Context-based path prediction for targets with switching dynamics

Kooij, Julian F.P.; Flohr, Fabian; Pool, Ewoud A.I.; Gavrila, Dariu M.

DOI
10.1007/s11263-018-1104-4

Publication date
2019

Document Version
Final published version

Published in
International Journal of Computer Vision

International Journal of Computer Vision
https://doi.org/10.1007/s11263-018-1104-4

Context-Based Path Prediction for Targets with Switching Dynamics

Julian F. P. Kooij[1] · Fabian Flohr[3] · Ewoud A. I. Pool[2] · Dariu M. Gavrila[1,2]

Received: 18 December 2017 / Accepted: 21 June 2018

Abstract

Anticipating future situations from streaming sensor data is a key perception challenge for mobile robotics and automated vehicles. We address the problem of predicting the path of objects with multiple dynamic modes. The dynamics of such targets can be described by a Switching Linear Dynamical System (SLDS). However, predictions from this probabilistic model cannot anticipate when a change in dynamic mode will occur. We propose to extract various types of cues with computer vision to provide context on the target's behavior, and incorporate these in a Dynamic Bayesian Network (DBN). The DBN extends the SLDS by conditioning the mode transition probabilities on additional context states. We describe efficient online inference in this DBN for probabilistic path prediction, accounting for uncertainty in both measurements and target behavior. Our approach is illustrated on two scenarios in the Intelligent Vehicles domain concerning pedestrians and cyclists, so-called Vulnerable Road Users (VRUs). Here, context cues include the static environment of the VRU, its dynamic environment, and its observed actions. Experiments using stereo vision data from a moving vehicle demonstrate that the proposed approach results in more accurate path prediction than SLDS at the relevant short time horizon (1 s). It slightly outperforms a computationally more demanding state-of-the-art method.

Keywords Intelligent vehicles · Path prediction · Situational awareness · Vulnerable road users · Intention estimation · Dynamic Bayesian Network · Probabilistic inference

1 Introduction

Anticipating how nearby objects will behave is a key challenge in various application domains, such as intelligent vehicles, social robotics, and surveillance. These domains concern systems that navigate trough crowded environments, that interact with their surroundings, or which detect potentially anomalous events. Predicting future situations requires understanding what the nearby objects are, and knowledge on how they typically behave. Object detection and tracking are therefore common first steps for situation assessment, and the past decade has seen significant progress in these fields. Still, accurately predicting the paths of targets with multiple motion dynamics remains challenging, since a switch in dynamics can result in a significantly different trajectory. People are an important example of such target. For instance, a pedestrian can quickly change between walking and standing.

To improve path prediction of objects with switching dynamics, we propose to exploit context cues that can be extracted from sensor data. Especially vision can provide measurements for a diverse set of relevant cues. But incorporating more observations in the prediction process also increases sensitivity to measurement uncertainty. In fact, uncertainty is an inherent property of any prediction on future events. To deal with uncertainties, we leverage existing probabilistic filters for switching dynamics, which are common for tracking maneuvering targets (Bar-Shalom

Communicated by Larry Davis.

✉ Dariu M. Gavrila
d.m.gavrila@tudelft.nl

Julian F. P. Kooij
j.f.p.kooij@tudelft.nl

Fabian Flohr
fabian.flohr@daimler.com

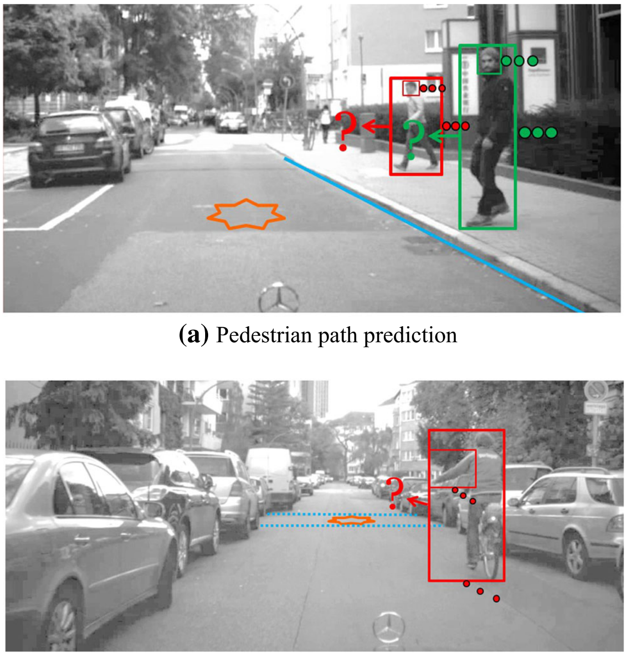
Ewoud A. I. Pool
e.a.i.pool@uva.nl

[1] Delft University of Technology, Mekelweg 2, 2628 CD Delft, The Netherlands

[2] AMLab, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands

[3] Department of Environment Perception, Daimler AG, Wilhelm-Runge-Str. 11, 89081 Ulm, Germany

Published online: 02 July 2018



(a) Pedestrian path prediction

(b) Cyclist path prediction

Fig. 1 Path prediction of vulnerable road users with switching dynamics. **a** The pedestrian can cross, or stop. Context for a crossing pedestrian's path includes vehicle trajectory, the pedestrian's awareness of the approaching vehicle, and pedestrian's position w.r.t. the curbside. **b** The cyclist approaching an intersection can cycle straight or turn left. Context includes the vehicle trajectory, the cyclist's expressed intent by raising an arm, and distance to the intersection

et al. 2001). Our proposed method therefore extends a Switching Linear Dynamical System (SLDS) with dynamic latent states that represent context. The resulting model is a Dynamic Bayesian Network (DBN) (Murphy 2002), where the latent states control the switching probabilities between the dynamic modes. We can utilize existing theory for approximate posterior inference in DBNs to efficiently compute predictive distributions on the future state of the target.

In this paper, we focus on applications in the Intelligent Vehicle (IV) domain. More specifically, we demonstrate our method on path prediction of pedestrians and cyclists, i.e. the so-called Vulnerable Road Users (VRUs). For automated vehicles, forecasting the future locations of traffic participants is a crucial input to plan safe, comfortable and efficient paths though traffic (Althoff et al. 2009; Paden et al. 2016). However, the 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. Small deviations in the prediction of, say, 30 cm in the estimated lateral position of VRUs can make all the difference, as this might place them just inside or outside the driving corridor. Better predictions can therefore warn the driver further ahead of time at the same false alarm rate, and more reliably initiate automatic braking and evasive steering (Keller et al. 2011; Köhler et al. 2013).

We evaluate our approach on two scenarios. The first scenario that we target considers a pedestrian intending to laterally cross the street, as observed by a stereo camera onboard an approaching vehicle, see Fig. 1a. Accident analysis shows that this scenario accounts for a majority of all pedestrian fatalities in traffic (Meinecke et al. 2003). We argue that the pedestrian's decision to stop can be predicted to a large degree from three cues: the existence of an approaching vehicle on collision course, the pedestrian's awareness thereof, and the spatial layout of the static environment. Likewise, the second scenario considers a cyclist driving on the same lane as the ego-vehicle, who may turn left at an upcoming crossing in front of the vehicle, see Fig. 1b. This scenario also has three predictive cues, namely the cyclist raising an arm to indicate intent to turn at the crossing, the cyclist's proximity to the crossing, and the existence of an approaching vehicle.

Our approach is general though, and can be extended with additional motion types (e.g. pedestrian crossing the road in a curved path), or to other application domains, such as robot navigation in human-inhabited environments. Our method also does not prohibit the use of other sensors or computer vision methods than the ones considered here.

2 Related Work

In this section we discuss existing work on state estimation and path prediction, especially for pedestrians and cyclists. We also present different context cues from vision that have been explored to improve behavior prediction.

2.1 Detection and Tracking

Object Detection The classical object detection pipeline first applies a sliding window on the input image to extract image features at candidate regions, and classify each region as containing the target object. In recent years, state-of-the-art detection and classification performance is instead achieved by deep ConvNets trained on large datasets. For online applications, ConvNet architectures are now also achieving real-time performance by combining detection and classification in a single forward pass, e.g. Single Shot Multibox Detector (Liu et al. 2016) or YOLO (Redmon et al. 2016).

There are many datasets for pedestrian detection, e.g. those presented in Enzweiler and Gavrila (2009), and Dollár et al. (2012). For an overview on vision-based pedestrian detection, see surveys from Enzweiler and Gavrila (2009), Dollár et al. (2012) and Ohn-Bar and Trivedi (2016). For cyclists, there is the Tsinghua-Daimler Cyclist Benchmark from Li et al. (2016). These datasets make it possible to create sophisticated models that require large amounts of training data, for instance for unified pedestrian and cyclist detection (Li et al. 2017), or recovering the 3D pose of vehicles and VRUs (Braun et al. 2016). Indeed, the IV domain is used in many challenging Computer Vision benchmarks, e.g. KITTI (Geiger et al. 2012; Menze and Geiger 2015) and ADE20K (Zhou et al. 2017), hence we expect VRU detection to improve even further in the near future.

State Estimation In the IV domain, state estimation is typically done in a 3D world coordinate system, where also information from other sensors (e.g. lidar, radar) is fused. Image detections can be projected to this world coordinates through depth estimation from monocular or stereo-camera setup (Hirschmüller 2008).

The per-frame spatial position of detections can then be incorporated in a tracking framework where the measurements are assigned to tracks, and temporally filtered. Filtering provides estimates and uncertainty bounds on the objects' true position and dynamical states. State estimation often models the state and measurements as a Linear Dynamical System (LDS), which assumes that the model is linear and that noise is Gaussian. In this case, the Kalman filter (KF) (Blackman and Popoli 1999) is an optimal filtering algorithm. In the intelligent vehicle domain, the KF is the most popular choice for pedestrian tracking (see Schneider and Gavrila 2013 for an overview). The Extended and Unscented KF (Meuter et al. 2008) can, to a certain degree, account for non-linear dynamical or measurement models, but multiple motion models are needed for maneuvering targets that alternate various dynamics.

The SLDS is a type of DBN which can model multiple possible dynamics. It extends the LDS with a top-level discrete Markov chain. At each time step, the state of this chain determines which of the various possible motion dynamics is applied to the underlying LDS, allowing to 'switch' the dynamics through discrete state transitions. Unfortunately, exact inference and learning in an SLDS becomes intractable, as the number of modes in the posterior distribution grows exponential over time in the number of the switching states (Pavlovic et al. 2000). There is however a large body of literature on approximate inference in such DBNs. One solution is to approximate the posterior by samples using some Markov Chain Monte Carlo method (Oh et al. 2008; Rosti and Gales 2004; Kooij et al. 2016). However, sampling is impractical for online real-time inference as convergence can be slow. Instead, Assumed Density Filtering (ADF) (Bishop 2006; Minka 2001) approximates the posterior at every time step with a simpler distribution. It has generally been applied to mixed discrete-continuous state spaces with conditional Gaussian posterior (Lauritzen 1992), and to discrete state DBNs, where it is also known as Boyen-Koller inference (Boyen and Koller 1998). ADF will be further discussed in Sect. 3.2.

The Interacting Multiple Model (IMM) KF (Blackman and Popoli 1999) is another popular algorithm to track a maneuvering target, mixes the states of several KF filters running in parallel. It has been applied for path prediction in the intelligent vehicle domain for pedestrian (Keller and Gavrila 2014; Schneider and Gavrila 2013), and cyclists (Cho et al. 2011) tracking. IMM can be seen as doing an alternative form of approximate inference in a SLDS (Murphy 2002).

2.2 Context Cues for VRU Behaviors

Even though SLDSs can account for changes in dynamics, a switch in dynamics will only be acknowledged after sufficient observations contradict the currently active dynamic model. If we wish to anticipate instead of reacting to changes in dynamics, a model should include possible causes for change.

Various papers provide naturalistic studies on pedestrians behavior, e.g. during encounters at unsignalized crossing (Chen et al. 2017), to predict when a pedestrian will cross (Völz et al. 2016), or to categorizing danger in vehicle-pedestrian encounters (Otsuka et al. 2017). Similar studies are also being performed for cyclists. Zernetsch et al. (2016) collected data at a single intersection for path prediction of a starting cyclists, and Hubert et al. (2017) used the same data to find indicators of cyclist starting behavior. Some studies have used naturalistic data to detect and classify critical vehicle-cyclist interactions at intersections (Sayed et al. 2013; Vanparijs et al. 2015; Cara and de Gelder 2015), while others use simulations to study bicycle motion at intersections (Huang et al. 2017; Zhang et al. 2017).

For online prediction of VRU behavior, cues must be extracted from sensor data. Especially computer vision provides many types of context cues, as the following subsections will discuss. From the extract features, behavior predicting can then be treated as a classification problem (Bonnin et al. 2014; Köhler et al. 2013). However, probabilistic methods integrate the inherent detection uncertainty directly into path prediction (Schulz and Stiefelhagen 2015a, b; Keller and Gavrila 2014; Kooij et al. 2014a).

Static Environment Cues The relation between spatial regions of an environment and typical behavior has been extensively researched in visual surveillance, where the viewpoint is static. For instance, different motion dynamics may frequently occur at specific space coordinates (Morris and Trivedi 2011; Kooij et al. 2016; Robicquet et al. 2016; Yi et al. 2016; Jacobs et al. 2017). Another approach is to interpret the environment, e.g. detect semantic regions and learn how these affect agent behavior (Kitani et al. 2012; Rehder and Kloeden 2015). Such semantics enable knowledge transfer to new scenes too (Ballan et al. 2016). In surveillance, agent models are also used to reason about intent (Bandyopadhyay et al. 2013), i.e. where the pedestrian intends to go.

In the IV domain, behavior is typically tied to road infrastructure (Oniga et al. 2008; Geiger et al. 2014; Kooij et al. 2014b; Sattarov et al. 2014; Pool et al. 2017). Road layout can be obtained from localization using GPS and INS sensors (Schreiber et al. 2013) to retrieve information map data on the surrounding infrastructure. SLAM techniques provide another means for accurate self-localization in a world coordinate frame, and are also used in automotive research (Geiger et al. 2012; Mur-Artal and Tardós 2017). Another approach is to infer local road layout directly from sensor data (Geiger et al. 2014; Yi et al. 2017). Here, too, semantic scene segmentation with ConvNets can be used to identify static and dynamic objects, and drivable road [c.f. Cityscapes benchmark (Cordts et al. 2016)].

Dynamic Environment Cues VRU behavior may also be influenced by other dynamic objects in their surrounding. For instance, social force models (Antonini et al. 2006; Helbing and Molnár 1995; Huang et al. 2017) expect agents to avoid collisions with other agents. Tamura et al. (2012) extended social force towards group behavior by introducing sub-goals such as "following a person". The related Linear Trajectory Avoidance model (Pellegrini et al. 2009) for short-term path prediction uses the expected point of closest approach to foreshadow and avoid possible collisions.

Neural nets can also learn how multiple agents move in each others presence (Alahi et al. 2016; Yi et al. 2016), even from a vehicle perspective (Karasev et al. 2016; Lee et al. 2017). In the IV domain, interaction of road users with the ego-vehicle is especially important. An often used indicator is the Time-To-Collision (TTC) which is the time that remains until a collision between two objects occurs if their course and speeds are maintained (Sayed et al. 2013). A related indicator is the minimum future distance between two agents, which like TTC assumes both travel with fixed velocity (Pellegrini et al. 2009; Cara and de Gelder 2015).

