Fast and Robust Object Detection Using Visual Subcategories

Eshed Ohn-Bar and Mohan M. Trivedi
Computer Vision and Robotics Research Laboratory
University of California, San Diego
eohnbar@ucsd.edu, mtrivedi@ucsd.edu

Abstract

Object classes generally contain large intra-class variation, which poses a challenge to object detection schemes. In this work, we study visual subcategorization as a means of capturing appearance variation. First, training data is clustered using color and gradient features. Second, the clustering is used to learn an ensemble of models that capture visual variation due to varying orientation, truncation, and occlusion degree. Fast object detection is achieved with integral image features and pixel lookup features. The framework is studied in the context of vehicle detection on the challenging KITTI dataset.

1. Introduction

Vision-based fast and robust detection of objects in varying orientation and occlusion levels is at the center of many computer vision applications, from robotics to surveillance. For many of the applications, in particular under mobile settings, the computational resources may be limited. This motivates the usage of lightweight feature extraction and classification algorithms, which are studied in this work.

We focus on vehicle detection from a mobile platform. To that end, detection of vehicles in the scene of all orientations are studied using the extensive KITTI dataset [9], which contains many challenges for vision-based vehicle detection. For instance, scenes contain significant illumination variation and are captured in a wide array of different driving environments. Furthermore, many of the ground truth vehicles are also occluded or truncated out of the camera view.

One common approach to address variability in appearance is the deformable parts model (DPM) [7]. The robustness of the DPM to appearance variations and truncation can be enhanced by employing motion features as in [18, 21]. In the DPM framework, the Latent SVM can be used in order to learn and refine subcategories [3], yet the framework has several speed bottlenecks (even with recent speedup attempts [24]). Instead, we adapt the real-time pedestrian detection framework in [4] for vehicle detection at multiple orientations and occlusion. This is not trivial, as pedestrian detection is commonly studied in monolithic settings (number of subcategories is $K = 1$).

We study unsupervised and discriminative techniques for visual categorization as well as the effect that the number of categories has on speed and performance. The proposed framework ranges in speed from 13-5 frames per second (fps) on full resolution images of size 1242×375 on a CPU. This is significantly faster than the classical DPM, which runs at about 0.1 fps. Furthermore, the detection scheme developed is shown to significantly outperform the DPM. The final detector is therefore fast while detecting objects at varying aspect ratio, orientation, and partial-occlusion. These attributes are essential for viable on-road driver assistance [20]. Furthermore, the proposed framework can be extended to detection of arbitrary objects [17].

2. Related Research Studies

A recent survey on different approaches for vehicle detection using monocular, stereo, and other vision-sensors can be found in [22]. Commonly, sliding window-based vehicle detection approaches employ either a variant of HOG+SVM (histogram of oriented gradients and a linear support vector machine), introduced by Dalal and Triggs [1], or cascade detectors [4]. Vehicle detection in mobile, on-road settings require fast detection under varying observation angles, occlusion, and truncation. Below, research studies addressing some of the challenges that may arise in our application are reviewed.

Vehicle detection with DPM: The DPM model [7] relies on HOG features and an SVM, which builds on HOG features and a latent SVM, has been applied successfully for vehicle detection [16, 25]. In [11], a variant of the DPM framework is used in order to detect vehicles under heavy occlusion and clutter. In [8], integrating scene information with vehicle detection was shown to improve both the detection and orientation estimation performance. Some approaches involve detection of discriminative parts of a vehicle first, and combine these to produce a detection [23]. In

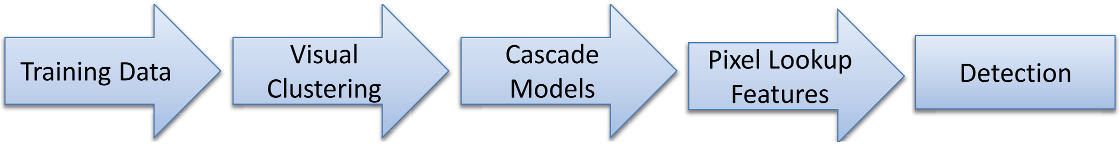


Figure 1: Overview of the components of the studied framework. Subclusters with visual homogeneity are extracted and used for training an ensemble of detection models. A set of pixel lookup color and gradient features is used for fast feature extraction.

