Pedestrian Path Prediction with Recursive Bayesian Filters: A Comparative Study

N. Schneider[1,2] and D. M. Gavrila[1,2]

[1]Environment Perception, Daimler R&D, Ulm, Germany
[2]Intelligent Systems Laboratory, Univ. of Amsterdam, The Netherlands

Abstract. In the context of intelligent vehicles, we perform a comparative study on recursive Bayesian filters for pedestrian path prediction at short time horizons ($< 2s$). We consider Extended Kalman Filters (EKF) based on single dynamical models and Interacting Multiple Models (IMM) combining several such basic models (constant velocity/acceleration/turn). These are applied to four typical pedestrian motion types (crossing, stopping, bending in, starting). Position measurements are provided by an external state-of-the-art stereo vision-based pedestrian detector. We investigate the accuracy of position estimation and path prediction, and the benefit of the IMMs vs. the simpler single dynamical models. Special care is given to the proper sensor modeling and parameter optimization. The dataset and evaluation framework are made public to facilitate benchmarking.

1 Introduction

Pedestrian path prediction is an important problem in several application contexts, such as architecture, social robotics and intelligent vehicles. Here we consider the intelligent vehicle context, in view of driver assistance and active pedestrian safety. Strong gains have been made over the years in improving computer vision-based pedestrian recognition performance. This has culminated in first active pedestrian safety systems reaching the market. For example, Mercedes-Benz introduces in its 2013 E- and S-Class models a novel stereo-vision based pedestrian system, which incorporates automatic full emergency braking.

A sophisticated situation assessment requires a precise estimation of the current and future position of the pedestrian with respect to the moving vehicle. A deviation of, say, 30 cm in the estimated lateral position of the pedestrian can make all the difference, as this might place the pedestrian just inside or outside the driving corridor. Current active pedestrian systems are typically designed conservatively in their warning and control strategy, emphasizing the current state rather than prediction, in order to avoid false system activations. Indeed, pedestrian path prediction is a challenging problem, due to the highly dynamic behavior of pedestrians. They can change their walking direction in an instance, or start/stop walking abruptly. As a consequence, sensible prediction horizons are typical short (we consider $< 2s$ in this paper).

Fig. 1: Four typical pedestrian motion types: bending in (top left), stopping (top right), crossing (bottom left) and starting (bottom right) with detection bounding boxes.

There has been surprisingly little analysis in previous work of the accuracy of pedestrian state estimation, let alone, that of prediction, in vehicle context. This paper addresses this by providing a quantitative comparative study of recursive Bayesian filters: we consider Extended Kalman Filters (EKF) based on single dynamical models and Interacting Multiple Models (IMM) combining several such basic models (constant velocity/acceleration/turn). These are applied to four typical pedestrian motion types (crossing, stopping, bending in, starting), see Fig. 1. Position measurements are provided by an external state-of-the-art stereo vision-based pedestrian detector. The rationale for focusing on recursive Bayesian filters in connection with modeling pedestrians as point targets is their relatively good performance and low computational cost (especially important in a vehicle context). We investigate the accuracy of position estimation and path prediction, and the benefit of the IMMs vs. the simpler single dynamical models. Special care is given to the proper sensor modeling and parameter optimization.

2 Previous Work

In this section, we focus on pedestrian state estimation based on parametric, recursive Bayesian filters. For an overview of vision-based pedestrian detection and tracking in more general context, see recent surveys (e.g. [7, 8]).

A popular choice for target state estimation is the Kalman Filter (KF). Its applicability in real-time systems has been proven over many years for different sensors and application domains [1, 3, 4, 18, 21]. State parameters (e.g. position, velocity, acceleration) of the tracked target can be estimated with appropriate dynamical and measurement models. The KF can further be used for prediction by propagating the current state with the dynamical model without the inclusion of new measurements. Work by [3] on FIR-based pedestrian tracking uses a constant acceleration (CA) model in image space. Working in image space, however, makes it difficult to incorporate prior knowledge on the dynamics of pedestrian motion. Therefore, [2] track pedestrians on the ground plane using a KF in an indoor, static stereo camera setup. The use of a linear KF in the

context of video-based pedestrian tracking in the world implies the use of 3D pseudo-measurements (i.e. back projection of 2D measurements); this does not account for the dependency of the longitudinal component of the noise on depth.

