Context-Based Pedestrian Path Prediction*

Julian Francisco Pieter Kooij[1,2], Nicolas Schneider[1,2], Fabian Flohr[1,2], and Dariu M. Gavrila[1,2]

[1] Environment Perception, Daimler R&D, Ulm, Germany
{nicolas.schneider,fabian.flohr}@daimler.com
[2] Intelligent Systems Laboratory, Univ. of Amsterdam, The Netherlands
{J.F.P.Kooij,D.M.Gavrila}@uva.nl

Abstract. We present a novel Dynamic Bayesian Network for pedestrian path prediction in the intelligent vehicle domain. The model incorporates the pedestrian situational awareness, situation criticality and spatial layout of the environment as latent states on top of a Switching Linear Dynamical System (SLDS) to anticipate changes in the pedestrian dynamics. Using computer vision, situational awareness is assessed by the pedestrian head orientation, situation criticality by the distance between vehicle and pedestrian at the expected point of closest approach, and spatial layout by the distance of the pedestrian to the curbside. Our particular scenario is that of a crossing pedestrian, who might stop or continue walking at the curb. In experiments using stereo vision data obtained from a vehicle, we demonstrate that the proposed approach results in more accurate path prediction than only SLDS, at the relevant short time horizon ($1\,s$), and slightly outperforms a computationally more demanding state-of-the-art method.

Keywords: intelligent vehicles, path prediction, situational awareness, visual focus of attention, Dynamic Bayesian Network, Linear Dynamical System.

1 Introduction

The past decade has seen a significant progress on video-based pedestrian detection. In the intelligent vehicle domain, this has recently culminated in the market introduction of active pedestrian systems that can perform automatic braking in case of dangerous traffic situations. An area that holds major potential for further improvement is situation assessment. Current active pedestrian systems are designed conservatively in their warning and control strategy, emphasizing the current pedestrian state (i.e. position) rather than prediction, in order to avoid false system activations. Indeed, pedestrian path prediction is a challenging problem, due to the highly dynamic nature of pedestrian motion, and systems need to react with limited computation time. Small deviations 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.

* Electronic supplementary material -Supplementary material is available in the online version of this chapter at http://dx.doi.org/10.1007/978-3-319-10599-4_40 Videos can also be accessed at http://www.springerimages.com/videos/978-3-319-10598-7

D. Fleet et al. (Eds.): ECCV 2014, Part VI, LNCS 8694, pp. 618–633, 2014.

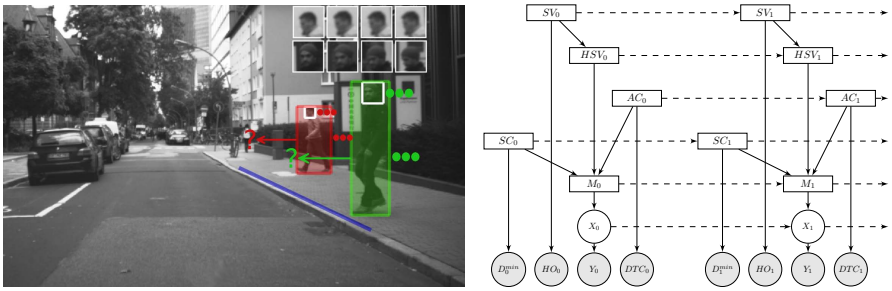

Fig. 1. Left: Pedestrian path prediction from an approaching vehicle, using situation criticality, pedestrian awareness thereof, and positioning vs. curbside. Right: DBN as directed graph, unrolled for two time slices. Discrete/continuous/observed nodes are rectangular/circular/shaded.

This paper focuses on the accurate path prediction of pedestrians intending to laterally cross the street, as observed by a stereo camera on-board an approaching vehicle (accident analysis shows that this scenario accounts for a majority of all pedestrian fatalities in traffic [23]). We argue that the pedestrian's decision to stop is for a large degree influenced by three factors: the existence of an approaching vehicle on collision course, the pedestrian's awareness thereof, and the spatial layout of the environment. We therefore propose a Dynamic Bayesian Network (DBN), which captures these factors as latent states on top of a Switching Linear Dynamical System (SLDS), thus controlling changes in the pedestrian dynamics. We estimate situation criticality by the distance between vehicle and pedestrian at the expected point of closest approach. Situational awareness assesses whether the pedestrian has seen the vehicle at some point up to now (whether the pedestrian currently sees the vehicle is estimated by means of the head orientation). Spatial layout is captured by the distance of the pedestrian to the road curbside. See Fig. 1 for an illustration of the scenario. The observables (shaded nodes in the graphical model), i.e. distance at closest approach, pedestrian location, head orientation, curbside location, are provided by external, state-of-the-art system components, for which we do not make novelty claims.

All DBN parameters are estimated from annotated training data. In the experiments, we collected data of pedestrians crossing in a supervised setting in traffic situations, where the vehicle has an implicit right-of-way. It would be straightforward to apply the approach to traffic situations where traffic lights or pedestrian crossings change the right-of-way, by adding an (observed) context variable to the DBN. Our approach can also be extended to additional motion types (e.g. pedestrian crossing the road in a curved path) or, more generally, to robot navigation in human-inhabited environments.

2 Previous Work

In this section, we focus on techniques for pedestrian state estimation and path prediction. For vision-based pedestrian detection, see recent surveys e.g. [10,12]. For pedestrian head/body orientation estimation, see e.g. [5,13,14].

State estimation in dynamical systems often involves the assumption that the underlying model is linear and that the noise is Gaussian, mainly due to the availability of the

Kalman filter (KF) [7] as an efficient inference algorithm for such Linear Dynamical Systems (LDS). In the intelligent vehicle domain, the KF is the most popular choice for pedestrian tracking (see [30] for an overview). The state distribution of a LDS can be propagated into the future without incorporating new observations to account for missing measurements, or to perform path prediction. The Extended and Unscented KF [24] can, to a certain degree, account for non-linear dynamical or measurement models, but Switching LDS (SLDS) are needed for maneuvering targets that alternate various motion types. A SLDS uses a top-level discrete Markov chain to select per time step the system dynamics of the underlying LDS. However, exact inference and learning becomes intractable as the number of modes in the posterior distribution grows exponential over time in the number of the switching states [27]. One solution is to approximate the posterior by samples using some Markov Chain Monte Carlo method [26,29]. Sampling can also be used when extending the SLDS hierarchy, e.g. to impose distributions on persistent state durations [26], or learn an SLDS mixture to cluster trajectories which exhibit similar switching behavior [21]. However, sampling is impractical for online real-time inference as convergence can be slow. Another solution is Assumed Density Filtering (ADF) [6,25], which approximates the posterior at every time step with a simpler distribution. ADF can be applied to discrete state DBNs, known as Boyen-Koller inference [8], and more generally to mixed discrete-continuous state spaces with conditional Gaussian posterior [22]. Interacting Multiple Model KF [7] is related to ADF for SLDS, as it mixes the states of several KF filters running in parallel, and has been applied for path prediction in the intelligent vehicle domain [18,30].

