Semi-automatic Training of a Vehicle Make and Model Recognition System

M.H. Zwemer[1,2(✉)], G.M.Y.E. Brouwers[2], R.G.J. Wijnhoven[2], and P.H.N. de With[1]

[1] Eindhoven University of Technology, Eindhoven, The Netherlands
m.zwemer@tue.nl
[2] ViNotion, Eindhoven, The Netherlands

Abstract. We propose a system for vehicle Make and Model Recognition (MMR) that automatically detects and classifies the make and model from a live camera mounted above the highway. Our system consists of a vehicle detection and MMR classification component. The vehicle detector is based on HOG features and can locate 98% of the vehicles with minimum false detections. We use a Convolutional Neural Network (CNN) for MMR classification on the vehicle locations. We propose a semi-automatic data-selection approach for the vehicle detector and the MMR classifier, by using an Automatic Number Plate Recognition engine for annotating new images, requiring minimal human annotation effort. In our results we show that our MMR classification has a top-1 accuracy of 98% for 500 vehicle models, where more than 500 training samples per model are desired to obtain accurate classification.

1 Introduction

There are thousands of surveillance cameras placed along highways which are mainly used for traffic management and law enforcement. Continuous manual inspection is not feasible, as it requires automatic visual interpretation. This enables detection and tracking of vehicles and classification into traffic classes. One specifically important concept is visual Make and Model Recognition (MMR). Make and model information of vehicles can be used to find stolen license plates when comparing the observed vehicle model with the model registered with the license plate. An additional application is to find specific vehicles after a crime when only a vehicle description is available (no license plate number). In such cases, make and model of the vehicle needs to be obtained visually.

Recognition of the vehicles in the above cases is now performed by an Automatic Number Plate Recognition (ANPR) system in a combination with a database lookup in the national vehicle registration database. Although this works for most cases, it is easy to circumvent this technique by altering the license plates. Moreover, it does not work for vehicles without a license plate, foreign vehicles or for motorcycles (when considering a frontal viewpoint).

We present an MMR system developed for the National Police, in which license plates are observed from a camera mounted in an overhead sign structure

S. Battiato et al. (Eds.): ICIAP 2017, Part II, LNCS 10485, pp. 321–332, 2017.
https://doi.org/10.1007/978-3-319-68548-9_30

on the highway where the focus is on a single lane (example video image in Fig. 3). The same camera is used to feed our recognition system. Due to bandwidth restrictions between the camera (online) and our training and testing facilities (offline), we have to optimize the gathering of training and testing samples. Therefore we propose a semi-automatic system to create a dataset.

The main contributions in this paper are the semi-automatic gathering of vehicle samples which are used for training our vehicle detector, an automatic procedure for acquiring make and model annotations for these samples and providing extensive insight in our MMR classification performance.

2 Related Work

Our vehicle recognition system consists of a detection and a classification stage, to localize and recognize vehicles in a full-frontal view. The first detection stage can be solved with different approaches. The full vehicle extent is detected using frame differencing by Ren and Lan [8] or background subtraction by Prokaj and Medioni [7]. Siddiqui *et al.* [12] and Petrović and Cootes [6] extend detections from a license-plate detector. Wijnhoven and de With [16] propose Histogram of Oriented Gradient (HOG) [2] to obtain contrast invariant detection. Recent work by Zhou *et al.* [18] reports on a Convolutional Neural Network (CNN) to obtain accurate vehicle detection. When the vehicle is detected, the vehicle region of the image is used as input for the classification task of MMR.

CNNs are state-of-the-art for image classification and originate by work from LeCun [5] and gained popularity by Krizhevsky [4] who used a CNN (AlexNet) to achieve top performance in the 1000-class ImageNet Challenge [9]. For MMR, Ren and Lan [8] propose a modified version of AlexNet to achieve 98.7% using 233 vehicle models in 42,624 images. Yang *et al.* [17] published a dataset which contains different car views, different internal and external parts, and 45,000 frontal images of 281 different models. They show that AlexNet [4] obtains comparable performance to the more recent Overfeat [10] and GoogLeNet [14] CNN models (98.0% vs. 98.3% and 98.4%, respectively). Siddiqui *et al.* [12] show that for small-scale classification problems, Bag of SURF features achieve an accuracy of 94.8% on the NTOU-MMR dataset[1] (containing 29 classes in 6,639 images).