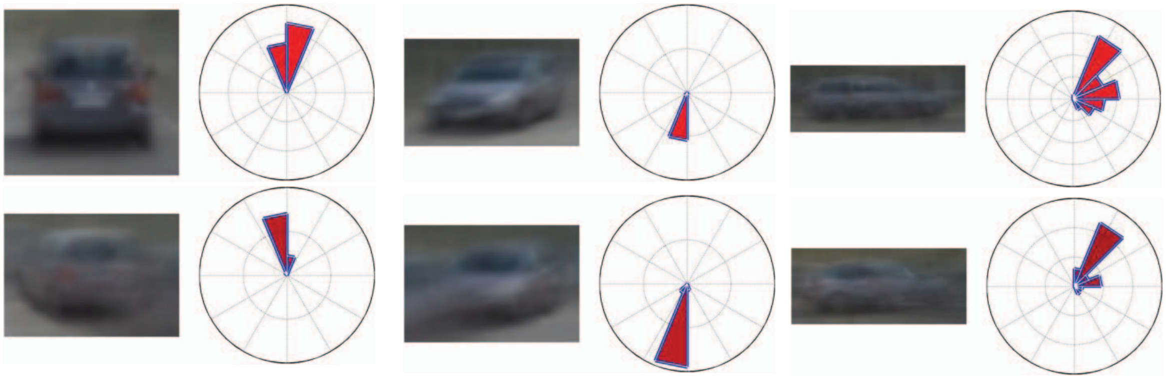


Figure 2: Visualization of the clustering output. Some example average images of clusters are shown, with a rose plot of the samples in the clusters showing ground truth 3D orientation and occlusion level (darker color implies more occluded samples in the cluster).

[19], occlusion patterns were mined in order to train DPM models that can reason over occluder-ocludee relationships for detection. Learning specific models for different appearance variations is therefore an important step towards achieving robust detectors.

Visual subcategory learning: A common approach for improving model generalization is by learning subcategories within an object class. For instance, these are commonly used with DPMs in order to detect objects from multiple viewing angles. In [13], visual subcategories corresponding to vehicle orientation are learned in an unsupervised manner using Locally Linear Embedding and HOG features. In [14], an exemplar SVM is learned for each positive example, and the learned weights are used in affinity propagation to generate visual subcategories. This exemplar-based step provides the initialization to Latent SVM clustering. Adding mixture components for occluded objects was shown to be promising in [19]. Vehicle orientation estimation is studied using supervised, semi-supervised, and unsupervised settings with DPM framework in [10], with supervised settings showing the best results. In this paper we only use visual data for the categorization.

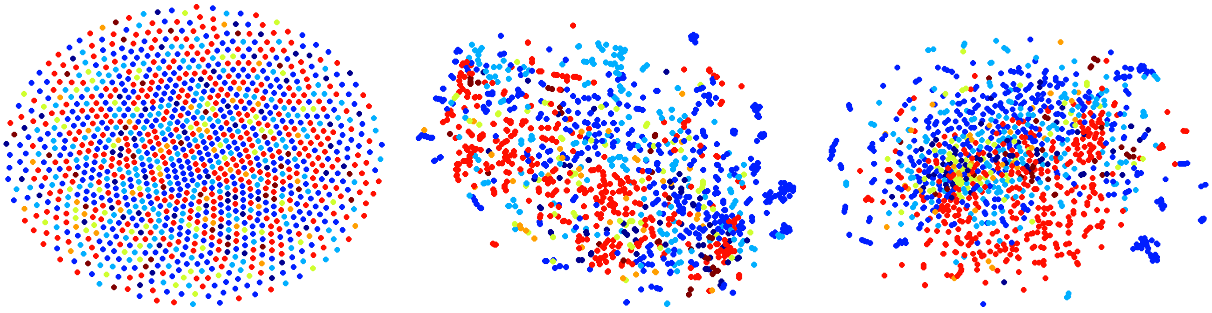
Recently, weakly supervised clustering has gained popularity to obtain discriminative subcategorization [12]. In [3], learning of visual subcategories for different objects on the PASCAL dataset was performed using a Latent SVM initialized with k-means. The authors motivate visual subcategorization over other forms of data partitioning by arguing that tighter clusters can be extracted from visual data, as semantic (human-based) subcategories are useful due to encoding visual consistency.

