IEEE 18th International Conference on Intelligent Transportation Systems, 2015 - To Appear

An Exploration of Why and When Pedestrian Detection Fails

Rakesh Nattoji Rajaram, Eshed Ohn-Bar, and Mohan M. Trivedi
Laboratory for Intelligent and Safe Automobiles
University of California, San Diego

***Abstract*—This paper undergoes a finer-grained analysis of current state-of-the-art in pedestrian detection, with the aims of discovering insights into why and when detection fails. Current pedestrian detection research studies are often measured and compared by a single summarizing metric across datasets. The progress in the field is measured by comparing the metric over the years for a given dataset. Nonetheless, this type of analysis may hinder development by ignoring the strengths and limitations of each method as well as the role of dataset-specific characteristics. For the experiments we employ two pedestrian detection datasets, Caltech and KITTI, and highlight their differences. The datasets are used in order to understand in what ways methods fail, and the impact of attributes, occlusion, and other challenges. Finally, the analysis is used to identify promising next steps for researchers.**

***Index Terms*—Pedestrian detection, fine-grained evaluation, Caltech pedestrians, KITTI pedestrians.**

I. Introduction

For researchers to build better pedestrian detectors, it is crucial to fully understand the strengths and limitations of current state-of-the-art methods. The goal of this paper is to perform an attribute-based study of failures of pedestrian detectors, attempting to better answer some of the questions shown in Fig. 1. In the process, tools for finer-grained analysis of pedestrian detectors are proposed. This also results in dataset-specific insights as to promising next research steps. As shown in Fig. 1, training a pedestrian detector is often done in a non-specific manner, oblivious to the underlying challenges that are specific to the datasets. Furthermore, detection is often measured in coarse performance metrics, such as a final ROC curve or area under the curve over the entire dataset. Such generic training and evaluation practices elude the strengths and limitations of each approach, thereby hindering progress in the field.

The contributions of this paper are as following. First, the extent to which dataset-specific attribute distribution can impact detection performance are studied. Two popular pedestrian datasets, Caltech [1] and KITTI [2], are used in the experiments. Dataset bias is highlighted for facilitating future progress. Second, Attribute sensitivity by detectors is shown to vary with the detection threshold dependent. This conclusion is important when comparing multiple methods, as the performance for certain attribute classes may degrade more gracefully as the detection threshold is varied for some methods. This improves our understanding of the method and its discriminative power at different points on the ROC or precision-recall curve, and provides insights not commonly found in existing literature. Furthermore, this study is applicable to certain applications which require fixing the final detection threshold.

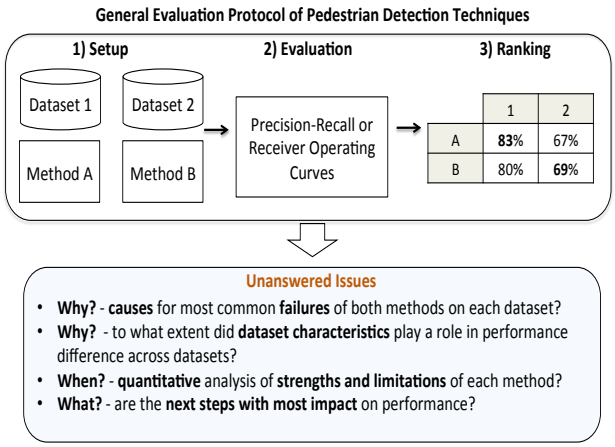


Fig. 1: Our goal is to gain insight into the strength and limitations of pedestrian detectors. Consider two methods and two datasets. Method A may be better at detecting certain pedestrians (e.g. occluded) and method B at other types (e.g. small pedestrians). Furthermore, the distribution of such instances is different in the two datasets. This scenario is common, but not easily identified in existing evaluation methodology which employs a single summarizing metric. This may hinder: 1) gaining full insights into the underlying causes of detection failures and successes, and 2) database-driven conclusions as to the most important next development steps.