Other work extends full-frontal recognition towards more unconstrained viewpoints. Sochor *et al.* [13] use a 3D box model to exploit viewpoint variation, Prokaj and Medioni [7] use structure from motion to align 3D vehicle models with images, and Dehghan *et al.* [3] achieve good recognition results but do not reveal details about their classification model.

In conclusion, detection methods involving background subtraction or frame differencing are sensitive to illumination changes and shadows. Therefore, we select Histogram of Oriented Gradients to obtain accurate detection. We have found that detection performance in this constrained viewpoint is sufficient whereas complex detection using CNNs [18] is considered too computationally

[1] NTOU MMR Dataset: http://mmplab.cs.ntou.edu.tw/mmplab/MMR/MMR.html.

expensive. We base our classification system on the AlexNet [4] classification model and focus on an extensive evaluation of the large-scale Make and Model Recognition problem. As shown by Yang *et al.* [17], AlexNet achieves state-of-the-art performance and the MMR problem does not benefit from more advanced CNN models such as GoogLeNet and Overfeat. Moreover, AlexNet is one of the fastest models at hand and suitable for a real-time implementation [1]. Our experiments are performed on our proprietary dataset, which contains 10x more images and car models than the public CompCar dataset [17]. We do not evaluate on the CompCar dataset because classification results are presented by Yang *et al.* [17] and we specifically focus on a large-scale evaluation.

3 System Description

The vehicle recognition system is shown in Fig. 1 and consists of two main components: detection and classification. The system is first trained offline and after training, it can fully automatically detect and recognize vehicles in a video stream from a camera mounted above the road. Both the detector and classification components are trained offline. To develop and train such a recognition system, it would be trivial to store long periods of raw video from the camera in the field and process this video data offline. However, this is not acceptable because typically only a low-bandwidth connection exists between the roadside setup and the backoffice. Therefore, the amount of video data transfer is rather limited. We obtain a low bandwidth when only transmitting a single image for each vehicle that passes the camera. To collect data for training the classification component, we use vehicle detection to select these images. However, we first need to train a vehicle detector. We start by downloading a limited amount of video (15 min) and manually annotating vehicles in these video frames. Using these images, we train our initial vehicle detector and then apply this detector to the roadside setup to collect images with vehicles and transmit these to our backoffice. We can now use these additional images to train an improved vehicle detector and to subsequently train our classification component. For both training purposes,

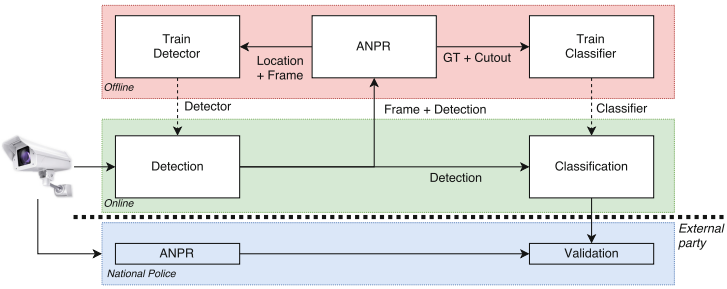

Fig. 1. System overview of the online make and model recognition and offline training of the detector and classifier. The blue bar is an additional validation stage carried out by an external party. (Color figure online)

we employ an Automatic Number Plate Recognition (ANPR)[2] engine. From the location of the number plate we will create additional vehicle annotations to improve the detector, while from the recognized license plate number we look up the vehicle make and model from a database. Next, we downscale each image to a lower resolution and only keep the make and model annotation and remove the license plate number to remove the identity of the vehicle. With this data, we train our vehicle recognition system, which has a privacy-friendly design because there is no identity information and license plates are not readable. Note that in the trained system, used in online operation, all images from the camera are directly down-sampled so that license plates are not readable, but the images are of sufficient resolution for classification. We will now discuss the detection and classification components in detail.

3.1 Detection: Vehicle Localization

Vehicle detection is performed by sliding a detection window over the image and classifying each window location into object/background. Our detector is trained on the grill of a vehicle covering the head lights and bumber, shown by the green rectangle in Fig. 3. Then, linear classification is realized, using HOG feature descriptions of the image as input. We compute HOG features of 12×5 cells of 4×4 pixels using 8 orientation bins ignoring the orientation sign, with L2 normalization of 1×1 blocks. For each cell, we add the gradient magnitude as an additional feature and train our linear classifier using Stochastic Gradient Descent [15]. The detection window is used on multiple, scaled versions of the input image and detections are merged by a mean-shift mode-finding merging algorithm. The detection process is performed every frame in the live video stream. Detections are tracked over time using Good Features to Track [11]. For each vehicle, the make and model classification is performed once when the vehicle is fully visible in the view.

