMANCE EVALUATION

To evaluate the performance of the developed system, 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 were used as the measure of the system's overall performance. A true positive (TP) represents a detected and correctly classified vehicle that has a corresponding manually tagged

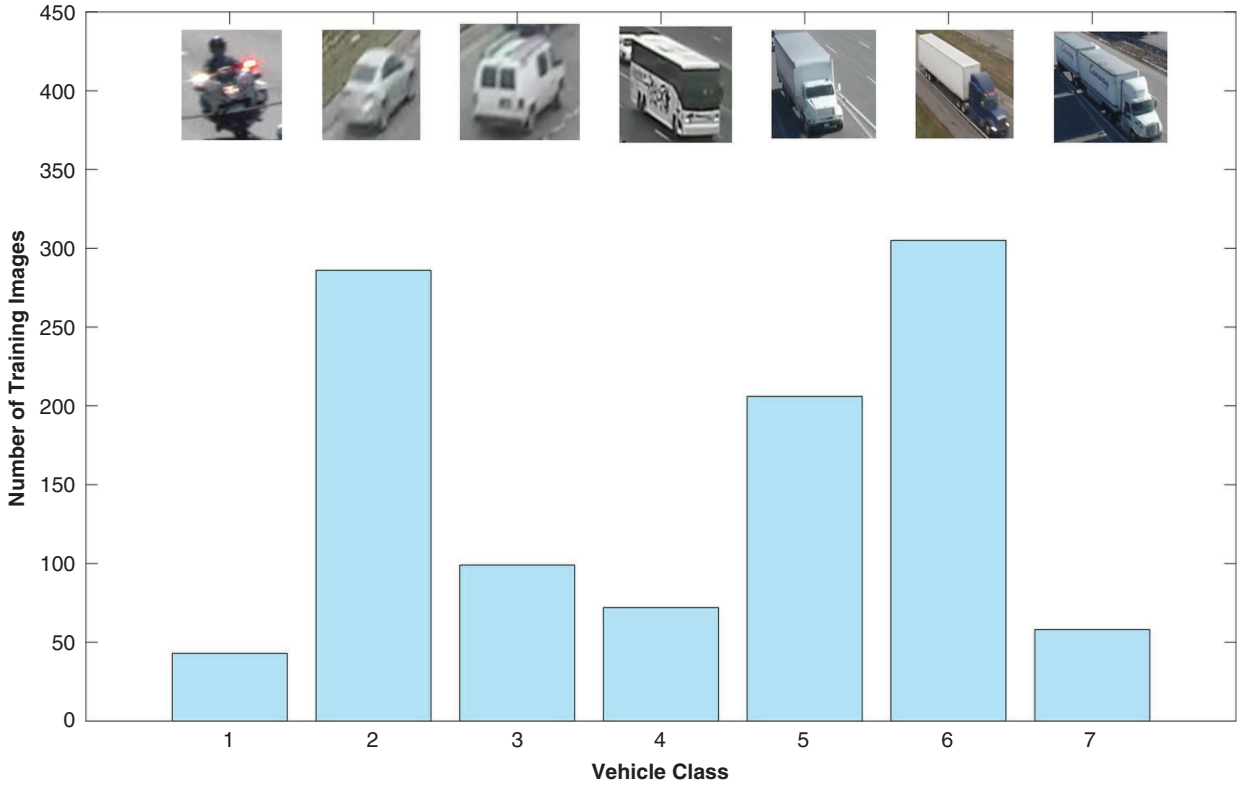


FIGURE 5 Histogram showing proportion of training image set per category class.

object in the test database. A false positive (FP) represents a detected and classified vehicle that has no corresponding manually tagged object in the test database. A detected but misclassified vehicle was denoted as a false positive even if it had a corresponding manually tagged object. A false negative (FN) represents objects that were missed by the vision system.

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

Figure 6 gives examples of positive, false, and missed calls from the vision system. Figure 6, *a* and *b*, gives examples of vehicles that the vision system correctly recognized. False detections are shown in Figure 6, *c* and *d*. In Figure 6*c*, a Class 5 vehicle towing a Class 3 vehicle is falsely classified as Class 3. In Figure 6*d*, the detected truck has no trailer and therefore could belong to either Class 6 or Class 7; however, the system classifies it as a Class 6–type vehicle. Missed objects as shown in Figure 6, *e* and *f*, were prominent in cameras with very poor resolution. Also, distance between the camera and the object may influence the accuracy of the vehicle classification. For example, a double-trailer truck may begin to look like a single-trailer truck as the vehicle moves away from the camera. Figure 7*a* illustrates the average precision and recall rates of the vision system for detecting all seven classes of objects from videos in the test database. On average, the developed vision system 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.

