2013 IEEE Intelligent Vehicles Symposium (IV)
June 23-26, 2013, Gold Coast, Australia

Vehicle Detection and Tracking at Intersections by Fusing Multiple Camera Views

Elias Strigel, Daniel Meissner, and Klaus Dietmayer
Institute of Measurement, Control, and Microtechnology
Ulm University
Ulm, Germany
Email:{elias.strigel, daniel.meissner, klaus.dietmayer}@uni-ulm.de

Abstract— **Intersections are challenging locations for drivers. Complex situations are common due to the variety of road users and intersection layouts. This contribution describes a real time method for detecting and tracking vehicles at intersections using images captured by a static camera network. After background subtraction, the foreground segments are projected on a common fusion map. Using this fusion map, the pose, width, and height of the vehicles can be determined. After that, the detected objects are tracked by a Gaussian-Mixture approximation of the Probability Hypothesis Density filter. Results of the intersection perception can further be communicated to equipped vehicles by wireless communication.**

I. INTRODUCTION

Almost 61,000 fatal accidents occurred at intersections in the years 1996-2004 representing 21 % of all traffic accident fatalities [1]. Complex situations are common due to the variety of road users and intersection layouts. Thus, intersections are challenging locations for drivers.

The aim of the joint initiative Ko-FAS [2] is to contribute to an increase of the road safety. To achieve this, the joint project Ko-PER, which is part of Ko-FAS, aspire a full perception of the vehicles local environment. Especially at intersections, with complex scenarios and extensive traffic density, the driver's field of view and the vehicle based perception systems are highly restricted.

Therefore, a public intersection has been equipped with a multi sensor network to capture all dynamic objects and generate a bird's eye view of the current scene. Vehicle-To-Vehicle (V2V) and Vehicle-To-Infrastructure (V2I) communication with the use of cooperative perception strategies is used to extend the vehicle's local perception.

This article presents a method for detecting and tracking moving vehicles in the inner part of a public intersection based on the images of eight monochrome CCD cameras. To overcome problems like occlusions and missing depth information, the information of the single cameras is fused. Other methods, like stereo or wide base line stereo are not applicable. They live from various feature correspondence between the different views, which is not given in our setup due to the systemic mounting positions of the cameras.

The moving vehicles are segmented using background subtraction. Inverse Perspective Mapping (IPM) transforms the different views to the ground plane of the intersection. The perspective distortions caused by the inverse mapping, which is only valid for world points on the ground plane can be removed by fusing the different views. This enables the detection of vehicles with their poses and dimensions.

This contribution is organized as follows. The related work is presented in the next section. Section 3 gives an overview of the perception system at the public intersection in Aschaffenburg (Germany). Section 4 describes the algorithm for detecting and tracking moving vehicles. Results are presented in Section 5. Section 6 concludes the article.

II. RELATED WORK

There are several published approaches in the field of object tracking for traffic applications, mainly with the scope of surveillance or traffic analysis. Some of them use complex three dimensional object models of the expected objects for detection and tracking [3]–[5]. Other approaches rely on the simple cuboid shape of the occurring objects [6]–[8]. Instead of fitting 3D models, region-based [9], [10] as well as feature-based [11]–[14] object detection and tracking in regions of interest of the image are common. A good overview can be found in [15].

A related algorithm to the proposed approach can be found in [16]. Here a Probability Fusion Map (PFM) is created based on three camera views, captured from a bridge to count vehicles on a highway. All three cameras were mounted at the same height, looking along the highway. In the proposed article however, we use eight cameras, mounted on different heights around an intersection and the traffic situation is very complex. Not all cameras have overlapping field of views, which motivates the use of the later proposed field of view map.

Another related approach is described in [17]. Khan et al. uses background subtraction and a planar homography constraint to fuse the different views into a arbitrary view. They aim at the detection and tracking of people in a crowded scene. A similar method is described in [18]. Xu et al. utilizes homography mapping to fuse the foreground region polygons of two camera views to detect vehicles and pedestrian. Therefor on camera serves as reference view.

The approach presented in this contribution was introduced in [19] and is extended by the tracking algorithm.

III. SYSTEM DESCRIPTION

In the context of the joint project Ko-PER, a complex public intersection (see Fig. 1) in an urban area is chosen

978-1-4673-2754-1/13/$31.00 ©2013 IEEE

to install the multi sensor network. This enables real-world test conditions with a medium traffic volume of 22,000 to 23,000 vehicles per day.