Beyond accounting for the presence of other road users, traffic participants also negotiate right of way to coordinate their actions. Rasouli et al. (2017) presents a study of such interactions between drivers and pedestrians.

Object Cues People may not always be fully aware of their surroundings, and inattentive pedestrians are an important safety case in the IV context. A study on pedestrian behavior prediction by Schmidt and Färber (2009) found that human drivers look for body cues, such as head movement and motion dynamics, though exactly determining the pedestrian's gaze is not necessary. Hamaoka et al. (2013) presents a study on head turning behaviors at pedestrian crosswalks regarding the best point of warning for inattentive pedestrians. They use gyro sensors to record head turning and let pedestrians press a button when they recognize an approaching vehicle. Continuous head estimation can be obtained by interpolating the results of multiple discrete orientation classifiers, adding physical constraints and temporal filtering to improve robustness (Enzweiler and Gavrila 2010; Flohr et al. 2015). Benfold and Reid (2009) uses a Histogram of Oriented Gradients (HOG) based head detector to determine pedestrian attention for automated surveillance. Ba and Odobez (2011) combines context cues in a DBN to model the influence of group interaction on focus of attention. Recent work uses ConvNets for real-time 2D estimation of the full body skeleton (Cao et al. 2017).

The full body appearance can also be informative for path prediction, e.g. to classify the object and predict a class-specific path (Klostermann et al. 2016), or to identify predictive poses. Köhler et al. (2013) rely on infrastructure-based sensors to classify whether a pedestrian standing at the curbside will start to walk. Keller and Gavrila (2014) estimates whether a crossing pedestrian will stop at the curbside using dense optical flow features in the pedestrian bounding box. They propose two non-linear, higher order Markov models, one using Gaussian Process Dynamical Models (GPDM), and one using Probabilistic Hierarchical Trajectory Matching (PHTM). Both approaches are shown to perform similar, and outperform the first-order Markov LDS and SLDS models, albeit at a large computational cost.

3 Proposed Approach

We are interested in predicting the path of an object with switching motion dynamics. We consider that non-maneuvering movement (i.e. where the type of motion is not changing) is well captured by a LDS with a basic motion model [e.g. constant position, constant velocity, constant turn rate (Blackman and Popoli 1999)]. An SLDS combines multiple of such motion models into a single model, using an additional switching state to indicate which of the basic motion model is in use at any moment. These probabilistic models can express the state probability given all past position measurements (i.e. online filtering), or given all past and future measurements (i.e. offline smoothing). Similarly, it is also possible to infer future state probability given only the current past measurements (i.e. prediction). Details on inference will be presented in Sect. 3.2.

While the SLDS can provide good predictions overall, we shall demonstrate that this unfortunately comes at the cost of bad predictions when a switch in dynamics occurs between the current time step and the predicted time step. To tackle the shortcomings of the SLDS, we propose an online filtering and prediction method that exploits context information on factors that may influence the target's motion dynamics. More specifically, for VRU path prediction we consider three types of context, namely interaction with the dynamic environment, the relation of the VRU to the static environment, and the VRU's observed behavior.

The presented work offers several contributions:

1. We present a generic approach to exploit context cues to improve predictions with a SLDS. The cues are represented as discrete latent nodes in a DBN that extends the SLDS. These nodes influence the switching probabilities between dynamic modes of the SLDS. An algorithm for approximate online inference and path prediction is provided.
2. We apply our approach to VRU path prediction. Various context cues are extracted with computer vision. The context includes the dynamic environment, the static environment, and the target's behavior. The proposed approach goes beyond existing work in this domain that has considered no or limited context. We show the influence of different types of context cues on path prediction, and the importance of combining them.
3. Our work targets online applications in real-world environments. We use stereo vision data collected from a moving vehicle, and compare computational performance to a state-of-the-art method in the IV domain.

We shall now formalize the SLDS, and demonstrate with a simple example how context can improve prediction quality when an actual switch in dynamics occurs. Afterwards, we discuss approximate inference, and specify how our general approach can be applied to VRU path prediction.

3.1 Contextual Extension of SLDS

Given noisy positional measurements Y_t of a moving target, the target's true dynamics can be modeled as a Linear Dynamical System (LDS) with a latent continuous state X_t. The process defines the next state as a linear transformation A of the previous state, with process noise $\epsilon_t \sim \mathcal{N}(0, Q)$ added through linear transformation B. Observation Y_t results from a linear transformation C of the true state X_t with also Gaussian noise $\eta_t \sim \mathcal{N}(0, R)$ added, referred to as the measurement noise.

A Switching LDS (SLDS) conditions the dynamics on a discrete switching state M_t. We shall consider that the switching state M_t selects the appropriate state transformation matrix $A^{(M_t)}$ for the process model, though generally other LDS terms could also be conditioned on M_t, if needed. Accordingly, the SLDS process is here defined as

$$X_t = A^{(M_t)} X_{t-1} + B\epsilon_t \quad \epsilon_t \sim \mathcal{N}(0, Q) \tag{1}$$

$$Y_t = C X_t + \eta_t \quad \eta_t \sim \mathcal{N}(0, R). \tag{2}$$

These equations can be reformulated as conditional distributions, i.e. $P(X_t|X_{t-1}, M_t) = \mathcal{N}(X_t|A^{(M_t)} X_{t-1}, BQB^\top)$ and $P(Y_t|X_t) = \mathcal{N}(Y_t|CX_t, R)$. The first time step is defined by initial distributions $P(X_0|M_0)$ and and $P(M_0)$. The former expresses our prior knowledge about the position and movement of a new target for each switching state, the latter expresses the prior on the switching state itself. Note that the SLDS describes a joint probability distribution over sequences of observations and latent states. It is therefore a particular instance of a DBN.

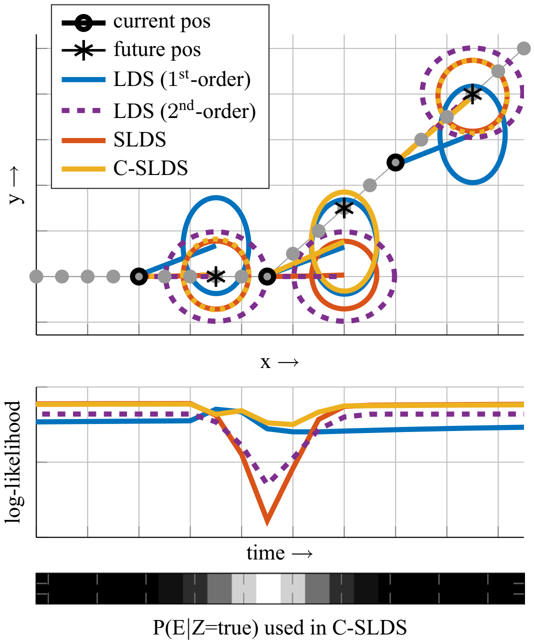


Fig. 2 Best viewed in color. Toy example of path prediction for a target (moving left to right) with two motion types. Four models are considered: a LDS with 1st- and 2nd-order state space, a SLDS, and the proposed Context-based SLDS (C-SLDS). Each model extrapolates the filtered dynamics three time steps ahead, resulting in a Gaussian distribution over the future state. Top: Spatial view where gray dots show the target's actual path. The noisy measurements are omitted for clarity. Black circles mark the target's position $t = 5$, 10 and 15. Stars mark its corresponding future position. Colored lines and uncertainty ellipses show the predicted path, and the distribution at the prediction horizon of each model. Middle: Log likelihood over time of the true future position under the predictive distributions. Bottom: The evidence from the spatial context over time, used by the C-SLDS. If this context is 'activated' (white) the state transition probabilities are high and the C-SLDS acts like the 1st-order LDS, otherwise (black) it acts like the SLDS

As an example, consider predicting the future position of a moving target which exhibits two types of motion, namely, moving in positive x direction (type A), and moving in positive x and y direction (type B). The target performs motion type A for 10 time steps, and then type B for another 10 time steps. The target's motion dynamics are known, and a LDS is selected to filter and predict its future position for three steps ahead. An LDS with a 1st-order state space only includes position in its state, $X_t = [x_t]$. The target velocity is assumed to be fixed. Each time step, this LDS adds the fixed velocity and random Gaussian noise to the position. For the considered target, the optimal fixed velocity of the LDS is an

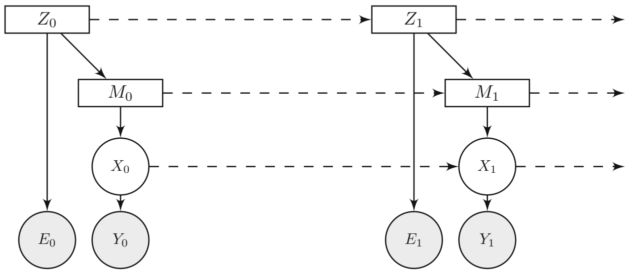


Fig. 3 Context-based SLDS as directed graph, unrolled for two time slices. Discrete/continuous/observed nodes are rectangular/circular/shaded

average of the two possible motion directions. Figure 2 illustrates this example, and shows predictions made using this LDS in blue. The LDS provides poor predictive distribution which do not adapt to the target motion.

An LDS with a 2nd-order state space also includes the velocity in the state, $X_t = [x_t, \dot{x}_t]^\top$. Through process noise on the velocity, this LDS can account for changes in the target direction. However, its spatial uncertainty grows rapidly when predicting ahead as the velocity uncertainty increases without bounds. The figure shows its predictions in purple.

An SLDS can instead combine multiple LDS instances, each specialized for one motion type. This example considers an SLDS combining two 1st-order LDSs, one with fixed horizontal, and one with fixed diagonal velocity. Less process noise is needed compared to the single LDS. The switching state is to $1/20$ chance of changing motion types. The predictions of this SLDS, shown in red in the figure are better during each mode. It exploits that changes between modes are rare, and the prediction uncertainty is therefore smaller. However, this notion leads to bad results when half-way the rare switch does occur, as the log-likelihood plots shows. The SLDS thus delivers good predictions for typical time steps where the dynamics do not change, at the cost of inaccurate predictions for rare moments where the dynamics switch. But a switch could be part of expected behavior for maneuvering targets. Preferably, the model should deliver good predictions for typical or 'normal' tracks, even if these switch, at the cost of inaccurate predictions during rare tracks with anomalous behavior.

We make a simple observation to tackle the poor SLDS performance during a switch. Consider having information that the target approaches a region with higher probability of switching than usual, i.e. spatial *context*. Outside this region the SLDS behaves as before. But inside, the switching probability is set to $1/2$, which makes every dynamic mode equally likely in the future such. The SLDS then behaves as the original 1st-order LDS. By selectively adapting the transition probabilities based on the spatial context, this model can ideally take best of both worlds, as the yellow log-likelihood plot in Fig. 2 confirms.

To obtain this behavior, the transition probability of the SLDS switching state is conditioned on additional discrete latent *context* variables (which will be specified in more detail later). These different context states can collectively be represented by a single discrete node Z_t. Each contextual configuration $Z_t = z$ defines a different motion model transition probability,

$$P(M_t = m_t | M_{t-1} = m_{t-1}, Z_t = z_t) = \mathcal{P}(M_t | M_{t-1}, Z_t) \tag{3}$$

$$P(Z_t = z_t | Z_{t-1} = z_{t-1}) = \mathcal{P}(Z_t | Z_{t-1}) \tag{4}$$

Here we use $\mathcal{P}(\cdot)$ to denote probability tables.

We also introduce a set of measurements E_t, which provide evidence for the latent context variables through conditional probability $P(E_t | Z_t)$. The bottom plot in Fig. 2 demonstrates this likelihood for the example. Even though the context Z_t is discrete, during inference the uncertainty propagates from the observables to these variables, resulting in posterior distributions that assign real-valued probabilities to the possible contextual configurations.

Like the SLDS, this extended model is also a DBN. Figure 3 shows all variables as nodes in a graphical representation of the DBN. The arrows indicate that child nodes are conditionally dependent on their parents. The dashed arrows show conditional dependency on the nodes in the previous time step.

3.2 Online 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 $|M|^T$ Gaussian modes after T time steps, hence exact online inference is intractable (Pavlovic et al. 2000). We therefore resort to Assumed Density Filtering (ADF) (Bishop 2006; Minka 2001) as an approximate inference technique. The filtering procedure consists of executing the three steps for each time instance: predict, update, and collapse. These steps will also be used for predicting the target's future path for a given prediction horizon, as described later in Sect. 3.4.

We will let $\overline{P}_t(\cdot) \equiv P(\cdot | Y_{1:t-1}, E_{1:t-1})$ denote a prediction for time t (i.e. before receiving observations Y_t and E_t), and $\widehat{P}_t(\cdot) \equiv P(\cdot | Y_{1:t}, E_{1:t})$ denote an updated estimate for time t (i.e. after observing Y_t and E_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 $t+1$. Figure 4 shows a flowchart of the computational performed in the steps, which will now be explained in more detail.

