MULTI-CAMERA PEDESTRIAN DETECTION BY MEANS OF TRACK-TO-TRACK FUSION AND CAR2CAR COMMUNICATION

Anselm Haselhoff, Lars Hoehmann, Anton Kummert
Communication Theory, University of Wuppertal, D-42119 Wuppertal, Germany

Christian Nunn, Mirko Meuter, Stefan Mueller-Schneiders
Delphi Electronics & Safety, D-42119 Wuppertal, Germany

Keywords: Pedestrian detection, AdaBoost, Object detection, Track-To-Track fusion, Car2Car communication.

Abstract: In this paper a system for fusion of pedestrian detections from multiple vehicles is presented. The application area is narrowed down to driver assistance systems, where single cameras are mounted in the moving vehicles. The main contribution of this paper is a comparison of three fusion algorithms based on real image data. The methods under review include *Covariance Fusion*, *Covariance Intersection*, and *Covariance Union*. An experimental setup is presented, with known ground truth positions of the detected objects. This information can be incorporated for the evaluation of the fusion methods.
The system setup consists of two vehicles equipped with LANCOM® wireless access points, cameras, inertial measurement units (IMU) and IMU enhanced GPS receivers. Each vehicle detects pedestrians by means of the camera and an AdaBoost detection algorithm. The results are tracked and transmitted to the other vehicle in appropriate coordinates. Afterwards each vehicle is responsible for reasonable treatment or fusion of the detection data.

1 INTRODUCTION

Image processing and machine learning enable technological progress in advanced driver assistance systems. State-of-the-art systems include vehicle detection with forward collision warning, lane detection with lane keep-assistance, traffic-sign recognition and pedestrian detection.

Another enabling technology is Car2Car and Car2Infrastructure communication. For example vehicles can transmit or receive information about the traffic perceived by other vehicles. Furthermore the environment perception of a vehicle can be enriched by information generated by an infrastructure like traffic lights.

The more information is gained about the environment of the host-vehicle the more precise decisions can be made. One possible application is the collective detection of road users by means of multiple vehicles. Thus, the vehicles can compensate for shortcomings of the individual vehicles. The advantages of fusing detection results can be summarized as follows (Kaempchen, 2007):

- improved precision of 3D information,
- enlarged field of view,
- increased availability,
- improved robustness,
- increased object detection accuracy (higher TP-rate and lower FP-rate).

The most important aspect for object detection from a single camera is the improved precision of 3D information. From a single camera the 3D location can only be computed using assumptions like a flat ground plane or a priori knowledge of the object dimensions (Ponsa et al., 2005). The calculation of the lateral object position in vehicle coordinates (VCOS) is relatively precise, but the depth component is very inaccurate e.g. due to pitching. A second camera could improve the calculation of the depth component similar to what is used in stereo-vision.

Especially for pedestrian detection the enlarged field of view is of importance. One example could be a pedestrian that is going to cross the street just in front of the host-vehicle, but the pedestrian is oc-

Haselhoff A., Hoehmann L., Kummert A., Nunn C., Meuter M. and Müller-Schneiders S..
MULTI-CAMERA PEDESTRIAN DETECTION BY MEANS OF TRACK-TO-TRACK FUSION AND CAR2CAR COMMUNICATION.
DOI: 10.5220/0003315603070312
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2011), pages 307-312
ISBN: 978-989-8425-47-8

cluded by some object. Another vehicle that is driving in the opposite direction can clearly perceive the pedestrian and communicate this information.

Besides the advantages there are some challenges that have to be considered:

- choice of the fusion method,
- data available just for a limited time period,
- corrupted data (e.g. delay due to communication),
- positioning of the vehicles.

For automotive applications a Track-To-Track fusion schema is most likely, since the automotive suppliers usually output processed, tracked object lists (Matzka and Altendorfer, 2008). Therefore it is focused on the three Track-To-Track fusion methods *Covariance Fusion*, *Covariance Intersection*, and *Covariance Union*. The choice of the method depends on the data at hand e.g. if the data is correlated. An advantage of the Track-To-Track fusion is that the data can be fused if available. Hence, if the field-of-view of one vehicle is not overlapping with that of the host-vehicle or the communication is interrupted, simply no fusion is applied.

