Driver and Pedestrian Mutual Awareness for Path Prediction and Collision Risk Estimation

Markus Roth, Jork Stapel, Riender Happee and Dariu M. Gavrila

***Abstract*—We present a novel method for vehicle-pedestrian path prediction that takes into account the awareness of the driver and the pedestrian towards each other. The method jointly models the paths of vehicle and pedestrian within a single Dynamic Bayesian Network (DBN). In this DBN, sub-graphs model the environment and entity-specific context cues of the vehicle and pedestrian (incl. awareness), which affect their future motion and allow to increase the prediction horizon. These sub-graphs share a latent state which models whether vehicle and pedestrian are on collision course; this accounts for a certain degree of motion coupling.**

The method was validated with real-world data obtained by on-board vehicle sensing (stereo vision, GNSS and proprioceptive). Data consist of 93 vehicle and pedestrian encounters, spanning various awareness conditions and dynamic characteristics of the participants. In ablation studies, we quantify the benefits of various components of our proposed DBN model for path prediction and collision risk estimation.

Results show that at a prediction horizon of 1.5 s, context-aware models outperform context-agnostic models in path prediction for scenarios with a dynamics change, while performing similarly otherwise. Results further indicate that driver attention-aware models improve collision risk estimation compared to driver-agnostic models.

***Index Terms*—Driver Awareness, Pedestrian Awareness, Path Prediction, Collision Risk Estimation**

I. Introduction

MORE than 1.35 million people are killed yearly in traffic worldwide, according to a much cited report of the World Health Organization [1]. Pedestrians make up 23% of this number. More than half of serious crashes between vehicles and pedestrians occur outside dedicated crossing locations (e.g. zebras, traffic lights) with marked right-of-way [2].

Despite the recent interest and effort spent on higher levels of automated driving (SAE level 3+), for the foreseeable future, the reality on the road (and the accident numbers) will largely be determined by assistance systems where the driver is still required to keep the eyes on the road. This especially holds for active pedestrian safety in urban traffic.

Pedestrians are highly manoeuvrable; they can stop walking or change direction in an instant. This makes it challenging to predict their paths. Current active pedestrian safety systems on the market provide driver assistance (SAE level 0-2). They are conservatively designed in their warning and control strategy, emphasizing the current pedestrian state (i.e. position) rather than prediction, in order to avoid false system activations (i.e. automatic braking and evasive steering [3]).

M. Roth, J. Stapel, R. Happee and D. M. Gavrila are with the Intelligent Vehicles Group at TU Delft. M. Roth is also with the Environment Perception Department at Mercedes-Benz AG. markus.r.roth@daimler.com

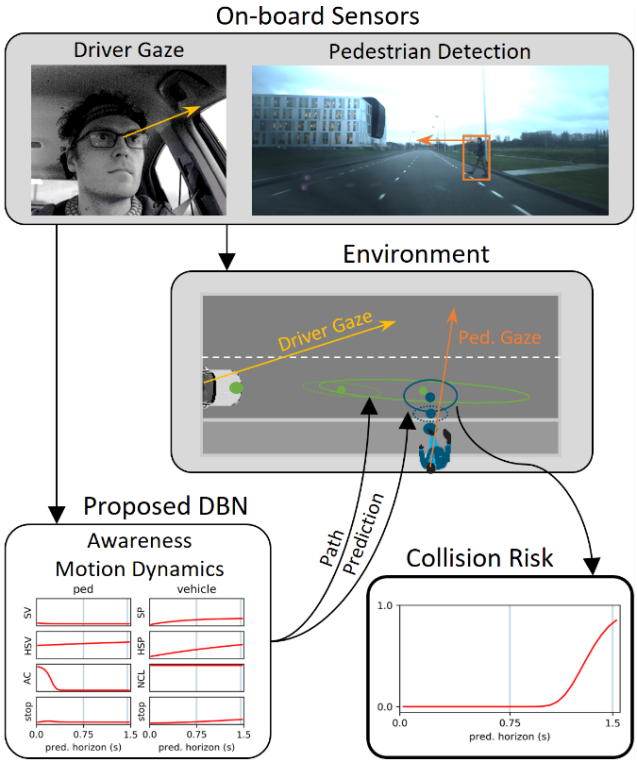


Fig. 1: The system assesses mutual awareness of pedestrian and driver in a scenario of a potentially crossing pedestrian. Cues about the driver, pedestrian and spatial environment are collected from on-board sensors. A probabilistic framework based on a Dynamic Bayesian Network (DBN) estimates latent states of awareness of the driver and pedestrian to predict their future motion. Based on the predicted paths, collision risk is estimated.