Classes 1 and 4

Despite limited training data for Class 1 and Class 4 vehicle types, motorcycles and buses are the simplest objects to recognize with the system developed and hence have a 100% precision rate. They are easily distinguishable from other classes, as shown in the confusion matrix in Figure 8. Because of the size of Class 1 vehicles, they are likely to be missed in poor-resolution videos or to be occluded by larger trucks and hence have a relatively low recall rate.

Class 2. Passenger Cars and SUVs

The system was able to recognize correctly vehicles belonging to Class 2 95% (precision) of the time. Class 2 vehicles constitute the largest proportion of vehicular traffic. Hence, this precision rate is appreciable. However, 89% (recall) of all Class 2 vehicles in the test database were recognized. The relatively lower recall rate was caused mainly by occlusions by trucks.

Class 3. Vans and Pickups

The system was least effective for recognizing Class 3 vehicles. It correctly recognized vehicles belonging to this class only 82% of the time, although 90% (recall) of all Class 3 vehicles in the test database were recognized. The box plot of precision rates for Class 3

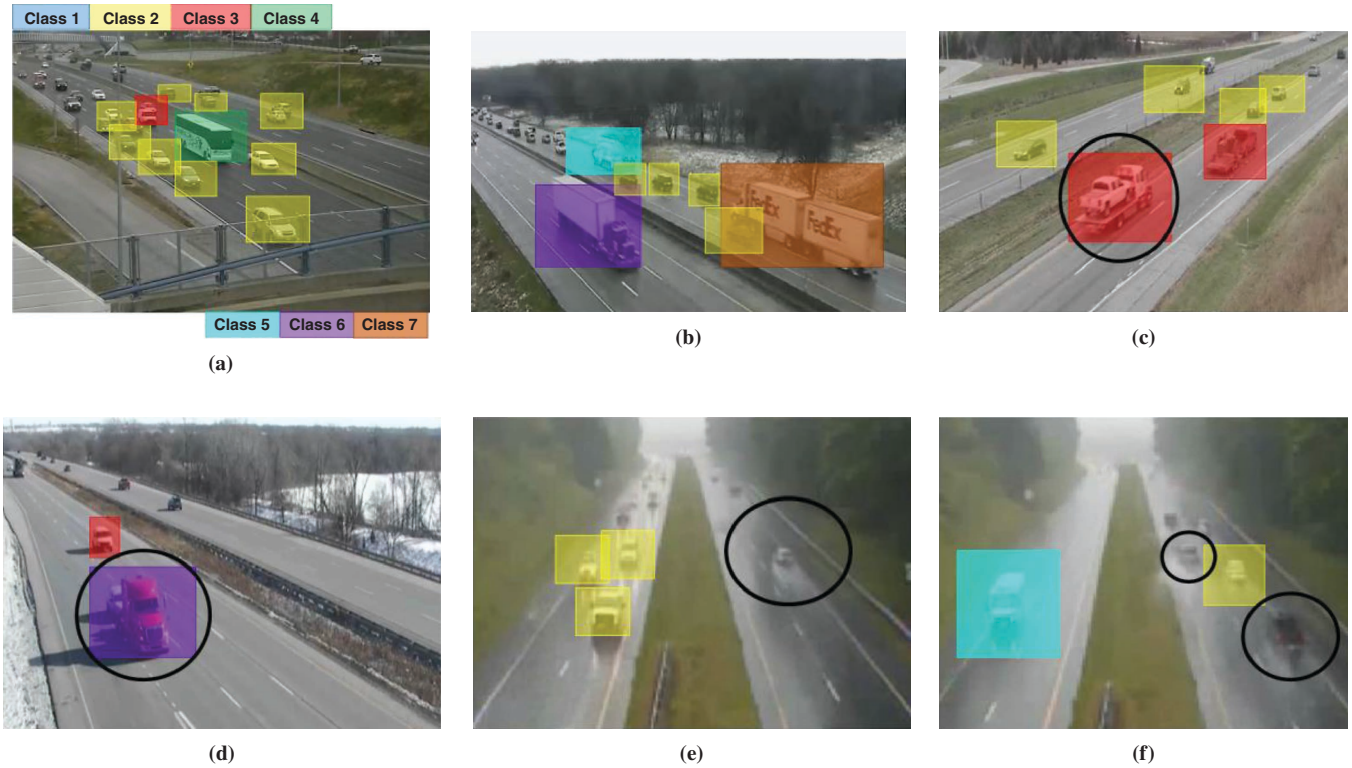
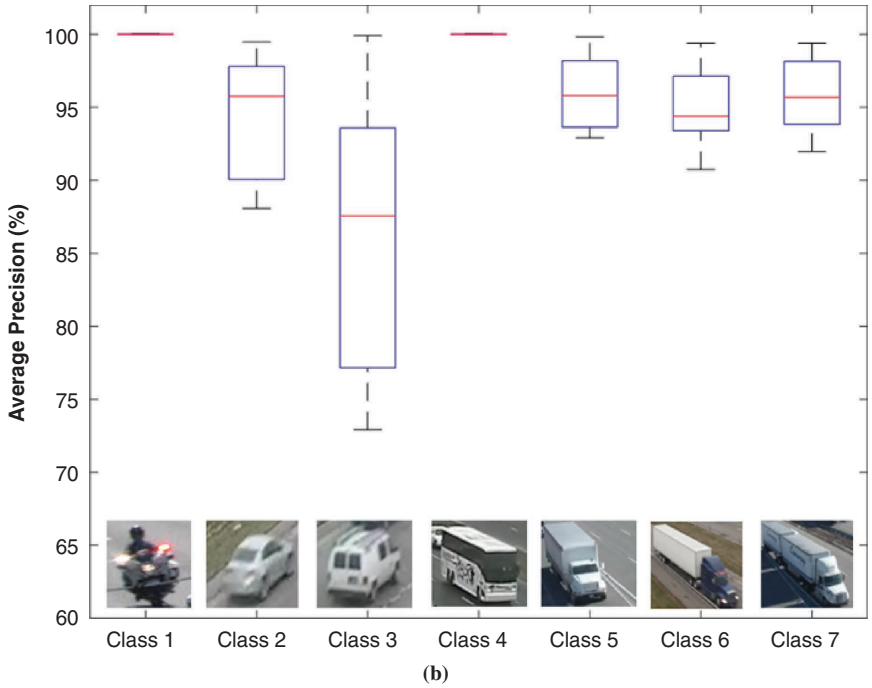