Detection speed: While the DPM detectors currently provide state-of-the-art detection on the KITTI dataset, they are computationally expensive and slow to evaluate. For on-road vehicle detection from one or multiple views, we require a significant reduction in computational demand of the algorithm. In detection pipelines, a main bottleneck in computation is in the feature extraction step. In particular, multi-scale detection requires repetitive feature extraction. We therefore employ the integral features studied in [4], which provides a significant speedup. The approach utilizes features that are efficiently computed using integral images and pixel lookups, as well as a fast soft cascade which is adopted from pedestrian detection in [4] to vehicle detection.

3. Visual Subcategorization

3.1. Features

The outline of the framework is shown in Fig. 1. We first inspect appropriate features for performing the visual clus-



(a) RGB-Color (b) LUV-Color (c) Gradient

Figure 3: Visualization of the feature space using t-SNE [2] with varying features on the entire KITTI training dataset (some samples were suppressed for visualization). The color of each point corresponds to an assigned bin according to annotated vehicle orientation. We note that both color and gradient cues provide useful information for detection and subcategory clustering.

tering. Commonly, HOG is used for capturing the shape of vehicle. As shown in Fig. 3, gradient features (histogram of oriented gradients at 6 bins and normalized gradient magnitude) are correlated with vehicle orientation (which was quantized to produce the coloring in Fig. 3). Interestingly, this is also the case with LUV color. One possible reason is due to tail lights and other parts of the vehicle that appears at certain orientations. On the other hand, RGB space color information is not very discriminative. The color and gradient features will be used to train and test the detectors. The total of 10 feature types described above can be extracted at more than 55 fps on a CPU (6 core, Intel Core i7 @ 3.30 GHz with 16 GB RAM) for full resolution images of size 1242 × 375.

3.2. Clustering

The color and gradient features are used to produce clusters in the data. We experiment with three techniques: k-means, spectral clustering followed by k-means [15], and the discriminative subcategorization framework of [12] (referred to as **DSC**). **DSC** differs from the previous two by incorporating negative instances into the clustering (these are obtained by three iterations of hard negatives mining).

Generally, spectral clustering was shown to perform better than k-means alone. In our implementation, a Gaussian kernel is employed as a similarity function between two samples $\mathbf{x}_i$ and $\mathbf{x}_j$, so that $W_{ij} = \exp \frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma^2}$ is the similarity matrix in the spectral clustering. We then compute the normalized graph Laplacian, $L = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$, where D is the diagonal degree matrix. Next, k-means is run on the L_2 normalized matrix of eigenvectors of L.

The performance of the unsupervised clustering techniques, k-means and spectral clustering, will be compared against the a discriminative framework, DSC. As in [12], our experiments showed DSC to be superior to Latent SVM in preventing degenerate clusters and overall cluster purity. DSC utilizes a block coordinate gradient-descent alternating between optimization of the SVM parameters and the cluster labels.

We found that the procedure of first training a linear SVM on the positive and negative instances to obtain a weight vector $\mathbf{w}$, and clustering the residual vectors of the positive samples after projection on $\mathbf{w}$ using $\mathbf{x} - \frac{1}{||\mathbf{w}||}(\mathbf{w}^T\mathbf{x})\mathbf{w}$ slightly improved the final detection results.

Some examples of the visual subcategorization step output are shown in Fig. 2, where we see how the clustering automatically separates occluded instances from non-occluded instances (compare top and bottom rows in Fig. 2). This is visualized in terms of color on the rose plot, where darker red corresponds to more occluded samples in the cluster. Furthermore, a separation occurs over orientation as well, which is intuitive as much of the appearance variation occurs due to orientation changes.

4. Detection Framework

AdaBoost [4] is learned using depth-2 decision trees as weak classifiers. Detection at multiple scales is handled using approximation of features at nearby scales [5]. The color and gradient image features are aggregated in 4 × 4 blocks in order to produce fast pixel lookup features. Bootstrapping was performed, with the first stage sampling 5000 random negative samples, and two additional stages of training using hard negatives.