Camera-based driver monitoring systems can detect fatigue, drowsiness, distraction, gestures, signs of being drunk and readiness to take over from automated driving. On-market systems for collision warning have been employed as early as 2007 (Toyota/Lexus) by monitoring head pose and eye opening. Recent releases allow for extended SAE level 2 capabilities while driving on specially mapped highways (Cadillac Super Cruise, 2018), in traffic jams with restricted velocity (BMW Extended Traffic Jam Assistant, 2018), or in single-lane cruising (Nissan ProPilot, 2019). Mercedes-Benz's latest S-Class features a driver camera that monitors driver's readiness to take over from automated driving mode on highways in an SAE level 3 system. This legally allows the driver to perform non-

driving related tasks for up to 10 s under specific conditions.

Active safety systems on the market stand to gain from improved path prediction capability of both ego-vehicle and other road users. Furthermore, they can benefit from more information regarding which specific parts of the scene have been perceived by the driver, to ascertain whether this includes the potential hazard. Ideally, a prediction horizon of 2.5 s is achieved, at which point the driver "feels no danger" [4]. For the pedestrian case, we will be hard pressed to achieve accurate predictions for a 1.5 s time horizon, as will become apparent.

In this paper, we consider the setting of a potentially crossing pedestrian and an approaching vehicle which has the right-of-way (i.e. no dedicated crossing location). We present a method which uses context cues about the spatial environment, driver-pedestrian mutual awareness and potential motion coupling to estimate the future paths of both participants and associated collision risk. See Figure 1 for an illustration of the overall system.

Specifically, we extend the Dynamic Bayesian Network (DBN) method from Kooij *et al*. [5], [6], which performs path prediction for an individual pedestrian, to the mutual vehicle-pedestrian case. As in [5], [6], we capture that pedestrian awareness of the on-coming vehicle will likely affect his/her future path. In our method we also model that driver awareness of the pedestrian will likely affect the future ego-vehicle path. We use head pose (pedestrian, driver) and eye gaze (driver) as proxies for awareness, as the latter cannot be determined directly.

There are several reasons for choosing a physics-based DBN approach for path prediction, as opposed to the popular neural networks. First, a DBN allows more easily to incorporate expert domain knowledge by means of its graphical model structure. Second, a DBN is interpretable, one can inspect the values of its latent variables and follow how it reaches its output. This is especially important for safety-critical applications. Third, one can expect a DBN to deal well with smaller datasets, as it has a comparatively small set of parameters, which will minimize the effects of over-training. Finally, recent work by Pool *et al*. [7] suggests that a DBN can deliver competitive path prediction results compared to a RNN, when its parameters are optimized by backpropagation as well.

The paper outline is as follows. Section II presents the related work. Section III describes the proposed context-based path prediction model for vehicle and pedestrian. Sections IV and V describe the collected dataset and the procedures for parameter estimation. Section VI describes the experimental results. Section VII provides a discussion and Section VIII lists the conclusions.

II. Related Work

Road user path prediction has attracted a lot of attention in recent years, see surveys regarding the ego-vehicle [8] and Vulnerable Road Users [9], [10]. Path prediction methods require positions as input. Ground plane positions relative to a vehicle coordinate system can be obtained from detections in various sensors (e.g. camera [11], radar [12], LiDAR [13], or a combination thereof [13], [14]). If ground plane positions relative to a global coordinate system are needed (e.g. this paper), then vehicle ego-motion compensation is necessary as an additional pre-processing step. For this, a combination of GNNS, INS and vehicle proprioceptive sensing can be used. Following sub-sections focus on context cues and motion models used for path prediction.

A. Context cues for path prediction

In the most rudimentary form, cues for path prediction consist of point kinematics, i.e. positions and velocities of the relevant object. It has however been well established that the use of additional "context" cues can improve path prediction performance [10]. These can be categorized into object cues, and static and dynamic environment cues.

Object context cues refer to cues pertaining to the object of interest itself. For example, Keller and Gavrila [15] improve pedestrian path prediction by using dense optical flow features extracted from a pedestrian bounding box. Kooij *et al*. [5] use relative head orientation as a "proxy" for the pedestrian's awareness of the oncoming ego-vehicle while crossing. Kooij *et al*. [6] and Pool *et al*. [7] incorporate the arm gesture of a cyclist to predict its turn at an intersection. Quintero *et al*. [16] recover a full 3D articulated pose of a pedestrian to better predict crossing action.