More accurate measurement models for the perspective projection of video sensors can be incorporated by means of non-linear Extended (EKF) or Unscented (UKF) Kalman filters. [15] use a UKF in a mono camera setup to track pedestrians on the ground plane (CV model). [19] apply the UKF to measurements from a stereo camera system comparing three different dynamical models (two CV and one constant position (CP) model) where two models have a state space in world coordinates and one in image coordinates.

KF-based approaches have also been used for pedestrian state estimation outside the video-only domain. [9] apply a CV model in a multi-sensor setup with an IR camera and laser scanner. In a previous paper [18], they used two different motion models (CA and CTRV), mentioning advantages of the latter model at near-zero pedestrian speeds. Work by [21] considers a setting where pedestrians wear electronic tags. It uses a KF with a turn motion model including orientation and velocity in polar coordinates (CTRV).

Maneuvering targets can be elegantly accounted for mathematically by means of the Interacting Multiple Model (IMM) framework [4, 13]. [11] use an IMM (CP/CV) for analyzing walking vs. stopping pedestrian motion types from a stereo vision sensor on-board a vehicle. [5] use an IMM combining eight CV models with fixed velocities in eight directions. It further contains an online adaption algorithm for the IMM transition probability matrix.

Within the class of non-parametric methods for pedestrian path prediction and action classification, [11] proposes a probabilistic trajectory matching method to estimate whether a pedestrian walking towards the curbside intends to cross or not, when viewed from a stereo vision system on-board a vehicle. [12] considers the complementary case of whether a pedestrian standing will start to walk using a SVM-based classification approach, albeit from a static monocular camera.

Quantitative evaluations of pedestrian state (position) estimation have been few and limited. [3, 5, 9, 15, 18, 21] do not include any such evaluation. [2] provides accuracy figures only related to its KF approach in indoor setting. [19] uses simulated data to compare CV and CP KFs. Our paper contribution is a broad quantitative study on pedestrian position estimation and path prediction using parametric Bayesian recursive filters in vehicle context. Compared to [11], we consider a wider range of pedestrian motion types. Whereas the IMM used by [11] uses 3D pseudo measurements and KFs, we use a more accurate stereo sensor modeling by EKFs.

3 Recursive Bayesian Filtering

3.1 Kalman Filter

The discrete-time KF estimates a state $\mathbf{x}(t)$ at time step t from measurement $\mathbf{z}(t)$ and previous state $\mathbf{x}(t-1)$ with the dynamical model

$$\mathbf{x}(t) = A\mathbf{x}(t-1) + Bu(t-1) + \boldsymbol{\omega}(t-1) \tag{1}$$

where the relation between measurement and state is given by

$$\mathbf{z}(t) = H\mathbf{x}(t) + \boldsymbol{\nu}(t). \tag{2}$$

A and B are transition matrices for the state $\mathbf{x}$ and the control input u, respectively, $\boldsymbol{\omega}(t-1)$ and $\boldsymbol{\nu}(t)$ are white, zero-mean, uncorrelated noise processes with covariances $\boldsymbol{\omega}(t) \sim (0, Q(t))$ and $\boldsymbol{\nu}(t) \sim (0, R(t))$. The filter process can be described as cycle of the two steps prediction (predicting the state $\mathbf{x}(t-1)$ to the next time step) and correction (updating the predicted state $\hat{\mathbf{x}}(t)$ with the current measurement) [20].

3.2 Interacting Multiple Model Kalman Filter

There are several KF extensions available to cover different motion types and maneuvers (see [13] for an overview), the most common is the Interacting Multiple Model KF (IMM). The IMM models that there is a probability of p_{ij} that the tracking target makes a transition from one type of motion (i) to another (j); these values are captured by the transition probability matrix (TPM). Each iteration of the IMM consists of the three steps: interaction, filtering and model probability update [4]. In the interaction step, the mixing probability $\boldsymbol{\mu}_{ij}(t-1)$ (cond. probability that the target changed its type of motion) is calculated based on model probabilities and the TPM to produce mixed state estimates $\hat{\mathbf{x}}_j^0(t-1)$ and covariances $\hat{P}_j^0(t-1)$ for all models j. The mixed states are used as input in the filtering step where each model is predicted and updated with the standard KF equations. In the last step, the model probabilities are updated based on the measurement likelihood.