Semi-automatic Training Data Collection for Detection
This approach is necessary because the detection performance of the initial detector is insufficient (missed cars and false detections). Because manual annotation of vehicles is cumbersome, we apply the initial vehicle detector at a low threshold to collect images which probably contain vehicles and validate these images using a ANPR engine. We assume that each real vehicle has a license plate and use a fixed extension of the license plate box as a new vehicle annotation. All images collected with the initial detector are now automatically annotated and the total set of annotations is used to train our improved vehicle detector.

3.2 Classification: Make and Model Recognition

Classification of make and model is performed once for each detected vehicle. The detection box is enlarged with a fixed factor to cover the grill, hood and

[2] ANPR Engine - CARMEN FreeFlow: http://www.arhungary.hu/.

windshield, shown as the blue rectangle in Fig. 3. This part of the image is scaled to a fixed low-resolution image of 256×256 pixels and used as the input of our MMR classifier combined with the make and model class label. We use the AlexNet classification model [4], which is a Convolutional Neural Network (CNN) consisting of 5 convolution layers and two fully-connected layers and a nonlinear operation between each layer. This large network is trained end-to-end by feeding our vehicle images and class labels and optimizing the network to predict the correct vehicle class for each image. Note that we predict the make and model combination, so that the number of classes equals the number of vehicle models. We use the AlexNet model pretrained on ImageNet and finetune it with our dataset. For each training image, multiple random subimages of 227×227 pixels are used. We train for 50,000 iterations using a batch size of 128. All other training parameters are equal to the original model [4].

Semi-automatic Make and Model Attribute Acquisition for Training Automatic finding of attributes is needed when classifying a large number of objects (order of 10^3–10^4), where sufficient samples are required for each class to distinguish intraclass variation from interclass variation. Moreover, not all vehicle models (classes) are equally popular, the distribution of models is extremely non-linear. To collect samples of rare vehicle models, it is required to on-the-fly annotate vehicles automatically. This is obtained when using our vehicle detector in the roadside setup. An ANPR engine processes every detection and the license-plate number is used to query a database with vehicle make and model information. Our setup is located in the Netherlands, enabling the use of the open-data interface of the Dutch Vehicle Authority (RDW), containing detailed information of all vehicles in the Netherlands[3]. This process allows for large-scale annotation of our dataset.

4 Evaluation of Proposed System

For evaluation, the vehicle detector is compared with the initial vehicle detector and furthermore, we provide insight in our make and model classification results.

4.1 Dataset

To train the initial vehicle detector, we have created a dataset by selecting frames which contain vehicles in 15 min of recorded video. This initial dataset contains 1,318 manually annotated vehicles. With our initial detector, we have collected images in the roadside setup over a period of four hours. The collected images are processed by the ANPR engine to remove false detections and correctly align vehicle boxes. In total, we collected 20,598 vehicle annotations. Half of this set is used to train the final detector, and half to evaluate the detector performance. The classification dataset was recorded during various weather conditions over a long interval of 34 days in which 670,706 images (100%) were collected. Examples

[3] RDW Open Data: https://opendata.rdw.nl/.

Fig. 2. Classification examples. Wind shield and license plates blurred for privacy.

of dark, strong shadows and rainy samples are shown in Fig. 2. All images are processed by the ANPR engine. In 649,955 of the images (97%), a license plate was found and the number could be extracted (other images contain too much noise for recognition). The make and model information was extracted from the database for 587,371 images (88%). Failure cases originate from non-Dutch license plates which are not registered in this database and license-plate numbers that are not read correctly (ANPR failure). In total we acquired 1,504 different vehicle models. The distribution of the number of samples per vehicle model is shown in Fig. 4, which follows a logarithmic behaviour. The top-500 models all have more than 30 samples. The last 700 models only have 1 or 2 samples and represent various high-end vehicles, old-timers and custom vehicles, such as modified recreational vehicles. The model which is seen most is the Volkswagen Golf, with a total of 20k samples (13% of the dataset).

