2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)
Windsor Oceanico Hotel, Rio de Janeiro, Brazil, November 1-4, 2016

Multi-Perspective Vehicle Detection and Tracking: Challenges, Dataset, and Metrics

Jacob V. Dueholm[1,2], Miklas S. Kristoffersen[1,2], Ravi K. Satzoda[1], Eshed Ohn-Bar[1], Thomas B. Moeslund[2] and Mohan M. Trivedi[1]

***Abstract*— The research community has shown significant improvements in both vision-based detection and tracking of vehicles, working towards a high level understanding of on-road maneuvers. Behaviors of surrounding vehicles in a highway environment is found as an interesting starting point, of why this dataset is introduced along with its challenges and evaluation metrics. A vision-based multi-perspective dataset is presented, containing a full panoramic view from a moving platform driving on U.S. highways capturing 2704x1440 resolution images at 12 frames per second. The dataset serves multiple purposes to be used as traditional detection and tracking, together with tracking of vehicles across perspectives. Each of the four perspectives have been annotated, resulting in more than 4000 bounding boxes in order to evaluate and compare novel methods.**

***Index Terms*— Vehicle detection, vehicle tracking, multi-perspective behavior analysis, autonomous driving.**

I. Introduction

Detecting and tracking vehicles in full surroundings of a vehicle are natural next steps given the improvements seen in monocular perspectives. This allows for the study of on-road behaviors [1], identifying behaviors [2], [3] and long-term prediction [4] for both active and passive safety applications. The motivation is clear: Bring down the number of fatal accidents happening every day. To this end, the research community has pushed forward the need for publicly available datasets and common benchmarks, as to strengthen the scientific methodology for development and evaluation of novel research.

Vision for Intelligent Vehicles & Applications (VIVA) [9] is a vision-based challenge set up to serve two major purposes. The first is to provide the research community with naturalistic driving data from looking-inside and looking-outside the vehicle, and thus to present the issues and challenges from real-world driving scenarios. The second purpose is to challenge the research community to highlight problems and deficiencies in current state-of-the-art approaches and simultaneously progress the development of future algorithms. Current challenges in VIVA include hands [10], traffic signs [11], [12], and traffic lights [13]. In this paper we introduce multi-perspective vehicle detection and tracking as a new part of the VIVA challenge, moving towards understanding trajectory based behaviors surrounding the ego-vehicle.

[1]Laboratory for Intelligent and Safe Automobiles, University of California, San Diego, USA
[2]Visual Analysis of People Laboratory, Aalborg University, Denmark

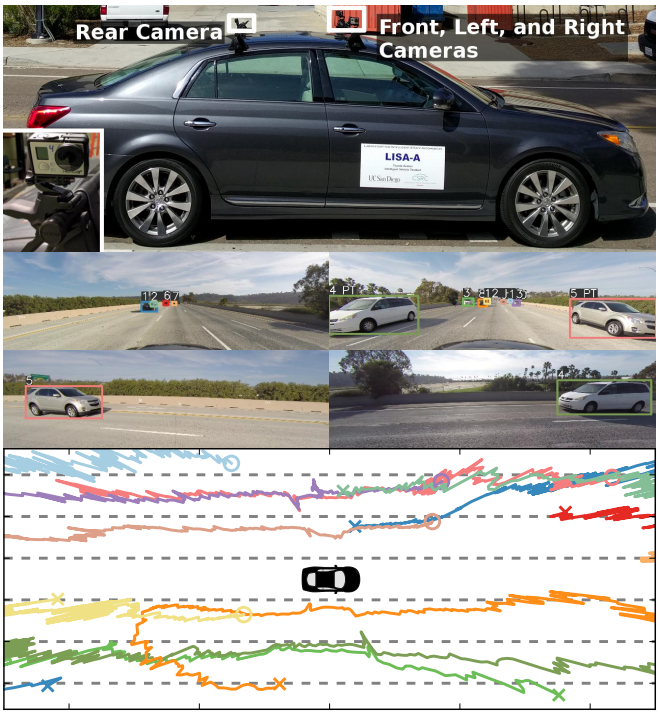

Fig. 1. Vehicle instrumented with four cameras (top), ground truth bounding boxes in the four perspectives (middle), and 3D ground truth trajectories (bottom). The trajectories are marked with a start position X and an end position O.