In general for multi-sensor fusion a calibration of the sensors is needed. In the multi-vehicle scenario the positions of the vehicles and the orientation is needed. This problem is encountered in two ways. For evaluation with ground truth data a map is used where the vehicles are registered and for online purposes an IMU enhanced GPS unit is in use.

The remainder of the paper proceeds as follows. It is started with the description of the overall system and the test vehicles in section 2. Afterwards a detailed describtion of the coordinate transformations and the fusion of pedestrian detections is presented. Finally, the evaluation and the conclusions are presented in section 5.

2 SYSTEM OVERVIEW

The system setup consists of two test vehicles equipped with LANCOM® wireless access points, cameras, inertial sensors, GPS, and a regular PC (Fig. 1). The monochrome camera is mounted at the position of the rear-view mirror and is connected to the PC. The vehicle bus enables the access to inertial sensors and GPS. The GPS is used to generate timestamps for the data that is subject to transmission. The GPS unit (AsteRxi system) delivers a position accuracy of a around 2cm and a heading accuracy of 1°. The position information is obtained relative to one vehicle that is defined to be the dedicated master. Finally, the LANCOM® unit is responsible for the data transmission.

Figure 1: Test vehicles.

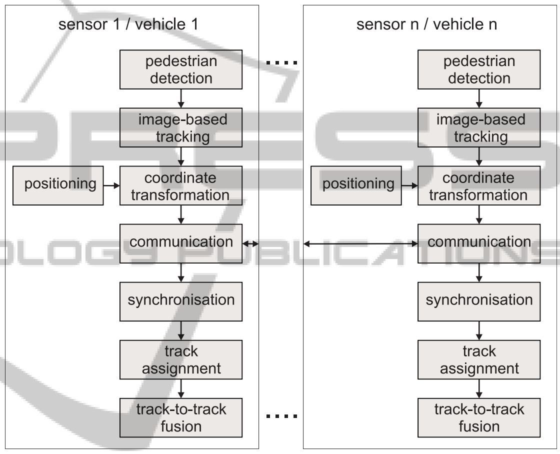


Figure 2: System overview: Pedestrian detection and fusion.

Fig. 2 illustrates the system that is used for multi-camera pedestrian detection and fusion. Firstly, the image is scanned via an AdaBoost detection algorithm. The pedestrian detection is based on the system presented in (Nunn et al., 2009). The detection results are then tracked using a Kalman filter that is working in image coordinates. The tracked detections are then transformed to appropriate coordinates that can be used for the fusion. These transformed detections, including their uncertainties, are then transmitted to the other vehicle as well as the vehicle position. The data is then synchronized using the GPS timestamps and passed to the track-assignment module. Corresponding tracks are finally fused by means of Track-To-Track fusion algorithms.

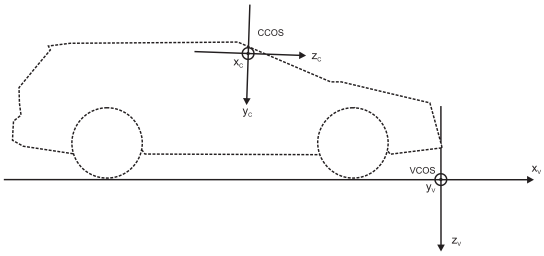


Figure 3: Definition of the VCOS and the CCOS.

3 COORDINATE TRANSFORMATIONS

The overall task of the coordinate transformation is to transform the detection results to a coordinate frame that can be used by both vehicles. Here it is focused on the positioning of the vehicles with the fixed map, since this map, including pedestrian ground truth data, is used for the evaluation in section 5. In general four coordinate systems are used:

1. Image Coordinate System (ICOS) $\mathbf{r} = (x, y, 1)^T$,
2. Camera Coordinate System (CCOS) $\mathbf{r}_C = (x_C, y_C, z_C, 1)^T$,
3. Vehicle Coordinate System (VCOS) $\mathbf{r}_V = (x_V, y_V, z_V, 1)^T$,
4. World Coordinate System (WCOS) $\mathbf{r}_W = (x_W, y_W, z_W, 1)^T$.