Whereas SLDSs can account for changes in dynamics, a switch in dynamics will only be detected after sufficient observations contradict the currently predominant dynamic model. If we wish to anticipate instead of react to changes in dynamics, a model should include possible causes for change. These influences on pedestrian behavior can be captured on an individual level using agent models, which have been used to reason about pedestrian intent [4,19] (i.e. where does observed agent want to go), account for preferences to move around certain regions of a static scene [19], and avoid collision with other agents, as is done in social force models [2,16]. [32] enhanced social force towards group behavior by introducing sub-goals such as “following a person”. The related Linear Trajectory Avoidance model [28] for short-term path prediction uses the expected point of closest approach to foreshadow and avoid possible collisions.

These agent-based models assume that pedestrians are fully aware of their environment [19,28]. However, this assumption does not hold when dealing with inattentive pedestrians in the intelligent vehicle context. [15] presented a study on head turning behaviors at pedestrian crosswalks regarding the best point of warning for inattentive pedestrians. They used gyro sensors to record head turning and let pedestrians press a button when they recognize an approaching vehicle. Apart from this sole study of Visual Focus of Attention (VFOA) in intelligent vehicle context we are aware of, VFOA has been investigated in other application contexts. For example, [5] used a HOG-based head detector to determine pedestrian attention for automated surveillance, and [3] combined contextual cues in a DBN to model influence of group interaction on VFOA.

Within the class of non-parametric methods for path prediction and action classification, [18] recently proposed two non-linear, higher order Markov models to estimate

whether a crossing pedestrian will stop at the curbside, one using Gaussian Process Dynamical Models (GPDM), and one using Probabilistic Hierarchical Trajectory Matching (PHTM). Both models use dense optical flow features in the pedestrian bounding box, in addition to the positional information. The first approach learns a GPDM of the dense flow for walking and stopping motion to predict future flow fields (and thereby lateral velocity). PHTM matches feature vectors of flow and position to a hierarchically organized tracklet database to extrapolate motion. Both approaches were shown to perform similar, and outperform the first-order Markov LDS and SLDS models, albeit at a large computational cost ([18] reports GPDM/PHTM is three/two orders of magnitude slower than KF). [20] considered the complementary case, whether a standing pedestrian will start to walk at the curbside. This only involved action classification and no path prediction, and an infrastructure-based sensor setup (no on-board vehicle sensing).

3 Proposed Approach

We are interested in modeling the motion dynamics of a pedestrian from the viewpoint of an approaching vehicle, in order to perform accurate path prediction. We consider that non-maneuvering pedestrian movement is well captured by a LDS with a basic motion model (e.g. constant position, constant velocity, constant turn rate) [7], and that maneuvering pedestrian movement can be suitably represented by means of an SLDS. Thus, the switching state indicates which basic motion model to use at any moment.

In this paper, we propose to condition the transition matrix of the SLDS switching state on latent factors that are likely going to influence the pedestrian's motion type. In a scenario of a lateral crossing pedestrian, we argue that the pedestrian's decision to continue walking or to stop is largely influenced by the existence of an approaching vehicle on collision course, the pedestrian's awareness thereof, and the position of the pedestrian with respect to the curbside.

Hence, we consider our main paper contribution a DBN which captures these three factors as latent states on top of an SLDS (see current section). The proposed approach goes beyond the state-of-the-art on pedestrian path prediction in vehicle context, which has considered the pedestrian in isolation, i.e. context free [18,20,30], and agent models that ignore a pedestrian's perception and resulting situational awareness [4,19,28].

3.1 Graphical Model

The proposed DBN is shown in Fig. 1. We distinguish two sets of variables: those relating to a SLDS (consisting of switching state M, latent position state X and associated observation Y) and those related to the scene context, i.e. spatial layout, situation criticality and the pedestrian's awareness (consisting of discrete latent variables $Z = \{SV, HSV, SC, AC\}$) that influence the SLDS switching state, and associated observables $E = \left\{HO, D^{min}, DTC\right\}$. These variables are now discussed in turn. Details on parameter estimation and computation of observables are given in Sec. 4.2.

SLDS. A SLDS contains a discrete switching state M_t, a continuous hidden state X_t, and a linear observation of the state Y_t with noise $\mathcal{N}(0, R)$ added. In our application, we

consider that any moment can exhibit one of two motion types, *walking* ($M_t = m_w$) and *standing* ($M_t = m_s$). While the velocity of any standing person is zero, different people can have different walking velocities, i.e. some people move faster than others. Let x_t denote a person's lateral position at time t (after vehicle ego-motion compensation) and v_t the corresponding velocity. Furthermore, v^{m_w} is the personal walking velocity of the pedestrian. The motion dynamics over a period Δt can then be described as,

$$x_t = x_{t-\Delta t} + v_t \Delta t + \epsilon_t \Delta t \qquad v_t = \begin{cases} 0 & \text{iff } M_t = m_s \\ v^{m_w} & \text{iff } M_t = m_w \end{cases} \tag{1}$$

Here $\epsilon_t \sim \mathcal{N}(0, Q)$ is zero-mean process noise that allows for deviations of the fixed velocity assumption. We will assume fixed time-intervals, and from here on set $\Delta t = 1$.