FIGURE 6 Vision system output examples: (*a* and *b*) true positives, (*c* and *d*) false positives, and (*e* and *f*) misses.

Class	Total Images	TP	FP	FN	Precision (%)	Recall (%)
Class 1	18	16	0	2	1.00	0.89
Class 2	8,542	7,250	410	882	0.95	0.89
Class 3	1,348	1,007	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	3,976	3,600	323	53	0.92	0.99
Class 7	165	152	10	3	0.94	0.98

(a)



(b)

FIGURE 7 Performance evaluations: (*a*) average precision and recall rates of system and (*b*) range of precision rates for all 30 videos in test database.

shows the most variation, ranging between 73% and 98%. The drastic drop in precision rates was caused by the inclusion of pickup trucks in this category class. Pickup trucks come in various forms: covered, uncovered, single or double cabin. The system confuses covered pickup trucks with SUVs, which belong to a different class. Also, single-cabin pickups are considered to be of 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 to 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. Most of the false positives (although few) in this category were related to single-unit trucks towing a Class 2 or Class 3 vehicle (Figure 6*c*), which confused the system and mostly led to misclassification. From the confusion matrix, this error happens only 4% of the time. Some trucks were missed if the system could see only a distant rear view and not the front of the truck. Truck-to-truck occlusions were also observed in some cases.

Classes 6 and 7. Single Trailer and Multitrailer

Single trailers and multitrailers also had appreciable precision and recall rates even in videos of very poor resolution and in congested conditions. The confusion matrix in Figure 8 shows that Classes 6 and 7 are easily distinguishable from the other five classes. However,