The position vectors are notated in homogeneous coordinates.

The definition of the VCOS and the CCOS is shown in Fig. 3. The transformation from the VCOS to CCOS can be expressed by the 4×4 external calibration matrix $\mathbf{K}_e$, which encodes the position of the camera $\mathbf{T}$ and the camera rotation $\mathbf{R}$

$$\mathbf{K}_e = \mathbf{VRT}.$$

The matrix $\mathbf{V}$ simply swaps the coordinates according to Fig. 3.

The transformation from CCOS to the ICOS is implemented with a pinhole camera model (Hartley and Zisserman, 2003). Thus, the effective focal length f_x in x- and f_y in y-direction are needed and the position of the principal point $\mathbf{p} = (c_x, c_y)$. The mapping is described by a 3×4 internal calibration matrix

$$\mathbf{K}_i = \begin{bmatrix} f_x & 0 & c_x & 0 \\ 0 & f_y & c_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

The combination of the internal and external calibration matrices enables the transformation from the VCOS to the ICOS by

$$s\mathbf{r} = \mathbf{K}_i \mathbf{K}_e \mathbf{r}_V. \quad (1)$$

For the WCOS that is uses in conjunction with the map, it is assumed that the map (WCOS) and the vehicle (VCOS) are located on the same plane. Therefore a point $\mathbf{r}_W = (x_W, y_W, 0, 1)$ can be mapped to $\mathbf{r}_V = (x_V, y_V, 0, 1)$ by

$$s\mathbf{r}_V = \mathbf{RTr}_W,$$

where $\mathbf{T}$ encodes the position of the vehicle in the WCOS and $\mathbf{R}$ describes the vehicle orientation.

3.1 Mapping Pedestrian Detections to the Vehicle Coordinate System

Equation 1 can be used to create the inverse function for mapping image to vehicle coordinates. The problem is that the depth information is lost and some assumptions have to be made. Firstly, it can be assumed that the detected objects are located at the same plane as the vehicle $z_W = 0$ (flat ground plane assumption (Ponsa et al., 2005)). This assumption holds for standard scenarios like highways and cross-ways. The problem is that pitching of the vehicle, due to uneven road, causes significant errors in the far distance. Another approach is to assume a standard object width in vehicle or world coordinates (Ponsa and Lopez, 2007). This approach can handle the pitching effects very well, but has a constant offset related to the standard width assumption. The mapping of a single image point $\mathbf{r}$ with the two methods is denoted by

$$\mathbf{r}_V = f_1(\mathbf{r}, \Phi),$$
$$\mathbf{r}_V = f_2(\mathbf{r}, w, w_V),$$

where Φ is the pitch angle and w is the object width in image coordinates and w_V is the fixed width assumption in vehicle coordinates. It is obvious that the first approach is very accurate in the near distance and the second approach is more accurate in the far distance. Thus, a combination of both approaches is used to calculate the pedestrian positions in the VCOS.

The tracked pedestrian detections are described by their image positions $\mathbf{r} = (x, y, 1)$ and the width w of the bounding box. In addition, the tracker delivers the uncertainties of the position and the width by means of a covariance matrix $\mathbf{C}$. Furthermore a fixed variance of the pitch angle Φ is assumed.

Using the detection result and the information about the uncertainties, the detections are then treated as two 3-D normal distributions $N(\boldsymbol{\mu}_1, \mathbf{C}_1)$ and $N(\boldsymbol{\mu}_2, \mathbf{C}_2)$, where $\boldsymbol{\mu}_1 = (x, y, \Phi)$ describes the object position and the pitch angle and $\boldsymbol{\mu}_2 = (x, y, w)$ describes the object position and the object width. $\mathbf{C}_1$ and $\mathbf{C}_2$ are the according covariance matrices.