Fig. 1. Public intersection in Aschaffenburg (Germany) used for the test system (Picture is kindly provided by the Wuerzburg Institut for Traffic Sciences GmbH, http://www.wivw.de).

The multi sensor network consists of 14 SICK LD-MRS research multilayer laserscanners (operating at 12.5 Hz), eight monochrome CCD cameras with VGA resolution operating at 25 Hz and two high definition cameras (2 megapixels, 50 Hz). To achieve temporal association of the sensor data, the different sensor types are synchronously triggered with pulses derived from the GPS time. Each sensor subsystem serves as a separate object detection system. The perceived objects of the different subsystems are then fused to improve perception results.

This contribution is focused on the VGA camera system. Prior to the installation of the camera sensors, the FOV of all cameras were simulated for different poses using 3D modeling software. This enabled a reduction of occlusions by stationary objects. Mounting the sensors at positions up to 12 m above the road level and overlapping field of views makes the camera system robust against occlusions by moving objects. Four cameras were equipped with wide-angle lenses for a comprehensive coverage of the central intersection area. The incoming lanes of the intersection are observed by three cameras and the last camera provides a top view of area, at a crossing of two crosswalks and a bicycle lane (see Fig. 2). More information about the KoPER intersection perception system can be found in [20].

IV. VEHICLE DETECTION AND TRACKING

A. Detection of Moving Vehicles

Each of the eight cameras captures a part of the intersection. To reduce occlusions, large parts of the intersection are captured by more than one camera. Segmentation of moving objects can be performed by state of the art background subtraction algorithms like the well-known Mixture of Gaussians (MoG) algorithm [21]. Therefore, most of the static image parts can be rejected. Only the moving vehicles

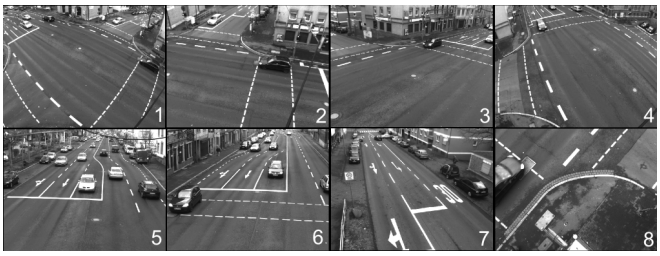


Fig. 2. Field of views of the cameras. Camera 1-4 cover the central area, camera 5-7 capture the incoming lanes and camera 8 serves as top view camera.

in the image domain of each camera remain. But the poses and dimensions of the moving vehicles in world coordinates are still unknown. Furthermore, overlapping vehicles lead to foreground segments containing more than one object. To overcome this problem, the information of the single camera views is combined. The fusion is based on Inverse Perspective Mapping (IPM) to a view plane which is common to all camera perspectives. IPM performs a back projection of the image plane to a previous chosen back projection plane. World points which are captured from the back projection plane will be correctly remapped. The scale of the remapped objects remains the same. World points captured above and below the back projection plane appear distorted. This can be seen in Fig. 3 where three of the captured perspectives are mapped onto the ground plane. For the algorithm, only the binary foreground images are mapped on the ground plane.

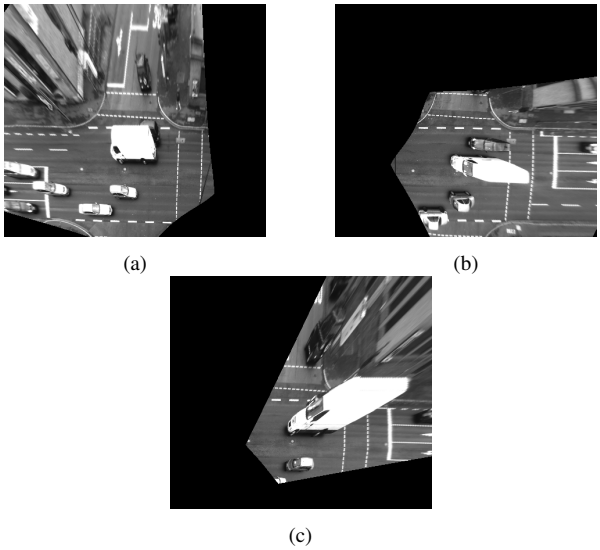


Fig. 3. Texture mapping onto the ground plane of three cameras. Objects which are above the ground plane are perspective distorted.