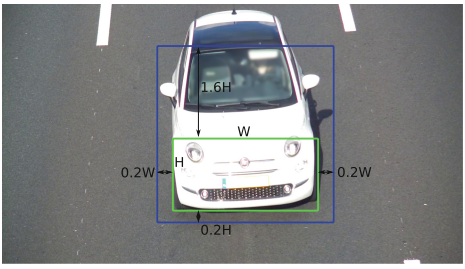

Fig. 3. Video frame, the detection box in green and classification ROI in blue. Windshield and license plate are blurred. (Color figure online)

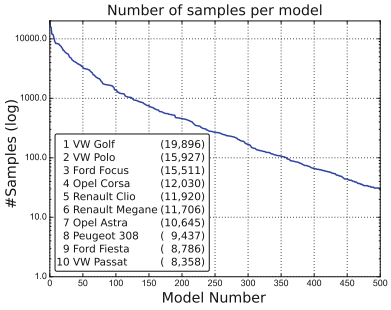

Fig. 4. Number of samples per model.

4.2 Evaluation Metrics

Detection performance is measured using recall and precision. A true positive TP rate is defined as a detection which has a minimum overlap (intersection over union) of 0.5 with the ground-truth box. Detections with lower overlap are

false positives FP. Missed ground-truth samples are denoted as false negatives FN. The recall and precision are then computed by:

$$recall = TP/(TP + FN), \quad precision = TP/(TP + FP). \tag{1}$$

We summarize the recall-precision curve by a single value as the Area Under Curve (AUC), where perfect detection has a value of 100%.

The classification performance is measured by the top-1 accuracy, in which the number of correct classifications is divided by the total number of classifications performed. As a second metric, the performance per vehicle model is measured using recall and precision as in Eq. (1). A classification is a true positive TP if the classification label is equal to the ground-truth label, otherwise it is a false positive FP. A sample is a false negative FN for a ground-truth vehicle model if it is not correctly classified. Note that a false negative for one class results into a false positive for another class.

4.3 Vehicle Detection

This section evaluates our initial vehicle detector based on manual annotations and our final detector trained with the automatically collected vehicle annotations (Sect. 3.1). Figure 5 portrays the recall-precision curves for these detectors. The blue curve shows the performance of our initial detector and the red curve shows the results of the final detector. The initial detector already shows good performance, but regularly generates false detections. The final detector clearly outperforms the initial detector and is almost perfect with an AUC of 99%. The operation point has been empirically chosen to detect 98% of the vehicles, having negligible false detections, which is sufficient for the MMR application.

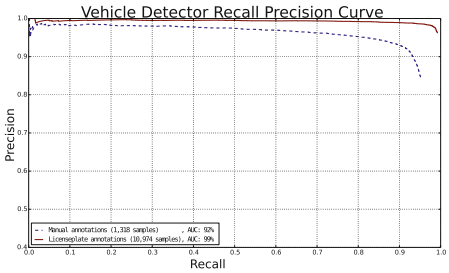

Fig. 5. Recall-precision curve of our initial and final vehicle detectors. (Color figure online)

Fig. 6. Average automatically annotated detection box (top) and average detected result (bottom).

In Fig. 6 the average images of our training set and our detector output are shown. The top image shows the average image of the annotations that are used to train the detector (the output of the ANPR detector). It can be clearly seen that the image is aligned on the license plate. The bottom image shows the

actual detections after training. Note that the detector does not focus only on the license plate but on the overall vehicle contour. This highlights that our process of automatic annotation is quite powerful and generalizes to the total vehicle characteristics.

4.4 Make and Model Classification

We investigate the classification performance in three main experiments. First, we investigate the relation between the number of classes and the classification performance. Second, we examine the effect of explicitly handling unknown classes. Third, we evaluate the per-class classification performance in relation to the amount of samples per class. Finally, miss-classifications are briefly discussed.

Training of our classification model is carried out by splitting up the vehicle classification dataset per day, to obtain a nice distribution of light and weather conditions. We randomly select 26 days (76%) for training and 6 days (18%) for testing and 2 days (6%) for monitoring the optimization process during training to avoid overfitting on the training set.

We investigate the classification performance when selecting an increasing number of classes in our model, incrementally adding the most frequent classes used. We distinguish the case where unconsidered classes are completely ignored during training ('no unknown' class) and where they are explicitly taken into account as an 'unknown' class.