Since $f_1(\mathbf{r}, \Phi)$ and $f_2(\mathbf{r}, w, w_V)$ are non-linear, it is proposed to use the scaled unscented transformation (SUT) (Merwe and Wan, 2003) to map the distribution from image to vehicle coordinates. For applying the SUT a set of deterministic sigma points is chosen and these points are then propagated using the non-linear function. The sigma points $\mathbf{x}_i$ and the weight values w_i^m and w_i^c are chosen according to (Merwe and Wan, 2003)

$$
\begin{aligned}
\mathbf{x}_0 &= \boldsymbol{\mu} \\
\mathbf{x}_i &= \boldsymbol{\mu} + \left(\sqrt{(L+\lambda)\,\mathbf{C}}\right)_i;\ \text{for } i = 1,..,L \\
\mathbf{x}_i &= \boldsymbol{\mu} - \left(\sqrt{(L+\lambda)\,\mathbf{C}}\right)_{i-L};\ \text{for } i = L+1,..,2L \\
w_i^m &= \frac{\lambda}{L+\lambda};\ \text{for } i = 0 \\
w_i^c &= \frac{\lambda}{L+\lambda} + \left(1-\alpha^2+\beta\right);\ \text{for } i = 0 \\
w_i^c &= w_i^m = \frac{1}{2(L+\lambda)};\ \text{for } i > 0,
\end{aligned}
$$

where L is the dimension (here $L = 3$) and $\lambda = \alpha^2(L+\kappa) - L$ is a scaling parameter. Moreover α defines the spread of the sigma points ($1e-2 \le \alpha \le 1$). κ is another scaling parameter that is usually set to 0 or $L-3$. Finally, β can be used to incorporate knowledge of the distribution ($\beta = 2$ for Gaussian distribution). $\left(\sqrt{(L+\lambda)\,\mathbf{C}}\right)_i$ is the i-th column of the matrix square root of the covariance.

After the determination of the sigma points, they are propagated using the non-linear functions f_1 and f_2

$$
\mathbf{y}_i = f(\mathbf{x}_i).
$$

The weight values and the sigma point can now be used to recover the statistics of the distribution after the non-linear transformation by means of

$$
\begin{aligned}
\tilde{\boldsymbol{\mu}} &= \sum_{i=0}^{2L} w_i^m \mathbf{y}_i \\
\tilde{\mathbf{C}} &= \sum_{i=0}^{2L} w_i^c (\mathbf{y}_i - \tilde{\boldsymbol{\mu}})(\mathbf{y}_i - \tilde{\boldsymbol{\mu}})^T.
\end{aligned}
$$

For both functions f_1 and f_2 an estimate of the position in the VCOS is obtained. Since it is known, that the first one is better in the near distance d and the latter is superior in the far distance, this information is taken into account by using a weighting function.

The weighting function involves a scaled sigmoid function and is 0 for $d \le d_{min}$ and 1 for $d \ge d_{max}$. For distance values $d_{min} < d < d_{max}$ the sigmoid function is used and scaled to values between 0 and 1. After the weight value for each estimate is determined, the covariance intersection equations (section 4) are used to determine the final result. Instead of optimizing $\omega = \arg\min[\det(\mathbf{C})]$, the described weight value is used. The approach can be summarized by

1. Describe the detections by two 3D vectors $\boldsymbol{\mu}_1 = (x, y, \Phi)$, $\boldsymbol{\mu}_2 = (x, y, w)$ and two 3×3 covariance matrices $\mathbf{C}_1$ and $\mathbf{C}_2$.
2. Calculate scaled sigma points for both distributions.
3. Propagate sigma points using f_1 and f_2.
4. Recover statistics $\tilde{\boldsymbol{\mu}}_1$, $\tilde{\boldsymbol{\mu}}_2$, $\tilde{\mathbf{C}}_1$, and $\tilde{\mathbf{C}}_2$.
5. Determine weight value ω, based on the distance of both estimates using the weighting function.
6. Fuse the results using covariance intersection with fixed weight value ω.