To determine the poses and dimensions of the vehicles, choosing the road surface of the intersection as common back projection plane is appropriate. In the central area of the intersection the flat world assumption is valid.

Consequently, parts of vehicles which are located on the road surface will be projected without perspective distortion. The remaining parts lead to distortions on the back projection

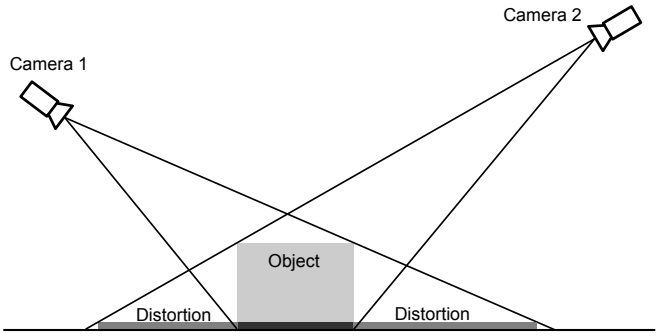
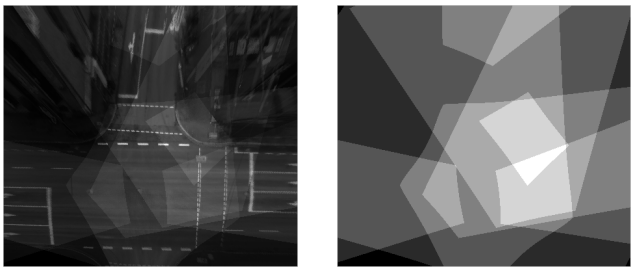


Fig. 4. Visualisation of the Inverse Perspective Mapping: In the view of a single camera, the object appears larger on the ground plane. By adding a second camera perspective, the real object size and position on the ground plane can be determined.

plane. With the use of the different camera views which capture the vehicles from different sides, a reduction of the distortions is possible. Having the correct base area of the vehicles on the road surface, determining the poses and dimensions can easily be done.

In Fig. 5(a) the IPM of all camera views onto the road surface is shown. The central area of the intersection is simultaneously captured by multiple cameras. For more details about IPM refer to [22].



(a) (b)

Fig. 5. a.) IPM of all camera views onto the road surface of the intersection. Objects like e.g. houses are distorted due to their height above the ground plane. The scale of the road markings is recovered correctly. Only at the pedestrian crossings, where the road has a light arching, the back projections match not exactly. b.) Field of View map $S'(x', y')$ of the intersection. The brighter the areas, the more cameras capture the area on the ground plane.

In the following, the vehicle detection algorithm is described in detail. To perform the IPM, an exact calibration of the whole camera sensor network is necessary. This includes the intrinsic camera parameters, the knowledge of the pose of each sensor and the pose of the ground plane. The intrinsic parameters of each camera were determined using a checkerboard and the software described in [23]. For the extrinsic calibration of the camera network, the whole central area of the intersection was covered with checkerboards of different sizes. The extrinsic parameters were determined by a parameter optimization which minimized the reprojection error of all boards in every view, simultaneously.

Fig. 6 describes the entire fusion algorithm. After image acquisition, the background of each captured image is subtracted. This can be done by several background subtraction approaches. A good overview is given in [24]. In this work, the MoG algorithm performs the background subtraction, followed by morphological filtering. The result of this step is a binary image $I_i(x, y)$ for each camera $i = 1 \dots 8$ which only contains the foreground segments. The pixel position in image coordinates is described by x and y. After the foreground segmentation, the IPM is performed to the foreground segments which leads to the perspective mapped image $I'_i(x', y')$. Here x' and y' are the coordinates on the common fusion map. To fuse the different mapped images, they are summed as followed:

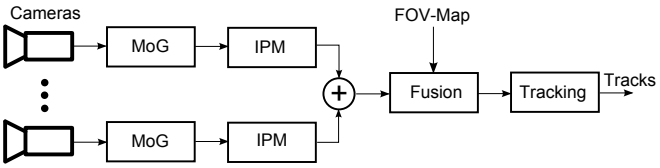


Fig. 6. Block diagram of the proposed algorithm. After image acquisition, the foreground is segemented using the MoG algorithm, followed by the IPM. Finally the results of the IPM for each camera are fused to determine the vehicles in world coordinates.