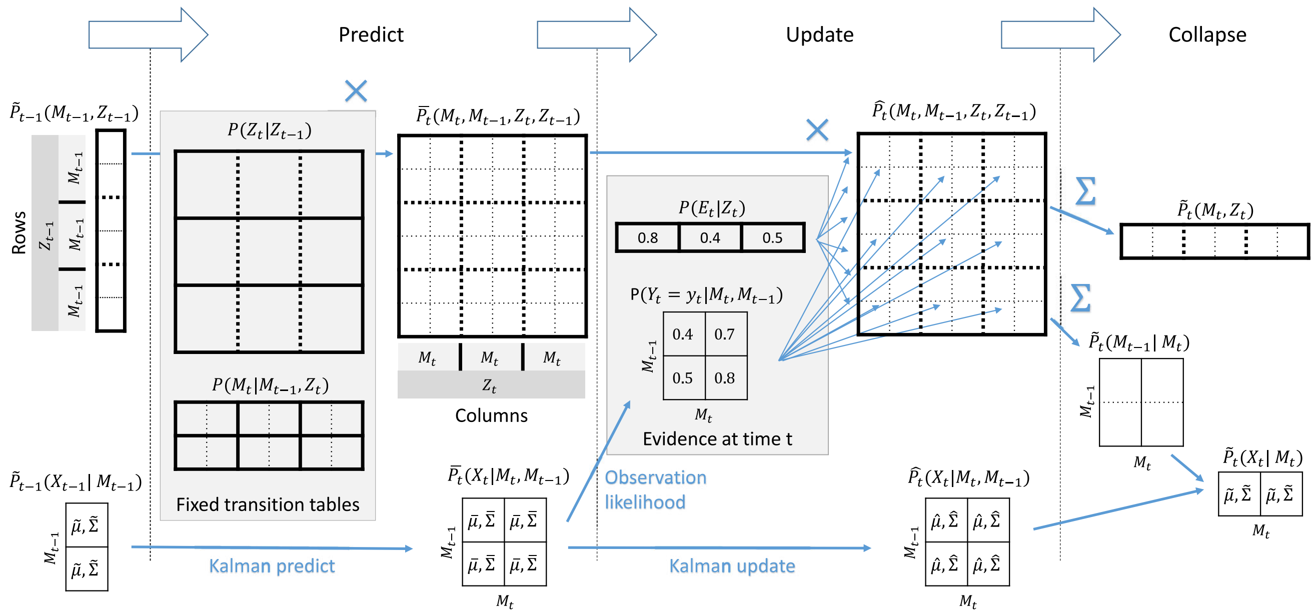

Predict
Update
Collapse
$\widetilde{P}_{t-1}(M_{t-1}, Z_{t-1})$
Rows
Z_{t-1}
M_{t-1}
M_{t-1}
M_{t-1}
$P(Z_t|Z_{t-1})$
$P(M_t|M_{t-1}, Z_t)$
Fixed transition tables
$\overline{P}_t(M_t, M_{t-1}, Z_t, Z_{t-1})$
M_t
M_t
M_t
Z_t
Columns
$P(E_t|Z_t)$
0.8
0.4
0.5
$P(Y_t = y_t|M_t, M_{t-1})$
M_{t-1}
0.4
0.7
0.5
0.8
M_t
Evidence at time t
$\widehat{P}_t(M_t, M_{t-1}, Z_t, Z_{t-1})$
Σ
$\widetilde{P}_t(M_t, Z_t)$
Σ
$\widetilde{P}_t(M_{t-1}| M_t)$
M_{t-1}
M_t
$\widetilde{P}_t(X_t| M_t)$
$\tilde{\mu}, \tilde{\Sigma}$
$\tilde{\mu}, \tilde{\Sigma}$
M_t
$\widetilde{P}_{t-1}(X_{t-1}| M_{t-1})$
M_{t-1}
$\tilde{\mu}, \tilde{\Sigma}$
$\tilde{\mu}, \tilde{\Sigma}$
Kalman predict
$\overline{P}_t(X_t|M_t, M_{t-1})$
M_{t-1}
$\bar{\mu}, \bar{\Sigma}$
$\bar{\mu}, \bar{\Sigma}$
$\bar{\mu}, \bar{\Sigma}$
$\bar{\mu}, \bar{\Sigma}$
M_t
Observation likelihood
Kalman update
$\widehat{P}_t(X_t|M_t, M_{t-1})$
M_{t-1}
$\hat{\mu}, \hat{\Sigma}$
$\hat{\mu}, \hat{\Sigma}$
$\hat{\mu}, \hat{\Sigma}$
$\hat{\mu}, \hat{\Sigma}$
M_t

Fig. 4 Flowchart of the three ADF steps in a single time instance. For simplicity, two motion models and three context states are assumed, and no example numbers are shown in the probability tables, except for the likelihoods of the context evidence E_t and observed position Y_t. Table rows correspond to the model M and/or context state Z of the previous time step $t-1$, columns correspond to the model M and/or context state Z of the time step t. Within each probability table, cell blocks with thick lines correspond to a single Z value, cell blocks with solid lines are normalized and therefore sum to one

3.2.1 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 into the conditional distribution and the dynamical state, $\widetilde{P}_{t-1}(X_{t-1}|M_{t-1}) = \mathcal{N}(X_{t-1}|\widetilde{\mu}_{t-1}^{(M_{t-1})}, \widetilde{\Sigma}_{t-1}^{(M_{t-1})})$.

First, the joint probability of the discrete nodes in the previous and current time steps is computed using the factorized transition tables of Eqs. (3) and (4),

$$\overline{P}_t(M_t, M_{t-1}, Z_t, Z_{t-1}) = P(M_t|M_{t-1}, Z_t) \times P(Z_t|Z_{t-1}) \times \widetilde{P}_{t-1}(M_{t-1}, Z_{t-1}). \tag{5}$$

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},

$$\overline{P}_t(X_t|M_t, M_{t-1}) = \int P(X_t|X_{t-1}, M_t) \times \widetilde{P}_{t-1}(X_{t-1}|M_{t-1}) \quad dX_{t-1}. \tag{6}$$

With the dynamics of Eq. (1), we find that the parametric form of (6) is the Kalman prediction step, i.e.

$$\overline{P}_t(X_t|M_t, M_{t-1}) = \mathcal{N}(X_t|\overline{\mu}_t^{(M_t,M_{t-1})}, \overline{\Sigma}_t^{(M_t,M_{t-1})}) \tag{7}$$

$$\overline{\mu}_t^{(M_t,M_{t-1})} = A^{(M_t)}\widehat{\mu}_{t-1}^{(M_{t-1})} \tag{8}$$

$$\overline{\Sigma}_t^{(M_t,M_{t-1})} = A^{(M_t)}\widehat{\Sigma}_{t-1}^{(M_{t-1})}A^{(M_t)^\top} + BQB^\top. \tag{9}$$

3.2.2 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

$$\begin{aligned} P(Y_t|M_t, M_{t-1}) &= \int P(Y_t|X_t) \times \overline{P}_t(X_t|M_t, M_{t-1})\, dX_t \\ &= \mathcal{N}(Y_t|C\overline{\mu}_t^{(M_t,M_{t-1})}, \overline{\Sigma}_t^{(M_t,M_{t-1})} + R), \end{aligned} \tag{10}$$

where we make use of Eq. (2). Combining this with the prediction [Eq. (5)] and context likelihood $P(E_t|Z_t)$, we obtain the posterior as one joint probability table

$$\widehat{P}_t(M_t, M_{t-1}, Z_t, Z_{t-1}) \propto P(Y_t|M_t, M_{t-1}) \times P(E_t|Z_t) \times \overline{P}_t(M_t, M_{t-1}, Z_t, Z_{t-1}). \tag{11}$$

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

$$\begin{aligned} \widehat{P}_t(X_t|M_t, M_{t-1}) &\propto P(Y_t|X_t) \times \overline{P}_t(X_t|M_t, M_{t-1}) \\ &= \mathcal{N}(X_t|\widehat{\mu}_t^{(M_t,M_{t-1})}, \widehat{\Sigma}_t^{(M_t,M_{t-1})}) \end{aligned} \tag{12}$$

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.

In case there is no observation for a given time step, there is no difference between the predicted and updated probabilities, which means both Eqs. 11 and 12 simplify to $\widehat{P}_t(\cdot) = \overline{P}_t(\cdot)$.

3.2.3 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.

$$\widetilde{P}_t(M_t, Z_t) = \sum_{M_{t-1}} \sum_{Z_{t-1}} \widehat{P}_t(M_t, M_{t-1}, Z_t, Z_{t-1}) \tag{13}$$

Similarly, $\widetilde{P}(M_{t-1}|M_t)$ is straightforward to obtain,

$$\widetilde{P}(M_{t-1}, M_t) \propto \sum_{Z_t} \sum_{Z_{t-1}} \widehat{P}_t(M_t, M_{t-1}, Z_t, Z_{t-1}) \tag{14}$$

$$\widetilde{P}(M_{t-1}|M_t) = \widetilde{P}(M_{t-1}, M_t) / \sum_{M_{t-1}} \widetilde{P}(M_{t-1}, M_t). \tag{15}$$

We approximate the $|M|^2$ Gaussian distributions from Eq. (12) by just $|M|$ distributions,

$$\begin{aligned}\widetilde{P}_t(X_t|M_t) &= \sum_{M_{t-1}} \widehat{P}_t(X_t|M_t, M_{t-1}) \times P(M_{t-1}|M_t) \\ &= \mathcal{N}(X_t|\widetilde{\mu}_t^{(M_t)}, \widetilde{\Sigma}_t^{(M_t)}).\end{aligned} \tag{16}$$

Here, the parameters $\left(\widetilde{\mu}_t^{(M_t)}, \widetilde{\Sigma}_t^{(M_t)}\right)$ are found by Gaussian moment matching (Lauritzen 1992; Minka 2001),

$$\widetilde{\mu}_t^{(M_t)} = \sum_{M_{t-1}} P(M_{t-1}|M_t) \times \widehat{\mu}_t^{(M_t,M_{t-1})} \tag{17}$$

$$\begin{aligned}\widetilde{\Sigma}_t^{(M_t)} = \sum_{M_{t-1}} P(M_{t-1}|M_t) \times \Bigg[& \widehat{\Sigma}_t^{(M_t,M_{t-1})} \\ & + \left(\widehat{\mu}_t^{(M_t,M_{t-1})} - \widetilde{\mu}_t^{(M_t)}\right) \cdot \left(\widehat{\mu}_t^{(M_t,M_{t-1})} - \widetilde{\mu}_t^{(M_t)}\right)^\top \Bigg].\end{aligned} \tag{18}$$

3.3 Context for VRU Motion

Until now the use of context in a SLDS has been described in general terms, but for VRU path prediction we distinguish a four binary context cues, $Z_t = \{Z_t^{\mathrm{DYN}}, Z_t^{\mathrm{STAT}}, Z_t^{\mathrm{ACT}}, Z_t^{\mathrm{ACTED}}\}$, which affect the probability of switching dynamics:

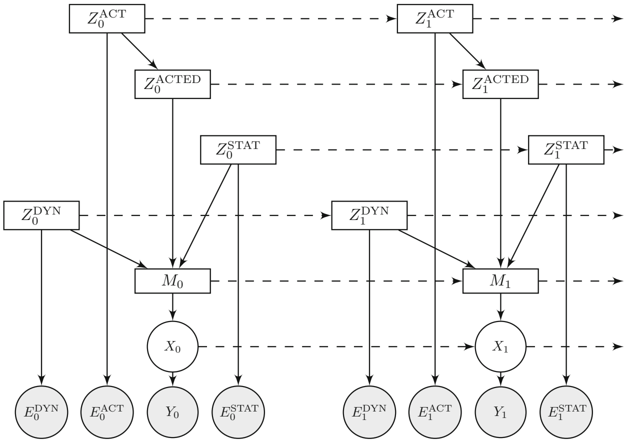


Fig. 5 DBN with context cues for VRU path prediction, unrolled for two time slices. Discrete/continuous/observed nodes are rectangular/circular/shaded. The binary context nodes represent interaction with the dynamic environment Z_t^{DYN}, relation to the static environment Z_t^{STAT}, and object behavior (i.e. how VRU acts, Z_t^{ACT}, or has acted, Z_t^{ACTED})

- Dynamic environment context: the presence of other traffic participants can deter the VRU to move too closely. In our experiments we only consider the presence of the ego-vehicle. Context indicator Z_t^{DYN} thus refers to a possible collision course, and therefore if the situation is potentially critical.
- Static environment context: the location of the VRU in the scene relative to the main infrastructure. Z_t^{STAT} is true iff the VRU is at the location where change typically occurs.
- Object context: Z_t^{ACT} indicates if the VRU's current actions provide insight in the VRU's intention (e.g. signaling direction), or awareness (e.g. line of gaze). The related context Z_t^{ACTED} captures whether the VRU performed the relevant actions in the past.

These cues are present in both the pedestrian and the cyclist scenario. The temporal transition of the context in Z is now factorized in several discrete transition probabilities,

$$\begin{aligned}P(Z_t|Z_{t-1}) = {} & \mathcal{P}\left(Z_t^{\mathrm{STAT}}|Z_{t-1}^{\mathrm{STAT}}\right) \times \mathcal{P}\left(Z_t^{\mathrm{DYN}}|Z_{t-1}^{\mathrm{DYN}}\right) \\ & \times \mathcal{P}\left(Z_t^{\mathrm{ACTED}}|Z_{t-1}^{\mathrm{ACTED}}, Z_t^{\mathrm{ACT}}\right) \times \mathcal{P}\left(Z_t^{\mathrm{ACT}}|Z_{t-1}^{\mathrm{ACT}}\right).\end{aligned} \tag{19}$$

The graphical model of the DBN obtained through this context factorization is shown in Fig. 5.

The latent object behavior variable, Z^{ACT}, indicates whether the VRU is currently exhibiting behavior that signals a future change in dynamics. The related object context Z^{ACTED} acts as a memory, and indicates whether the behavior has occurred in the past. For instance, the behavior of a crossing pedestrian is affected by the pedestrian's awareness, i.e. whether the pedestrian has seen the vehicle approach at any moment in the past, $Z_{t'}^{\mathrm{ACT}} = \text{true}$ for some $t' \leq t$. The transition probability of Z_t^{ACTED} encodes simply a logical OR between the Boolean Z_{t-1}^{ACTED} and Z_t^{ACT} nodes:

$$\mathcal{P}\left(Z_t^{\mathrm{ACTED}} | Z_{t-1}^{\mathrm{ACTED}}, Z_t^{\mathrm{ACT}}\right) = \begin{cases} \text{true} & \text{iff } \left(Z_{t-1}^{\mathrm{ACTED}} \vee Z_t^{\mathrm{ACT}}\right) \\ \text{false} & \text{otherwise.} \end{cases} \tag{20}$$

The context states have observables associated to them, except Z^{ACTED} which is conditioned on the Z^{ACT} state only. Hence, there are only three types of context observables, $E = \left\{E^{\mathrm{DYN}}, E^{\mathrm{STAT}}, E^{\mathrm{ACT}}\right\}$, which are assumed to be conditionally independent distributed given the context states. This yields the following factorization,

$$P(E_t|Z_t) = P\left(E_t^{\mathrm{DYN}}|Z_t^{\mathrm{DYN}}\right) \times P\left(E_t^{\mathrm{STAT}}|Z_t^{\mathrm{STAT}}\right) \times P\left(E_t^{\mathrm{ACT}}|Z_t^{\mathrm{ACT}}\right). \tag{21}$$

3.4 VRU Path Prediction

The goal of probabilistic path prediction is to provide a useful distribution $\overline{P}_{t_p|t}$ on the future target position,

$$\overline{P}_{t_p|t}(X_{t+t_p}) \equiv \overline{P}(X_{t+t_p}|Y_{1:t}). \tag{22}$$

Here t_p is the *prediction horizon*, which defines how many time steps are predicted ahead from the current time t. This formulation can reuse the steps from approximate online inference of Sect. 3.2, treating the unknown observations of the future time steps as 'missing' observations. Iterative application of these steps creates a predictive distribution of each moment in the future path, until the desired prediction horizon is reached.

However, the static environment context Z^{STAT} exploits the relation between VRU's position and the static environment. Since the expected position is readily available during path prediction, we *can* estimate the future influence of the static environment on the predicted continuous state of the VRU. For instance, while predicting a walking pedestrian's path, we can also predict the decreasing distance of the pedestrian to the static curbside.

Accordingly, to obtain prediction $\overline{P}(X_{t+t_p}|Y_{1:t})$ at time t for t_p time steps in the future, we use the current filtered state distribution and iteratively apply the *Predict*, *Update* and *Collapse* steps as before. However, the *Update* step now only includes measurements for the future static environment context using the expected VRU position. It does not have measurements for the object and dynamic environment indicators, thereby effectively skipping these context cues. Thus, to predicting future time steps, we replace Eq. (11) by

$$\widehat{P}_t(M_t, M_{t-1}, Z_t, Z_{t-1}) \propto P\left(E_t^{\mathrm{STAT}}|Z_t^{\mathrm{STAT}}\right) \times \overline{P}_t(M_t, M_{t-1}, Z_t, Z_{t-1}). \tag{23}$$

This enables the method to predict *when* the change in dynamics will occur, *if* the VRU is inclined to do so.