Object context cues can also refer to properties derived from the driver of the ego-vehicle, when interested in predicting the future ego-vehicle path. Typical such cues are driver head orientation or gaze, or performed driver actions, as inferred from accelerator pedal position, braking force and steering wheel angle. For example, Roth *et al*. [17] employ driver head pose to capture the driver's awareness of a crossing pedestrian.

Static environment context cues refer to elements of the static traffic infrastructure which will likely influence road user motion, such as road topology [7], [18], road markings and traffic lights.

Dynamic environment context cues capture the presence and motion properties of other road users (including that of the ego-vehicle itself) that may influence the target road user's behavior, i.e. to avoid hazards or to minimize hindrance. For example, [5], [6], [17], [19], [20] use basic kinematics properties, such as relative distances and velocities, and the expected point of closest approach.

B. Motion models

Models for human motion trajectory estimation can be subdivided into physics-based, pattern-based and planning-based methods [10]. As motivated earlier, we focus here on physics-based methods, which represent motion by explicitly defined dynamic equations of one or more underlying dynamical models. Simple motion dynamics can be modeled by Linear Dynamical Systems (LDS), which commonly assume a linear relationship between states and measurements with Gaussian noise. Under these assumptions, the Kalman Filter (KF) [21] is an optimal filtering algorithm, which has been widely applied for pedestrian and vehicle tracking [8], [22].

In the scope of collision analysis, motion models play a role for predicting paths of targets such as a potentially

crossing pedestrian and the ego-vehicle. The probabilistic models described here allow to extrapolate observed behaviors into the future while accounting for uncertainties in the assumed dynamics and observations.

Since traffic behavior may change at any time, a common approach is to treat the complex dynamics by switching between or combining multiple motion models at each prediction step, e.g., by using Switching LDS (SLDS). SLDS can be extended by dynamical models to incorporate contextual cues for path prediction [6], [16]. Li *et al*. [23] combine the path prediction output of Kooij *et al*. [6] with a sequence-to-sequence trajectory generation method to leverage the complementary advantages of hand-crafted models and data-driven methods.

Different methods have been introduced to predict the paths of multiple interacting road users, e.g., Social Force models for human-human interactions [24]. For pedestrian-vehicle encounters, e.g., Kooij *et al*. [6] assume that the vehicle does not change motion dynamics, while Braeuchle *et al*. [25] use a Bayesian Network to find an appropriate vehicle motion model which minimizes pedestrian injury risk. The pedestrian motion model is fixed based on initial velocity. Gupta *et al*. [26] simulate actions (speed up, slow down) of a self-driving vehicle within a negotiation cycle with a crossing/yielding pedestrian to optimize traffic throughput.

III. Joint Vehicle and Pedestrian Path Prediction

A. Overview and Main Contributions

Kooij *et al*. [6] note that a pedestrian's decision to continue walking or to stop in a crossing scenario is mainly influenced by the presence of an approaching vehicle on collision course, the pedestrian's awareness thereof, and the position of the pedestrian with respect to the curbside. This knowledge is encoded in a context-based SLDS (a special DBN), where latent discrete states control the switching probabilities between the continuous states dynamics of walking and standing.

In this work, we are interested in vehicle-pedestrian collision risk, thus we extend the prediction component to the ego-vehicle. We analogously argue that the vehicle's outcome of continue moving or stopping is mainly influenced by the presence of an approaching pedestrian on collision course, the driver's awareness thereof and the distance of the vehicle to the pedestrian's crossing location. We model pedestrian and vehicle motion with two SLDSes which are linked to each other by a shared latent state, which captures the motion coupling between the two objects. The proposed DBN is shown in Figure 2 (see Table I for the corresponding node descriptions).

Our main contributions are:

- We present a method for joint path prediction and collision risk estimation of vehicle and pedestrian using observed kinematics, mutual awareness, and environment cues.
- We provide an ablation study of the effect of various context cues on situations where an intervention of either road user is needed to avoid a collision.
- We apply our method on real sensor data from a vehicle.

Compared to our earlier work [17], we add collision risk analysis and perform more extensive evaluations (incl. *estimated* head pose and *estimated* eye gaze in addition to invasively measured head pose [17]) on a new and larger dataset.