Monitoring surround vehicles is found using both passive and active sensors. Active sensors include radar and lidar to obtain 3D point clouds of nearby objects. The advantage of using a passive vision-based approach is an easier interpretation compared to 3D point clouds and a general lower cost. The downside of a vision-based approach is the increased computational complexity. Several benchmarks have emerged for vision-based vehicle detection and tracking. These fields together push towards a higher semantic scene understanding, of which this dataset is aimed at. Related datasets are summarized in Table I.

Detection datasets are among the most researched fields, and also the first step towards a complete framework. Early detection datasets such as the Pascal VOC Challenge [14] used to detect cars at various viewpoints in static images among 20 other object classes. The KITTI Vision Benchmark Suite [8] dataset comprises a large collection of on-road data.

Many applications using video feeds require consecutive detections such as driving on-road. The main interest so far

TABLE I

RELATED VISION-BASED VEHICLE DATASETS.

Dataset	Sensors	Detection Tracking Trajectory	Moving Platform	Full Surround	Environment
TME [5]	Two 1024×768 front faced color cameras at 20 FPS, IBEO 4 layer laser scanner	✓/✗/✗	✓	✗	Northern Italian Highways
PKU POSS [6]	Ladybug3 at 1 FPS	✓/✗/✗	✓	✓	Chinese Highways
Ford Campus vision and lidar data set [7]	Ladybug3 at 8 FPS, lidars, IMU	✗/✗/✗	✓	✓	U.S. Urban
KITTI [8]	Front faced stereo camera capturing 1392×512 images at 10 FPS, lidar, GPS	✓/✓/✗	✓	✗	German Urban, Rural and Highway
LISA Trajectory	Four GoPro cameras capturing 2704×1440 resolution images at 12 FPS	✓/✓/✓	✓	✓	U.S. Highways

has been in a frontal perspective as found in the TME [5]. A full surround view is found in the PKU POSS dataset [6], gathered by the omni-directional Ladybug3 camera at highways, and later divided into four images of equal size for detecting vehicles at different viewpoints. The Ford Campus vision and lidar data set [7] also use the Ladybug3 but in rural areas for the purpose of SLAM (Simultaneous Localization and Mapping) thereby no vehicle annotations.

Tracking by detection can be described as associating detections across frames often denoted by an identification number. Most notably, is the KITTI dataset [8] with its comprehensive annotations which not only allows object detection but also tracking in a frontal perspective. Trajectories allow for behavioral studies as seen in traffic applications with a surveillance perspective as the one found in [15], using both real and simulated data with the focus of clustering similar trajectories.

This vision-based dataset consists of on-road full surround images captured at high resolution at 12 FPS. These data are suitable for both detection, tracking, and estimating trajectories around the ego-vehicle introduced as multi-perspective 3D tracking. The contributions of this dataset are as follows: 1) Full surround using four independent cameras with slightly overlapping perspectives; 2) Image data from U.S. Highways recorded in high resolution of 2704×1440 at 12 frames per second; 3) More than 4000 bounding box annotations with id, occlusion, and truncation tags.

II. DATASET AND CHALLENGES

A. Data Acquisition

The database is collected on U.S. Highways in southern California over 2.5 hours of driving. The drives are divided into sequences consisting of challenging behaviors found around the ego-vehicle including e.g. overtaking, cut-ins, and cut-outs. The data capturing vehicle is equipped with four GoPro HERO3+ achieving a full surround view with limited overlap as seen from Figure 1. Each camera is recording at a resolution of 2704×1440 at 12 FPS, and post-processed offline to synchronize and correct distortion. All four cameras are calibrated to a common road plane using homographies estimated from recordings at a parking lot, where the relative world positions of points used for the estimation are known.