4 VRU Scenarios

The previous section explained the general approach of using a DBN to incorporate context cues, infer current and future use of dynamics, and ultimately perform future path prediction. This section now specifies the dynamics and context used for the two VRU scenarios of interest.

4.1 Crossing Pedestrian

The first scenario concerns the pedestrian wanting to cross the road, and approaching the curb from the right, as illustrated in Fig. 1a.

Motion Dynamics In this scenario, we consider that the pedestrian can exhibit at any moment 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 $\dot{x}_t$ the corresponding velocity. Furthermore, $\dot{x}^{m_w}$ is the preferred walking speed of this particular pedestrian. The motion dynamics over a period Δt can then be described as,

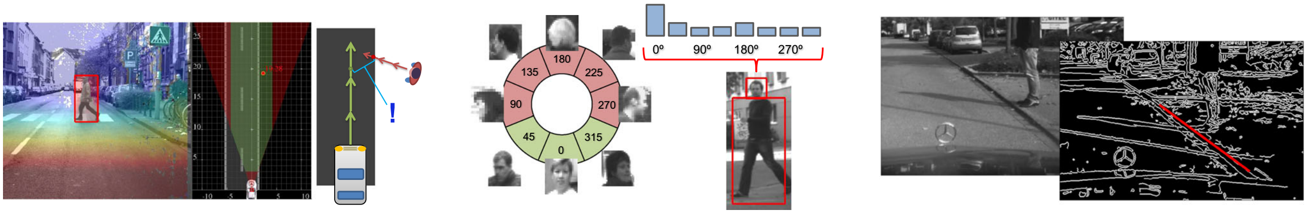
$$x_t = x_{t-\Delta t} + \dot{x}_t \Delta t + \epsilon_t \Delta t \quad \dot{x}_t = \begin{cases} 0 & \text{iff } M_t = m_s \\ \dot{x}^{m_w} & \text{iff } M_t = m_w \end{cases} \tag{24}$$

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 $\Delta t = 1$.

Since the latent $\dot{x}^{m_w}$ is constant over the duration of a single track, $\dot{x}_t^{m_w} = \dot{x}_{t-1}^{m_w}$. Still, it varies between pedestrians. We include the velocity $\dot{x}^{m_w}$ in the state of an SLDS together with the position x_t such that we can filter both. The prior on $\dot{x}_0^{m_w}$ represent walking speed variations between pedestrians. By filtering, the posterior on $\dot{x}_t^{m_w}$ converges to the preferred walking speed of the current track.

The switching state M_t selects the appropriate linear state transformation $A^{(M_t)}$, and the matrices from Eq. (1) become

$$X_t = \begin{bmatrix} x_t \\ \dot{x}_t^{m_w} \end{bmatrix}, \; A^{(m_s)} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \; A^{(m_w)} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \; B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{25}$$



(a) Minimum distance of approach, D^{min} **(b)** Head Orientation class confidences, HO **(c)** Distance To Curb, DTC

Fig. 6 Context observables used in the crossing pedestrian scenario. **a** The closest distance between pedestrian and the ego-vehicle. **b** The pedestrian head orientation. **c** The distance to the curb

The observations $Y_t \in \mathbb{R}$ from Eq. (2) are the observed lateral position. The observation matrix is defined as $C = [1\ 0]$. The initial distribution on the state X_0 and both the process and measurements noise are estimated from the training data (see Sec. 5.4).

Context Following the study on driver perception (Schmidt and Färber 2009), the context cues in the pedestrian scenario are collision risk, pedestrian head orientation, and where the pedestrian is relative to the curb. The context observations E_t for this scenario are illustrated in Fig. 6. The related Fig. 7 shows the empirical distributions of the context observations estimated on annotated training data from a pedestrian dataset. The dataset will be discussed in more detail in Sect. 5.1.

The dynamic environment context Z^{DYN} indicates whether the current trajectories of the ego-vehicle and pedestrian creates a critical situation. Namely, if there is a possible collision when both pedestrian and vehicle continue with their current velocities. For the interaction cue, we consider the minimum distance D^{min} between the pedestrian and vehicle if their paths would be extrapolated in time with fixed velocity (Pellegrini et al. 2009), see Fig. 6a. 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. As part of our model, will thereby assist to make path prediction more accurate. We define a Gamma distribution over D^{min} conditioned on the latent interaction state Z^{DYN}, parametrized by shape a and scale b,

$$P\left(E_t^{\text{DYN}} | Z_t^{\text{DYN}} = z\right) = \Gamma(D_t^{min} | a_z, b_z). \tag{26}$$

This distribution is illustrated in Fig. 7a.

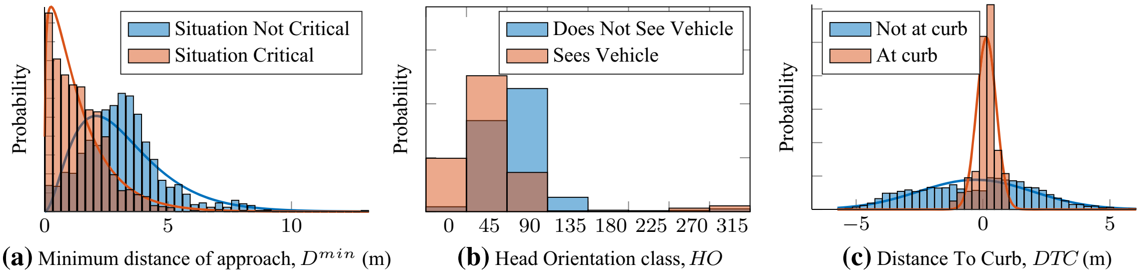
The object behavior context Z^{ACT} describes if the pedestrian is seeing the approaching vehicle. Z^{ACTED} indicates whether this was the case at any moment in the past, i.e. if the pedestrian did see the vehicle. It therefore indicates the pedestrian's awareness: a pedestrian will likely stop when he is aware of the fact that it is dangerous to cross. The *Head-Orientation* observable HO_t serves as evidence E_t^{ACT} for the behavior. A head orientation estimator is applied to the head image region. It consists of multiple classifiers, each trained to detect the head in a particular looking direction, and HO_t is then a vector with the classifier responses, see Fig. 6b. The values in this vector form different unnormalized distributions over the classes, depending on whether the pedestrian is looking at the vehicle or not, see Fig. 7b. 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 Z_t^{ACT}, thus with parameter vectors p_{true} and p_{false} for $Z^{\text{ACT}} = \text{true}$ and $Z^{\text{ACT}} = \text{false}$ respectively,

$$P\left(E_t^{\text{ACT}} | Z_t^{\text{ACT}} = z\right) = \text{Mult}(HO_t | p_z). \tag{27}$$

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 classifier outputs, HO_t will have equal likelihood for any value of Z_t^{ACT}.

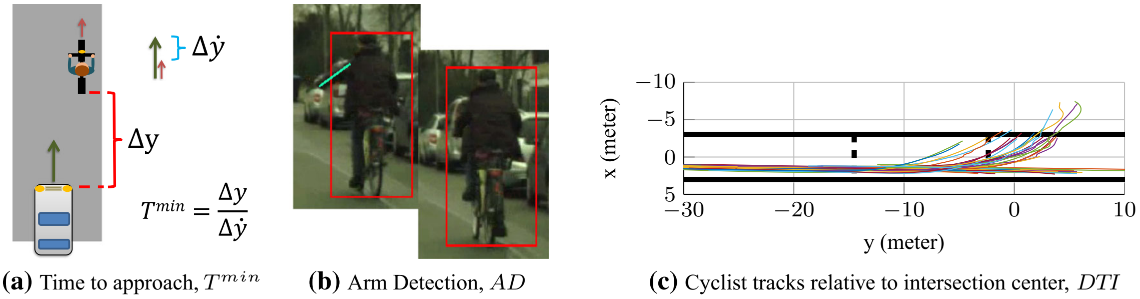
The static environment context Z^{STAT} indicates if the pedestrian is currently at the position next to the curb where a person would normally stop if they wait for traffic before crossing the road. The relative position of the pedestrian with respect to the curbside therefore serves as observable for this cue. As shown in Fig. 6c, we detect the curb ridge in the image. It is then projected to world coordinates with the stereo disparity to 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, $\mathbf{E}[x_t^{\text{curb}}]$. Note that for path prediction we can estimate DTC even at future time steps, using predicted pedestrian positions. The distribution over DTC_t given Z^{STAT} is modeled as a Normal distribution, see Fig. 7c,

$$P\left(E_t^{\text{STAT}} | Z_t^{\text{STAT}} = z\right) = \mathcal{N}(DTC_t | \mu_z, \sigma_z). \tag{28}$$



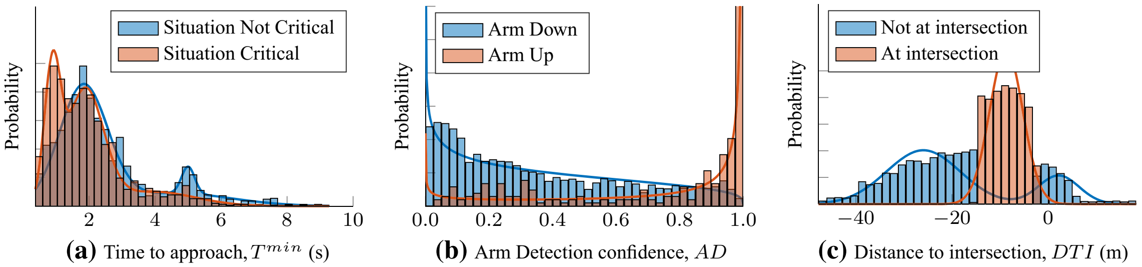
(a) Minimum distance of approach, D^{min} (m) **(b)** Head Orientation class, *HO* **(c)** Distance To Curb, *DTC* (m)

Fig. 7 Histograms and the fitted distributions of the context observations E_t for the pedestrian scenario, conditioned on their GT context states Z_t: **a** the minimum future distance between pedestrian and ego-vehicle, conditioned on Z_t^{DYN} (critical vs non-critical situation). **b** The head orientation conditioned on Z_t^{ACT} (pedestrians sees vs does not see the ego-vehicle). **c** The pedestrian's distance to the curb, conditioned on conditioned on Z_t^{STAT} (not at curb vs at the curb)



(a) Time to approach, T^{min} **(b)** Arm Detection, *AD* **(c)** Cyclist tracks relative to intersection center, *DTI*

Fig. 8 Context observables used in the cyclist scenario. **a** The extrapolated time till ego-vehicle reaches cyclist. **b** Detection of the cyclist's raised arm. **c** An static environment map is built offline through SLAM. The map's coordinate system is aligned with the intersection center. By projecting tracked cyclist positions to this coordinate system, their longitudinal distance to the intersection is obtained



(a) Time to approach, T^{min} (s) **(b)** Arm Detection confidence, *AD* **(c)** Distance to intersection, *DTI* (m)

Fig. 9 Histograms and the fitted distributions of the context observations E_t for the cyclist scenario, conditioned on their GT context states Z_t: **a** the time until the ego-vehicle would approach the cyclist, if both kept moving at the same speed, conditioned on Z^{DYN} (critical vs non-critical). **b** The arm detector's confidence conditioned on Z^{ACT} (cyclists has arm up vs arm down). **c** The cyclists' longitudinal position conditioned on Z^{STAT} (cyclist not at intersection vs at intersection)

4.2 Cyclist Approaching Intersection

The second scenario concerns the ego-vehicle driving behind a cyclist, and approaching an intersection. As illustrated in Fig. 1b, the cyclist may or may not turn left at the intersection, but can indicate intent to turn by raising an arm in advance. In our training data, the cyclist always does this when turning in a critical situation where the ego-vehicle is quickly approaching. But in non-critical situations, cyclists may turn even without raising an arm. The context observables of this scenario are illustrated in Figs. 8, and 9 shows the empirical distributions of the observables on the cyclist dataset that will be presented later in Sect. 5.2.

Motion Dynamics The cyclist can switch between the motion types cycling straight, m_{st}, and turning left, m_{tu}. Since a turning cyclist changes velocity in both lateral x and longitudinal y direction, we now include both spatial dimensions in the cyclist's dynamic state. While the pedestrian model included only a latent velocity for the preferred walking speed, our cyclist models includes latent velocities for both motion types. The matrices from Eq. (1) are as follows:

$$X_t = \left[x_t \; y_t \; \dot{x}_t^{m_{tu}} \; \dot{y}_t^{m_{tu}} \; \dot{x}_t^{m_{st}} \; \dot{y}_t^{m_{st}}\right]^\top$$
$$A^{(m_{tu})} = \begin{bmatrix} I & I & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} A^{(m_{st})} = \begin{bmatrix} I & 0 & I \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} \quad B = \begin{bmatrix} I \\ 0 \\ 0 \end{bmatrix}. \tag{29}$$

Here, I defines as a 2×2 identity matrix, and 0 represents a 2×2 matrix of zeros. The observations $Y_t \in \mathbb{R}^2$ from Eq. (2) are the observed lateral and longitudinal position with observation matrix $C = [I \; 0 \; 0]$.

Context In this scenario, the dynamic environment context Z^{DYN} indicates whether a situation would be critical if the cyclist would decide to turn left at that moment. Here we consider time T^{min} it would take for the vehicle to reach the cyclist, if both the vehicle and the cyclist would keep moving at the same speed. This is represented schematically in Fig. 8a. This is not a perfect prediction of the criticality of the situation, because the speed of the cyclist is not constant when turning left. But, similar as the pedestrian case, it still conveys useful information and will therefore improve the prediction. Figure 9a shows that the empirical distribution over T^{min} has multiple modes. We therefore define a mixture of m Gaussians over T^{min}, conditioned the dynamic context state Z^{DYN}. From the data we find that $m = 3$ for $Z^{\text{DYN}} =$ true (the situation is critical), and that $m = 2$ for $Z^{\text{DYN}} =$ false. The Gaussian mixture is parametrized by means $\mu_{zdyn}^{(k)}$, covariances $\sigma_{zdyn}^{(k)}$ and the mixture weights $\phi_{zdyn}^{(k)}$,

$$P\left(E_t^{\text{DYN}} | Z_t^{\text{DYN}} = z\right) = \sum_{k=1}^{m} \phi_{zdyn}^{(k)} \mathcal{N}\left(T_t^{min} | \mu_{zdyn}^{(k)}, \sigma_{zdyn}^{(k)}\right). \tag{30}$$

The object context Z^{ACT} captures whether the cyclist raises an arm to indicate intent to turn left, or not (see Fig. 8b). Accordingly, Z^{ACTED} represents whether the cyclist did raise an arm in the past. For evidence E^{ACT}, an *Arm-Detector* provides a classification score AD_t in the [0, 1] range, where a high score is indicative of a raised arm. The Beta distribution is a suitable likelihood function for this domain, see Fig. 9b. The distribution is parameterized by α_z and β_z,

$$P\left(E_t^{\text{ACT}} | Z_t^{\text{ACT}} = z\right) = \text{Beta}(AD_t | \alpha_z, \beta_z). \tag{31}$$

The static environment context Z^{STAT} represents if the cyclist has reached the region around the intersection where turning becomes possible. Figure 8c shows that cyclist tracks have some variance in their turn angle and the location of onset. Rather than an exact spot, this region is a bounded range on the relative longitudinal distance of the cyclist to the center of the intersection. The dashed lines in the figure mark this region. We rely on map information and ego-vehicle localization to estimate the longitudinal distance of the cyclist to the next intersection, the *Distance-To-Intersection*, DTI_t. As shown in Fig. 9c, the distribution over DTI_t given Z^{STAT} is also modeled as a mixture of m Gaussians, using $m = 1$ for $Z^{\text{STAT}} =$ true, and $m = 2$ for $Z^{\text{STAT}} =$ false:

$$P\left(E_t^{\text{STAT}} | Z_t^{\text{STAT}} = z\right) = \sum_{k=1}^{m} \phi_{zstat}^{(k)} \mathcal{N}\left(DTI_t | \mu_{zstat}^{(k)}, \sigma_{zstat}^{(k)}\right). \tag{32}$$

For path prediction we can estimate DTI_t using the predicted cyclist positions and static map information.