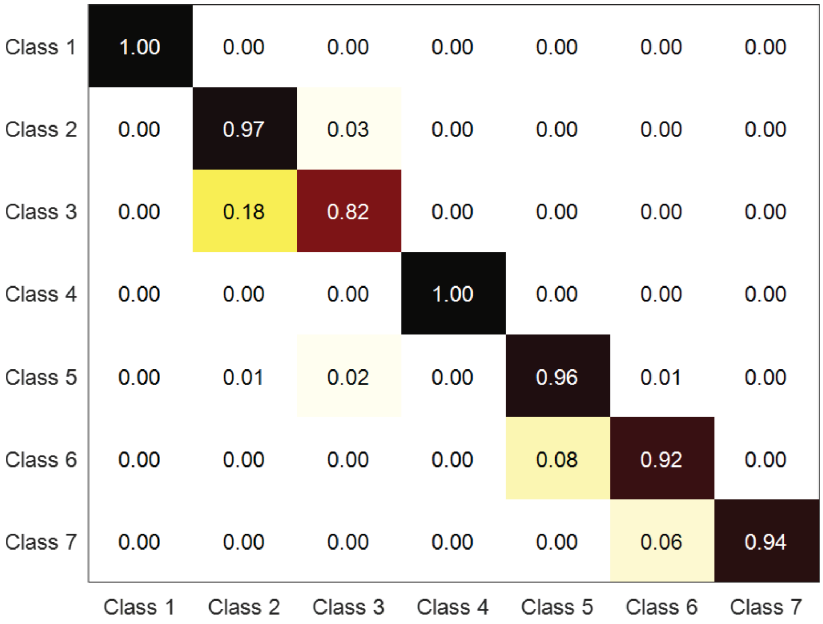


FIGURE 8 Confusion matrix for vehicle classes. (The color scale corresponds to the magnitude of the values: the darker the color, the higher the value.)

a multitrailer begins to look like a single trailer as the truck moves away from the camera, a reason for a relatively higher false positive rate for Class 6. Another source of false positive detections is trucks whose trailer has been removed (Figure 6*d*). Such situations mostly confused the vision system. The system missed some trucks belonging to this category if it could see only a distant rear view and not the front of the truck. Truck-to-truck occlusions were observed in some cases.

Sensitivity Analysis

Finally, conditions and configurations that could influence the performance of the developed vision system were investigated. The sensitivity of the proposed system to three key conditions was evaluated: the influence of time of day on vehicle recognition (day and night), the prevailing traffic conditions (free flow and congested, stop-and-go traffic), and camera resolution (blurring, sampling rates, rain, and snow). To evaluate the sensitivity of the system, the test database was partitioned into eight subgroups according to the combination of factors influencing the effectiveness of the vision system.

The vision system was used to process videos from each of these subgroupings. Receiver operating characteristics (ROC) curves were then used to compare the performance of the system per each subgroup according to true positive versus false positive rates. True positive and false positive rates are defined in Equations 3 and 4:

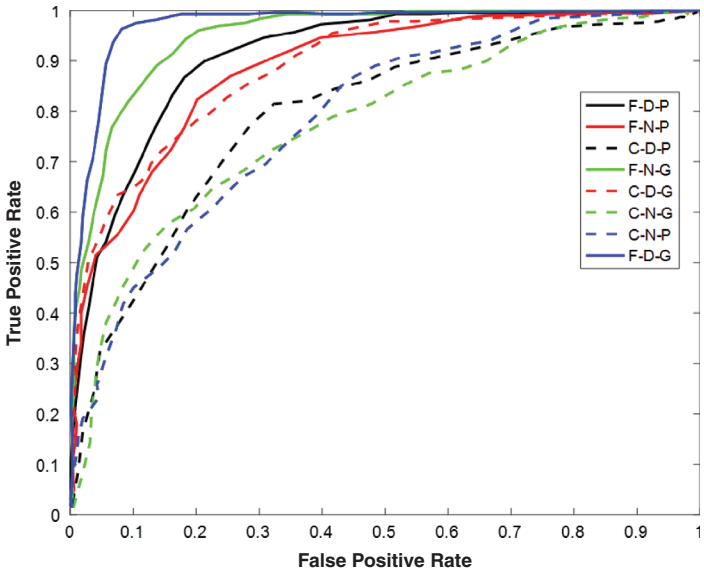
$$\mathrm{TPr} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})} \tag{3}$$

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

where TPr is the true positive rate and FPr is the false positive rate, and TN is the total number of nonvehicular objects that were not classified as vehicles. For a poorly constructed vision system, as its sensitivity (true positive rate) increases, it loses the ability to discriminate between vehicular and nonvehicular objects such as shadows, buildings, or trees. As a result, the true positive and false positive rates are almost directly proportional. Conversely, the mark of a good vision system is that its true positive rates are marginally higher than the corresponding false positive rates. Figure 9*a* shows the ROC curves for each subgrouping, and Figure 9*b* gives the calculated area under each curve.