B. Ground Truth Annotation

The ground truth is obtained for each of the four perspectives by manually annotating bounding boxes in the format as seen in (1), where (x_1, y_1) denotes the top-left corner, and (x_2, y_2) denotes the lower-right corner.

$$[\textit{frame}, \textit{id}, \textit{occlusion}, \textit{truncation}, x_1, y_1, x_2, y_2] \quad (1)$$

Each vehicle is assigned an identification number, *id*, to evaluate tracking. Note the *id* is consistent between perspectives giving the option of multi-perspective tracking. The occlusion and truncation tags are both divided into three levels being *No*, *Partial*, and *Heavy*. Here, *No* equals 0%, *Partial* includes vehicles up 50%, while *Heavy* covers 50%+ for both occlusion and truncation. Three levels are chosen to simplify the annotation workload, while maintaining a certain division for analysis purposes. Annotation examples are shown in Fig. 2, and histograms of the data are shown in Fig. 3. Note that almost 50% of the vehicles are present in three perspectives, which motivates the challenge of observing vehicles as they move around the ego-vehicle. Though, vehicles tend to transition between perspectives, it is also clear that the front and rear perspectives dominate in the number of visible vehicles.

The bounding box annotations allow for evaluation of both detection and tracking. In order to evaluate 3D tracking, we use the homographies to transform each ground truth trajectory to the road plane. The middle of the bottom of the bounding box, $(x_1 + 0.5(x_2 - x_1), y_2)$, is used as vehicle position. Using this position sets high demands to the precision of the bounding box, especially the bottom. This also means the trajectories will show the closest point approximately, instead of the more intuitive center of the surrounding vehicle. The road plane annotations have four entries:

$$[\textit{frame}, \textit{id}, x, y] \quad (2)$$

Examples of ground truth trajectories in the road plane are shown in Fig. 1. The ground truth trajectories appear noisy

Fig. 2. Sample images from one of the sequences in the dataset with overlaid ground truth annotations at six different time instances. The bounding boxes are color coded by *id* and labled with *PO*, *HO*, *PT*, and *HT* denoting partial occlusion, heavy occlusion, partial truncation and heavy truncation, respectively.