As a first experiment, we investigate the 'no unknown' case. The classification performance for a low number of classes is constrained by the distribution of the data in the test set, e.g. for one model (VW Golf), the best possible accuracy is 13% because there are only so many samples in the dataset. The results are shown at the left in Fig. 7, where the classification performance (red, solid) approaches the theoretical boundary (green, dotted). Even for 500 classes, we

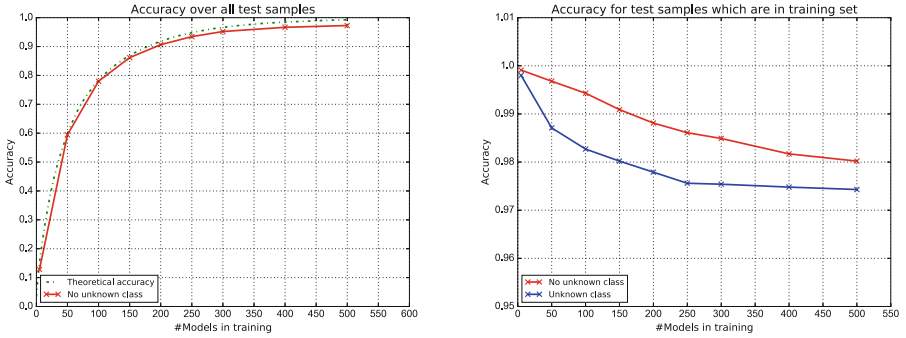

Fig. 7. Make and model classification accuracy. The left diagram evaluates over all test samples, the right diagram only over vehicle models that are incorporated in our classification model. The green dotted line denotes the theoretical accuracy (left), the red solid line is trained without 'unknown' label and the blue (bottom curve) with 'unknown' label. (Color figure online)

achieve an accuracy of 98%, showing that the classification model is able to handle this large-scale classification task. We now explicitly remove the bias from our results by normalizing to the theoretical accuracy, so that we only measure the performance of the classification model and completely ignore the statistics of the dataset. This is obtained by dividing the actual performance by the theoretical optimal performance. For example, for one model (VW Golf) the best possible accuracy is now 100%. The results are shown at the right in Fig. 7 by the red solid line. Although the accuracy over the complete range is high (>98%), it continuously decreases for a growing number of classes.

In a second experiment, we investigate the effect of explicitly taking the ignored classes into account. We expect a lower accuracy because the model has to deal with an extra class with a high amount of intraclass variation (contains all other vehicle classes). However, this case is very interesting because it learns the system to better classify the known models and also learns when a model is not recognizable. We compare this model to our 'no unknown' classification model without statistical bias in Fig. 7 by the blue bottom line. The model considering 'unknown' classes has a high accuracy, not much lower than the 'no unknown' model. More interestingly, the performance seems to saturate at 97.4%. However, with an increasing number of classes in our classification model, only a few samples are sorted into the additional (both for known and unknown) classes. Note that when training with all classes in the classification model, there is no 'unknown' class and both curves will have the same performance. In the third experiment, we provide more insight in the influence of the amount of training samples per model. We show the recall and precision per vehicle model versus the number of training samples available for that model in Fig. 8 for the top-500 vehicle models. Note that the plot is zoomed-in at sample sizes below 2,000. For the 66 models having more than 2,000 samples, both recall and precision approach unity (perfect classification). For classes with more than 500 samples, recall and precision are both exceed 95%. Using less than 200 model samples results in a performance drop. Outliers to this trend are annotated in the figure, which are further investigated and some examples are shown in Fig. 9. This figure shows an example TP classification and the highest FP classifications. We

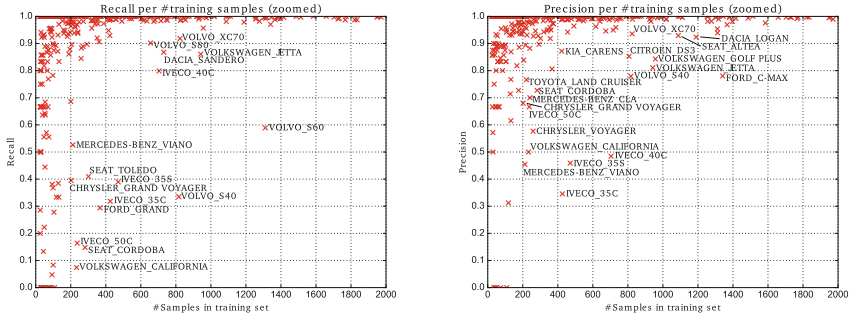

Fig. 8. Recall and precision for the amount of training samples available.