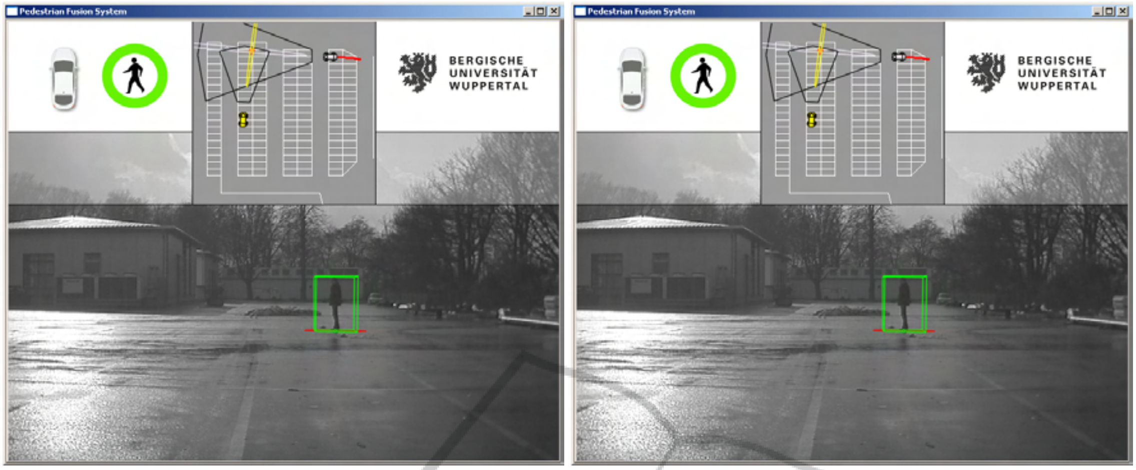
4 FUSION OF PEDESTRIAN DETECTIONS

Each vehicle sends the tracked detection results to the other vehicle and receives the tracked detection results of the opponent. The detections are described by their position and their uncertainties in the WCOS. The state estimates of each track are denoted by $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, whereas the fused state is $\boldsymbol{\mu}$. The according covariance matrices are denoted by $\mathbf{C}_1$, $\mathbf{C}_2$, and $\mathbf{C}$. The fusion is subdivided into two components, namely the track assignment and the Track-To-Track fusion.

4.1 Track-to-Track Fusion Algorithms

Once the track assignment is completed, the tracks are subject to fusion. As aforementioned the choice of the fusion method depends on the data at hand and to what extent the data or sensors are correlated. Therefore, three well known methods are analyzed. A comparison of fusion methods on simulated data is given in (Matzka and Altendorfer, 2008), a survey is given in (Smith and Singh, 2006) and a general treatment on data fusion can be found in (Bar-Shalom and Blair, 2000).

The first method for fusion is based on the Kalman equations and is presented in (Smith and Cheeseman, 1986). In this work a different notation is used, that is described in (Bar-Shalom and Blair, 2000), but that result is equivalent to (Smith and Cheeseman, 1986). For the evaluation this method it is denoted by *Covariance Fusion* (CF) for uncorrelated data, which



(a) View from vehicle 1.

(b) View from vehicle 2.

Figure 4: Results of pedestrian detection fusion from two vehicles. Fig. 4(a) and 4(b) are recorded at the same timestamp. The ego vehicle is gray and the remote vehicle is yellow. The trajectory of the vehicles is denoted by red dots.

can be defined by the following equations

$$\begin{aligned} \mathbf{C} &= \mathbf{C}_1 (\mathbf{C}_1 + \mathbf{C}_2)^{-1} \mathbf{C}_2 \\ \boldsymbol{\mu} &= \mathbf{C}_2 (\mathbf{C}_1 + \mathbf{C}_2)^{-1} \boldsymbol{\mu}_1 + \mathbf{C}_1 (\mathbf{C}_1 + \mathbf{C}_2)^{-1} \boldsymbol{\mu}_2. \end{aligned}$$