$$I'_F(x', y') = \sum_i I'_i(x', y') \tag{1}$$

If a vehicle was captured by two cameras and the IPM $I'_i(x'_0, y'_0)$ projects a part of the foreground to the same position (x'_0, y'_0) in both projections, then $I'_F(x'_0, y'_0) = 2$. The maximum value of I'_F depends on the number of the overlapping field of views. An area on the ground plane, which can be observed by e.g. four cameras, can lead to a maximum value of 4 in the fusion map I'_F. The more different views of the observed vehicles are available, the better are the estimated poses and orientations. Hence the achievable reduction of the reprojection error depends on the viewing direction and assembly positions of the cameras.

In the next step the vehicle detections are calculated based on the fusion map I'_F. Therefore, a Field of View (FOV) map $S'(x', y')$ is calculated. The sizes and resolutions of $S'(x', y')$ and the fusion map $I'_F(x', y')$ are equal. They describe the same ground plane grid. $S'(x', y')$ contains the number of cameras, which observe the ground plane at each map point (x', y'). Fig. 5(b) shows the Field of View map $S'(x', y')$ for the intersection in Aschaffenburg.

For an vehicle moving on the back projection plane, the $I_F(x', y')'$ has the same value as $S'(x', y')$ for the base area of the vehicle. Consequently:

$$I'_F = S', \text{ Imagepixel lies in ground plane} \tag{2}$$
$$I'_F < S', \text{ Imagepixel lies above the ground plane} \tag{3}$$

Depending on the object constellation and the assembly of the camera, the projective distortions cannot be fully removed. However, using different perspectives and a sufficient number of cameras, the vehicles can be detected robustly.

Fig. 4 shows the mapping process to a common ground plane. Seen from a single camera, the object appears larger

on the ground plane. When a second camera is added to the scene, the object dimensions on the ground plane can be limited to the real object base.

In the last step, connected regions in the fusion map, which fulfill 2, are segmented. For each segment the minimal area rectangle, containing the segment, is determined. Each rectangle describes the detection of a vehicle and serves as input for the tracking module, which will be described in the next section.

Like all methods based on the assumption of static background, the proposed algorithm is afflicted by sudden light changes, moving shadows, insufficient light conditions and stop-and-go handling.

B. Tracking the Detected Vehicles

The aim of the tracking is the estimation of the number of objects and their states based on a sequence of noisy and cluttered vehicle detections. Since the number of objects as well as the object states are random variables, the multi-object state X_k of N_k objects to time k and the multi-object measurement Z_k are represented by a random finite set (RFS) [25]. This leads to the multi-object Bayes filter [25] to solve the tracking problem. The multi-object Bayes filter is in general computational intractable, therefore the Gaussian-Mixture approximation of the Probability Hypothesis Density (GM-PHD) filter is used [26]. This section summarizes some details of the GM-PHD filter. For more details about PHD filters, refer to [25] and [26]. In case of PHD filters, the intensity (PHD) of the RFS is used to represent the multi-object state X_k. The PHD function is defined over the state space and the integral of the PHD function over a region of the state space gives the number of elements of X_k that are in this region. Another advantage of PHD filters is the missing, time consuming data association step. The PHD is updated with the whole multi-object measurement. Assuming linear Gaussian process and measurement models as well as state independent detection and survival probabilities, the PHD filter can be implemented using GM methods [26]. The approximation of the PHD function with a weighted sum of Gaussian distributions results in a closed form solution of the multi-object tracking problem using a Kalman filter. To track the vehicle detections of the camera system, a linear constant velocity (CV) model is used. Due to the GM representation, the sum of the Gaussian weights is an estimate for the number of objects $\hat{N}_k$. Thus, the $\hat{N}_k$ Gaussians with the highest weight determine the current tracks.

V. RESULTS

The calculation of the IPM was performed by creating a lookup table for each view with a resolution of 451x401 cells. Each cell represents a ground plane square with an edge length of 0.1 m. Using a lookup table enables a fast computation of the transformation between the image and the fusion map including the lens distortions of each camera.