3.3 Measurement Model

Measurements come from a pedestrian detector applied on sequences recorded with a stereo camera system. A measurement vector (dropping time index t in the following) $\mathbf{z} = (u, d)$ is derived from the footpoint $p_f = (u, v)$ and the median disparity d of a pedestrian bounding box. The relation of a point in the image $p_i = (u, v)$ and its disparity d to a point $p_c = (x_c, y_c, z_c)$ in the camera coordinate system is given by the perspective camera model [1]:

$$\begin{pmatrix} u \\ v \\ d \end{pmatrix} = \begin{pmatrix} h_1(p_c) \\ h_2(p_c) \\ h_3(p_c) \end{pmatrix} = \begin{pmatrix} u_0 + \frac{f_u x_c}{z_c} \\ v_0 + \frac{-f_v y_c}{z_c} \\ \frac{f_u b}{z_c} \end{pmatrix} \tag{3}$$

where $f_u = \frac{f}{s_u}$ and $f_v = \frac{f}{s_v}$ with focal length f, baseline b, horizontal and vertical pixel width s_u and s_v, respectively. Eq. (3) leads to the nonlinear measurement function h. For a position $p_c^g = (x_c, z_c)$ on the groundplane h_2 can be ignored. To predict a measurement at time step t, the predicted state vector $\hat{\mathbf{x}}$ (camera coordinates) has to be projected into the measurement (image) space with $\hat{\mathbf{z}} = h(\hat{\mathbf{x}})$. For the EKF we further need to calculate the Jacobian $H = \frac{\partial h}{\partial \mathbf{x}}$.

Table 1: Mean sojourn times of different target dynamics in the training set (diagonals $P_{i,i}$ of the TPM based on a cycle time of $T \approx 60\,ms$). "Straight Walking" consists of the straight walking segments of starting, stopping and bending in sequences as well as complete crossing sequences. "Maneuver" relates to all other segments. "Turning", a subset of "Maneuver", relates to the turning segments within the bending in sequences.

Motion type	**mean sojourn time τ_i (s)**	$P_{ii}(T) = 1 - T/\tau_i$
Straight Walking	6.66	0.99
Maneuver	1.67	0.96
Turning	2.50	0.98

3.4 Dynamical Models

Several discretized continuous-time dynamical models are considered in this study: the popular constant velocity (white noise acceleration) model (CV), the constant acceleration (Wiener process acceleration) model (CA) and the constant turn model (CT) with Cartesian state vector. These characterized by their state vectors $\mathbf{x}$, transition matrices A and process noise matrices Q. The CV model state vector holds position and velocity ($\mathbf{x} = [x, z, v_x, v_z]$), the CA model further has acceleration ($\mathbf{x} = [x, z, v_x, v_z, a_x, a_z]$) and the CT model turn rate ($\mathbf{x} = [x, z, v_x, v_z, \omega]$) variables. For details, such as transition and process noise matrices, see [4, 14].

Several approaches can be taken to specify the TPM. There is the ad-hoc approach to fill the diagonals with values close to one. [4, 13] discuss the use of the mean sojourn time (the mean time a target stays in a motion type) for the TPM. Lastly, one could perform parameter optimization of the entries of the TPM directly. In preliminary experiments, we obtained similar best performance with the second and third approaches, thus we selected the sojourn time approach to specify the TPM, derived from a training set, see Section 4 and Table 1.