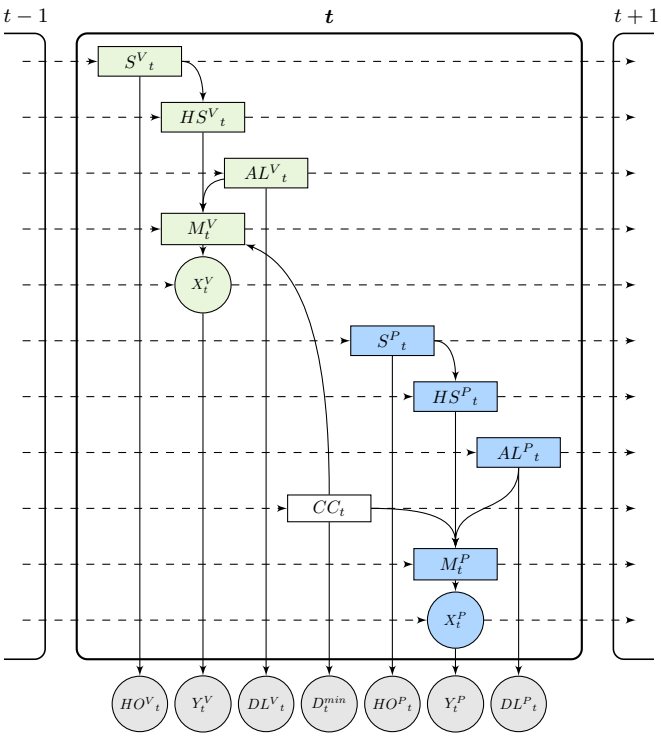


Fig. 2: Graphical model representation of the Dynamic Bayesian Network (DBN). Discrete nodes are rectangular, continuous nodes are circular. Grey nodes represent observable variables while the other nodes represent latent states. Dashed lines depict temporal connections between latent context states in subsequent time instances. Driver-related nodes are shaded in green while pedestrian-related nodes are shaded in blue. Context state description and purpose are provided in Table I.

B. DBN

The DBN consists of two sub-graphs, one for the pedestrian and one for the vehicle. The pedestrian sub-graph is congruent with the DBN of Kooij *et al*. [6]. The vehicle sub-graph displays analogous behavior for the vehicle, by encoding driver awareness by driver gaze and braking manifestation by being close to the crossing location of the pedestrian.

1) Pedestrian-related context states: The pedestrian P can exhibit one of two motion types: *walking* ($M_t^P = m^P_{\text{move}}$, constant velocity) and *standing* ($M_t^P = m^P_{\text{stop}}$, constant position). The motion state of the pedestrian contains two-dimensional positions and velocities: $X_t^P = [x_t, y_t, \dot{x}_t, \dot{y}_t]^T$. This results in the linear state transformation matrices:

$$A^{(m^P_{\text{move}})} = \begin{bmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, A^{(m^P_{\text{stop}})} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (1)$$

The vehicle observes pedestrian world positions $Y_t^P \in \mathbb{R}^2$ without velocities, resulting in the corresponding observation matrix $C^P = \left[\begin{smallmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{smallmatrix}\right]$.

For the context-based SLDS, the switching state M_t^P of the pedestrian motion model is encoded in the DBN as a categorical distribution $M^P_{t+1} = \text{Cat}\left(M_t^P, {AL^P}_{t+1}, {HS^P}_{t+1}, CC_{t+1}\right)$ as shown in Figure 2. The pedestrian awareness context ${S^P}_t$

TABLE I: Latent context states, their associated observation and the purpose within the DBN structure. States are grouped by vehicle/driver (common superscript V), pedestrian (superscript P) and shared contexts.

Latent State	Abbr.	Observation	Abbr.	Purpose
driver-sees-pedestrian	S^V	driver-head-orientation (gaze)	HO^V	encodes driver's awareness of the pedestrian
driver-has-seen-pedestrian	HS^V	-	-	memorizes driver's (past) awareness of the pedestrian
vehicle-at-location	AL^V	vehicle-distance-to-location	DL^V	manifests typical location of braking (ped. crossing location)
vehicle-motion-model	M^V	-	-	switches between *driving* and *braking* LDS
vehicle-position-state	X^V	vehicle-position	Y^V	LDS for vehicle state estimation
pedestrian-sees-vehicle	S^P	pedestrian-head-orientation	HO^P	encodes pedestrian's awareness of the driver/vehicle
pedestrian-has-seen-vehicle	HS^P	-	-	memorizes pedestrian's (past) awareness of the driver/vehicle
pedestrian-at-location	AL^P	pedestrian-distance-to-location	DL^P	manifests typical location of stopping (curb)
pedestrian-motion-model	M^P	-	-	switches between *walking* and *standing* LDS
pedestrian-position-state	X^P	pedestrian-position	Y^P	LDS for pedestrian state estimation
collision-course	CC	minimum-future-distance	D^{min}	separates early crossings from critical crossing

models whether the pedestrian sees the approaching vehicle. Head orientation ${HO^P}_t$ forms the evidence. The context variable ${HS^P}_t$ memorizes whether the pedestrian has seen the vehicle in the past, acting as a logical *OR* between previous ${HS^P}_{t-1}$ and current ${S^P}_t$. The environment context ${AL^P}_t$ models whether the pedestrian is near the curb, thus encoding where a pedestrian would normally stop to yield for oncoming traffic.