We include the velocity v^{m_w} in the state of an SLDS, together with the position x_t, such that we can filter both as we obtain observations over time, i.e. $X_t = [x_t, v_t^{m_w}]^\top$,

$$X_t = A^{(M_t)} X_{t-1} + \begin{bmatrix} \epsilon_t \\ 0 \end{bmatrix} \qquad \epsilon_t \sim \mathcal{N}(0, Q) \tag{2}$$

$$Y_t = C X_t + \eta_t \qquad \eta_t \sim \mathcal{N}(0, R) \tag{3}$$

where the switching state M_t selects the appropriate linear state transformation $A^{(m)}$,

$$A^{(m_s)} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad A^{(m_w)} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}. \tag{4}$$

$Y_t \in \mathbb{R}$ is the observed lateral position with observation matrix $C = [1\ 0]$. The initial distribution on the state X_0 expresses our prior beliefs about a pedestrian's position and walking speed, as learned from the training data (see Sec. 4.2). From the definition of the SLDS, we obtain the following conditional probability distributions for the graphical model, $P(X_t|X_{t-1}, M_t) = \mathcal{N}(X_t|A^{(M_t)}X_{t-1}, Q)$ and $P(Y_t|X_t) = \mathcal{N}(Y_t|CX_t, R)$.

Context. The transition probability of the SLDS switching state is conditioned on the Boolean latent context variables Z. Although all these variables are discrete, during inference the uncertainty propagates from the observables to these variables (and over time), resulting in posterior distributions that contain values between 0 and 1. Each contextual configuration $Z_t = z$ is associated with a motion model transition probability $\mathcal{P}_z$, where $\mathcal{P}(\cdot)$ indicates that the distribution is represented by a probability table, and the subscript here denotes a table for each value z, such that

$$P(M_t|M_{t-1}, Z_t = z) = \mathcal{P}_z(M_t|M_{t-1}). \tag{5}$$

The temporal transition of the context in Z is factorized by the probability tables

$$\begin{aligned} P(Z_t|Z_{t-1}) = \mathcal{P}(HSV_t|HSV_{t-1}, SV_t) \times \mathcal{P}(SV_t|SV_{t-1}) \\ \times \mathcal{P}(SC_t|SC_{t-1}) \times \mathcal{P}(AC_t|AC_{t-1}). \end{aligned} \tag{6}$$

The latent *Sees-Vehicle* (SV) variable indicates whether the pedestrian is currently seeing the vehicle. *Has-Seen-Vehicle* (HSV) indicates whether the pedestrian is aware

of the vehicle, i.e. whether $SV_{t'} = \text{true}$ for some $t' \leq t$. The transition probability of HSV_t encodes simply a logical OR between the Boolean HSV_{t-1} and SV_t nodes:

$$\mathcal{P}(HSV_t|HSV_{t-1}, SV_t) = \begin{cases} 1 & \text{iff } HSV_t = (HSV_{t-1} \vee SV_t) \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

The latent variable *Situation-Critical* (SC) indicates whether a situation is critical when both, pedestrian and vehicle, continue with their current velocities. *At-Curb* (AC) indicates if the pedestrian is currently at the distance from the curbside (as found in the training data) where a person would stop if they choose to wait and postpone crossing the road. The SV, SC and AC nodes furthermore depend on their value in the preceding time step, which improves the temporal consistency of these latent variables.

Next we discuss observations E_t which provide evidence for the latent context Z_t,

$$P(E_t|Z_t) = P(HO_t|SV_t) \times P(D_t^{min}|SC_t) \times P(DTC|AC_t). \quad (8)$$

The *Head-Orientation* observable HO_t serves as evidence for the *Sees-Vehicle* (SV_t) variable. We apply multiple classifiers to the head image region, each trained to detect the head in a particular looking direction (for details, see Section 4.1), and HO_t is then a vector with the classifier responses. The values in this vector form different unnormalized distributions over the classes, depending on whether the pedestrian is looking at the vehicle or not. However, if the head is not clearly observed (e.g. it is too far, or in the shadow), all values are typically low, and the observed class distribution provides little evidence of the true head orientation. We therefore model HO_t as a sample from a Multinomial distribution conditioned on SV_t, with parameter vector p_{sv},

$$P(HO_t|SV_t = sv) = \text{Mult}(HO_t|p_{sv}). \quad (9)$$

As such, higher classifier outputs count as stronger evidence for the presence of that class in the observation. In the other limit of all zero outputs, HO_t will have equal likelihood for any value of SV_t.

For *Situation-Critical* (SC), we consider the minimum distance D^{min} between the pedestrian and vehicle, if their paths would be extrapolated in time with fixed velocity [28]. While this indicator makes naive assumptions about the vehicle and pedestrian motion, it is still informative as a measure of how critical the situation is, and thereby, as part of our model, will lead to more accurate pedestrian path prediction. We define a Gamma distribution over D^{min} given SC, parametrized by shape a and scale b,

$$P(D_t^{min}|SC_t = sc) = \Gamma(D_t^{min}|a_{sc}, b_{sc}). \quad (10)$$

To obtain evidence for *At-Curb* (AC_t), we detect the curb ridge in the image, and measure its lateral position near the pedestrian. These noisy measurements are filtered with a constant position Kalman filter with zero process noise, such that we obtain an accurate estimate of the expected curb position, x_t^{curb}. *Distance-To-Curb*, DTC_t, is then calculated as the difference between the expected filtered position of the pedestrian, $\mathbf{E}[x_t]$, and of the curb, x_t^{curb}. Note that for path prediction we can estimate DTC even at future time steps, using predicted pedestrian positions, and accordingly predict AC too. The distribution over DTC_t given AC is modeled as a Normal distribution,

$$P(DTC_t|AC_t = ac) = \mathcal{N}(DTC_t|\mu_{ac}, \sigma_{ac}). \quad (11)$$

3.2 Inference

The DBN is used in a forward filtering procedure to incorporate all available observations of new time instances directly when they are received. We have a mixed discrete-continuous DBN where the exact posterior includes a mixture of $|M|^T$ Normal modes after T time steps, hence exact inference is intractable. We therefore resort to Assumed Density Filtering [22,25] for approximate inference, where after each time step the found posterior is approximated by a simpler distribution. The procedure consists of executing the following three steps for each time instance: predict, update, and collapse.

We will let $\overline{P}_t(\cdot) \equiv P(\cdot|O_{1:t-1})$ denote a prediction for time t (i.e. before receiving the observation O_t), and $\widehat{P}_t(\cdot) \equiv P(\cdot|O_{1:t})$ denote an updated estimate for time t (i.e. after observing O_t). Finally, $\widetilde{P}_t(\cdot)$ is the collapsed or approximated updated distribution that will be carried over to the predict step of the next time instance $t+1$.

Predict. To predict time t we use the posterior distribution of $t-1$, which is factorized into the joint distribution over the latent discrete nodes $\widetilde{P}_{t-1}(M_{t-1}, Z_{t-1})$ and the conditional Normal distribution $\widetilde{P}_{t-1}(X_{t-1}|M_{t-1}) = \mathcal{N}(X_{t-1}|\widetilde{\mu}_{t-1}^{(M_{t-1})}, \widetilde{\Sigma}_{t-1}^{(M_{t-1})})$.