3.5 Ego Motion Compensation

At each time step, the filter state is projected from the previous camera coordinate system to the current one using the inertial motion matrix M_v (vehicle coordinates) based on velocity and yaw rate measured by on-board sensors. The inverse ego motion homography matrix is given by $M_c = D^{-1} M_v D$ (where D defines the relation between camera and vehicle coordinate system). Translational ego compensation is done using t_{M_c} as control vector u (Eq. (1) with $B = I_{2x2}$), the ego rotation is integrated into the transition matrix A_e (exemplary for the CV model) [16]:

$$A_e = \begin{bmatrix} R_{M_c} & 0_{2x2} \\ 0_{2x2} & R_{M_c} \end{bmatrix} A \tag{4}$$

Table 2: Sequences in our dataset recorded with standing and moving vehicle.

Sequences	**veh. stand.**	**veh. mov.**	**total**
Bending in	5	18	23
Stopping	5	13	18
Starting	0	9	9
Crossing	3	15	18

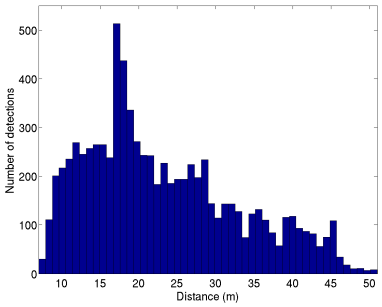

Fig. 2: Pedestrian distance distribution over entire dataset.

4 Experiments[1]

Dataset. Image sequences were recorded with a stereo camera system (baseline $22\,cm$, $16\,fps$, 1176×640 pixels) mounted behind the windshield of a vehicle. They contain four typical pedestrian motion types: (walking laterally towards the street and) crossing, (walking laterally towards the street and) stopping, (standing at the curbside and) starting (to walk laterally) and (walking alongside the street,) bending in (and crossing). See Fig. 1. The dataset consists of 68 sequences of which 12485 images contain (single) pedestrians. 55 sequences were recorded at vehicle speeds of $20-30\,km/h$, the others involved a standing vehicle. See Table 2 and Figs. 2 and 6 (bottom left) for further data statistics. The dataset is splitted evenly in a training and test set. The latter was splitted 5-folds for evaluation in Fig.6 (bottom right); parameter optimization was quite time consuming, so we preferred not vary the training set as part of cross-validation.

Ground Truth (GT) is obtained by manual labeling of pedestrian bounding boxes and computating the median disparity over the rough upper pedestrian body area. The position of the pedestrian in the vehicle coordinate system is calculated with Eq. (3) and the camera-to-vehicle homography matrix. The transformed positions are fitted with a curvilinear model. The GT locations are obtained by longitudinal projections on the fitted curve. Sequences are further labeled with event tags and time-to-event (TTE in frames) values. For stopping

[1] The dataset and evaluation framework are made public for non-commercial research purposes. Follow the links from http://isla.science.uva.nl/ or contact the 2nd author.

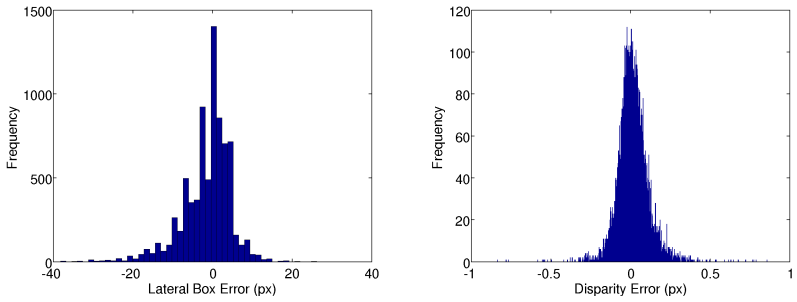

Fig. 3: Measurement error distribution: (left) lateral, u (right) longitudinal, d.

Table 3: Optimized process noise parameters (σ_v, σ_a, $\{\sigma_v, \sigma_\omega\}$) for the different filters.

Filter	EKF CV	EKF CA	EKF CT	IMM(CV,CA)	IMM(CV,CT*)
Process noise	0.77	0.44	$\{0.95, 0.90\}$	$(0.70, 0.80)$	$(0.75, \{0.40, 0.90\})$

pedestrians the last placement of the foot on the ground at the curbside is labeled as TTE = 0. For crossing pedestrians, the closest point to the curbside (before entering the roadway), for pedestrians bending in and starting to walk the first moment of visually recognizable body turning or leg movements are labeled with TTE = 0. All frames previous to an event have TTE values > 0, therefore all frames following the event have TTE values < 0.