5 Datasets and Feature Extraction

The experiments in this paper used two stereo-camera datasets of VRU encounters recorded from a moving vehicle, one for the crossing pedestrian and one for the cyclist at intersection scenario. Due to the focus on potentially critical situations, both driver and pedestrian/cyclist were instructed during recording sessions. A sufficient safety distance between vehicle and VRU was applied in all scenarios recorded. In the following sections, ‘critical situation’ thus refers to a theoretic outcome where both the approaching vehicle and pedestrian would not stop.

5.1 Pedestrian Dataset

For pedestrian path prediction, we use a dataset (c.f. Kooij et al. 2014a) consisting of 58 sequences recorded using a stereo camera mounted behind the windshield of a vehicle (baseline 22 cm, 16 fps, 1176 × 640 12-bit color images). All sequences involve single pedestrians with the intention to cross the street, but feature different interactions (Critical vs. Non-critical), pedestrian situational awareness (Vehicle seen vs. Vehicle not seen) and pedestrian behavior (Stopping at the curbside vs. Crossing). The dataset contains four different male pedestrians and eight different locations. Each sequence lasts several seconds (min / max / mean: 2.5 s / 13.3 s / 7.2 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 (Hirschmüller 2008). These positions are then corrected for vehicle ego-motion provided by GPS and IMU, and projected to world coordinates. From this correction we obtain the pedestrian’s GT lateral position, and use the temporal difference as the GT lateral speed.

Table 1 Breakdown of the number of tracks in the pedestrian dataset (c.f. Kooij et al. 2014a) for the four normal sub-scenarios (above the line), and in the anomalous one (below the line)

Pedestrian scenario (58 tracks)			
Sub-scenario			Occurences
Non-critical	Vehicle not seen	Crossing	9
Non-critical	Vehicle seen	Crossing	14
Critical	Vehicle not seen	Crossing	11
Critical	Vehicle seen	Stopping	14
Critical	Vehicle seen	Crossing	10

Table 2 Breakdown of the number of tracks in the cyclist dataset for the normal (above the line) and anomalous (below the line) sub-scenarios

Cyclist scenario (42 tracks)			
Sub-scenario			Occurrences
Non-critical	Arm not raised	Straight/Turn	6/6
Non-critical	Arm raised	Turn	6
Critical	Arm not raised	Straight	10
Critical	Arm raised	Turn	7
Critical	Arm not raised	Turn	7

The GT for context observations is obtained by labeling the head orientation of each pedestrian. The 16 labeled discrete orientation classes were reduced to 8 GT orientation bins by merging three neighbored orientation classes (c.f. Flohr et al. 2015) together.

This 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’ with respect to the normal sub-scenarios, since the pedestrian crosses even though he is aware of the critical situation.

5.2 Cyclist Dataset

A new dataset was collected for the cyclist scenario, in a similar fashion to the pedestrian dataset. This new dataset contains 42 sequences with another stereo camera setup in the vehicle (baseline 21 cm, 16 fps, 2048 × 1024 12-bit color images). The cyclist and vehicle are driving on the same road, such that the cyclist is observed from the back, and they approach an intersection with an opportunity for the cyclist to turn left.

The cyclist GT positions are obtained similarly to the pedestrian scenario from stereo vision. To obtain information about the road layout further ahead, intelligent vehicles can rely on map information and self-localization. Since the cyclist scenario was collected in a confined road area, we use Stereo ORB-SLAM2 (Mur-Artal and Tardós 2017) on all collected stereo video to build a 3D map of the environment for our experiments. This results in a fixed world coordinate system shared by all tracks. The spatial layout of the crossing (road width and intersection point) is expressed in these world coordinates, and the detected cyclist positions can be projected to this global coordinate system too. In a pre-processing step GT cyclist tracks are smoothed to compensate for the estimation noise for stereo vision, which especially affects the longitudinal position. The aligned road layout and cyclist tracks are shown in Fig. 8c.

This dataset is also divided into several sub-scenarios, with the number of recordings for each sub-scenario listed in Table 2. We consider that initially the cyclist intent is unknown, i.e. whether he will turn or go straight at the intersection. By raising an arm, he can give a visual indication of the intent to turn left. However, the cyclist might not always properly raise an arm in non-critical situations. Therefore, in non-critical situations without raising an arm, our data contains an equal number of tracks with turning and going straight. In summary, the normal sub-scenarios reflect situations where the cyclist must indicate intent in critical situations with the approaching ego-vehicle, but could neglect to do this in non-critical cases. The additional anomalous sub-scenario contains a turning cyclist in a critical situation, without having raised an arm.

5.3 Feature Extraction

Both cyclist and pedestrian are detected by using neural networks with local receptive fields (Wöhler and Anlauf 1999), 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 position in world coordinates is then used as positional observation Y_t.

For an estimation of the pedestrian head orientation HO_t, the method described in Flohr et al. (2015) is used. 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, i.e. $f_0, \cdots, f_{315}$, using again neural networks with local receptive fields. 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. 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. Flohr et al. 2015). 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 orientation confidences of the selected region.

The expected minimum distance D^{min} between pedestrian and vehicle is calculated as in Pellegrini et al. (2009) 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 (Duda and Hart 1972). Though other approaches are available, e.g. stereo (Oniga et al. 2008) or scene segmentation (Cordts et al. 2016), this simple approach was already sufficient for our experiments. The image region of interest is determined by the specified accuracy of the vehicle localization using typical on-board sensors (GPS+INS) and map data (Schreiber et al. 2013). Y_t^{curb} is then the mean lateral position of the detected line back-projected to world coordinates.

To determine whether the cyclist raises an arm ($AD_t = 1$), or not ($AD_t = 0$), we apply the chamfer matching approach from Gavrila and Giebel (2002). First, a binary foreground segmentation of the cyclist is generated from the disparity values in the tracked cyclist bounding box r. The foreground consists of all pixels with a disparity in the range of $[\tilde{d}_r - \epsilon, \tilde{d}_r + \epsilon]$. Here $\tilde{d}_r$ is the median disparity value in region r. We set $\epsilon_D = 1.5$ to account for disparity errors. The binary segmentation is then matched against multiple rectangular contour templates near the expected shoulder location in the bounding box. These arm templates vary in length, width and angle. The arm detector AD_t is the output of a Naive Bayesian Classifier which integrates several likelihood terms over all templates: a Gamma distribution for the chamfer matching score, and a Gaussian mixture for both the intensity and disparity values in the segmented foreground. This classifier uses a uniform prior.

5.4 Parameter Estimation

Estimating the parameters of the conditional distributions is straightforward, if the values of the latent variables are known. We have therefore annotated the dataset with ground truth (GT) labels for all latent variables in the sequences. During training, the distributions are then fitted on the training data using maximum likelihood estimation. The Expectation-Maximization (Dempster et al. 1977) algorithm is used to fit the Gaussian mixtures. We now explain for both scenarios how the GT labels were obtained.

5.4.1 Pedestrian Scenario

Sequences where potentially critical 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. 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 $M_t = m_w$.

Considering head observation HO, we assume pedestrians recognize an approaching vehicle (GT label $Z_t^{\text{ACT}} = \text{true}$) when the GT head direction is in a range of ±45° around angle 0° (head is pointing towards the camera), and do not see the vehicle ($Z_t^{\text{ACT}} = \text{false}$) for angles outside this range (future human studies could allow a more precise threshold, or provide an angle distribution, the study in Hamaoka et al. (2013) 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 Z_t^{DYN} labels ($\forall_t\ Z_t^{\text{DYN}} = \text{true}$ for trajectories with critical situations, $\forall_t\ Z_t^{\text{DYN}} = \text{false}$ otherwise), and fit 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 the spacial Z_t^{STAT} labels, where $Z_t^{\text{STAT}} = \text{true}$ only at time instances where $-1 \leq \text{TTE} \leq 1$ when crossing, and TTE ≥ -1 when stopping.

The histogram of the GT distributions and the estimated fits can be seen in Fig. 7.

5.4.2 Cyclist Scenario

The turning cyclists have TTE $= 0$ defined at the frame where it is first visible that they are turning. For the cyclists going straight, TTE $= 0$ is defined as the first frame where they pass the point at which 25% of all turning cyclists have passed their TTE $= 0$. For all turning cases, the GT switching state is defined as $M_t = m_{tu}$ at moments with TTE ≥ 0. All other moments, and all straight cases have their GT state defined as $M_t = m_{st}$ These average turning velocity of a track is estimated on its frames where $M_t = m_{tu}$. The prior for the speed of the turning cyclist is estimated on these average turning velocities.

The GT for Z_t^{ACT} is taken from annotated GT arm angles. When the arm is raised further than 30°, $Z_t^{\text{ACT}} = \text{true}$. Below 30°, the arm is considered down, or $Z_t^{\text{ACT}} = \text{false}$.

We define one value for all Z_t^{DYN} labels of a specific track. It is set as true if the ego-vehicle would overtake the cyclist in two seconds at TTE $= 0$, assuming the cyclist has no more longitudinal speed. This can be interpreted as the worst-case scenario, where a cyclist would make an instant 90-degree turn.

Finally, the spatial extent of the turning region is determined from smallest and largest longitudinal position of all turning cyclists at TTE $= 0$. The GT label for Z^{STAT} is set to true whenever the cyclist is in this region.

5.4.3 Both Scenarios

In both scenarios, we compute maximum likelihood estimates of the parameters for the priors and noise distributions using the GT position and speed profiles. For each pedestrian track, we take the average speed during the walking motion type as GT preferred walking speed $\dot{x}^{m_w}$. Similarly, for each cyclist track we take averages of their GT speeds during each motion type as GT of the preferred speeds $\dot{x}_t^{m_{tu}}$, $\dot{y}_t^{m_{tu}}$, $\dot{x}_t^{m_{st}}$ and $\dot{y}_t^{m_{st}}$.

With all continuous states X_t fully defined for all tracks, we can compute the ϵ_t using Eq. (1), i.e. $B\epsilon_t = X_t - A^{(M_t)}X_{t-1}$. From these ϵ_t the process noise covariance Q is estimated. Likewise, observation noise covariance R is estimated from the differences between GT and measured positions, since $\eta_t = Y_t - CX_t$ from Equation(2). Finally, the mean and covariance parameters of the state priors can be estimated from the X_0 of all training tracks.

The prior and transition probability tables for the discrete context states Z^{ACT}, Z^{STAT} are obtained by counting and normalizing the occurrences in the GT labels. The same applies to the dynamic switching state M, conditioned on Z^{ACTED}, Z^{DYN} and Z^{STAT}. The transition probability for Z^{ACTED} is a logical OR, as described in Sect. 3.3. Since we only got one Z^{DYN} label per track, we fix the Z^{DYN} transition probability to $1/100$ for changing state.

6 Experiments

In the experiments, we compare the proposed DBNs using all context cues to variants using less cues, and to baseline approaches. We also compare the use of visual detections to using GT annotations as measurements, and we investigate computational performance.

In all experiments, *leave-one-out* cross-validation is used to separate training and test sequences. Sequences from anomalous sub-scenarios are always excluded from the training data. For each time t with state X_t, we create a predictive distribution $\widetilde{P}_{t_p|t}(X_{t+t_p})$ for prediction horizon t_p. Two performance metrics are used to evaluate a sequence, namely the Euclidean distance between predicted expected position Y_{t+t_p} and GT position G_{t+t_p}, and the log likelihood of G under the predictive distribution:

$$error(t_p|t) = |\mathbb{E}\left[\widetilde{P}_{t_p|t}(Y_{t+t_p})\right] - G_{t+t_p}| \tag{33}$$

$$predll(t_p|t) = \log\left[\widetilde{P}_{t_p|t}(G_{t+t_p})\right]. \tag{34}$$

6.1 Pedestrian Scenario

We first investigate the influence of context on the pedestrian scenario. The proposed DBN with full context is compared to model variants that only uses the head orientation, to one that only uses the minimal distance to the pedestrian, and to one that combines these two contexts. We refer to these variants after the used measurements: E^{ACT}, E^{DYN} and E^{DYN}+E^{ACT} respectively. We also include a common 2nd-order LDS as additional baseline. The dynamic process of this LDS for the lateral pedestrian position can be described as

$$\begin{bmatrix} x_t \\ \dot{x}_t \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_{t-1} \\ \dot{x}_{t-1} \end{bmatrix} + \epsilon_t \qquad \epsilon_t \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, Q\right). \tag{35}$$

where the process noise covariance is also estimated from the GT position and speed.

6.1.1 Comparison of Model Variations

The results in Table 3 show the predictive log likelihood $predll$ for $t_p = 16$ time steps (~ 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, 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 sub-scenario, only the proposed model results in *lower* likelihood than for normal behavior, which is a useful property for anomaly detection as mentioned in Sect. 3.1. A future driver warning strategy could benefit from the more accurate path prediction of our full model in high likelihood situations, whereas falling back to simpler models/strategies when anomalies are detected.