The second method is the *Covariance Intersection* (CI) (Matzka and Altendorfer, 2008). This method is known to work well in situations where signals are correlated, but the correlation is unknown. The CI can be implemented using

$$\begin{aligned} \mathbf{C}^{-1} &= \omega \mathbf{C}_1^{-1} + (1-\omega) \mathbf{C}_2^{-1} \\ \boldsymbol{\mu} &= \mathbf{C} \left[\omega \mathbf{C}_1^{-1} \boldsymbol{\mu}_1 + (1-\omega) \mathbf{C}_2^{-1} \boldsymbol{\mu}_2 \right] \\ \omega &= \arg\min \left[\det(\mathbf{C}) \right]. \end{aligned}$$

ω defines the influence of each estimate and is determined by an optimization procedure that minimizes e.g. $\det(\mathbf{C})$. Consistent estimates are guaranteed for $\omega \in [0,1]$.

The third method is named *Covariance Union* (CU) (Matzka and Altendorfer, 2008) and is defined by

$$\begin{aligned} \tilde{\mathbf{C}}_1 &= \mathbf{C}_1 + (\boldsymbol{\mu} - \boldsymbol{\mu}_1)(\boldsymbol{\mu} - \boldsymbol{\mu}_1)^T \\ \tilde{\mathbf{C}}_2 &= \mathbf{C}_2 + (\boldsymbol{\mu} - \boldsymbol{\mu}_2)(\boldsymbol{\mu} - \boldsymbol{\mu}_2)^T \\ \mathbf{C} &= \max\left(\tilde{\mathbf{C}}_1, \tilde{\mathbf{C}}_2\right) \\ \boldsymbol{\mu} &= \arg\min \left[\det(\mathbf{C}) \right]. \end{aligned}$$

Just like for CI, an optimization procedure has to be applied for CU to determine $\boldsymbol{\mu}$. The advantage of the CU is that this method is able to resolve statistically inconsistent states. This problem is faced by determining a new state estimate that can exceed the covariance indicated by at least one track (Matzka and Altendorfer, 2008).

5 EVALUATION

The evaluation of the three fusion methods is performed on real data (see Fig. 4). The positioning of the vehicles is performed by means of image registration in conjunction with an environment map. This map is used to determine the starting position and orientation of the vehicles. The vehicle movement is then calculated using the IMU and the motion model presented in (Meuter et al., 2008).This setup is used for evaluation since it is easy to generate ground truth data by using the map. The inaccuracies due to the movement of the vehicle are negligible compared to the errors induced by the object distance calculation from a single camera.

The system is tested on various video-sequences and different scenarios. The used scenarios are based on typical cross-way situations. In the first scenario both vehicles are approaching the object with an angle difference of 90° and in the second scenario the vehicles are placed at an angle difference of 180°. The distance of the vehicles to the objects at the starting position goes up to around $50m$.

To demonstrate the advantage of the fusion, firstly the detection results of the individual vehicles are evaluated. The results of the RMSE are shown in Table 1. As expected the lateral position of the objects can be measured precisely, whereas the depth information is inaccurate.

The results in Table 2 reveal that the fusion dramatically improves the overall precision. Whatever fusion method is used, the RMSE of d_w (distance of ground truth object and prediction) gets improved. For the first scenario the CI performs best and for the

Table 1: RMSE: single vehicles.

scenario	x_w [m]	y_w [m]	d_w [m]
scenario 1	0.68	4.18	4.28
scenario 2	3.72	0.98	4.05
overall performance	1.89	2.90	4.19

Table 2: RMSE: fusion of two vehicles.