A state-of-the-art HOG/linSVM pedestrian detector [6] provides measurements, given region-of-interests supplied by an obstacle detection component using dense stereo data [10]. The resulting bounding boxes are used to calculate a median disparity over the upper pedestrian body area based on the disparity maps. The measurement vector $\mathbf{z} = (u, d)$ is derived using the central lateral position of the bounding box and this median disparity value.

Evaluation Setup. Sequences have an average of 121 measurements (min. 39, max. 274), they start with a minimum of three consecutive measurements, and contain no missing detections longer than five consecutive frames. Evaluation is done with respect to the lateral localization error in TTE interval $[10, -50]$, corresponding to an evaluation interval from $0.60\,s$ before to $3.0\,s$ after the event. At each time step, predictions of up to 32 frames ($1.9\,s$) are made ahead. We use the predict and update functions of the EKF/UKF MATLAB toolbox [17].

Parameter Setting. The measurement noise $\boldsymbol{\nu}$ (see Eq. (2)) has been derived statistically on the training set, in terms of lateral bounding box error σ_u and disparity error σ_d of the pedestrian detections w.r.t. the GT. See Fig. 3. The measurement noise matrix $R = diag(\sigma_u^2, \sigma_d^2)$ is thus set to $\sigma_u = 6.15$ and $\sigma_d = 0.32$.

Process noise $\boldsymbol{\omega}$ is determined by $Q(t) = Q^0(t)q$, where $q \in \{\sigma_v^2, \sigma_a^2\}$ and for the CT model, $Q_{5,5}(t) = \frac{\sigma_\omega^2 Q^0_{5,5}(t)}{q}$ [4]. It has been optimized for each filter

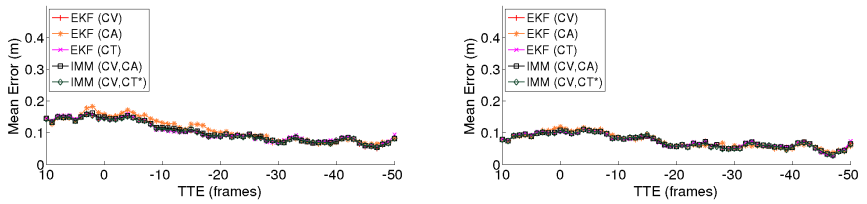

Fig. 4: Position error at current time ($t = 0$) averaged over all sequences: lateral and longitudinal combined (left) and only lateral (right).

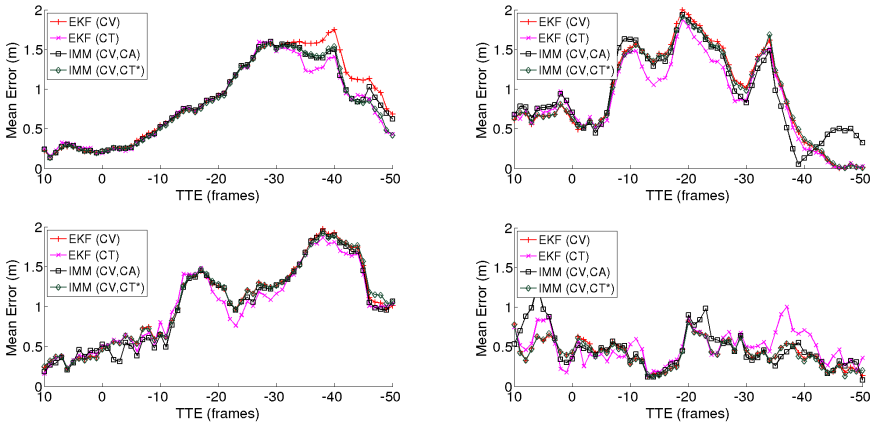

Fig. 5: Mean lateral position error when predicting 32 frames ($t = 1.9\,s$) ahead.