The emphasis of these three contributions is to perform an attribute-based study of the underlying reasons for detection performance improvement. Additionally, it provides clearer tools for identifying dataset bias. This emphasis is in contrast to common related studies [3]–[11] involving improved feature space design or learning procedures. Although the improvement due to novel features and their fusion is remarkable, it often falls under 'trying and seeing what works', leaving most of the questions in Fig. 1 unresolved. For instance, perhaps a paper introduces a novel shape, motion feature, or fusion with promising results. Readers are left to wonder about the underlying cause of the improvement-are partially-occluded pedestrians better detected? Or perhaps fully visible

[1]AP is defined as the area under Precision-Recall curve times 100. This metric is chosen to facilitate comparison between the metrics in this paper and KITTI benchmark results.

(a) Caltech (b) KITTI

Fig. 2: Sample images with annotations from the two studied datasets.

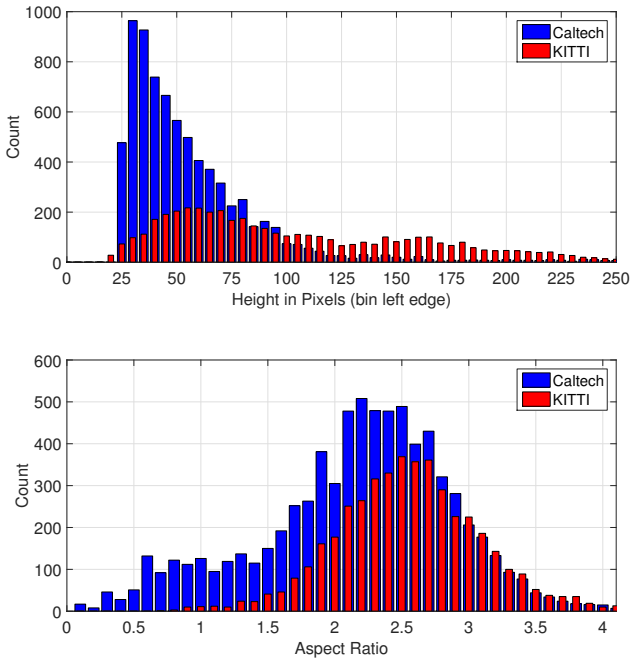


Fig. 3: Comparison of pedestrian height and aspect ratio across datasets. Compared to KITTI, Caltech has a larger fraction of pedestrians smaller than 50 pixels in height. Caltech has a wider range of aspect ratios as well. This is an important observation as most of the state-of-the art methods use only single model with fixed height and aspect ratio. Furthermore, the official evaluation metrics for the two datasets handle aspect ratio variation differently.

pedestrians are better handled by removal of false positives or deformation handling? Most importantly, the reader is left uncertain as to what should be the next steps. Were there any steps that could have been done better in the careful dataset-specific features and learning tuning? Is the specific attribute-class address now resolved, and researchers should divert attention to other types of visually challenging instances of pedestrians? Did the approach make assumptions at a certain trade-off cost for certain attribute-classes of pedestrians? An example for such an assumption would be the exclusion of occluded pedestrians in the training process, which may improve performance for non-occluded cases but reduce performance on occluded pedestrians. Dataset bias towards either one of these classes could potentially hinder informative insights towards resolving such issues. The above is also true, but to a lesser extent, in methods improving modeling capacity [12]–[16] as these may explicitly address a specific class of pedestrians. Nonetheless the study of common failure modes is useful in designing such methods. The failure analysis may also be useful for choosing methods for specific applications [17], [18].

Therefore, the goal in this paper is not to have best detector but make sense out of a state-of-the-art object detector. For the experiments, the Aggregate Channel Features (ACF) [19] detector is used, due to its simplicity, speed, and effectiveness on the Caltech pedestrian dataset.

II. DATASETS

A. Caltech Pedestrians

The common evaluation split is performed, where the first six out of the 10 available sets of data are split into training and the remaining for testing. Each video clip has a resolution of 640x480 and is recorded at 30 frames per second. By periodically sampling at 1 second, we extract about 60 frames per video clip. This translates to 4250 and 4024 images in the training and testing set, respectively. Sample images with annotation are shown in Fig. 2. The distribution of pedestrian height and aspect ratio[2] are shown in Fig. 3, and compared against pedestrians from the KITTI dataset.