Pooling detectors: The trained models for each subcategory are evaluated on a test image. Overlapping detections are merged using a greedy non-maximum suppression (NMS) procedure; once a bounding box is suppressed by the overlap criterion, it can no longer suppress weaker detections. We experimented with the alternative NMS threshold proposed in [6], where instead of the PASCAL overlap

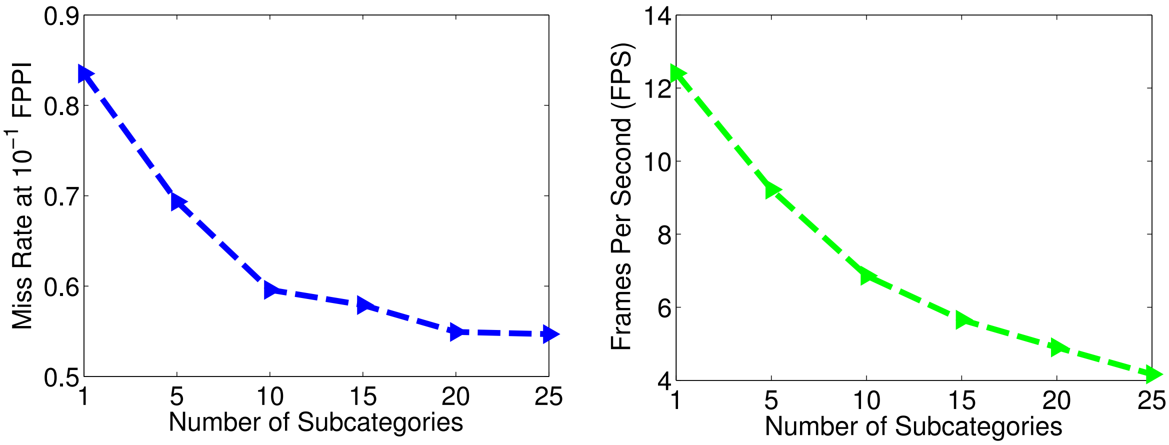


Figure 4: Left: Results of varying K, the number of subcategories. $K = 20$ is shown to work well. Right: Effects of increasing K on detection speed of the entire detection pipeline.

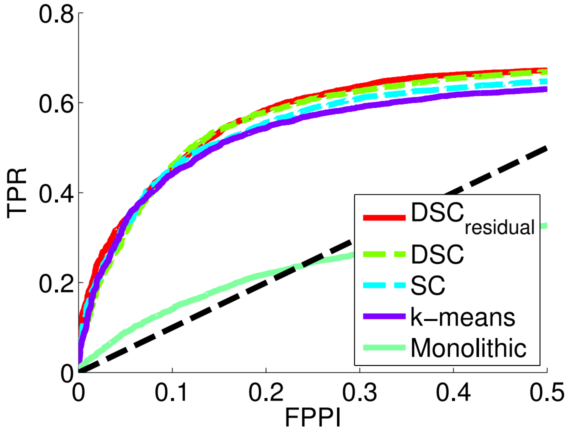


Figure 5: Analysis of different clustering methods for obtaining the subcategory clusters. SC refers to spectral clustering, and DSC to discriminative subclustering [12].

criteria of intersection-over-union, the union denominator is replaced by min(area(b_1),area(b_2)), but the results were inferior. We did not find it necessary to carefully calibrate the models as in [3].

5. Experimental Settings

The experiments are performed on the KITTI training dataset, which compromises a total of 7481 training images and over 20,000 vehicle instances. The first half of the dataset is used for training, and the second for testing. The curves and analysis is shown on 'moderately' difficult test settings, which include fully visible and partially-occluded vehicles bigger than 25 pixels in height with up to 30% truncation out of the camera view.

We note that the official benchmark evaluation requires a challenging 70% overlap (intersection-over-union of a detection and a ground truth box) for a detection to count as a true positive, and so this threshold is used in the experiments. Nonetheless, it has a significant impact on performance when it is compared to the commonly used 50% threshold.

6. Experimental Evaluation

The model dimensions appear to have significant effect on performance. We experimented with varying model sizes. Using grid optimization, the base height of the models that was found to work best is 32 pixels. From this height, the other dimension can be obtained by taking the median aspect ratio in each cluster (this was shown to work better than the mode, which was used in [3]). Model padding was also grid optimized, and $1/8$ of the model size worked well. Using a different aspect ratio for each subcategory, as opposed to a fixed one, resulted in a significant improvement mainly due to better localization of the vehicle (tighter bounding boxes).

The effect of increasing K, the number of subcategories, on performance and speed is shown in Fig. 4. We observed a plateau in detection performance after $K = 20$. At $K = 20$, the framework runs at about 5 fps on full resolution images, providing a large speed up compared to the DPM counterpart while matching its detection performance.