6.1.2 Detailed Analysis on a Single Track

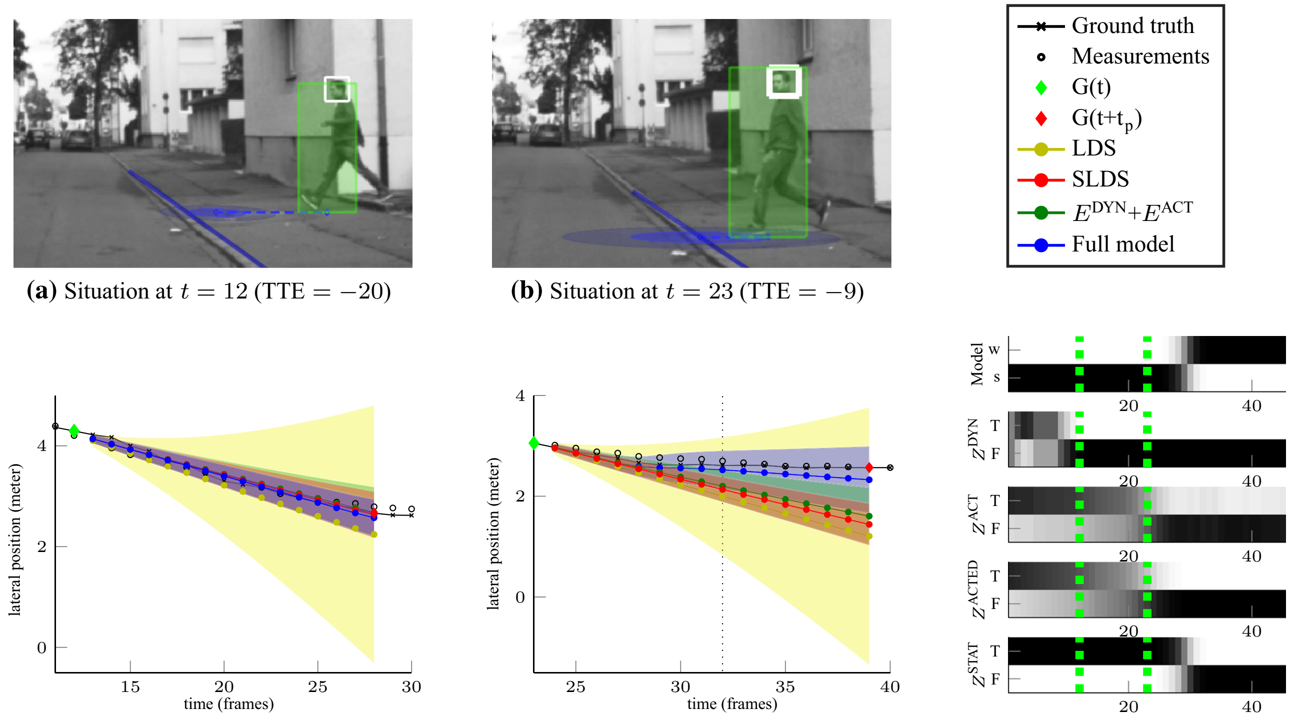
Fig. 10 illustrates a sequence from the stopping sub-scenario (fourth row in Table 3), with a snapshot just *before* (TTE = – 20) and *after* (TTE = – 9) the pedestrian becomes aware of the critical situation. At TTE = – 20, the predicted distributions of all models are close together and indicate that

Table 3 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 $\in [-15, 0]$

Sub-scenario				Full model	$E^{\text{DYN}}+E^{\text{ACT}}$	E^{ACT}	E^{DYN}	SLDS	LDS
Normal	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
		Over all normal sub-scenarios		**− 0.51**	− 0.52	− 0.58	− 0.61	− 0.66	− 1.90
Anomalous	Critical	Vehicle seen	Crossing	**− 0.90**	− 0.27	− 0.15	− 0.25	− 0.13	− 1.88

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, no curb ($E^{\text{DYN}}+E^{\text{ACT}}$), only head ($E^{\text{ACT}}$), only criticality ($E^{\text{DYN}}$), no context (SLDS), KF (LDS)

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 $E^{\text{DYN}}+E^{\text{ACT}}$ 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



(a) Situation at $t = 12$ (TTE $= -20$)

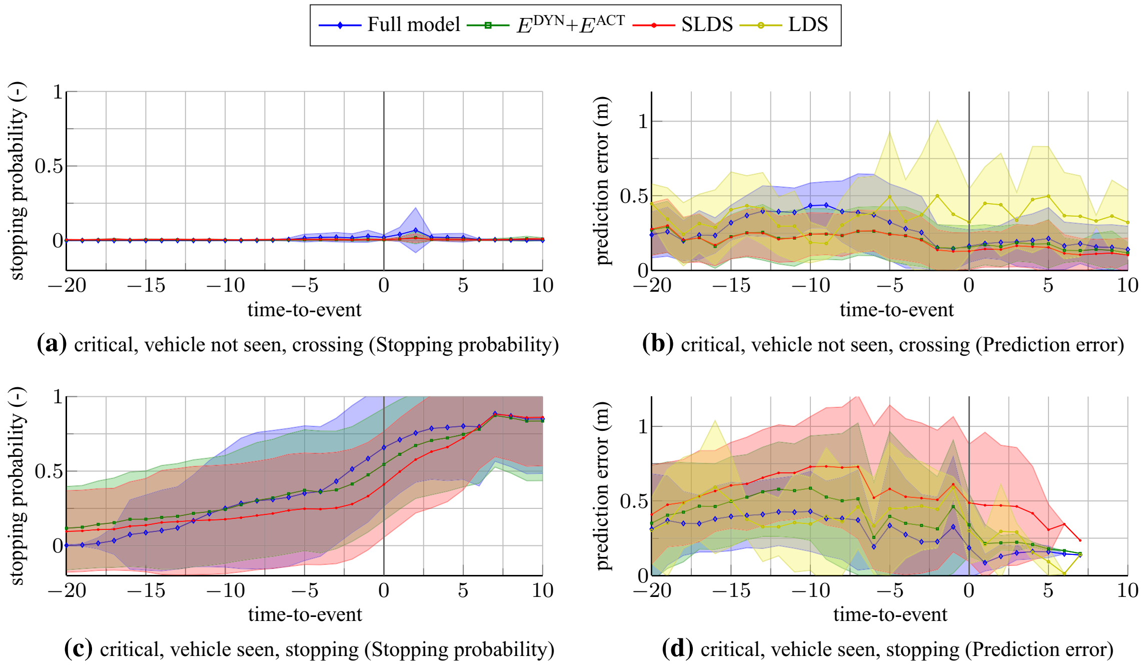
(b) Situation at $t = 23$ (TTE $= -9$)

(c) Predicted position at $t = 12$ (TTE $= -20$)

(d) Predicted position at $t = 23$ (TTE $= -9$)

(e) Inferred marginal distributions

Fig. 10 Best viewed in color. Example of a pedestrian stopping at the curb after becoming aware of a critical situation, see also the marginal distributions in **e**. Predictions are made $t_p = 16$ ($\sim$ 1 s) time steps ahead from different times t. **a** Pedestrian with head detection bounding box (white), tracking bounding box (green), collapsed predicted distribution of the full model (blue ellipses show one and two std. dev.) and curb detection (blue line). **b** The pedestrian became aware of the critical situation. **c** Predictions [mean and std. dev. (shaded)] made at $t = 12$ (green 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. **d** Predictions of the lateral position at $t + t_p$ made from $t = 23$. **e** Marginal distributions of the full model latent variables over time [probabilities range from 0 (black) to 1 (white)]. The dotted green lines are set on the time of the two snapshots, $t = 12$ and $t = 23$. Variable labels are **T**rue and **F**alse, and **w**alking and **s**tanding



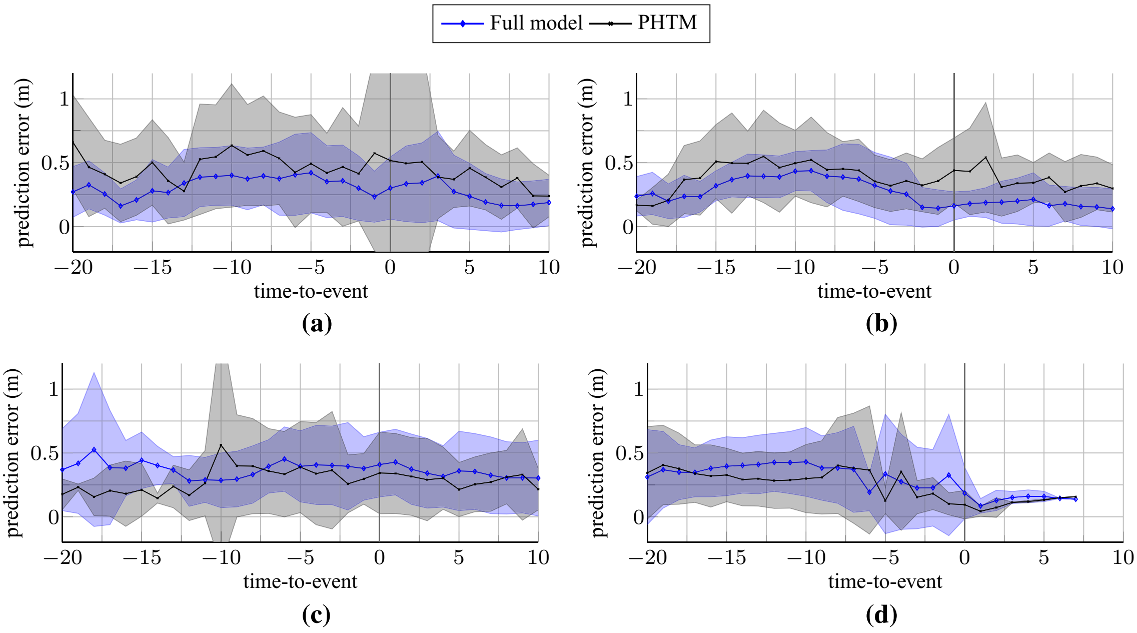
(a) critical, vehicle not seen, crossing (Stopping probability)

(b) critical, vehicle not seen, crossing (Prediction error)

(c) critical, vehicle seen, stopping (Stopping probability)

(d) critical, vehicle seen, stopping (Prediction error)

Fig. 11 Best viewed in color. Pedestrian scenario: stopping probability (left) and lateral prediction error (right) when predicting 16 time steps (∼ 1 s) ahead in the two critical sub-scenarios. **a**, **b** Pedestrian is not aware of the critical situation and crosses. **c**, **d** 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 full model, an intermediate model without spatial layout information (E^{DYN}+E^{ACT}), the baseline SLDS model without context cues, and a LDS



(a)

(b)

(c)

(d)

Fig. 12 Best viewed in color. Pedestrian scenario: 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. **a** Non-critical, vehicle seen, crossing. **b** Critical, vehicle not seen, crossing. **c** Non-critical, vehicle not seen, crossing. **d** Critical, vehicle seen, stopping

full model makes the best predictions as it also anticipates where the pedestrian will stop, namely at the curbside.

For instance, at $t = 23$, Z_t^{ACTED} starts to change as it becomes more probable that the pedestrian has seen the vehicle, and indeed around $t = 29$ the pedestrian stops at the curb.

6.1.3 Comparison Over Time

In the context of action classification, Fig. 11 shows for various model variations, the standing probability $\tilde{P}_t(M_t = m_s)$, and 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 (Fig. 11a), 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 (Fig. 11b) 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 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 (Fig. 11c) 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 pedestrian awareness, interaction, 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 (Fig. 11d). Averaged over the sequences, it outperforms the baseline SLDS model by up to 0.39 m (at TTE $= 1$) and the E^{DYN}+E^{ACT} model by up to 0.16 m (at TTE $= -10$).

6.1.4 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: the use of GT as measurements does not notably improve the results. Ideal head measurements alter the five sub-scenario scores of the full model w.r.t. Table 3 to − 0.57, − 1.08, − 0.32, − 0.12 (“normal” cases), and to − 3.67 (anomalous case). Predictions became more accurate for critical sub-scenarios. However the second sub-scenario (Non-critical, Vehicle seen, Crossing) became less accurate, as some moments were deemed critical, and seeing the vehicle implied stopping, Still, the likelihood of the anomalous fifth sub-scenario is much lower than all other sub-scenarios.

Table 4 Pedestrian scenario. Computational costs for the different models per frame (avg. per frame, in *ms*)

Approach	Observables	State est. & pred.	Total
Full model	160	40	200
SLDS	60	10	70
LDS	60	0.4	60
PHTM	70	600	670

6.1.5 Comparison with PHTM and Computational Cost

Fig. 12 shows a comparison of the mean prediction error of our proposed model with the state-of-the-art PHTM model (Keller and Gavrila 2014) which uses optical flow features and an exemplar database, on the four “normal” sub-scenarios. On two of these sub-scenarios (Fig. 12a, b) 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 other two sub-scenarios (Fig. 12c, d) PHTM performs slightly better.

The computational costs of the various approaches were assessed on standard PC hardware (Intel Core i7 X990 CPU at 3.47 GHz), see Table 4. We differentiate between the computational cost for obtaining the observables and that for 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 full 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 4, 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 (Keller and Gavrila 2014)], and it is three times more efficient in total.

6.2 Cyclist Scenario

To demonstrate that our approach is not specific for the crossing pedestrian scenario, we compare the full model to SLDS and LDS baselines on the cyclist scenario. Like the pedestrian scenario, we use an LDS baseline with the dynamic model from Eq. (35), except that the state is extended to $[x_t, \dot{x}_t, y_t, \dot{y}_t]$.

6.2.1 Comparison with Baselines

Unlike stopping pedestrians, turning cyclists remain in the vehicle's view and driving corridor for a longer time. Also, the path during a turn slowly diverges from the straight path. Therefore, we extend the duration over which we summarize predictions with 15 time steps (frames). Another difference between the cyclist and pedestrian scenario is that cyclist's predictive distributions are 2D multivariate Gaussians for lateral and longitudinal position, instead of 1D Gaussians on the lateral position only. Hence, the reported likelihood values cannot be compared directly between scenarios, as the underlying domains have different dimensions.

The log likelihoods of the one-second-ahead prediction (16 time steps) are shown in Table 5, averaged over TTE $\in [-15, 15]$, i.e. starting with prediction for the moment when the cyclist either starts to turn, or keeps moving straight.

Since the cyclist tracks are long, and mostly consist of moving straight, the LDS predictions reflect the common behavior of straight motion with little variance. As a result, its predictions are accurate when the cyclist indeed does not turn. As expected, this comes at the cost of inaccurate predictions in normal and anomalous sub-scenarios where the cyclist does turn.

Compared to the LDS, the switching models demonstrate the benefit of having separate dynamics for straight and turning motion, as the predictions for turning sub-scenarios are considerably more accurate. Furthermore, our model outperforms the SLDS in almost all but one normal sub-scenarios. On the non-critical sub-scenario where there is ambiguity on whether the cyclist will turn or go straight, the SLDS performs best. Still, the proposed method performs best overall on all normal sub-scenarios, demonstrating that context does improve prediction accuracy generally compared to the LDS and SLDS baselines.

We also note that all models obtain lower predictive likelihood for turning than for moving straight. We find that this is a result from the large variance in how cyclists execute the turn. The data shows the cyclists vary in when they initiate the turn, and in used turning speed and angle. This variance is also reflected in the predictive distributions, which show larger uncertainty for turning than for moving straight.

During the evaluation period of Table 5 around $TTE = 0$, the cyclist is always near to the intersection. We note that our model outperforms the baselines in all sub-scenarios when including all earlier predictions ($TTE < -15$). When the cyclists are still far from the intersection, our full model benefits from the static environment context which predicts that turning is not feasible yet.

The final row of the table illustrates the predictive likelihood for the anomalous sub-scenario where the cyclist turns in a critical situation without raising an arm. Here, our model performs more similar to the LDS, as both expect the cyclist to continue moving straight. The next session will show a more detailed analysis of all sub-scenarios.

6.2.2 Comparison Over Time

Figure 13 compares the prediction error over time around $TTE = 0$ for all normal sub-scenarios. For the critical sub-scenario where the cyclist maintains its straight path, Fig. 13a, all models demonstrate low prediction errors. The LDS shows lowest Euclidean error, but it has larger prediction uncertainty as Table 5 showed. In the critical and non-critical sub-scenarios where the cyclist raises an arm, our model anticipates the turn, and predicts some lateral motion to the left as soon as the cyclist is near the intersection. As a result, the model has slightly larger prediction error than the baselines before the turn is initiated, but yields lower errors as soon as turnings starts, see Fig. 13b, c. It keeps an advantage over SLDS for almost a full second after the turn started, until approximately $TTE = 13$. On the critical sub-scenario, the largest difference in average error of 0.41 m occurs at $TTE = 5$. On the non-critical sub-scenario without the cyclist raising an arm, Fig. 13d, the context cannot uniquely distinguish if the cyclist will turn or continue moving straight. Indeed, here the SLDS and our model also show similar performance.