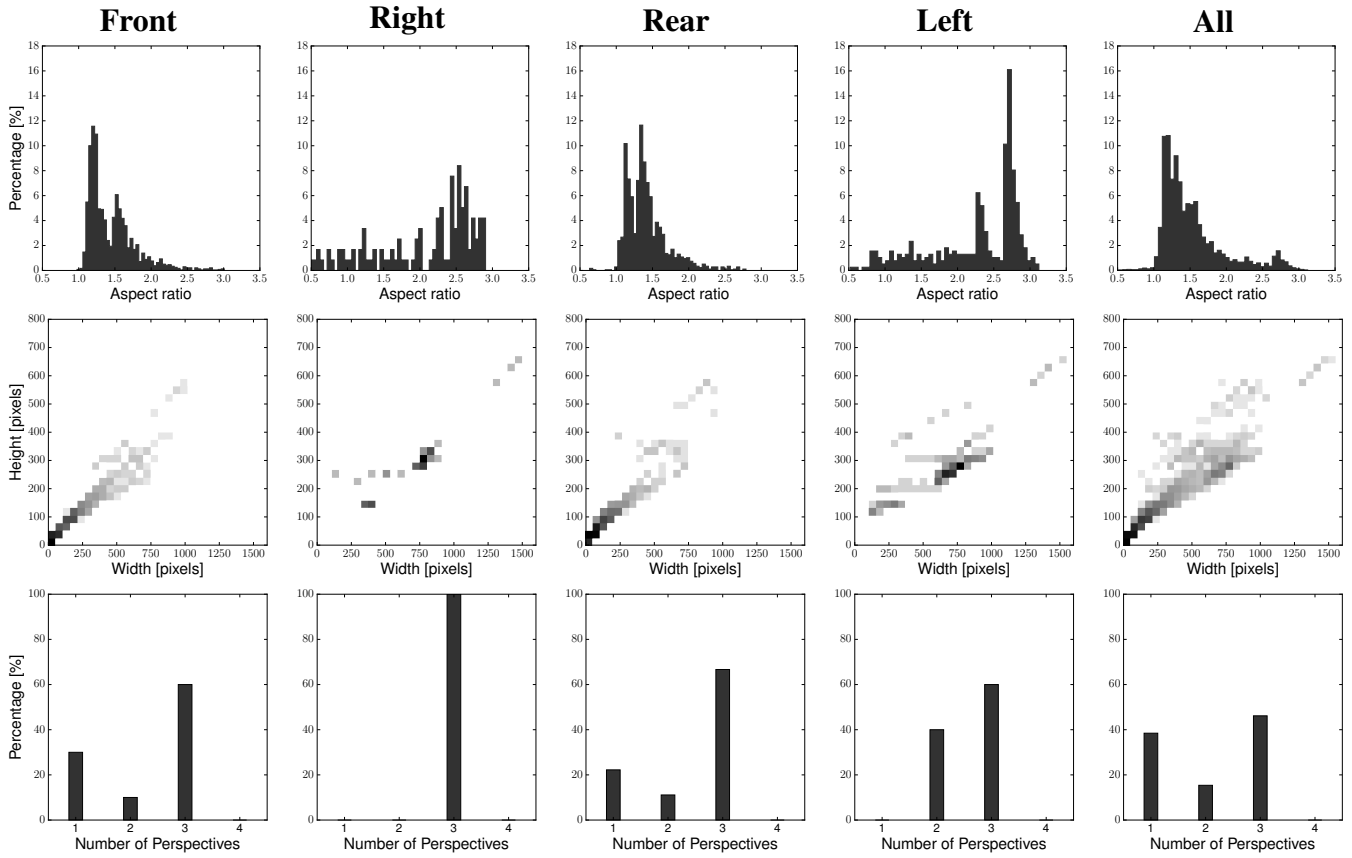

Fig. 3. Histograms of bounding box annotation measures. Each column represents a perspective, and the rightmost column is the total. The rows are from top to bottom: Bounding box aspect ratio, bounding box width and height in pixels, and the number of perspectives in which the visible vehicles are present during the sequence.

due to vibrations caused by the moving platform, which are amplified by the road transformation. Note the applied computer vision algorithms will be affected by the same noise as the ground truth.

C. Evaluation Metrics

Well-established metrics exist for both detection and single-perspective tracking of vehicles. Evaluating vehicle trajectories is less commonly used, in this work referred to as multi-perspective 3D tracking, and is explained in further detail.

Vehicle Detection: The average precision (AP) is commonly used in vehicle detection, calculated as the area under the precision-recall curve. For a ground truth bounding box to be matched with a detected bounding box, an overlap is defined as the intersection over union. Traditionally an overlap of 0.5 is used, or 0.7 for higher precision which is important in this work since homographies are utilized, where a small deviation in the image plane can have a large effect in road plane coordinates.

Vehicle tracking: The single-perspective multiple-object tracking is evaluated using the CLEAR MOT metrics (MOTA, MOTP) [16], together with metrics including fragmentations (Frag) and ID switches (IDS) [17], mostly tracked (MT), and mostly lost (ML) [18]. A ground truth trajectory is counted as mostly tracked if it is associated more than 80% of the time by definition. Likewise is a ground truth trajectory accounted as a ML if associated in less than 20% of the time. A fragmentation is added every time a ground truth trajectory is split. An ID switch is added if a ground truth trajectory is matched with another ID than the one that is currently associated.

Multi-Perspective 3D tracking: The field of multi-perspective 3D tracking is less explored without any common metrics. The problem, however, is not so different from tracking in single perspectives. For this reason, we propose to use similar metrics in the road plane. Instead of association between ground truth bounding boxes and tracked bounding boxes, this will require association between points in the road plane. To this end, we use a weighted euclidean distance from ground truth trajectory points. This does however mean that the cost of matching a candidate to ground truth is not normalized (as the bounding box overlap definition), and thus MOTP is not well-defined ([19] suggests to normalize the distance with the matching threshold). The original definition of MOTP says that it is the average dissimilarity between all true positives and their corresponding ground truth targets. With bounding box overlap a score as close to 1.0 is ideal. Using euclidean distance changes the ideal score to be as close to zero as possible, as it is now defined as the average distance between true positives and their corresponding ground truth targets. To reduce the confusion, we refer to this score as multiple object tracking euclidean precision (MOTEP):

$$\text{MOTEP} = \frac{\sum_{t,i} d_{t,i}}{\sum_t c_t} \tag{3}$$

Where c_t denotes the number of matches in frame t, and $d_{t,i}$ is the weighted euclidean distance between target i in frame t and the corresponding ground truth target. A direct transfer to the road plane domain would be a static distance allowed from each ground truth trajectory point. However, as a nature of inverse perspective mapping, small variations close to the ego-vehicle will not be as severe as small variations further away. We propose to make the matching criterion a function of the x-distance from the ego-vehicle. A target is matched to a ground truth target if it fulfills:

$$d_{t,i} < a|x| + b \tag{4}$$

where $|x|$ is the absolute x-coordinate of the ground truth target, a is the gradual increase in allowed distance, and b is the allowed distance at $|x| = 0$. From inspection of the ground truth trajectories, we define $a = 0.04$ and $b = 2$. It should however be noted that variations in the y-direction are more likely to cause erroneous matches. For this reason, we define the weighted euclidean distance as:

$$d_{t,i} = \sqrt{(gt_x - tr_x)^2 + 4\,(gt_y - tr_y)^2} \tag{5}$$

where $gt = [gt_x \; gt_y]^{\mathrm{T}}$ is the 2D ground truth position and $tr = [tr_x \; tr_y]^{\mathrm{T}}$ is the position of the target. Thus, distances in the y-direction have a double weight.

Though, ID switches are counted, the metrics suggested above do not encapsulate the importance of ID switches that happen specifically in the transition between perspectives. For this reason, we intend to include trajectory similarity measures in the future, which could e.g. be longest common subsequence (LCS).

III. Experimental Evaluation

In this section we evaluate methods for each of the tasks; detection, tracking, and 3D tracking. Vehicle detection is performed for each of the four inputs of the cameras. The detections in each perspective are used by a vehicle tracker to associate detections between frames for each of the four perspectives. The positions of the tracked vehicles are transformed to the road surface, where the trajectories are connected between perspectives. A detailed description of the implementation is found in [20].

In this baseline a bounding box overlap criterion of 0.7 is used for both detection and tracking. A partial truncation level is used to evaluate up to 50% truncated vehicles. Heavily truncated vehicles are ignored i.e. not included even if it is correctly detected or not. Also, vehicles with a height less than 35 pixels are ignored, since these are far away from the ego vehicle (approximately 50 meters) and therefore not of interest. Lastly, an ignore region is defined to ignore oncoming traffic on the other side of the crash barrier. Ignored vehicles are not included in the evaluation, even if they are correctly detected/tracked or not.

Detection Evaluation: Both the DPM [21], [22], [23] and the SubCat [24] are tested on the proposed dataset for detection of vehicles in the four different perspectives. All detectors are trained on the KITTI dataset [8], using

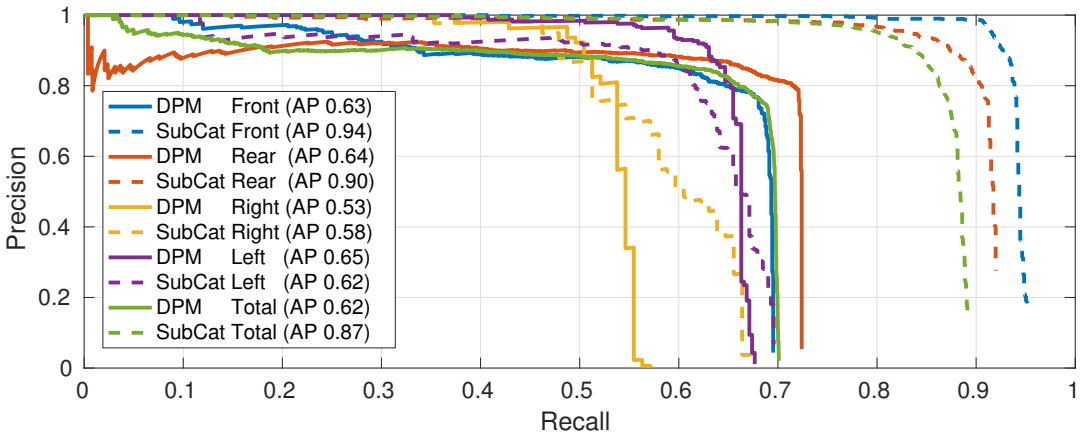

Fig. 4. Precision-Recall curve of DPM and SubCat detections with the average precision denoted in the label. Note the lower precisions in the side perspectives. Evalutated with a minimum bounding box overlap 0.7 and up to 50% occlusion and 50% truncation level.