Figure 9 shows that the true positive rates are marginally higher than the corresponding false positive rate irrespective of traffic condition or camera configuration. However, it is evident that the prevailing traffic conditions have 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%. During free-flow conditions, the system generally incurs between 5% and 30% false positives to reach a high true positive rate of 90% or more. Figure 9*b* shows an observable difference between the areas under the curve for free flow and that for congested conditions. The influence of video quality on system performance is minimal during free-flow conditions. Marginal effects of poorly resolved videos, however, are observed when traffic conditions are congested and the time of day is nighttime. Generally, the system is relatively more effective at processing daytime videos. 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 the traffic condition is congested.

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 the traffic condition is congested and at the same time video resolution is poor and the time of day is nighttime, the false positive rate reaches 45% for a true positive rate greater than 85%. To get the best results out of the system in such conditions, the use of a camera with frame rates greater than



(a)

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

(b)

FIGURE 9 ROC curves comparing influence of combination of factors that influence performance of vision system.

30 frames per second is suggested. Also, areas where video cameras are mounted should be well illuminated, which 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 systems purposes has stagnated in recent years. The best-performing systems can report only up to three or four classes, compared with the 13 classes required by FHWA's vehicle classification scheme. This study took 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 those of other nonintrusive recognition systems.

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

This performance was achieved through two main tasks. First, the selective search algorithm was used to generate class-independent region proposals to localize and segment objects. Second, 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 and stranded-vehicle detection. A comparison with existing automated video-based vehicle recognition systems also would 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, Saint Paul, 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. Transportation Statistics Office, Florida Department of Transportation. *Traffic Monitoring Handbook.* http://www.dot.state.fl.us/planning/statistics/tmh/tmh.pdf. Accessed July 2, 2016.
4. Zhang, G., R. P. Avery, and Y. Wang. Video-Based Vehicle Detection and Classification System for Real-Time Traffic Data Collection Using Uncalibrated Video Cameras. *Transportation Research Record: Journal of the Transportation Research Board,* No. 1993, 2007, pp. 138–147. https://doi.org/10.3141/1993-19.
5. *Highway Capacity Manual 2010.* Transportation Research Board of the National Academies, Washington, D.C., 2010.
6. *AASHTO Guide for Design of Pavement Structures.* AASHTO, Washington, D.C., 1993.
7. FHWA, U.S. Department of Transportation. *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. In *Proceedings of the IEEE Intelligent Transportation Systems Conference,* Oakland, Calif., 2001, pp. 201–206. https://doi.org/10.1109/ITSC.2001.948656.
9. Avery, R. P., Y. Wang, and G. S. Rutherford. Length-Based Vehicle Classification Using Images from Uncalibrated Video Cameras. In *Proceedings of the 7th International IEEE Conference on Intelligent Transportation Systems,* 2004, pp. 737–742. https://doi.org/10.1109/ITSC.2004.1398994.

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. https://doi.org/10.1109/6979.994794.
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, No. 4, 1989, pp. 541–551. https://doi.org/10.1162/neco.1989.1.4.541.
14. Krizhevsky, A., I. Sutskever, and G. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *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. https://doi.org/10.1007/s11263-013-0620-5.
16. Abramov, K.V., P. Skribtsov, and P.A. Kazantsev. Image Segmentation Method Selection for Vehicle Detection Using Unmanned Aerial Vehicle. *Modern Applied Science,* Vol. 9, No. 5, 2015. https://doi.org/10.5539/mas.v9n5p295.
17. Felzenszwalb, P.F., R.B. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively Trained Part Based Models. *TPAMI,* Vol. 32, No. 9, 2010, pp. 1627–1645. https://doi.org/10.1109/TPAMI.2009.167.
18. Collobert, R., K. Kavukcuoglu, and C. Farabet. Torch7: A Matlab-Like Environment for Machine Learning. Presented at 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. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* IEEE, New York, 2014, pp. 580–587. https://doi.org/10.1109/CVPR.2014.81.
21. Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, and F.-F. Li. *ImageNet Large Scale Visual Recognition Challenge.* 2015. https://arxiv.org/pdf/1409.0575.pdf.
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. In *Dr. Dobb's Journal of Software Tools for the Professional Programmer,* 2000.

The Standing Committee on Artificial Intelligence and Advanced Computing Applications peer-reviewed this paper.