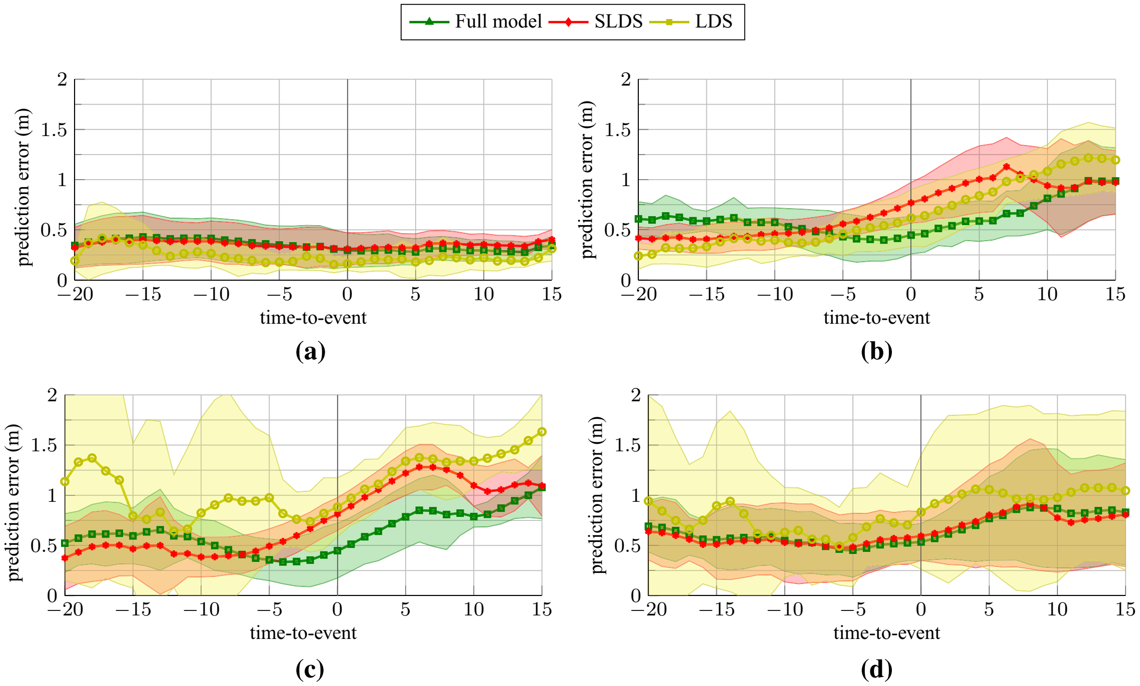
Figure 14 shows the prediction log likelihood for two sub-scenarios where the cyclist turns in critical situations, namely the normal sub-scenario where the turn happens after raising an arm (Fig. 14a), and the anomalous sub-scenario where the turn happens without raising an arm (Fig. 14b). In both cases, the LDS likelihood drops fast, as it predicts continuation of the past motion instead of turning. Our full model also expects moving straight in case the arm was not raised, therefore its predictive likelihood declines too for the anomaly. Interestingly, the model does adapt after $TTE = 0$ when the turning behavior becomes apparent. Later, at $TTE = 10$, the predictive log likelihood of all models drops due to the variation in turning behavior.

Finally, we note that the LDS shows larger errors for the non-critical sub-scenarios. We observe that in these cases, where the vehicle is generally further away, the longitudinal estimates from stereo-vision are more unstable. The LDS is more sensitive to such noise as it adapts to all observed speed changes.

Table 5 Cyclist scenario

Sub-scenario				Full model	SLDS	LDS
Normal	Critical	Arm not raised	Straight	**0.05**	– 0.23	0.00
	Non-critical	Arm raised	Turn	– **2.36**	– 3.19	– 22.75
	Critical	Arm raised	Turn	– **2.38**	– 2.77	– 16.66
	Non-critical	Arm not raised	Straight/Turn	– 2.28	– **2.22**	– 14.88
		Over all normal sub-scenarios		– **1.93**	– 2.22	– 14.60
Anomalous	Critical	Arm not raised	Turn	– 14.33	– 4.62	– 31.91

Log likelihood at 16 time steps (∼ 1 s) into the future for TTE ∈ [– 15, 15]. Below sub-scenarios indicate the expected behavior of a cyclist. Best model performance are given in bold



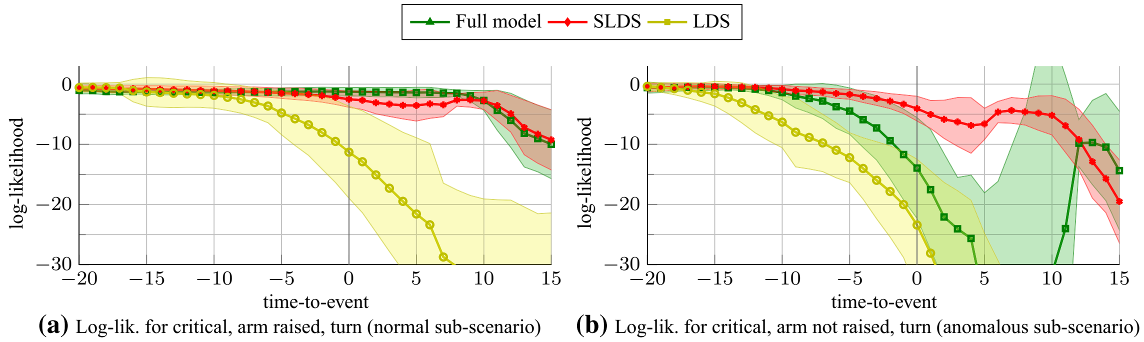
(a)

(b)

(c)

(d)

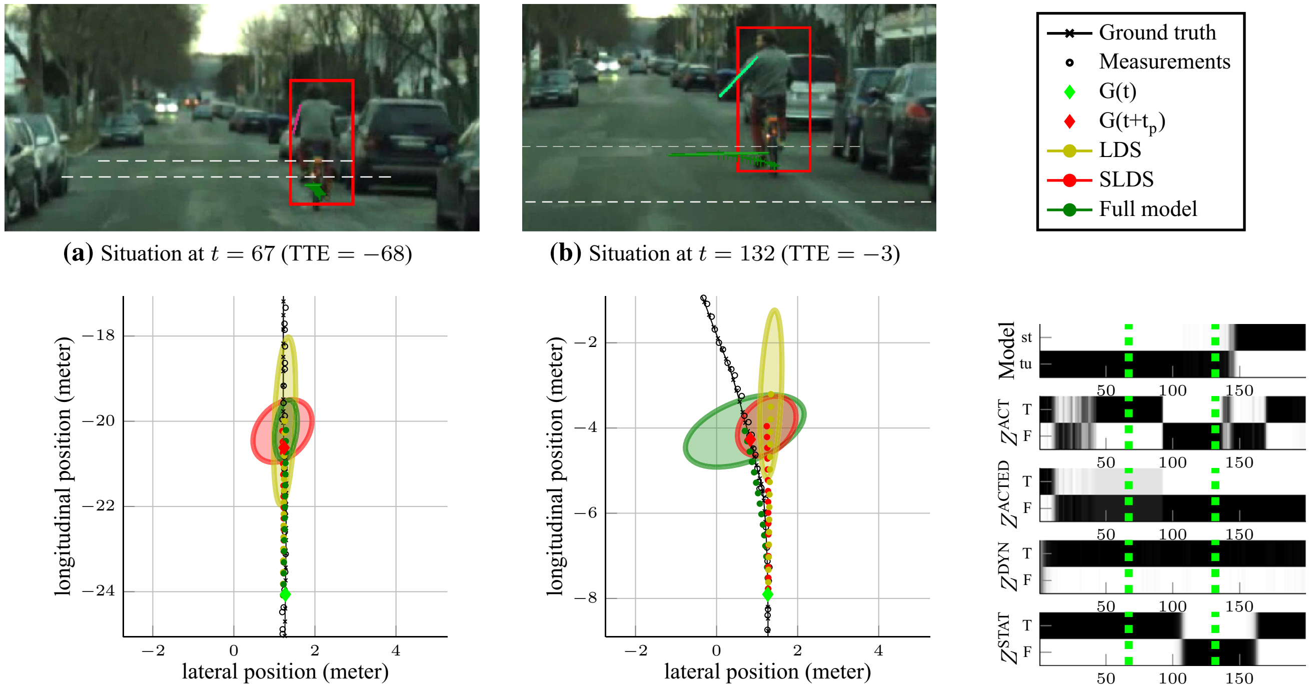
Fig. 13 Cyclist scenario. Prediction error on the normal sub-scenarios when predicting 16 time steps (∼ 1 s) ahead. **a** Critical, arm not raised, straight. **b** Critical, arm raised, turn. **c** Non-critical, arm raised, turn. **d** Non-critical, arm not raised, straight/turn



(a) Log-lik. for critical, arm raised, turn (normal sub-scenario)

(b) Log-lik. for critical, arm not raised, turn (anomalous sub-scenario)

Fig. 14 Best viewed in color. Cyclist scenario: Prediction likelihood when predicting 16 time steps (∼ 1 s) ahead on critical sub-scenarios where the cyclist turns. **a** Normal sub-scenario where the arm was raised. **b** Anomalous sub-scenario where the cyclist did not raise an arm to express intent



(a) Situation at $t = 67$ (TTE $= -68$)

(b) Situation at $t = 132$ (TTE $= -3$)

(c) Predicted position at $t = 67$ (TTE $= -68$)

(d) Predicted position at $t = 132$ (TTE $= -3$)

(e) Inferred marginal distributions

Fig. 15 Best viewed in color. Example of a cyclist turning in after holding his arm up, in a non-critical situation. Predictions are made sixteen time steps (∼ 1 s) into the future. **a** Cyclist with tracking bounding box (red), arm detection (red line), collapsed predicted distribution of the full model (green ellipsoid shows one standard deviation), and "at intersection" region (white dotted line). **b** the cyclist enters the intersection region, which results in a wider predictive distribution shifted to the left. **c** Predictions (mean, and shaded std. dev.) made at $t = 67$ (green diamond) for the lateral position up to 16 time steps into the future. The red diamond indicates the corresponding future GT position. **d** Predictions made at $t = 132$, the moment where the cyclist starts turning. **e** Marginal posterior distributions of the latent variables in our full model over time. Probabilities range from 0 (black) to 1 (white). Variable labels are **T**rue and **F**alse, and **st**raight and **tu**rning

6.2.3 Detailed Analysis on a Single Track

Figure 15 shows two snapshots of the prediction over time for a cyclist who is turning at the intersection after holding his arm up while the situation is not critical. Before the cyclist arrives at the intersection, at $t = 67$, the prediction of the full model is very specific. Even though it was already detected that the cyclist raised his arm, he is expected to keep moving straight as he is not yet close enough to the intersection to turn. This prediction is done with less uncertainty than the baseline LDS because the process noise on the LDS must also account for the parts where the cyclist is turning left. At $t = 132$, when the cyclist is at the intersection, there is an increased probability that he could turn left. As a result, the expected future position also shifts left, and the uncertainty region of the prediction increases. The one standard deviation area of the prediction reflects the possibility that the cyclist may still move straight before turning.

7 Discussion

Our DBN provides predictive distributions of the future position of the tracked objects, reflecting uncertainty on possible trajectories. But as our results show, different scenarios give rise to different sources of uncertainty. For instance, if the system anticipates a change in dynamics, the future stopping position near the curb can be determined quite accurately for the pedestrian scenario. However, in the cyclist scenario there is more variance in turning behavior than in moving straight ahead, making accurate path predictions for the anticipated switch in dynamics more challenging. The available context may also be insufficient to unambiguously predict the behavior, as was seen in a cyclist sub-scenario. Here, our model's performance approximated that of the SLDS baseline.

Another property of our model is that it is not a black box, but explicitly defines the relation between context observables, states, motion modes, and their dynamics. This ensures that a designer can inspect the system, investigate how it assesses the context, and determine causes for failure or success. The explicit formulation also facilitates adding additional cues, or improve individual parts (e.g. using different dynamical models) while keeping existing parts unchanged. The generic formulation also ensures that many forms of information can be included, from up-to-date map data, to past image classification results. Accordingly, our method and the PHTM approach do not stand directly in competition, as they use different sources of information that could conceivably be combined.

Anomalous sub-scenarios contain situations not represented by the training data. In anomalous sub-scenarios, our method predicts position with lower likelihood than for the normal sub-scenarios. Therefore, low-likelihood predictions can be used to detect anomalous behavior, for instance to switch to an emergency control strategy of the vehicle. Of course, anomalous situations are expected to be rare, since the training data is expected to be representative of 'normal' behavior. Larger realistic datasets should therefore provide better estimates of 'normal' behavior, though the principle demonstrated on our example scenarios should remain the same.

Future 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 in different directions). Indeed, initial work has already begun on extending our earlier conference (Kooij et al. 2014a) submission, which only considered the pedestrian crossing scenario. For instance, (Roth et al. 2016) incorporated driver attention in the same framework, and (Hashimoto et al. 2016) tackled additional pedestrian scenarios with similar DBNs.

8 Conclusions

This paper investigated the use of DBNs for path prediction of maneuvering objects. As a toy example illustrated, SLDSs can make good overall predictions, but prediction accuracy suffers when an actual change in dynamics occurs. We therefore proposed to condition the switching probability on dynamic context variables. Measurements of various visual cues inform the model if and when changes in dynamics are likely to occur. An efficient approximate inference method was presented for online path prediction. Parameters are estimated on annotated training data.

We validated our approach on two use cases of VRU path prediction in the IV domain, a pedestrian and a cyclist scenario. Combining several context cues proved to improve overall prediction accuracy, namely the VRU's interaction with the ego-vehicle as dynamic environment context, the VRU's location in the static environment, and the VRU behavior indicating awareness or intent. The experiments demonstrate that the expected benefits of the context also occur with real-world vehicle measurements, and when using features extracted from vision. Compared to the SLDS, when predicting up to $\sim 1\,\text{s}$ ahead, our method improved up to $0.39\,\text{m}$ for the pedestrian scenario, and up to $0.41\,\text{m}$ for the cyclist scenario. It also slightly outperforms the PHTM approach at less than a third of computational cost.

We are encouraged that the presented context-based models can play an important role in saving lives for the future intelligent vehicles.

Acknowledgements The research leading to the results of this work has received funding from the European Communitys Eighth Framework Program (Horizon2020) under Grant Agreement No. 634149, the PROSPECT project.

References

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., & Savarese, S. (2016). Social LSTM: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 961–971).

Althoff, M., Stursberg, O., & Buss, M. (2009). Model-based probabilistic collision detection in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, *10*(2), 299–310.

Antonini, G., Martinez, S. V., Bierlaire, M., & Thiran, J. P. (2006). Behavioral priors for detection and tracking of pedestrians in video sequences. *International Journal of Computer Vision*, *69*(2), 159–180.

Ba, S., & Odobez, J. (2011). Multiperson visual focus of attention from head pose and meeting contextual cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(1), 101–116.