B. KITTI Pedestrians

The KITTI object dataset [2] was introduced more recently, with higher resolution images (1242×375 and careful calibration with other sensor modalities as well. Like Caltech, it is meant for automotive environments, and it exhibits specific visual challenges for object detectors. Along with Pedestrian, it also has annotations for other objects, such as cars and cyclists.

[2]Aspect Ratio is defined as the ratio of bounding box height to bounding box width.

The detection benchmark provides 7481 training images and 7518 test images (without available annotations). To evaluate our performance we created a validation set as follows. The split has to resemble Caltech datasets as closely as possible. The first step is to apply the inverse mapping to remove the random permutation of images and then separate them into corresponding video clips. This can be automatically achieved, thanks to the mapping sequence and video sequence information provided in KITTI's development kit. It turns out that the training images were extracted from 144 video sequences, each with an average of 52 frames. Next step is to divide these video sequences into train and validation sets such that they have almost same number of images and also pedestrians. This is achieved by sorting the video according to number of pedestrians and manually assigning each video to either test or validation set. The final training and validation split contains 3742 and 3739 images, respectively. Sample images with annotation are shown in Fig. 2. Fig 3 compares the distribution of pedestrian height and aspect ratio with Caltech.

The dataset performs separate evaluation on three categories, 'easy', 'moderate', and 'hard', referring to increasingly challenging settings. The 'easy' category contains fully visible pedestrians higher than 40 pixels in height and truncation percentage below 15%, 'moderate' settings allow for partial occlusion, minimum height of 25 pixels, and truncation up to 30%, and 'hard' include heavy occlusion and truncation up to 50%.

III. EXPERIMENTAL SETUP

Ideally we are tempted to look at the hardest settings because this is where the detectors fail. But according to [2] under hard settings as described by KITTI, around 2% of the pedestrians were not recognized by humans. Furthermore, methods are ranked according to moderate difficulty, and so moderate settings were chosen for the analysis.

First, for performing the experimental analysis, we seek to standardize evaluation among KITTI and Caltech. For instance, Caltech does not annotate truncation, hence the moderate settings constraints cannot be enforced directly on Caltech. Furthermore, KITTI has qualitatively annotated occlusion as integers from 0 to 2 in the increasing order of occlusion, whereas Caltech has annotated occlusion by providing bounding box of occluded portion of the ground truth.

A qualitative mapping between the evaluation constraints across datasets is described below. By observing the occluded samples from KITTI, a minimum visibility of 65% was enforced on Caltech. Truncation was handled using the following algorithm to avoid manual annotation. First, a canonical aspect ratio, a, was assumed. Next, to get the truncation value for each pedestrian, a subset of all pedestrians who are very close to the left or right or bottom boundaries was collected. Using bounding box height h (or width w if truncation is at the bottom), the expected width $w_e = ha$ (height $h_e = w/a$) is calculated. Finally, truncation is estimated as $t = \frac{w_e - w}{w}$ ($t = \frac{h_e - h}{h}$). All pedestrians with truncation less than 0.3 are ignored. Under these constraints, Caltech has 3106 and 2425 pedestrians in training and testing sets respectively. On the other hand, KITTI has 1785 and 1784 pedestrians in training and validation sets respectively. So with comparable number of images in both Caltech and KITTI, Caltech has significantly higher number of pedestrians.

KITTI follows the general object evaluation literature, where object detectors must match a ground truth annotation as much as possible, irrelevant of the aspect ratio of the ground truth or prediction boxes. Caltech on the other hand reduces impact due to aspect ratio by standardizing all pedestrians to a certain aspect ratio. Nonetheless, as aspect ratio corresponds to a visual challenge (associated with pedestrian deformation), we would like to study its impact on detection performance. Furthermore, the requirement was introduced in [1] in order to evaluate many detectors that were trained in different settings, and this is not the case in this paper as we have full control over the training parameters. Hence, no aspect ratio standardization is performed in the experiments.