First, the joint probability of the discrete nodes in the previous and current time steps is computed using the factorized transition tables of Eq. (5) and (6),

$$\overline{P}_t(M_t, M_{t-1}, Z_t, Z_{t-1}) = P(M_t|M_{t-1}, Z_t)P(Z_t|Z_{t-1})\widetilde{P}_{t-1}(M_{t-1}, Z_{t-1}). \quad (12)$$

Then for the continuous latent state X_t we predict the effect of the linear dynamics of all possible models M_t on the conditional Normal distribution of each M_{t-1},

$$\overline{P}_t(X_t|M_t, M_{t-1}) = \int P(X_t|X_{t-1}, M_t) \times \widetilde{P}_{t-1}(X_{t-1}|M_{t-1})\, dX_{t-1}. \quad (13)$$

Applying Eq. (2), we find that the parametric form of (13) is the Kalman prediction step

$$\begin{aligned}\mathcal{N}(X_t|\overline{\mu}_t^{(M_t,M_{t-1})}, \overline{\Sigma}_t^{(M_t,M_{t-1})}) = \\ \int \mathcal{N}(X_t|A^{(M_t)}X_{t-1}, Q) \times \mathcal{N}(X_{t-1}|\widehat{\mu}_{t-1}^{(M_{t-1})}, \widehat{\Sigma}_{t-1}^{(M_{t-1})})\, dX_{t-1}.\end{aligned} \quad (14)$$

Update. The update step incorporates the observations of the current time step to obtain the joint posterior. For each joint assignment (M_t, M_{t-1}), the LDS likelihood term is

$$\begin{aligned}P(Y_t|M_t, M_{t-1}) &= \int P(Y_t|X_t) \times \overline{P}_t(X_t|M_t, M_{t-1})\, dX_t \\ &= \mathcal{N}(Y_t|C\overline{\mu}_t^{(M_t,M_{t-1})}, \overline{\Sigma}_t^{(M_t,M_{t-1})} + R),\end{aligned} \quad (15)$$

where we make use of Eq. (3). Combining this with the prediction (Eq. (12)) and contextual likelihood (Eq. (8)), we obtain the posterior as one joint probability table

$$\widehat{P}_t(M_t, M_{t-1}, Z_t, Z_{t-1}) \propto P(Y_t|M_t, M_{t-1})P(E_t|Z_t)\overline{P}_t(M_t, M_{t-1}, Z_t, Z_{t-1}) \quad (16)$$

where we normalize the r.h.s. over all possible $(M_t, Z_t, M_{t-1}Z_{t-1})$ combinations to obtain the distribution on the l.h.s. The posterior distribution over the continuous state,

$$\begin{aligned}\widehat{P}_t(X_t|M_t, M_{t-1}) &\propto P(Y_t|X_t) \times \overline{P}_t(X_t|M_t, M_{t-1}) \\ &= \mathcal{N}(X_t|\widehat{\mu}_t^{(M_t,M_{t-1})}, \widehat{\Sigma}_t^{(M_t,M_{t-1})})\end{aligned} \tag{17}$$

has parameters $\left(\widehat{\mu}_t^{(M_t,M_{t-1})}, \widehat{\Sigma}_t^{(M_t,M_{t-1})}\right)$ for the $|M|^2$ possible transition conditions, which are obtained using the standard Kalman update equations.

Collapse. In the third step, the state of the previous time step is marginalized out from the joint posterior distribution, such that we only keep the joint distribution of variables of the current time instance, which will be used in the predict step of the next iteration.

$$\widetilde{P}_t(M_t, Z_t) = \sum_{M_{t-1}} \sum_{Z_{t-1}} \widehat{P}_t(M_t, M_{t-1}, Z_t, Z_{t-1}) \tag{18}$$

Likewise, we approximate the $|M|^2$ Normal distributions by just $|M|$ distributions,

$$\widetilde{P}_t(X_t|M_t) = \sum_{M_{t-1}} \widehat{P}_t(X_t|M_t, M_{t-1}) \times P(M_{t-1}|M_t) = \mathcal{N}(X_t|\widetilde{\mu}_t^{(M_t)}, \widetilde{\Sigma}_t^{(M_t)}) \tag{19}$$

Here, the parameters $\left(\widetilde{\mu}_t^{(M_t)}, \widetilde{\Sigma}_t^{(M_t)}\right)$ are found by Gaussian moment matching [22,25], and $P(M_{t-1}|M_t)$ through marginalizing and normalizing $\widehat{P}_t(M_t, M_{t-1}, Z_t, Z_{t-1})$.

4 Experiments

4.1 Dataset and Observations

Our dataset consists of 58 sequences recorded using a stereo camera (baseline $22\,cm$, $16\,fps$, 1176×640 pixels) mounted behind the windshield of a vehicle[1]. All sequences involve single pedestrians with the intention to cross the street, but feature different situation criticalities (critical[2] vs. non-critical), pedestrian situational awareness (vehicle seen vs. vehicle not seen) and pedestrian behavior (stopping at the curbside vs. crossing). Due to the focus on potentially dangerous situations, both driver and pedestrian were instructed during recording sessions. The dataset contains four different male pedestrians and eight different locations. Each sequence lasts several seconds (min / max / mean: $2.53\,s$ / $13.27\,s$ / $7.15\,s$), and pedestrians are generally unoccluded, though brief occlusions by poles or trees occur in three sequences.

Positional ground truth (GT) is obtained by manual labeling of the pedestrian bounding boxes and computing the median disparity over the upper pedestrian body area using dense stereo [17]. Analysis of crossing trajectories shows an mean gait cycle of

[1] The dataset, including annotations, will be made available for non-commercial, research purposes within a year after publication. Please contact the last author.

[2] N.B. None of the experiments exposed pedestrians to danger; "critical situation" refers to a theoretic outcome where both the approaching vehicle and pedestrian would not stop.

17.3 frames ($1.0\,s$) with 1.6 frames ($0.1\,s$) standard deviation. GT for contextual observations is obtained by labeling head orientation (16 discrete clock-wise increasing orientation angles). Sequences where potentially dangerous situations occur, i.e. when either pedestrian or vehicle should stop to avoid a collision, have been labeled as critical. Sequences are further labeled with event tags and time-to-event (TTE, in frames) values. For stopping pedestrians, TTE $= 0$ is when the last foot is placed on the ground at the curbside, and for crossing pedestrians at the closest point to the curbside (before entering the roadway). Frames before/after an event have negative/positive TTE values.