TABLE II

SINGLE PERSPECTIVE TRACKING RESULTS.

Methods	MOTA↑	MOTP↑	IDS↓	Frag↓	MT↑	ML↓	Recall↑	Precision↑
Front								
SubCat-MPMDP	0.82	0.83	1	1	0.80	0.00	0.83	1.00
DPM-MPMDP	0.71	0.78	0	0	0.80	0.10	0.81	0.89
Rear								
SubCat-MPMDP	0.82	0.85	0	9	0.75	0.00	0.87	0.94
DPM-MPMDP	0.87	0.80	1	4	0.75	0.00	0.87	1.00
Left								
SubCat-MPMDP	0.76	0.77	0	1	0.40	0.20	0.76	1.00
DPM-MPMDP	0.77	0.80	0	1	0.40	0.40	0.77	1.00
Right								
SubCat-MPMDP	0.55	0.83	0	0	0.33	0.33	0.55	1.00
DPM-MPMDP	0.62	0.82	0	0	0.67	0.33	0.62	1.00
Total								
SubCat-MPMDP	0.81	0.84	1	11	0.65	0.08	0.83	0.97
DPM-MPMDP	0.79	0.79	1	5	0.69	0.15	0.83	0.95

TABLE III

MULTI-PERSPECTIVE 3D TRACKING RESULTS.

Methods	MOTA↑	MOTEP↓	IDS↓	Frag↓	MT↑	ML↓	Recall↑	Precision↑
SubCat-MPMDP-3D	0.64	1.23	5	93	0.46	0.15	0.79	0.85
DPM-MPMDP-3D	0.42	1.23	3	83	0.38	0.15	0.74	0.70

the same model for all four perspectives. The detectors are evaluated using the average precision calculated as the area under the precision-recall curve found in Fig. 4. SubCat is found to outperform the DPM, especially in the front and rear perspectives, with mixed results for the side perspectives. This shows that the side views introduce new challenges compared to the traditional front and rear perspective. DPM employs parts, which implies more robustness to distortion in appearance due to side views. SubCat learns many models, so on normal settings it operates better on rigid or quasi-rigid objects like vehicles.

Tracking Evaluation: The detections are used by a modified version of the MDP tracker presented in [25] to track the vehicles between frames. This modified tracker is referred to as multi-perspective MDP (MPMDP) tracker. The tracking evaluation is presented in Table II. As expected, better detections of the SubCat detector also lead to better scores in tracking, though the overall difference in score is reduced by the tracker.

Multi-Perspective 3D Tracking Evaluation: Lastly, the trajectories are transformed from image planes to road plane and compared to ground truth. The multi-perspective 3D tracking performance is listed in Table III. The choice of detector proves to be of importance to the score, and is in particular impacting the precision and thus also the MOTA.

IV. CONCLUDING REMARKS

This paper introduces a novel dataset that builds upon the advances within monocular vision-based on-road detection and tracking of vehicles, presenting multiple perspectives for a full surround analysis. A total of four high resolution cameras are used to capture the surroundings of a vehicle

in a highway environment. Furthermore, ground truth annotations enable evaluation of detection, tracking, and multi-perspective 3D tracking. This means that methods designed for observing vehicles can be evaluated in a full surround looking framework, ultimately exposing and solving challenges introduced in a multi-perspective camera setup that potentially could reinforce current sensor suites of intelligent vehicles.

ACKNOWLEDGMENT

The authors would like to thank the support of associated industry partners, especially Toyota CSRC. We also thank our colleagues at the Laboratory for Intelligent and Safe Automobile (LISA), University of California, San Diego, for assisting with the data gathering and providing invaluable discussions and comments.

REFERENCES

[1] S. Sivaraman and M. Trivedi, "Looking at Vehicles on the Road: A Survey of Vision-Based Vehicle Detection, Tracking, and Behavior Analysis," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 4, 2013.

[2] M. S. Kristoffersen, J. V. Dueholm, R. K. Satzoda, M. M. Trivedi, A. Møgelmose, and T. B. Moeslund, "Towards Semantic Understanding of Surrounding Vehicular Maneuvers: A Panoramic Vision-Based Framework for Real-World Highway Studies," *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016.