IV. EXPERIMENTAL EVALUATION

The ACF detector [19] is employed in order to study how pedestrian attributes and dataset characteristics impact detection performance. First, the parameters are optimized independently for both the datasets to obtain maximum AP, and consequently failure reasons are analyzed. The main parameter tuned is model size, which is the size to which every possible window in the image is re-sized before feature extraction. Therefore, it plays a significant role in both training and test-time. If it is too large, the detector will have difficulty in detecting small pedestrians. On the other hand if it is too small, the quality of features will be degraded, reducing discriminative power between pedestrian and background instances. We sweep through template height and aspect ratio and choose the best performing parameter. Other parameters were also optimized, however they have more to do with the model capacity and to some extent are dataset independent.

A. Parameter sweep on Caltech

The common parameters from [19] are optimal parameters for easy settings (not moderate) on Caltech. However, these parameters may not be the best when we include smaller, occluded and truncated pedestrians. The aspect ratio is kept fixed at the default value (2.439). The template height is swept from 25 to 65 pixels in increments of 5. Unlike the default difficulty settings, we now have pedestrians as small as 25 px. in height. In order to include these small pedestrians at a larger template size, the parameter number of octave up (henceforth referred as $\boldsymbol{\delta}$) was set to both 0 and 1 the during testing stage. AP at $\delta = 1$ was 10% higher than at $\delta = 0$. Best AP was obtained with model size 50 by 20.5. Aspect ratio turned out to be a very sensitive parameter. By introducing 1% change from the default value resulted in 2.5% decrease in detection AP.

B. Parameter sweep on KITTI

The ACF detector was optimized for KITTI as following. The template height is swept from 40 to 60 in increments of 5. Similar to Caltech, AP plateaued around model size 50 by 20. In principle, setting $\delta = 1$ during test time should have resulted in increase in AP because, as the detector will now try to detect pedestrians smaller than the template size. However, AP did not improve. One probable reason is the

TABLE I: Evaluation on KITTI and Caltech using the standardized training and testing settings. δ is the number of octaves test image is up-sampled. θ is case where only pedestrians taller than 50 pixels are considered, as opposed to 25 pixels.

Training Set \ Testing Set	Caltech			KITTI		
	$\delta = 0$	$\delta = 1$	θ	$\delta = 0$	$\delta = 1$	θ
Caltech	37.00	**47.44**	**68.25**	50.10	45.31	56.35
KITTI	17.23	17.84	34.05	**57.76**	57.62	**65.32**

difference in height distribution among the two datasets. While Caltech has 53% of their pedestrians smaller that 50 pixels tall, KITTI has only about 21%. It could also be due to small pedestrians in KITTI being very difficult to detect. So the impact on Recall is more severe in Caltech than in KITTI.

C. Overall training parameters

On both the datasets, ACF Detector was trained with 4096 depth-4 decision trees using AdaBoost. Four rounds of hard negative mining were performed. In each round, 25,000 negatives were randomly mined and upto 50,000 of the hardest negatives were employed. Only pedestrian windows taller than 50 pixels were consider for positive samples. Horizontally flipped versions of pedestrians windows were also included as positive samples. Each window was scaled to respective model size (50 by 20.5 for Caltech and 50 by 20 for KITTI) and padded such that the final size is 64 by 32. While ACF on Caltech trained all the decision trees, on KITTI training stopped early with only 1865 decision trees being trained.

D. Performance Summary

Table [I] summarizes the detection performance for different parameter settings and for cross-dataset detection performance as well. While δ impacts AP significantly on Caltech, AP on KITTI essentially remains the same. Notice that cross datasets training and testing results in significant reduction in AP. The impact is more severe when testing on Caltech using a detector trained on KITTI. This is likely due to the following two reasons. (1) KITTI pedestrians may not generalize well with Caltech pedestrians. (2) KITTI images were acquired at a higher resolution. This could lead to detector favoring sharper features. When the detector fails to identify sharp features in Caltech images, it tends to discard those windows. We believe reason 2 is more likely, and methods such as pre-smoothing may improve generalization. We also report AP under moderate difficulty settings, but considering only pedestrians taller than 50 pixels (reported under θ). Now AP on Caltech and KITTI differ only by 3%. Under this difficulty settings both the datasets could be considered equally difficult.