A HOG/linSVM pedestrian detector [9] provides measurements, given region-of-interests supplied by an obstacle detection component using dense stereo data. The resulting bounding boxes are used to calculate a median disparity over the upper pedestrian body area. The vehicle ego-motion compensated lateral position in world coordinates is then used as positional observation Y_t.

For the observed head orientation HO_t, the angular domain of $[0°, 360°)$ is split into eight discrete orientation classes of $0°, 45°, \cdots, 315°$. We trained a detector for each class [13], i.e. $f_0, \cdots, f_{315}$, such that the detector response $f_o(I_t)$ is the strength for the evidence that the observed image region I_t contains the head in orientation class o. For each detector we used neural networks with local receptive fields [33] trained in a one vs. rest manner. We used a separate training set with 9300 manually contour labeled head samples from 6389 gray-value images with a min./max./mean pedestrian height of 69/344/122 pixels (c.f. [14]). For additional training data, head samples were mirrored and shifted, and 22109 non-head samples were generated in areas around heads and from false positive pedestrian detections. For detection, we generate candidate head regions in the upper pedestrian detection bounding box from disparity based image segmentation. The most likely head image region $I^\star$ is selected from all candidates based on disparity information and detector responses. Before classification, head image patches are rescaled to $16 \times 16\,px$. The head observation $HO_t = [f_0(I_t^\star), \cdots, f_{315}(I_t^\star)]$ contains the confidences of the selected region.

The expected minimum distance D^{min} between pedestrian and vehicle is calculated as in [28] for each time step based on current position and velocity. Vehicle speed is provided by on-board sensors, for pedestrians the first order derivative is used and averaged over the last 10 frames. For DTC, the curbside is detected with a basic Hough transform [11]. The image region of interest is determined by the specified accuracy of a state-of-the-art vehicle localization approach (GPS+INS) using map data [31]. Y_t^{curb} is then the mean lateral position of the detected line back-projected to world coordinates.

4.2 Parameter Estimation

All distribution parameters are estimated from annotated training data. For stopping sequences, the GT switching state is defined as $M_t = m_s$ at moments with TTE $>= 0$, and as $M_t = m_w$ at all other moments, crossing sequences always have $M_t = m_w$. From the GT at time $t = 0$ we estimate the position and walking speed prior for X_0. Process noise Q is estimated from the differences of the estimated mean walking speed and a pedestrian's true walking speeds, and observation noise $\mathcal{N}(0, R)$ is estimated by the difference between GT and measured positions.

Considering head observation HO, we assume pedestrians recognize an approaching vehicle (GT label SV_t = true) when the GT head direction is in a range of $\pm 45^\circ$ around angle 0° (head is pointing towards the camera), and do not see the vehicle (SV_t = false) for angles outside this range (future human studies could allow a more precise threshold, or provide an angle distribution, the study in [15] only reported the frequency of head turning). For each ground truth label sv, we estimate the orientation class distributions p_{sv} by averaging the class weights in the corresponding head measurements. For the observation D^{min}, we define per trajectory one value for all SC_t labels ($\forall_t SC_t$ = true for trajectories with critical situations, $\forall_t SC_t$ = false otherwise), and estimate the distributions $\Gamma(D^{min}|a_{sc}, b_{sc})$. The distributions $\mathcal{N}(DTC_t|\mu_{ac}, \sigma_{ac})$ are estimated from GT curb positions and *At-Curb* labels, which are set to AC_t = true only at time instances where $-1 \leq$ TTE ≤ 1 when crossing, and TTE ≥ -1 when stopping. Finally, it is straightforward to estimate prior and transition probability tables for the discrete contextual quantities SV, AC from their GT labels. The same applies to the dynamic switching state M, conditioned on HSV, SC and AC. The transition probability for HSV is a logical OR, as described in 3.1. Since we only set SC labels once per sequence, we fix the SC transition probability to $1/100$ for changing state.

4.3 Evaluation

The dataset is divided into five sub-scenarios, listed in Table 1. Four sub-scenarios represent "normal" pedestrian behaviors (e.g. the pedestrian stops if he is aware of a critical situation and crosses otherwise). The fifth sub-scenario is anomalous, since the pedestrian crosses even though he is aware of the critical situation. We compare our proposed DBN with full context, referred to as *SC+HSV+AC*, to model variations with less context, and to a fixed velocity Kalman Filter with acceleration noise (see caption Table 1).

Leave-one-out cross-validation is used to separate training and test sequences, though sequences from the anomalous sub-scenario are excluded from the training data. For each time t with state X_t, we create a predictive distribution for X_{t+t_p} at t_p time steps in the future by iteratively applying the *Predict* and *Collapse* steps (see Sec. 3.2), and only *Update* with the DTC likelihood (Eq. (11)) using the predicted positions,

$$\overline{P}_{t_p|t}(X_{t+t_p}) \equiv \overline{P}(X_{t+t_p}|Y_{1:t}). \tag{20}$$

We define two performance metrics for a sequence, namely the Euclidean distance between lateral predicted expected position x_{t+t_p} and lateral GT position G_{t+t_p}, and the log likelihood of G under the predictive distribution:

$$error(t_p|t) = |\mathbb{E}\left[\overline{P}_{t_p|t}(x_{t+t_p})\right] - G_{t+t_p}| \tag{21}$$

$$predll(t_p|t) = \log\left[\overline{P}_{t_p|t}(G_{t+t_p})\right] \tag{22}$$

Note that the predictive log likelihood of [1] corresponds to $predll(0|t)$.

Comparison of Model Variations. The results in Table 1 show the predictive log likelihood $predll$ for $t_p = 16$ time steps ($\sim 1\ s$) in the future, averaged over the second up to TTE = 0 when the pedestrian reaches the curb. In the first three normal sub-scenarios,

Table 1. Prediction log likelihood of the GT pedestrian position for $t_p = 16$ frames ($\sim 1\ s$) ahead, for different sub-scenarios (rows) and models (columns), for TTE $\in [-15, 0]$. The first four sub-scenarios contain "normal" pedestrian behavior. The fifth case is anomalous (*lower* likelihood is better). Model variations (best SLDS variant marked in bold): full context (SC+HSV+AC), no curb (SC+HSV), only head (HSV), only criticality (SC), no context (SLDS), KF (LDS).