Because of the lack of ground truth trajectories, the evaluation is done based on back projection of the result into two different camera views. Thus, a validation is possible by comparing the tracked objects with the actual position of the vehicles in the two representative views. Additionally, for the evaluation of the tracking stage, a part of the evaluation sequences is presented with respect to the intersection map.

To get valid detections, we assume that the vehicles were captured by a minimum of two cameras from different sides. This assumption is only valid in the central intersection area. The incoming lanes are observed by only one camera each. Fig. 7 shows two vehicles, crossing the intersection side by side. The view of camera 1 is depicted in Fig. 7a, the perspective of camera 4 in Fig. 7b. After the background subtraction and IPM, the fusion map is created (see. Fig. 7c). Applying the Field of View map (see. 2) and the requirement, that a minimum of two cameras contribute to the detection, leads to the fusion results depicted in Fig. 7d. The dimensions and poses of the vehicles are shown by the minimum area rectangle. The height of the objects can not be determined using the 2D fusion map. For visualization, a fixed height is assumed.

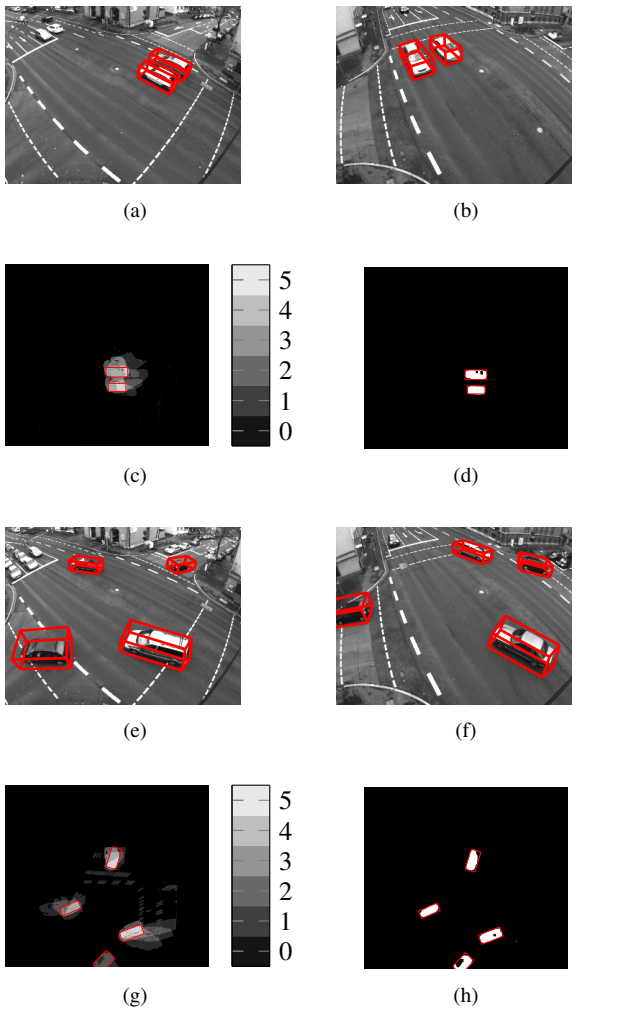


Fig. 7. a,b) Back projected detection results into the images of camera 1 and camera 4. c) Fusion map: The grayscale values show the number of camera perspectives for each map cell. d) Shows the Fusion map after applying the the FOV map, morphological postprocessing and connected component analysis. e-h) Detection results analogous to a-c).

Note that even though the vehicles overlap in the foreground image of camera 1, the fusion algorithm is able to split those cars using the remaining perspectives. Fig. 8 shows further detection results.