V. FAILURE ANALYSIS

In order to improve any detector, it is crucial to understand where it fails. We propose to perform this analysis using the ROC plot. Most likely reasons for detector misses can be visualized in Fig. 4. Colored area under each reason corresponds to the fraction of miss rate most likely contributed by them. The detector failure cases are categorized into several cases. Localization error is defined as the fraction of missed pedestrians that would have been detected if the minimum overlap threshold criteria was reduced from 50% to 20%. Miss due to height of the pedestrian is defined as the fraction of missed pedestrians that were not detected and are smaller than 50 pixels in height. This is definitely relevant to KITTI, due to the detector AP being better at $\delta = 0$. Aspect ratio failure corresponds to the cases where the ground truth boxes have aspect ratio that largely deviates from the average aspect ratio. Specifically, if a missed ground truth box has an aspect ratio larger than 3 or smaller than 2, it is considered a failure due to aspect ratio. Some cases which can not be resolved using the aforementioned attributes are marked into 'others'. Priority when generating the plot is in the order of truncation, occlusion, height, aspect ratio, localization, and finally others. Occlusion exhibits significant correlation with localization error especially at higher false positives per image rate, and hence a category of occlusion and localization occurring together was added (this was not the case for truncated samples). Examples corresponding to each miss case are visualized for both datasets in Fig. 5.

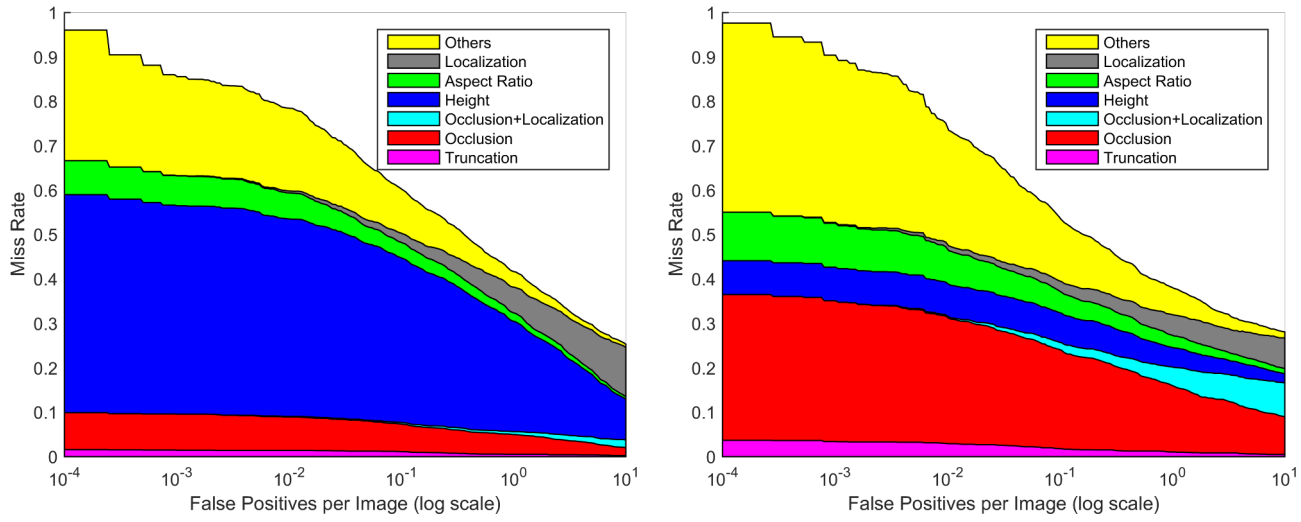
A. Failure on Caltech

When it comes to pedestrian detection on Caltech datasets, failure can mainly be attributed to the datasets bias with respect to pedestrian size. Since it has more than double the fraction of small pedestrians compared to KITTI datasets, the failure seems to heavily favor this reason. At fppi higher than 10^0, we see that the detector detects these small pedestrians, but with a small score. Occluded and Truncated pedestrians seem to be almost never detected. Interestingly, the plots in Fig. 4, show a large difference in failure case distribution among the datasets.