[3] D. Kasper, G. Weidl, T. Dang, G. Breuel, A. Tamke, A. Wedel, and W. Rosenstiel, "Object-Oriented Bayesian Networks for Detection of Lane Change Maneuvers," *Intelligent Transportation Systems Magazine, IEEE*, vol. 4, no. 3, pp. 19–31, Fall 2012.

[4] J. Schlechtriemen, A. Wedel, J. Hillenbrand, G. Breuel, and K.-D. Kuhnert, "A lane change detection approach using feature ranking with maximized predictive power," in *Intelligent Vehicles Symposium Proceedings, 2014 IEEE*, 2014.

[5] C. Caraffi, T. Vojir, J. Trefny, J. Sochman, and J. Matas, "A System for Real-time Detection and Tracking of Vehicles from a Single Car-mounted Camera," in *IEEE Conference on Intelligent Transportation Systems*, 2012.

[6] C. Wang, Y. Fang, H. Zhao, C. Guo, S. Mita, and H. Zha, "Probabilistic Inference for Occluded and Multiview On-road Vehicle Detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 1, 2016.

[7] G. Pandey, J. R. McBride, and R. M. Eustice, "Ford Campus vision and lidar data set," *The International Journal of Robotics Research*, vol. 30, no. 13, 2011.

[8] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[9] Laboratory for Intelligent and Safe Automobiles, UCSD, "Vision for Intelligent Vehicles and Applications (VIVA)." [Online]. Available: http://cvrr.ucsd.edu/vivachallenge/

[10] N. Das, E. Ohn-Bar, and M. M. Trivedi, "On Performance Evaluation of Driver Hand Detection Algorithms: Challenges, Dataset, and Metrics," in *IEEE Conference on Intelligent Transportation Systems*, 2015.

[11] A. Møgelmose, M. M. Trivedi, and T. B. Moeslund, "Vision-Based Traffic Sign Detection and Analysis for Intelligent Driver Assistance Systems: Perspectives and Survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 4, 2012.

[12] A. Møgelmose, D. Liu, and M. M. Trivedi, "Detection of U.S. Traffic Signs," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 6, 2015.

[13] M. B. Jensen, M. P. Philipsen, A. Møgelmose, T. B. Moeslund, and M. M. Trivedi, "Vision for Looking at Traffic Lights: Issues, Survey, and Perspectives," *IEEE Transactions on Intelligent Transportation Systems*, vol. PP, no. 99, 2016.

[14] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[15] B. T. Morris and M. M. Trivedi, "Trajectory Learning for Activity Understanding: Unsupervised, Multilevel, and Long-Term Adaptive Approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, 2011.

[16] K. Bernardin and R. Stiefelhagen, "Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics," *Journal on Image and Video Processing*, 2008.

[17] A. Milan, L. Leal-Taixé, I. D. Reid, S. Roth, and K. Schindler, "MOT16: A Benchmark for Multi-Object Tracking," *CoRR*, vol. abs/1603.00831, 2016.

[18] B. Wu and R. Nevatia, "Tracking of Multiple, Partially Occluded Humans based on Static Body Part Detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2006.

[19] A. Milan, K. Schindler, and S. Roth, "Challenges of Ground Truth Evaluation of Multi-target Tracking," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013.

[20] J. V. Dueholm, M. S. Kristoffersen, R. K. Satzsoda, T. B. Moeslund, and M. M. Trivedi, "Trajectories and Maneuvers of Surrounding Vehicles with Panoramic Camera Arrays," *IEEE Transactions on Intelligent Vehicles*, 2016.

[21] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part Based Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, 2010.

[22] H. Zhang, A. Geiger, and R. Urtasun, "Understanding High-Level Semantics by Modeling Traffic Patterns," in *IEEE International Conference on Computer Vision*, 2013.

[23] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun, "3D Traffic Scene Understanding From Movable Platforms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, 2014.

[24] E. Ohn-Bar and M. M. Trivedi, "Learning to Detect Vehicles by Clustering Appearance Patterns," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 5, 2015.

[25] Y. Xiang, A. Alahi, and S. Savarese, "Learning to Track: Online Multi-Object Tracking by Decision Making," in *IEEE International Conference on Computer Vision*, 2015.