Method	x_w [m]	y_w [m]	d_w [m]
scenario 1			
CF	0.35	0.55	0.69
CI	0.35	0.55	0.69
CU	2.87	2.67	3.94
scenario 2			
CF	0.15	1.00	1.01
CI	0.14	1.07	1.08
CU	0.15	3.72	3.73
overall performance on both scenarios			
CF	0.25	0.78	0.85
CI	0.24	0.81	0.89
CU	1.51	3.19	3.83

second scenario the CF outperforms the other algorithms. In contrast to the results presented in (Matzka and Altendorfer, 2008) the CU has the worst performance in all scenarios. Based on these results it is proposed to use the CF as a general fusion method, since the algorithm delivers precise results in all scenarios. One could imagine a combination of CF and CI to get the best results in all situations. It is not surprising that the CF and CI deliver similar good results as long as the position vectors $\mathbf{r}_w$ of the detections are relatively accurate. For each vehicle the lateral positions of the detections are very accurate. Thus one could get a fused result of two cameras by calculating the intersection of the two rays on which the detections are located. This is similar to stereo vision. The fused position vector of CF and CI is close to the result that would be obtained by this ray intersection. The main difference of CI and CU is determined by the fused covariance.

This leads to the scenario where CU could outperform CF and CI. Inconsistent states are obtained if lateral position errors occur due to erroneous data of the detection algorithm or an erroneous vehicle position. These states can then be handled by a CU algorithm. It would make sense to include an algorithm that can detect where inconsistent states occur and then change the fusion algorithm to CU.

ACKNOWLEDGEMENTS

The developed system is part of the Active Safety Car project which is co-funded by the European Union and the federal state NRW.

REFERENCES

Bar-Shalom, Y. and Blair, W. D. (2000). *Multitarget-Multisensor Tracking: Applications and Advances*. Artech House Inc, Norwood, USA.

Hartley, R. and Zisserman, A. (2003). *Multi View Geometry in Computer Vision (2nd Edition)*. Cambridge University Press, Cambridge, United Kingdom.

Kaempchen, N. (2007). Feature-level fusion of laser scanner and video data for advanced driver assistance systems. Technical report, Fakultaet fuer Ingenieurwissenschaften und Informatik, Universitaet Ulm.

Matzka, S. and Altendorfer, R. (2008). A comparison of track-to-track fusion algorithms for automotive sensor fusion. In *Proc. of International Conference on Multisensor and Integration for Intelligent Systems , 2008 IEEE*, pages 189–194.

Merwe, R. V. D. and Wan, E. (2003). Sigma-point kalman filters for probabilistic inference in dynamic state-space models. In *In Proceedings of the Workshop on Advances in Machine Learning*.

Meuter, M., Iurgel, U., Park, S.-B., and Kummert, A. (2008). The unscented kalman filter for pedestrian tracking from a moving host. In *Proc. of Intelligent Vehicles Symposium, 2008 IEEE*, pages 37–42.

Nunn, C., Kummert, A., Muller, D., Meuter, M., and Muller-Schneiders, S. (2009). An improved adaboost learning scheme using lda features for object recognition. In *Proc. of Intelligent Transportation Systems, 2009. ITSC '09. 12th International IEEE Conference on*, pages 1–6.

Ponsa, D. and Lopez, A. (2007). Vehicle trajectory estimation based on monocular vision. In Marti, J., Benedi, J., Mendonca, A., and Serrat, J., editors, *Pattern Recognition and Image Analysis*, volume 4477 of *Lecture Notes in Computer Science*, pages 587–594. Springer Berlin / Heidelberg.

Ponsa, D., Lopez, A., Lumbreras, F., Serrat, J., and Graf, T. (2005). 3d vehicle sensor based on monocular vision. In *Proceedings of the 8th International IEEE Conference on Intelligent Transportation Systems*, Vienna, Austria.

Smith, D. and Singh, S. (2006). Approaches to multisensor data fusion in target tracking: A survey. *Knowledge and Data Engineering, IEEE Transactions on*, 18(12):1696 –1710.

Smith, R. C. and Cheeseman, P. (1986). On the representation and estimation of spatial uncertainly. *Int. J. Rob. Res.*, 5(4):56–68.