B. Failure on KITTI

Detector on KITTI seems to fail largely due to different reasons from Caltech. Occluded pedestrians are very hard to detect on KITTI, and this highlights the difference in the kind of bias that exists between datasets. KITTI has less number of pedestrians who are small, however, there are significant number of them that are occluded. Another interesting aspect to consider is the correlation between occlusion and localization error especially at fppi higher than 10^0 (light blue colored area). This typically happens when there is a vertical pole or another pedestrian nearby.

It would be interesting to know objects that are occluding pedestrians. Thanks to the rich annotation provided for the KITTI datasets, we can analyze to some extent the distribution of occludee. With the help of depth information we can find a set of annotated objects that are in front of the occluded pedestrian. For this subset, we calculate the overlap area of the missed pedestrian with the occludes. Ocludees causing overlap above 5% are added to the distribution. Fig. 6 plots the distribution.



(a) Caltech (b) KITTI

Fig. 4: Failure analysis at various false positives per image (fppi) rates. It is useful to not only look at overall miss rate vs fppi curve, but also to see the fractional contribution by various elements in the datasets. Truncation and occlusion have their usual meaning under moderate difficulty settings. Localization error is defined as the fraction of missed pedestrian who would have been detected if the minimum overlap threshold criteria was reduced from 50% to 20%. Miss due to height is the fraction of missed pedestrian who were smaller than 50 pixels in height. Aspect ratio corresponds to the case where the annotated ground truth box exhibits large deviation from the model aspect ratio.

Fig. 5: Examples of different possible reason for missed detections: Green boxes indicate the ground truth. Blue boxes are the detections which fails the 50% overlap criteria. These failure are associated as localization error. Top row are the samples taken from Caltech while the bottom row are samples from KITTI.

C. Failure Analysis Summary

In order to analyze where the detector fails, let us take a look at the miss rate slice around 10^{-1} fppi. At this setting, irrespective of the datasets, the reasons for the failure can be majorly attributed to occlusion and lack of sufficient resolution (i.e. small pedestrians). Although truncated pedestrians are hardly detected, the number of instances is small. This implies that although challenging, addressing other challenges than truncation may provide more significant improvement in performance on the datasets. Some of the high scoring false positives are visualized in Fig. 7. One particular aspect common to all these images is presence of strong horizontal gradient. This suggests that the detector favors gradient information over color information and gets very confident about the presence of pedestrian under the influence of strong horizontal gradient component. Some of the false positive detections do not lie on the ground plane, while some that are far away from camera, are very tall. By estimating ground plane and depth from

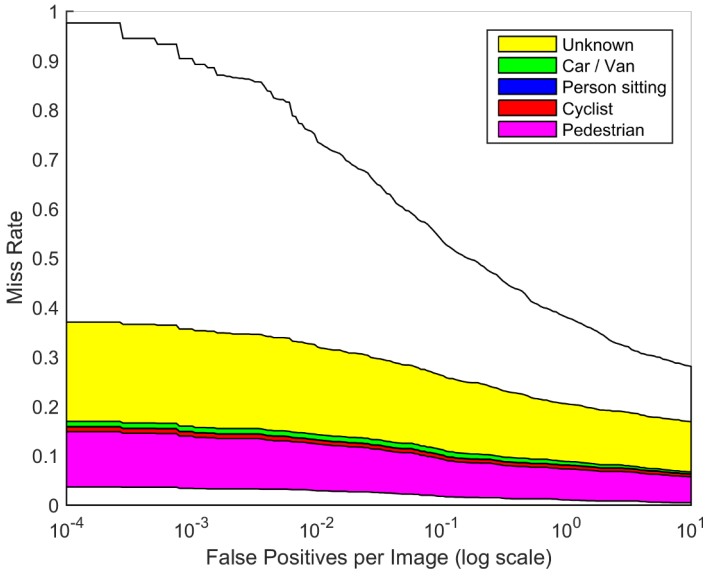


Fig. 6: Miss analysis of occluded pedestrians at various False Positives per Image on KITTI datasets. Since occlusion is a major contributor of misses on KITTI datasets, we further analyze the occludee distribution. Color shaded area corresponds to the fraction of occludee objects causing occlusion.