in terms of positions mean-squared-error (MSE) including all N position state estimates $\mathbf{x}_i(t = 0)$ $(i = 1, ..., N)$ and the corresponding P predictions $\mathbf{x}_i(t = 1, ..., P)$ $(P = 32)$ with the objective function:

$$\arg\min_{\sqrt{q}} \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{t=0}^{P} \frac{MSE(\mathbf{x}_i(t))}{P+1} \right) \tag{5}$$

using a discrete parameter search on the training set. The outcome is shown in Table 3. Search space has been discretized using 60 steps for single models with one noise parameter (CV, CA) and 18 steps each for models with two noise parameters (CT, 324 parameter combinations). In a coarse-to-fine fashion discretization for the IMM(CV,CA) could be reduced based on single model results to 9 steps (81 parameter combinations). Including the TPM with a discretization of 12 values per diagonal in the optimization results in 11664 parameter combinations. Furthermore, the CT model has been optimized using only segments of bending in sequences around the labeled turn event (TTE $\in [10\ -50]$). The resulting "turn expert" filter will be termed CT* in the remainder.

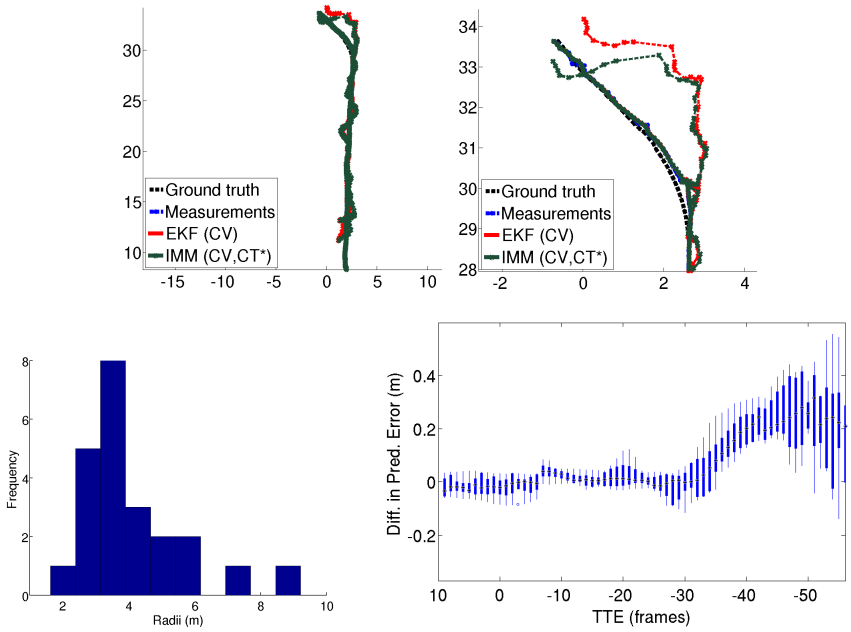

Fig. 6: Bending in: Bird's eye view of an example trajectory showing predictions (32 steps ahead) for various filters (top row, far view and close up, left and right). Turn radius distribution over the bending in sequences (bottom left). Distribution of the prediction error improvement by IMM(CV,CT*) vs. EKF(CV), based on cross-validation.

Results. Fig. 4 shows the position error at current time ($t = 0$), averaged over all sequences. One observes a similar performance for all filters. Performance differences become more evident when predicting 32 frames (1.9 s) ahead, see Fig. 5 (the CA model was removed from the plots since its predictions were far off, i.e. velocities accumulate over the prediction horizon to implausible values). In terms of the single motion models (CV vs. CT), one observes benefits for the CV on the crossing sequences, and benefits for the CT on the others. For example, CT predictions during the turning of the bending in sequence are more accurate by up to 36 cm, compared to CV. The IMM(CV,CT*) combines the best of both worlds, it shows an improvement of up to 30 cm vs. CV. One further observes that IMM(CV,CA) does not outperform CV and lags IMM(CV,CT*), overall. A more detailed analysis of the bending in case is given in Fig. 6.