For a fixed $K = 20$, different clustering approaches are analyzed in Fig. 5. We note how a monolithic classifier performs poorly in vehicle detection settings, yet this is mitigated by the framework studied in this paper. Interestingly, the improvement of DSC is mild over spectral clustering. We note that unlike the original implementation of [12], where initialization is done using k-means, we found it better to initialize using labels provided by a spectral clustering step (the results in Fig. 5 are shown for DSC with spectral clustering). Furthermore, we experimented with other initializations to DSC, such as with ground truth 3D orientations, but this did not significantly improve the final detection performance.

7. Concluding Remarks

In this work, the role of visual subcategories was studied for vehicle detection at varying orientation and occlusion levels. An emphasis was put on fast detection, while maintaining results comparable to the successful DPM.

Future work would introduce further speedups; for instance, not all detectors need be evaluated on each block. Furthermore, occluded vehicle detection can still be improved as certain types of occlusions (vehicles parked closely at a side view) are still not handled well. Similarly, truncated vehicle is also challenging. Finally, there is still a need for better clustering as some of the clusters are not clearly defined. Poorly defined clusters often hinder the performance of the associated model.

References

[1] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[2] L. V. der Maaten and G. Hinton. Visualizing data using t-SNE. *JMLR*, 9(85):2579–2605, 2008.

[3] S. K. Divvala, A. A. Efros, and M. Hebert. How important are deformable parts in the deformable parts model? In *ECCVW*, 2012.

[4] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *PAMI*, 2014.

[5] P. Dollár, S. Belongie, and P. Perona. The fastest pedestrian detector in the west. In *BMVC*, 2010.

[6] P. Dollár, Z. Tu, P. Perona, and S. Belongie. Integral channel features. In *BMVC*, 2009.

[7] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9):1627–1645, 2010.

[8] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun. 3D traffic scene understanding from movable platforms. *PAMI*, 2014.

[9] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *IJRR*, 32(11):1231–1237, 2013.

[10] C. Gu and X. Ren. Discriminative mixture-of-templates for viewpoint classification. In *ECCV*, 2010.

[11] M. Hejrati and D. Ramanan. Analyzing 3D objects in cluttered images. In *NIPS*, 2012.

[12] M. Hoai and A. Zisserman. Discriminative subcategorization. In *CVPR*, 2013.

[13] C.-H. Kuo and R. Nevatia. Robust multi-view car detection using unsupervised sub-categorization. In *WACV*, 2009.

[14] T. Lan, M. Raptis, L. Sigal, and G. Mori. From subcategories to visual composites: A multi-level framework for object detection. In *ICCV*, 2013.

[15] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, 2001.

[16] H. T. Niknejad, A. Takeuchi, S. Mita, and D. McAllester. On-road multivehicle tracking using deformable object model and particle filter with improved likelihood estimation. *IEEE TITS*, 13(2):748–758, 2012.

[17] E. Ohn-Bar, S. Martin, and M. M. Trivedi. Driver hand activity analysis in naturalistic driving studies: Issues, algorithms and experimental studies. 22:1–10, 2013.

[18] E. Ohn-Bar, S. Sivaraman, and M. M. Trivedi. Partially occluded vehicle recognition and tracking in 3D. *IEEE Intell. Veh. Symp.*, 2013.

[19] B. Pepik, M. Stark, P. Gehler, and B. Schiele. Occlusion patterns for object class detection. In *CVPR*, 2013.

[20] A. Ramirez, E. Ohn-Bar, and M. M. Trivedi. Panoramic stitching for driver assistance and applications to motion saliency-based risk analysis. In *IEEE Conf. Intell. Transp. Syst.*, 2013.

[21] A. Ramirez, E. Ohn-Bar, and M. M. Trivedi. Integrating motion and appearance for overtaking vehicle detection. *IEEE Intell. Veh. Symp.*, 2014.

[22] S. Sivaraman and M. M. Trivedi. Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking and behavior analysis. *IEEE TITS*, 14(4):1773–1795, 2013.

[23] S. Sivaraman and M. M. Trivedi. Vehicle detection by independent parts for urban driver assistance. *IEEE TITS*, 14(4):1597–1608, 2013.

[24] J. Yan, Z. Lei, L. Wen, and S. Z. Li. The fastest deformable part model for object detection. In *CVPR*, 2014.

[25] J. J. Yebes, L. M. Bergasa, R. Arroyo, and A. Lázaro. Supervised learning and evaluation of kittis cars detector with dpm. *IEEE Intell. Veh. Symp.*, 2014.