Sub-scenario	SC+HSV+AC	SC+HSV	HSV	SC	SLDS	LDS
non-critical, vehicle not seen, crossing	-0.61	-0.53	**-0.52**	-0.59	-0.59	-1.90
non-critical, vehicle seen, crossing	-0.53	**-0.45**	-0.46	-0.47	-0.49	-1.93
critical, vehicle not seen, crossing	-0.48	-0.34	**-0.17**	-0.59	-0.33	-1.88
critical, vehicle seen, stopping	**-0.33**	-0.70	-1.13	-0.80	-1.26	-1.88
critical, vehicle seen, crossing	**-0.90**	-0.27	-0.15	-0.25	-0.13	-1.88

all five SLDS-based models perform similarly, clearly outperforming the LDS (which has similar low likelihoods across the board, i.e. it is unspecific for any sub-scenario). However, in the fourth sub-scenario (pedestrian sees the vehicle in a critical situation and stops), the simpler DBNs have low predictive likelihoods, except for our proposed model. Without the full context, the other models are not capable to predict *if*, *where* and *when* the pedestrian will stop. For the anomalous fifth sub-scenario, only the proposed model results in *lower* likelihood than for normal behavior, which is a useful property for anomaly detection. A future driver warning strategy could benefit from the more accurate path prediction of our SC+HSV+AC model in high likelihood situations, whereas falling back to simpler models/strategies when anomalies are detected.

Fig. 2 illustrates a sequence from the stopping sub-scenario (fourth row in Table 1), with a snapshot just *before* (TTE $= -20$) and *after* (TTE $= -9$) the pedestrian becomes aware of the critical situation. At TTE $= -20$, the predicted distributions of all models are close together and indicate that the pedestrian continues walking (the LDS does so with high uncertainty). At TTE $= -9$, the mean position predictions of the LDS are furthest away from the GT (still within one std.dev. because of high uncertainty). The SLDS-only prediction shows a comparatively low uncertainty, but the predicted means have a high distance to the GT (not within one std.dev.). Predictions of the SC+HSV model are closer to the true positions, since it captures the situational awareness of the pedestrian and therefore assigns a higher probability, compared to SLDS, to switch to the standing model m_s. The SC+HSV+AC model makes the best predictions as it also anticipates where the pedestrian will stop, namely at the curbside.

In the context of action classification, Fig. 3 shows for various model variations, (left) the standing probability $\tilde{P}_t(M_t = m_s)$, and (right) the $error(t_p|t)$ for predictions made $t_p = 16$ frames ahead, plotted against the TTE. In the first sub-scenario (top row), the pedestrian crosses in a critical situation without seeing the approaching vehicle. All models have a very low stopping probability, but since a few sequences have ambiguous head observations, our proposed model does not exclude the possibility that the vehicle has been seen. This translates to a higher stopping probability near the curb, and to a higher error of the average prediction for a short while. Still, the model recuperates as the pedestrian approaches the curb and shows no sign of slowing down, which informs the model that the pedestrian did not see the vehicle (i.e. joint inference also means that observed motion dynamics can disambiguate low-level head orientation estimation). In

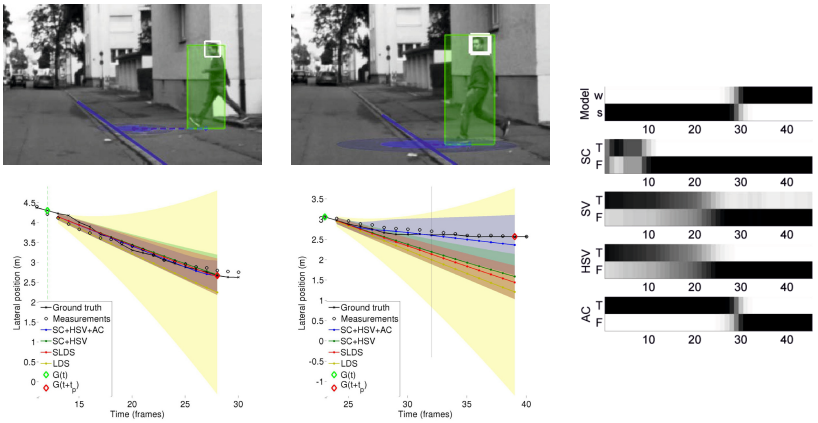

Fig. 2. Example of a pedestrian that will stop at the curb after becoming aware of a critical situation. Predictions are made $t_p = 16$ ($\sim 1\ s$) time steps ahead from different times t. Top left: Pedestrian with head detection bounding box (white), tracking bounding box (green), collapsed predicted distribution of the SC+HSV+AC model (blue ellipses show one and two std.dev.) and curb detection (blue line) made at time $t = 12$ (TTE $= -20$). Top center: The pedestrian became aware of the critical situation, shown is time step $t = 23$ (TTE $= -9$). Bottom left: Predictions (mean and std.dev.(shaded)) made at $t = 12$ (dashed green line and diamond) for the lateral position at time $t + t_p$ (red diamond indicates the GT at $t + t_p$) . Vertical black line denotes the event. Black dots indicate position measurements, the black line the GT positions. Colored lines are predicted positions by different models. Bottom center: Predictions of the lateral position at $t + t_p$ made from $t = 23$. Right: Inferred marginal distributions for the latent binary variables in the SC+HSV+AC model, using gray scale coded probability from 0 (black) to 1 (white). Horizontal axis is time. Variable labels are **T**rue and **F**alse, and **w**alking and **s**tanding.

the second sub-scenario (bottom row), the pedestrian is aware of the critical situation and stops at the curb. Now, all models show an increasing stopping probability towards the event point. In a few scenarios, the SLDS switches too early to the standing state, reacting to perceived de-acceleration (noise) of the pedestrian walking, hence the high std. dev. of the SLDS over all sequences early on. However, on average the SLDS assigns a higher probability to standing (> 0.5) than walking after the pedestrian has already reached the curb (TTE > 0). It can only react to changing dynamics, but not anticipate it. Our proposed model, on the other hand, gives the best action classification (highest stopping probability at TTE $= 0$). It anticipates the change in motion dynamics a few frames earlier as the SLDS, benefiting from the combined knowledge about situation criticality and spatial layout. Further, the knowledge about the spatial layout helps to keep the standing probability low while the pedestrian is still far away from the curb. The model with limited context information ends up in between proposed model and SLDS. Accordingly, our proposed model has the lowest prediction error (bottom right plot). Averaged over the sequences, it outperforms the baseline SLDS model by up to $0.39\ m$ (at TTE $= 1$) and the SC+HSV model with up to $0.16\ m$ (at TTE $= -10$).