5 Conclusions

In this paper, we studied several single dynamical models (CV, CA, CT) and IMMs combining such basic models for pedestrian position estimation and path prediction, in vehicle context. Results show no significant performance gain of the more sophisticated IMMs considered vs. the simpler CV, for current position estimation. We attribute this to the high sampling rate and the low measurement

error for this application. For path prediction ($1.9\,s$ ahead), an IMM(CV,CT*) involving a constant velocity and a "turn expert" model, is shown to provide an improvement in the lateral position estimation of up to $30\,cm$ during maneuvers. Future work involves extending the database, both in terms of motion types considered and in terms of their sample count.

References

1. Barth, A., Franke, U.: Estimating the driving state of oncoming vehicles from a moving platform using stereo vision. IEEE Trans. ITS 10(4), 560–571 (2009)
2. Bertozzi, M., et al.: Pedestrian localization and tracking system with Kalman filtering. In: IEEE Intell. Veh. pp. 584–589 (2004)
3. Binelli, E., et al.: A modular tracking system for far infrared pedestrian recognition. In: IEEE Intell. Veh. pp. 759–764 (2005)
4. Blackman, S., Popoli, R.: Design and Analysis of Modern Tracking Systems. Artech House Norwood, MA (1999)
5. Burlet, J., et al.: Pedestrian tracking in car parks: An adaptive Interacting Multiple Models based filtering method. In: Proc. of the IEEE ITSC. pp. 462–467 (2006)
6. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. CVPR. vol. 1, pp. 886–893. IEEE (2005)
7. Dollár, P., et al.: Pedestrian detection: An evaluation of the state of the art. IEEE PAMI 34(4), 743–761 (2012)
8. Enzweiler, M., Gavrila, D.: Monocular pedestrian detection: Survey and experiments. IEEE PAMI 31(12), 2179–2195 (2009)
9. Fardi, B., Scheunert, U., Wanielik, G.: Shape and motion-based pedestrian detection in IR images: a multisensor approach. In: IEEE Intell. Veh. pp. 18–23 (2005)
10. Hirschmüller, H.: Stereo processing by semiglobal matching and mutual information. IEEE PAMI 30(2), 328–341 (2008)
11. Keller, C.G., Hermes, C., Gavrila, D.M.: Will the pedestrian cross? Probabilistic path prediction based on learned motion features. Proc. DAGM pp. 386–395 (2011)
12. Köhler, S., et al.: Early detection of the pedestrians intention to cross the street. In: Proc. of the IEEE ITSC. pp. 1759–1764 (2012)
13. Li, X.R., Jilkov, V.P.: Survey of maneuvering target tracking. Part V. Multiple-model methods. IEEE Trans. Aerosp. Electron. Syst. 41(4), 1255–1321 (2005)
14. Li, X.R., Jilkov, V.: Survey of maneuvering target tracking. Part I. Dynamic models. IEEE Trans. Aerosp. Electron. Syst. 39(4), 1333–1364 (2003)
15. Meuter, M., et al.: Unscented Kalman filter for pedestrian tracking from a moving host. In: IEEE Intell. Veh. pp. 37–42 (2008)
16. Rabe, C.: Detection of Moving Objects by Spatio-Temporal Motion Analysis. Ph.D. thesis, University of Kiel, Kiel, Germany (2011)
17. Särkkä, S., Hartikainen, J., Solin, A.: EKF/UKF toolbox for Matlab v1.3 (2011), http://becs.aalto.fi/en/research/bayes/ekfukf/
18. Scheunert, U., et al.: Multi sensor based tracking of pedestrians: a survey of suitable movement models. In: IEEE Intell. Veh. pp. 774–778 (2004)
19. Tao, J., Klette, R.: Tracking of 2d or 3d irregular movement by a family of Unscented Kalman filters. JICCE 10(3), 307–314 (2012)
20. Welch, G., Bishop, G.: An introduction to the Kalman filter. In: Proc. of the ACM SIGGRAPH. ACM Press, Addison-Wesley, Los Angeles, CA, USA (2001)
21. Westhofen, D., et al.: Transponder- and camera-based advanced driver assistance system. In: IEEE Intell. Veh. pp. 293–298 (2012)