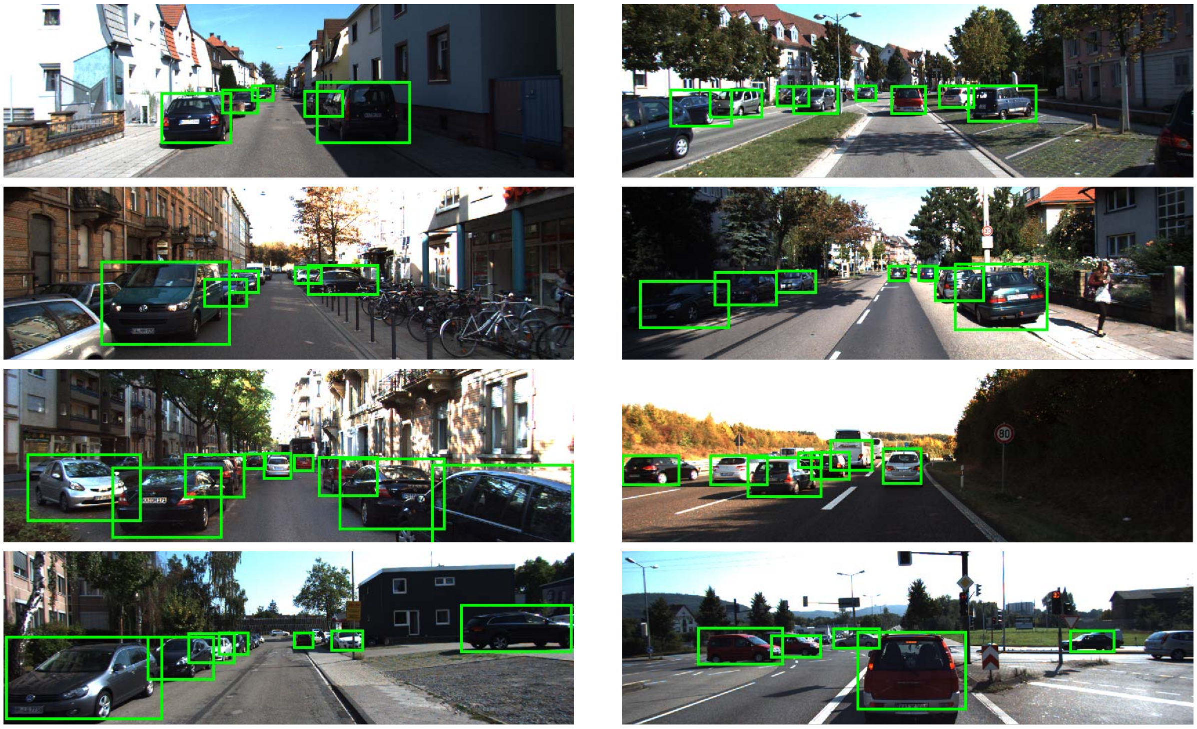


Figure 6: Detection results on the KITTI dataset.

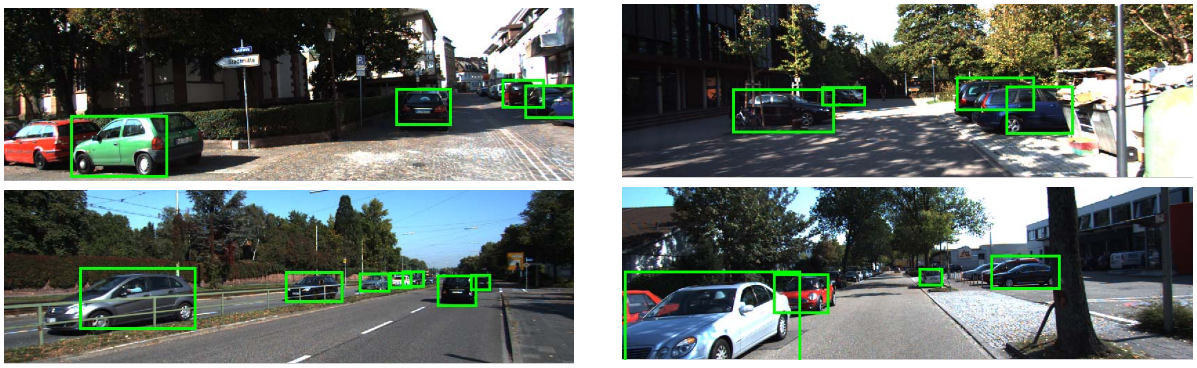


Figure 7: For future work, handling truncation, certain types of occlusion, and localization tightness can be further improved.