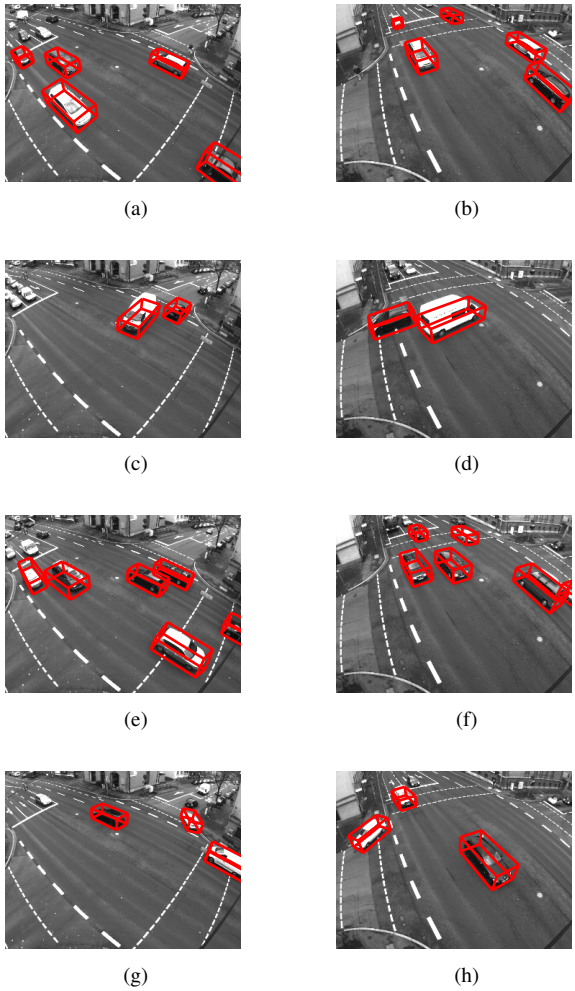


Fig. 8. Further back projected detection results into the images of camera 1 and camera 4.

The results of tracking the detected vehicles during crossing the central area of the intersection are depicted in Fig. 9. The figure shows a part of the evaluation sequences. All the vehicles are persistently tracked on their way through the intersection.

Detection and tracking runs real-time with 25 fps standard desktop PC. The background subtraction of the eight video streams was calculated using the GPU MoG algorithm of the well-known OpenCV library [27].

VI. CONCLUSION

In this paper we have presented an algorithm for detection and tracking of vehicles, based on a multi camera network. This was achieved by fusing the different image detection in a common map. The vehicles were detected and tracked including pose, width and length. Even occlusions can be handled by the proposed algorithm. The system was tested at a complex public intersection of the Ko-PER-project. We are currently working on the integration of improvements for shadow and stop-and-go handling.

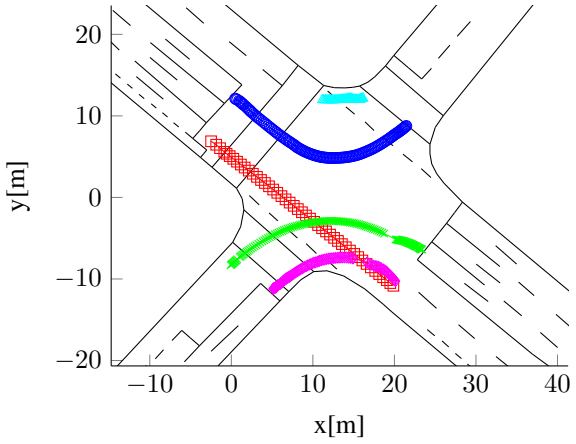


Fig. 9. Tracking results of a part of the evaluation sequences. All the vehicles could be tracked during crossing the central area of the intersection.

ACKNOWLEDGMENT

This work partially results from the joint project Ko-PER, which is part of the project initiative Ko-FAS, and has been funded by the German *Bundesministerium für Wirtschaft und Technologie* (Federal Department of Commerce and Technology) under grant number 19S9022G.

REFERENCES

[1] European Road Safety Observatory, "Traffic safety basic facts 2006 junctions," European Road Safety Observatory, Tech. Rep., January 2007, safetyNet, Project co-financed by the European Commission. [Online]. Available: http://ec.europa.eu/transport/road_safety/specialist/statistics/care_reports_graphics/index_en.htm

[2] "Forschungsinitiative Ko-FAS," http://www.ko-fas.de, Juli 2012. [Online]. Available: http://www.ko-fas.de

[3] H. Dahlkamp, A. Ottlik, and H.-H. Nagel, "Comparison of edge-driven algorithms for model-based motion estimation," in *Proc. First International Workshop on Spatial Coherency for Visual Motion Analysis*. Springer, 2004, pp. 38–50.

[4] M. Leotta and J. Mundy, "Vehicle surveillance with a generic, adaptive, 3d vehicle model," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 7, pp. 1457–1469, 2010.

[5] N. Buch, J. Orwell, and S. Velastin, "Urban road user detection and classification using 3d wire frame models," *Computer Vision, IET*, vol. 4, no. 2, pp. 105–116, 2008.