Ballan, L., Castaldo, F., Alahi, A., Palmieri, F., & Savarese, S. (2016). Knowledge transfer for scene-specific motion prediction. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 697–713). Springer.

Bandyopadhyay, T., Won, K., Frazzoli, E., Hsu, D., Lee, W., & Rus, D. (2013). Intention-aware motion planning. In E. Frazzoli, T. Lozano-Perez, N. Roy, & D. Rus (Eds.), *Algorithmic foundations of robotics X* (pp. 475–491). Berlin: Springer.

Bar-Shalom, Y., Li, X., & Kirubarajan, T. (2001). *Estimation with applications to tracking and navigation*. Hoboken: Wiley-Interscience.

Benfold, B., & Reid, I. (2009). Guiding visual surveillance by tracking human attention. In *Proceedings of the British machine vision conference (BMVC)*

Bishop, C. M. (2006). *Pattern recognition and machine learning* (Vol. 1). Berlin: Springer.

Blackman, S., & Popoli, R. (1999). *Design and analysis of modern tracking systems*. Norwood: Artech House Norwood.

Bonnin, S., Weisswange, T. H., Kummert, F., & Schmuedderich, J. (2014). General behavior prediction by a combination of scenario-specific models. *IEEE Transactions on Intelligent Transportation Systems*, *15*(4), 1478–1488.

Boyen, X., & Koller, D. (1998). Tractable inference for complex stochastic processes. In *Proceedings of uncertainty in artificial intelligence (UAI)* (pp. 33–42). Morgan Kaufmann Publishers Inc.

Braun, M., Rao, Q., Wang, Y., & Flohr, F. (2016). Pose-RCNN: Joint object detection and pose estimation using 3d object proposals. In

Proceedings of the IEEE intelligent transportation systems conference (pp. 1546–1551).

Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.

Cara, I., & de Gelder, E. (2015). Classification for safety-critical car-cyclist scenarios using machine learning. In *Proceedings of the IEEE intelligent transportation systems conference* (pp. 1995–2000).

Chen, B., Zhao, D., & Peng, H. (2017). Evaluation of automated vehicles encountering pedestrians at unsignalized crossings. In *Proceedings of the IEEE intelligent vehicles symposium*

Cho, H., Rybski, P. E., & Zhang, W. (2011). Vision-based 3D bicycle tracking using deformable part model and interacting multiple model filter. In *Proceedings of the international conference on robotics and automation (ICRA)* (pp. 4391–4398). IEEE.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., et al. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3213–3223).

Dempster, A., Laird, N., & Rubin, D. B. (1977). Maximum-likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, *39*, 1–38.

Dollár, P., Wojek, C., Schiele, B., & Perona, P. (2012). Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(4), 743–761.

Duda, R. O., & Hart, P. E. (1972). Use of the Hough transformation to detect lines and curves in pictures. *Communications of ACM*, *15*(1), 11–15.

Enzweiler, M., & Gavrila, D. M. (2009). Monocular pedestrian detection: Survey and experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(12), 2179–2195.

Enzweiler, M., & Gavrila, D.M. (2010). Integrated pedestrian classification and orientation estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 982–989). IEEE.

Flohr, F., Dumitru-Guzu, M., Kooij, J. F. P., & Gavrila, D. M. (2015). A probabilistic framework for joint pedestrian head and body orientation estimation. *IEEE Transactions on Intelligent Transportation Systems*, *16*(4), 1872–1882.

Gavrila, D. M., & Giebel, J. (2002). Shape-based pedestrian detection and tracking. In *Proceedings of the IEEE intelligent vehicles symposium* (Vol. 1, pp. 8–14).

Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.

Geiger, A., Lauer, M., Wojek, C., Stiller, C., & Urtasun, R. (2014). 3d traffic scene understanding from movable platforms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(5), 1012–1025.

Hamaoka, H., Hagiwara, T., Tada, M., & Munehiro, K. (2013). A study on the behavior of pedestrians when confirming approach of right/left-turning vehicle while crossing a crosswalk. In *Proceedings of the IEEE intelligent vehicles symposium* (pp. 106–110).

Hashimoto, Y., Gu, Y., Hsu, L. T., Iryo-Asano, M., & Kamijo, S. (2016). A probabilistic model of pedestrian crossing behavior at signalized intersections for connected vehicles. *Transportation Research Part C*, *71*, 164–181.

Helbing, D., & Molnár, P. (1995). Social force model for pedestrian dynamics. *Physical Review E*, *51*(5), 4282.

Hirschmüller, H. (2008). Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(2), 328–341.

Huang, L., Wu, J., You, F., Lv, Z., & Song, H. (2017). Cyclist social force model at unsignalized intersections with heterogeneous traffic. *IEEE Transactions on Industrial Informatics*, *13*(2), 782–792.

Hubert, A., Zernetsch, S., Doll, K., & Sick, B. (2017). Cyclists' starting behavior at intersections. In *IEEE intelligent vehicles symposium (IV)* (pp. 1071–1077). IEEE.

Jacobs, H., Hughes, O., Johnson-Roberson, M., & Vasudevan, R. (2017). Real-time certified probabilistic pedestrian forecasting. *IEEE Robotics and Automation Letters*, *2*, 2064–2071.

Karasev, V., Ayvaci, A., Heisele, B., & Soatto, S. (2016). Intent-aware long-term prediction of pedestrian motion. In *Proceeding of the international conference on robotics and automation (ICRA)* (pp 2543–2549). IEEE.

Keller, C. G., & Gavrila, D. M. (2014). Will the pedestrian cross? A study on pedestrian path prediction. *IEEE Transactions on Intelligent Transportation Systems*, *15*(2), 494–506.

Keller, C. G., Dang, T., Fritz, H., Joos, A., Rabe, C., & Gavrila, D. M. (2011). Active pedestrian safety by automatic braking and evasive steering. *IEEE Transactions on Intelligent Transportation Systems*, *12*(4), 1292–1304.

Kitani, K. M., Ziebart, B. D., Bagnell, J. A., & Hebert, M. (2012). Activity forecasting. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 201–214). Springer.

Klostermann, D., Osep, A., Stückler, J., & Leibe, B. (2016). Unsupervised learning of shape-motion patterns for objects in urban street scenes. In *Proceedings of the British machine vision conference (BMVC)*.

Köhler, S., Schreiner, B., Ronalter, S., Doll, K., Brunsmann, U., & Zindler, K. (2013). Autonomous evasive maneuvers triggered by infrastructure-based detection of pedestrian intentions. In *Proceedings of the IEEE intelligent vehicles symposium* (pp. 519–526).

Kooij, J. F. P., Schneider, N., Flohr, F., & Gavrila, D. M. (2014a). Context-based pedestrian path prediction. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 618–633). Springer International Publishing.

Kooij, J. F. P., Schneider, N., & Gavrila, D. M. (2014b). Analysis of pedestrian dynamics from a vehicle perspective. In *Proceedings of the IEEE intelligent vehicles symposium* (pp. 1445–1450).

Kooij, J. F. P., Englebienne, G., & Gavrila, D. M. (2016). Mixture of switching linear dynamics to discover behavior patterns in object tracks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(2), 322–334.

Lauritzen, S. L. (1992). Propagation of probabilities, means, and variances in mixed graphical association models. *Journal of the American Statistical Association*, *87*(420), 1098–1108.

Lee, N., Choi, W., Vernaza, P., Choy, C. B., Torr, P. H., & Chandraker, M. (2017). DESIRE: Distant future prediction in dynamic scenes with interacting agents. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.

Li, X., Flohr, F., Yang, Y., Xiong, H., Braun, M., Pan, S., et al. (2016). A new benchmark for vision-based cyclist detection. In *Proceedings of the IEEE intelligent vehicles symposium* (pp. 1028–1033). IEEE

Li, X., Li, L., Flohr, F., Wang, J., Xiong, H., Bernhard, M., et al. (2017). A unified framework for concurrent pedestrian and cyclist detection. *IEEE Transactions on Intelligent Transportation Systems*, *18*(2), 269–281.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y. et al. (2016). SSD: Single shot multibox detector. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 21–37). Springer.

Meinecke, M. M., Obojski, M., Gavrila, D. M., Marc, E., Morris, R., Töns, M., et al. (2003). Strategies in terms of vulnerable road user protection. In *EU project SAVE-U*, Deliverable D6.

Menze, M., & Geiger, A. (2015). Object scene flow for autonomous vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.

Meuter, M., Iurgel, U., Park, S. B., & Kummert, A. (2008). Unscented Kalman filter for pedestrian tracking from a moving host. In *Proceedings of the IEEE intelligent vehicles symposium* (pp. 37–42).

Minka, T. P. (2001). Expectation propagation for approximate Bayesian inference. In *Proceedings uncertainty in artificial intelligence (UAI)* (pp. 362–369). Morgan Kaufmann Publishers Inc.

Morris, B. T., & Trivedi, M. M. (2011). Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(11), 2287–2301.

Mur-Artal, R., & Tardós, J. D. (2017). ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Transactions on Robotics*, *33*(5), 1255–1262.

Murphy, K. P. (2002). *Dynamic bayesian networks: Representation, inference and learning*. PhD thesis, University of California, Berkeley.

Oh, S. M., Rehg, J. M., Balch, T., & Dellaert, F. (2008). Learning and inferring motion patterns using parametric segmental switching linear dynamic systems. *International Journal of Computer Vision*, *77*(1–3), 103–124.

Ohn-Bar, E., & Trivedi, M. M. (2016). Looking at humans in the age of self-driving and highly automated vehicles. *IEEE Transactions on Intelligent Vehicles*, *1*(1), 90–104.

Oniga, F., Nedevschi, S., & Meinecke, M. M. (2008). Curb detection based on a multi-frame persistence map for urban driving scenarios. In *Proceedings of the IEEE intelligent transportation systems conference* (pp. 67–72).

Otsuka, K., Hara, K., Suzuki, T., & Aoki, Y. (2017). Danger level modeling and analysis of vehicle-pedestrian encounter using situation dependent topic model. In *Proceedings of the IEEE intelligent vehicles symposium* (pp. 251–256).

Paden, B., Čáp, M., Yong, S. Z., Yershov, D., & Frazzoli, E. (2016). A survey of motion planning and control techniques for self-driving urban vehicles. *IEEE Transactions on Intelligent Vehicles*, *1*(1), 33–55.

Pavlovic, V., Rehg, J. M., & MacCormick, J. (2000). Learning switching linear models of human motion. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems (NIPS)* (pp. 981–987). Massachusetts, US: MIT Press.

Pellegrini, S., Ess, A., Schindler, K., & Van Gool, L. (2009). You'll never walk alone: Modeling social behavior for multi-target tracking. In *Proceedings of the international conference on computer vision (ICCV)* (pp. 261–268).

Pool, E. A. I., Kooij, J. F. P., & Gavrila, D. M. (2017). Using road topology to improve cyclist path prediction. In *Proceedings of the IEEE intelligent vehicles symposium* (pp. 289–296). IEEE.

Rasouli, A., Kotseruba, I., & Tsotsos, J. K. (2017). Agreeing to cross: How drivers and pedestrians communicate. In *Proceedings of the IEEE intelligent vehicles symposium*.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 779–788).

Rehder, E., & Kloeden, H. (2015). Goal-directed pedestrian prediction. In *Proceedings of IEEE international conference on computer vision workshops* (pp. 50–58).

Robicquet, A., Sadeghian, A., Alahi, A., & Savarese, S. (2016). Learning social etiquette: Human trajectory understanding in crowded scenes. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 549–565). Springer.

Rosti, A. V. I., & Gales, M. J. F. (2004). Rao-Blackwellised Gibbs sampling for switching linear dynamical systems. In *Proceedings of IEEE international conference on acoustics, speech, and signal processing (ICASSP)* (Vol. 1, pp. 809–812).

Roth, M., Flohr, F., & Gavrila, D. M. (2016). Driver and pedestrian awareness-based collision risk analysis. In *Proceedings of the IEEE intelligent vehicles symposium* (pp. 454–459).

Sattarov, E., Gepperth, A., & Reynaud, R., et al. (2014). Context-based vector fields for multi-object tracking in application to road traffic. In *Proceedings of the IEEE intelligent transportation systems conference* (pp. 1179–1185).

Sayed, T., Zaki, M. H., & Autey, J. (2013). Automated safety diagnosis of vehicle–bicycle interactions using computer vision analysis. *Safety Science*, *59*, 163–172.

Schmidt, S., & Färber, B. (2009). Pedestrians at the kerb—Recognising the action intentions of humans. *Transportation Research Part F: Traffic Psychology and Behaviour*, *12*(4), 300–310.

Schneider, N., & Gavrila, D. M. (2013). Pedestrian path prediction with recursive Bayesian filters: A comparative study. In J. Weickert, M. Hein, & B. Schiele (Eds.), *Lecture notes in computer science* (Vol. 8142, pp. 174–183). Berlin, Heidelberg: Springer-Verlag.

Schreiber, M., Knöppel, C., & Franke, U. (2013). LaneLoc: Lane marking based localization using highly accurate maps. In *Proceedings of the IEEE intelligent vehicles symposium* (pp. 449–454).

Schulz, A. T., & Stiefelhagen, R. (2015a). A controlled interactive multiple model filter for combined pedestrian intention recognition and path prediction. In *Proceedings of the IEEE intelligent transportation systems conference* (pp. 173–178).

Schulz, A. T., & Stiefelhagen, R. (2015b) Pedestrian intention recognition using latent-dynamic conditional random fields. In *Proceedings of the IEEE intelligent vehicles symposium* (pp. 622–627).

Tamura, Y., Le, P. D., Hitomi, K., Chandrasiri, N., Bando, T., Yamashita, A., et al. (2012). Development of pedestrian behavior model taking account of intention. In *Proceedings IEEE international conference on intelligent robots and systems (IROS)* (pp. 382–387).

Vanparijs, J., Panis, L. I., Meeusen, R., & de Geus, B. (2015). Exposure measurement in bicycle safety analysis: A review of the literature. *Accident Analysis & Prevention*, *84*, 9–19.

Völz, B., Mielenz, H., Siegwart, R., & Nieto, J. (2016). Predicting pedestrian crossing using quantile regression forests. In *Proceeding of the IEEE intelligent vehicles symposium*, pp. 426–432.

Wöhler, C., & Anlauf, J. K. (1999). A time delay neural network algorithm for estimating image-pattern shape and motion. *Image and Vision Computing*, *17*(3–4), 281–294.

Yi, S., Li, H., & Wang, X. (2016). Pedestrian behavior understanding and prediction with deep neural networks. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 263–279). Springer.

Yi, Y., Hao, L., Hao, Z., Songtian, S., Ningyi, L., & Wenjie, S. (2017). Intersection scan model and probability inference for vision based small-scale urban intersection detection. In *Proceedings of the IEEE intelligent vehicles symposium* (pp. 1393–1398).

Zernetsch, S., Kohnen, S., Goldhammer, M., Doll, K., & Sick, B. (2016). Trajectory prediction of cyclists using a physical model and an artificial neural network. In *Proceedings of the IEEE intelligent vehicles symposium* (pp. 833–838).

Zhang, R., Wu, J., Huang, L., & You, F. (2017). Study of bicycle movements in conflicts at mixed traffic unsignalized intersections. *IEEE Access*, *5*, 10108–10117.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., & Torralba, A. (2017). Scene parsing through ADE20K dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.