(a) High Scoring FP windows from Caltech datasets

(b) High Scoring FP windows from KITTI datasets

Fig. 7: Most likely false positive windows generated by the detector at 10^{-2} false positives per image. Notice the similar strong horizontal gradient component in all these samples.

camera one could remove some of the false positive detections.

VI. Concluding Remarks

The analysis in this paper highlights dataset bias and common failure modes in pedestrian detection. While each researcher attempts to generate a non-biased dataset, two common datasets were used to exemplify inherent dataset-bias. The evaluation strategies among the two datasets had to be standardized and optimal parameters studied. The suitable choice of detector parameters already hinted the existing dataset bias, and the two datasets generated very different distributions of failure cases, as shown in Fig. 4. The plot also provides take-aways for researchers studying the two datasets. While occlusion is a scene problem, height is not. By capturing data at higher resolution, we can eliminate this problem completely. As demonstrated quantitatively by our study, occlusion-handling methods could reduce a significant portion of failed detection cases.

References

[1] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, 2012.

[2] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *CVPR*, 2012.

[3] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Pedestrian detection with spatially pooled features and structured ensemble learning," in *arXiv*, 2014.

[4] W. Nam, P. Dollár, and J. Han, "Local decorrelation for improved pedestrian detection," in *NIPS*, 2014.

[5] S. Zhang, R. Benenson, and B. Schiele, "Filtered channel features for pedestrian detection," in *CVPR*, 2015.

[6] R. Benenson, M. Omran, J. Hosang, , and B. Schiele, "Ten years of pedestrian detection, what have we learned?" in *ECCV, CVRSUAD workshop*, 2014.

[7] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *BMVC*, 2009.

[8] J. Hosang, M. Omran, R. Benenson, and B. Schiele, "Taking a deeper look at pedestrians," in *CVPR*, 2009.

[9] C. Long, X. Wang, G. Hua, M. Yang, and Y. Lin, "Accurate object detection with location relaxation and regionlets relocalization," in *ACCV*, 2014.

[10] S. J. Krotosky and M. M. Trivedi, "On color-, infrared-, and multimodal-stereo approaches to pedestrian detection," *IEEE Transactions on Intelligent Transportation Systems*, Dec 2007.

[11] C. G. Keller, M. Enzweiler, M. Rohrbach, D. F. Llorca, C. Schnörr, and D. M. Gavrila, "The benefits of dense stereo for pedestrian detection," *IEEE Trans. Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1096–1106, 2011.

[12] E. Ohn-Bar and M. M. Trivedi, "Learning to detect vehicles by clustering appearance patterns," *IEEE Trans. Intelligent Transportation Systems*, 2015.

[13] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, "Data-driven 3D voxel patterns for object category recognition," in *CVPR*, 2015.

[14] E. Ohn-Bar and M. M. Trivedi, "Fast and robust object detection using visual subcategories," *CVRPW*, 2014.

[15] A. Prioletti, A. Mgelmose, P. Grisleri, M. M. Trivedi, A. Broggi, and T. Moeslund, "Part-based pedestrian detection and feature-based tracking for driver assistance: Real-time, robust algorithms and evaluation," *IEEE Trans. Intelligent Transportation Systems*, vol. 14, no. 3, pp. 1346–1359, Sep. 2013.

[16] E. Ohn-Bar and M. M. Trivedi, "Can appearance patterns improve pedestrian detection?" *IV*, 2015.

[17] E. Ohn-Bar, A. Tawari, S. Martin, and M. M. Trivedi, "On surveillance for safety critical events: In-vehicle video networks for predictive driver assistance systems," *Computer Vision and Image Understanding*, vol. 134, pp. 130–140, 2015.

[18] A. Tawari, A. Mgelmose, S. Martin, T. Moeslund, and M. M. Trivedi, "Attention estimation by simultaneous analysis of viewer and view," *IEEE Intelligent Transportation Systems Conference*, pp. 130–140, Oct 2014.

[19] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast Feature Pyramids for Object Detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2014.