[6] X. Song and R. Nevatia, "Detection and tracking of moving vehicles in crowded scenes," in *Motion and Video Computing, 2007. WMVC '07. IEEE Workshop on*, 2007, pp. 4–4.

[7] B. Johansson, J. Wiklund, P.-E. Forssén, and G. Granlund, "Combining shadow detection and simulation for estimation of vehicle size and position," *Pattern Recogn. Lett.*, vol. 30, no. 8, pp. 751–759, June 2009. [Online]. Available: http://dx.doi.org/10.1016/j.patrec.2009.03.005

[8] S. Atev and N. Papanikolopoulos, "Multi-view 3d vehicle tracking with a constrained filter," in *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, 2008, pp. 2277–2282.

[9] H. Veeraraghavan, O. Masoud, and N. Papanikolopoulos, "Computer vision algorithms for intersection monitoring," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 4, no. 2, pp. 78–89, 2003.

[10] S. Messelodi, M. Modena, and M. Zanin, "A computer vision system for the detection and classification of vehicles at urban road intersections," *Pattern Anal. Appl.*, vol. 8, no. 1, pp. 17–31, Sept. 2005. [Online]. Available: http://dx.doi.org/10.1007/s10044-004-0239-9

[11] N. Saunier and T. Sayed, "A feature-based tracking algorithm for vehicles in intersections," in *Computer and Robot Vision, 2006. The 3rd Canadian Conference on*, 2006, pp. 59–59.

[12] D. Beymer, P. McLauchlan, B. Coifman, and J. Malik, "A real-time computer vision system for measuring traffic parameters," in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, 1997, pp. 495–501.

[13] H. Veeraraghavan, P. Schrater, and N. Papanikolopoulos, "Adaptive geometric templates for feature matching," in *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*, 2006, pp. 3393–3398.

[14] N. Kanhere, S. Pundlik, and S. Birchfield, "Vehicle segmentation and tracking from a low-angle off-axis camera," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2, 2005, pp. 1152–1157 vol. 2.

[15] N. Buch, S. Velastin, and J. Orwell, "A review of computer vision techniques for the analysis of urban traffic," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 12, no. 3, pp. 920–939, 2011.

[16] F. Lamosa, Z. Hu, and K. Uchimura, "Vehicle detection using probability fusion maps generated by multi-camera systems," *Journal of Information Processing*, vol. 17, pp. 1–13, 2009.

[17] S. M. Khan and M. Shah, "A multiview approach to tracking people in crowded scenes using a planar homography constraint," in *In European Conference on Computer Vision*, 2006.

[18] M. Xu, J. Ren, D. Chen, J. Smith, and G. Wang, "Real-time detection via homography mapping of foreground polygons from multiple cameras," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*, 2011, pp. 3593–3596.

[19] E. Strigel, M. Schuster, and K. Dietmayer, "Detektion von Fahrzeugen an Straßenkreuzungen durch Fusion mehrerer Kameraansichten," in *8. Workshop Fahrerassistenzsysteme FAS 2012*, 2012, pp. 119 – 128. [Online]. Available: http://www.uni-das.de/de/Veranstaltungen/fas2012.php

[20] M. Goldhammer, E. Strigel, D. Meissner, U. Brunsmann, K. Doll, and K. Dietmayer, "Cooperative multi sensor network for traffic safety applications at intersections," in *15th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, sept. 2012, pp. 1178 –1183.

[21] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, 1999, pp. 637–663.

[22] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, ISBN: 0521540518, 2004.

[23] J. Y. Bouguet, "Camera calibration toolbox for Matlab," 2008. [Online]. Available: http://www.vision.caltech.edu/bouguetj/calib_doc/.

[24] S.-C. S. Cheung and C. Kamath, "Robust techniques for background subtraction in urban traffic video," S. Panchanathan and B. Vasudev, Eds., vol. 5308, no. 1. SPIE, 2004, pp. 881–892. [Online]. Available: http://dx.doi.org/10.1117/12.526886

[25] R. P. Mahler, *Statistical Multisource-Multitarget Information Fusion*. Artech House Inc., Norwood, 2007.

[26] B.-N. Vo and W.-K. Ma, "The gaussian mixture probability hypothesis density filter," *Signal Processing, IEEE Transactions on*, vol. 54, no. 11, pp. 4091 –4104, nov. 2006.

[27] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.