Fig. 9. TP classification (green) and the strongest FP classifications (red) for several models with low precision. The rate represents the distribution of the classifications. (Color figure online)

observe that for these cases, either the class labels are inconsistent (for example 'Citroen DS3' and 'DS 3'), or the classes are visually similar. For example, the Iveco model number relates to the wheel base and payload capacity which cannot be visually observed from the front of the vehicle and the more visually similar sedan models versus estate versions of a vehicle model. The proposed MMR system has been deployed as a live system in the Netherlands for the National Police. An evaluation of our system has been carried out by an external party (see blue region in Fig. 1). This independent evaluation uses our top-500 classification model and processed 4 different time periods with a total duration of 8 hours, under different weather and light conditions. A top-1 accuracy of 92.4% was measured. These results include very low-light conditions, where the vehicle is barely visible, which was not incorporated in our training process.

5 Conclusions

We have proposed a system for vehicle Make and Model Recognition (MMR) that automatically detects and classifies the make and model of each vehicle from a live camera mounted above the road. We have shown that with minimal manual annotation effort we can train an accurate vehicle detector (99% AUC), by using an Automatic Number Plate Recognition (ANPR) engine. Using this concept, we automatically detect vehicles and by extracting the license plate number we acquire the make and model information from a national database. We use a CNN for classification and experiment with the AlexNet model, leading to an MMR classifier with a top-1 accuracy of 98% for 500 vehicle models. The resulting classifier requires at least 500 training samples per model for accurate classification. An explicit unknown model class only leads to a small drop in performance (∼0.5%), but makes the model aware of unrecognizable vehicles. This approach can be used to gather automatically more samples of rare vehicle models and new models.

Our choice of classifying the front of the vehicle has limitations. Differences between certain models are not always visible from the vehicle front. These models should be joined in a combined model description, or additional input data (e.g. a side view) is required to solve this classification task.

References

1. Canziani, A., Paszke, A., Culurciello, E.: An analysis of deep neural network models for practical applications. CoRR abs/1605.07678 (2016). http://arxiv.org/abs/1605.07678
2. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), vol. 1, pp. 886–893, June 2005
3. Dehghan, A., Masood, S.Z., Shu, G., Ortiz, E.G.: View independent vehicle make, model and color recognition using convolutional neural network. CoRR abs/1702.01721 (2017). http://arxiv.org/abs/1702.01721
4. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (2012)
5. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural Comput. **1**(4), 541–551 (1989)
6. Petrovic, V.S., Cootes, T.F.: Analysis of features for rigid structure vehicle type recognition. In: BMVC, vol. 2, pp. 587–596 (2004)
7. Prokaj, J., Medioni, G.: 3-D model based vehicle recognition. In: Workshop on Applications of Computer Vision (WACV), pp. 1–7, December 2009
8. Ren, Y., Lan, S.: Vehicle make and model recognition based on convolutional neural networks. In: Proceedings of the IEEE ICSESS, pp. 692–695, August 2016
9. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
10. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: integrated recognition, localization and detection using convolutional networks. CoRR abs/1312.6229 (2013). http://arxiv.org/abs/1312.6229
11. Shi, J., Tomasi, C.: Good features to track. In: 1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 593–600, June 1994
12. Siddiqui, A.J., Mammeri, A., Boukerche, A.: Real-time vehicle make and model recognition based on a bag of surf features. IEEE Trans. Intell. Transp. Syst. **17**(11), 3205–3219 (2016)
13. Sochor, J., Herout, A., Havel, J.: BoxCars: 3D boxes as CNN input for improved fine-grained vehicle recognition. In: IEEE Conference on CVPR, June 2016
14. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015
15. Wijnhoven, R.G.J., de With, P.H.N.: Fast training of object detection using stochastic gradient descent. In: Proceedings of the IEEE ICPR, pp. 424–427. IEEE Computer Society, Washington, D.C. (2010)

16. Wijnhoven, R.G.J., de With, P.H.N.: Unsupervised sub-categorization for object detection: finding cars from a driving vehicle. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 2077–2083, November 2011
17. Yang, L., Luo, P., Change Loy, C., Tang, X.: A Large-Scale Car Dataset for Fine-Grained Categorization and Verification. ArXiv e-prints, June 2015
18. Zhou, Y., Liu, L., Shao, L., Mellor, M.: DAVE: a unified framework for fast vehicle detection and annotation. CoRR abs/1607.04564 (2016). http://arxiv.org/abs/1607.04564