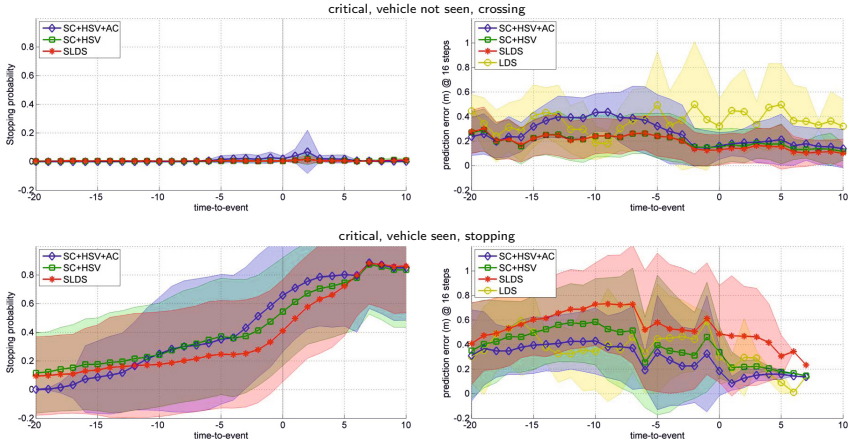

Fig. 3. Stopping probability (left) and lateral prediction error (right) when predicting 16 time steps ($\sim 1\ s$) ahead in the two critical sub-scenarios. Top: Pedestrian is not aware of the critical situation and crosses. Bottom: Pedestrian is aware of the critical situation and stops. Shown are mean and standard deviation (shaded) of each measure over all corresponding sequences, for our proposed model (SC+HSV+AC), an intermediate model without spatial layout information (SC+HSV), the baseline SLDS model without contextual cues, and a LDS.

Idealized Vision Measurements. To investigate how the vision components affect performance, we train and test using GT as idealized measurements for pedestrian location, curb location, and head orientation. We find that the lateral pedestrian and curb measurements are sufficiently accurate: GT does not notably change the results. Ideal head measurements alter the five sub-scenario scores of the SC+HSV+AC model w.r.t. Table 1 to -0.57, -1.08, -0.32, -0.12 ("normal" cases), and to -3.67 (anomalous case). Note that predictions became more accurate for critical sub-scenarios, less accurate in the second sub-scenario (non-critical, vehicle seen, crossing) at moments that are deemed critical since seeing the vehicle implies stopping, and that the likelihood of the anomalous fifth sub-scenario is still the lower than all other sub-scenarios, as expected.

Comparison with PHTM. Fig. 4 shows a comparison of the mean prediction error of our proposed model with the state-of-the-art PHTM model [18] which uses optical flow features and an exemplar database, on the four "normal" sub-scenarios. On two of these sub-scenarios (upper right and lower left plots) the proposed model outperforms PHTM slightly, both in terms of mean and variance, in particular on the arguably most important sub-scenario for a pedestrian safety application: critical, vehicle not seen, crossing. On the last sub-scenario (lower right plot) PHTM performs slightly better.

Computational Costs. The computational costs of the various approaches were assessed on standard PC hardware (Intel Core i7 X990 CPU at 3.47 GHz), see Table 2. We differentiate between the computational cost for obtaining the observables and that for

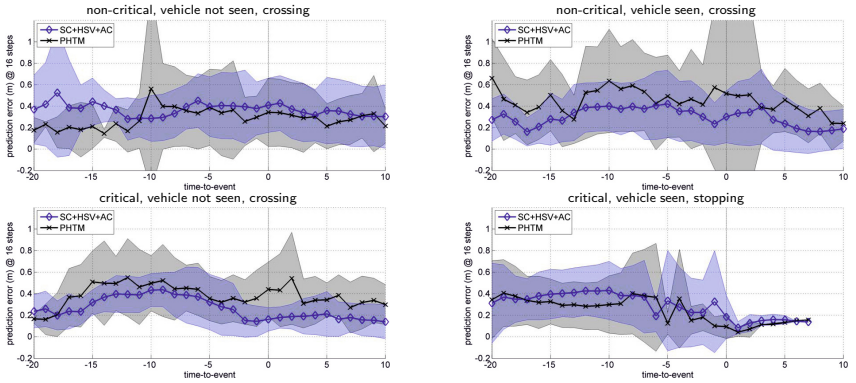

Fig. 4. The plots show the lateral prediction error of our proposed model and the PHTM model in various sub-scenarios. The lines show the avg. error over all sequences in a sub-scenario, after aligning the results by their TTE values, and the shaded region shows the std. dev. of the error.

Table 2. Computational costs for the different models per frame (avg. per frame, in ms)

Approach	Observables	State est. & pred.	Total
SC+HSV+AC	160	40	200
SLDS	60	10	70
LDS	60	0.4	60
PHTM	70	600	670

performing state estimation and prediction. In terms of observables, all approaches used positional information derived from a dense stereo-based pedestrian detector (about $60\,ms$). The additional observables used in our proposed SC+HSV+AC model (e.g. head orientation and curb detection) cost an extra $100\,ms$ to compute. PHTM on the other hand requires computing dense optical flow within the pedestrian bounding box (about $10\,ms$). But, as seen in Table 2, the proposed model is *one order* of magnitude more efficient than PHTM when considering only the state estimation and prediction component (this even though PHTM implements its trajectory matching by an efficient hierarchical technique [18]), and it is three times more efficient in total.

5 Conclusions

We presented a novel model for pedestrian path prediction in the intelligent vehicle domain. The model, a DBN, incorporated the pedestrian situational awareness, situation criticality and spatial layout of the environment (curbside) as latent states on top of an SLDS, thus controlling changes in the pedestrian dynamics. The proposed model overall outperformed simpler models with or without partial contextual cues by predicting GT pedestrian positions more accurate (up to $0.39\,m$ compared to the SLDS when predicting $\sim 1\,s$ ahead) and with *higher* likelihood in situations similar to those in the training set. In atypical situations, it predicted GT pedestrian position with a *lower* likelihood, a desirable property for anomaly detection.

We show that the proposed approach even slightly outperformed a state-of-the-art PHTM approach at less than *a third* of computational cost. These two approaches do not stand directly in competition, however, as they use different sources of information that could conceivably be combined. Further work involves the incorporation of additional scene context (e.g. traffic light, pedestrian crossing) and the extension of the basic motion types of the SLDS (e.g. turning). We are encouraged that the presented context-based models can play an important role in future generation driver warning and vehicle control strategies that save pedestrian lives.

References

1. Abbeel, P., Coates, A., Montemerlo, M., Ng, A.Y., Thrun, S.: Discriminative training of Kalman filters. In: Robotics: Science and Systems, pp. 289–296 (2005)
2. Antonini, G., Martinez, S.V., Bierlaire, M., Thiran, J.P.: Behavioral priors for detection and tracking of pedestrians in video sequences. IJCV 69(2), 159–180 (2006)
3. Ba, S., Odobez, J.: Multiperson visual focus of attention from head pose and meeting contextual cues. IEEE PAMI 33(1), 101–116 (2011)
4. Bandyopadhyay, T., Won, K., Frazzoli, E., Hsu, D., Lee, W., Rus, D.: Intention-aware motion planning. In: Algorithmic Foundations of Robotics X, pp. 475–491. Springer (2013)
5. Benfold, B., Reid, I.: Guiding visual surveillance by tracking human attention. In: Proc. BMVC (2009)
6. Bishop, C.M.: Pattern Recognition and Machine Learning, vol. 1. Springer (2006)
7. Blackman, S., Popoli, R.: Design and Analysis of Modern Tracking Systems. Artech House Norwood (1999)
8. Boyen, X., Koller, D.: Tractable inference for complex stochastic processes. In: Proc. of UAI, pp. 33–42. Morgan Kaufmann Publishers Inc. (1998)
9. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. CVPR, pp. 886–893. IEEE (2005)
10. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. IEEE PAMI 34(4), 743–761 (2012)
11. Duda, R.O., Hart, P.E.: Use of the Hough transformation to detect lines and curves in pictures. Commun. ACM 15(1), 11–15 (1972)
12. Enzweiler, M., Gavrila, D.M.: Monocular pedestrian detection: Survey and experiments. IEEE PAMI 31(12), 2179–2195 (2009)
13. Enzweiler, M., Gavrila, D.M.: Integrated pedestrian classification and orientation estimation. In: Proc. CVPR, pp. 982–989. IEEE (2010)
14. Flohr, F., Dumitru-Guzu, M., Kooij, J.F.P., Gavrila, D.M.: Joint probabilistic pedestrian head and body orientation estimation. In: IEEE Intell. Veh. (2014)
15. Hamaoka, H., Hagiwara, T., Tada, M., Munehiro, K.: A study on the behavior of pedestrians when confirming approach of right/left-turning vehicle while crossing a crosswalk. In: IEEE Intell. Veh., pp. 106–110 (2013)
16. Helbing, D., Molnár, P.: Social force model for pedestrian dynamics. Phys. Rev. E 51(5), 4282 (1995)
17. Hirschmüller, H.: Stereo processing by semiglobal matching and mutual information. IEEE PAMI 30(2), 328–341 (2008)
18. Keller, C.G., Gavrila, D.M.: Will the pedestrian cross? A study on pedestrian path prediction. IEEE Trans. ITS 15(2), 494–506 (2014)

19. Kitani, K.M., Ziebart, B.D., Bagnell, J.A., Hebert, M.: Activity forecasting. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part IV. LNCS, vol. 7575, pp. 201–214. Springer, Heidelberg (2012)
20. Köhler, S., Schreiner, B., Ronalter, S., Doll, K., Brunsmann, U., Zindler, K.: Autonomous evasive maneuvers triggered by infrastructure-based detection of pedestrian intentions. In: IEEE Intell. Veh., pp. 519–526 (2013)
21. Kooij, J.F.P., Englebienne, G., Gavrila, D.M.: A non-parametric hierarchical model to discover behavior dynamics from tracks. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 270–283. Springer, Heidelberg (2012)
22. Lauritzen, S.L.: Propagation of probabilities, means, and variances in mixed graphical association models. Journal of the American Statistical Association 87(420), 1098–1108 (1992)
23. Meinecke, M.M., Obojski, M., Gavrila, D.M., Marc, E., Morris, R., Töns, M., Lettelier, L.: Strategies in terms of vulnerable road user protection. In: EU Project SAVE-U, Deliverable D6 (2003)
24. Meuter, M., Iurgel, U., Park, S.B., Kummert, A.: Unscented Kalman filter for pedestrian tracking from a moving host. In: IEEE Intell. Veh., pp. 37–42 (2008)
25. Minka, T.P.: Expectation propagation for approximate Bayesian inference. In: Proc. of UAI, pp. 362–369. Morgan Kaufmann Publishers Inc. (2001)
26. Oh, S.M., Rehg, J.M., Balch, T., Dellaert, F.: Learning and inferring motion patterns using parametric segmental switching linear dynamic systems. IJCV 77(1-3), 103–124 (2008)
27. Pavlovic, V., Rehg, J.M., MacCormick, J.: Learning switching linear models of human motion. In: Advances in NIPS, pp. 981–987 (2000)
28. Pellegrini, S., Ess, A., Schindler, K., Van Gool, L.: You'll never walk alone: Modeling social behavior for multi-target tracking. In: Proc. ICCV, pp. 261–268 (2009)
29. Rosti, A.V.I., Gales, M.J.F.: Rao-Blackwellised Gibbs sampling for switching linear dynamical systems. In: Proc. of the IEEE ICASSP, vol. 1, pp. 809–812 (2004)
30. Schneider, N., Gavrila, D.M.: Pedestrian path prediction with recursive Bayesian filters: A comparative study. In: Weickert, J., Hein, M., Schiele, B. (eds.) GCPR 2013. LNCS, vol. 8142, pp. 174–183. Springer, Heidelberg (2013)
31. Schreiber, M., Knöppel, C., Franke, U.: LaneLoc: Lane marking based localization using highly accurate maps. In: IEEE Intell. Veh., pp. 449–454 (2013)
32. Tamura, Y., Le, P.D., Hitomi, K., Chandrasiri, N., Bando, T., Yamashita, A., Asama, H.: Development of pedestrian behavior model taking account of intention. In: IEEE IROS, pp. 382–387 (2012)
33. Wöhler, C., Anlauf, J.K.: A time delay neural network algorithm for estimating image-pattern shape and motion. IVC 17(3-4), 281–294 (1999)