2) Vehicle-related context states: The vehicle motion state is $X_t^V = [x_t, y_t, \dot{x}_t, \dot{y}_t]^T$. It uses a constant velocity model while driving, and a velocity decay model for braking:

$$A^{(m^V_{\text{move}})} = \begin{bmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, A^{(m^V_{\text{stop}})} = \begin{bmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & d & 0 \\ 0 & 0 & 0 & d \end{bmatrix} \quad (2)$$

The decay parameter $d = \sqrt[10]{0.5} \approx 0.93$ is empirically chosen to represent a velocity half-life of 0.5 s, i.e., the velocity becomes $d^{10} = 0.5$ of its initial value after 10 discrete time steps (0.5 s). This results in a mean initial deceleration of $\sim 4.2\,\mathrm{m/s^2}$ over the first second, thus reflecting moderate braking. Also, the vehicle V observes its own velocity, resulting in the observation matrix $C^V = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$.

For the vehicle, the context-based SLDS' switching state M^V is encoded as a categorical distribution $M^V_{t+1} = \text{Cat}\left(M^V_t, {AL^V}_{t+1}, {HS^V}_{t+1}, {CC}_{t+1}\right)$. The driver awareness context ${S^V}_t$ models the driver's awareness of the pedestrian. It is inferred from the attention eccentricity ${HO^V}_t$, i.e., the absolute visual angle difference between the driver's center of gaze (or head direction) and the pedestrian. The context variable ${HS^V}_t$ memorizes whether the driver has seen the pedestrian analogous to ${HS^P}_t$. The static environment context ${AL^V}_t$ indicates whether the vehicle is at a distance from the pedestrian's crossing location where the driver can be expected to yield, assuming he/she has the intention to do so.

3) Shared context state: Both pedestrian and vehicle dynamics depend on CC_t, which indicates whether pedestrian and vehicle are on a collision course. It uses the minimum distance D_t^{min} obtained when linearly extrapolating the trajectories with their momentary estimated velocities [20].

C. Inference

During inference the DBN states are propagated over time by incorporating observations in a forward filtering procedure (predict, update) following [6]. At each time step t, the entire state of the DBN is represented by the 9 discrete latent states (4 vehicle, 4 pedestrian, 1 shared) and two partially observable continuous latent states (X_t^V, X_t^P), see Figure 2. During the predict step, the value of each discrete latent state changes according to a fixed transition table, based on the values of its input states, i.e., each state's input nodes in Figure 2, including the state from the previous time step $t-1$ (dashed line). During the update step, observations are incorporated based on the context likelihood distributions, see Figure 3. The intermediate goal is to have the motion model switching states for both vehicle (M_t^V) and pedestrian (M_t^P) which represent the switching probability of the SLDS of each road user. The two continuous latent states X_t^V, X_t^P are propagated over time using observations (Y_t^V, Y_t^P) by standard LDS means, i.e., Kalman filter. Prediction into future without observation follows the same procedure, but without the update steps. Overall, this results in predicted motion states including uncertainties for both vehicle and pedestrian. To keep inference tractable, we apply Assumed Density Filtering [27], resulting in the probability distributions of X_t^V, X_t^P to be each modeled by a Gaussian Mixture (K=2).

IV. Parameter Estimation

We set the DBN model parameters by performing a data-driven initialization step, followed by a gradient-based optimization step, using the dataset we introduce in Section V.

A. Model parameter initialization

Model parameters relate to motion dynamics and context. They are initialized similar to Kooij *et al.* [6].

Motion Dynamics: The underlying motion models of M^V and M^P are represented by LDSes which model process noise Q and observation uncertainty R. Process noise Q of vehicle and pedestrian are set for both position and velocity states and are limited to diagonal matrix entries. Values were selected to reflect model uncertainty under typical velocity

changes of drivers and pedestrians [28], [29]. Observation noise R is set to reflect typical variance of measurement noise for pedestrian detection and vehicle movement observed on-board our vehicle, see Section V. The motion state transition matrices were obtained as follows. The vehicle motion state M^V was categorized as *braking* when such activity was detected, analogous to AL^V, and as *driving* otherwise. The pedestrian motion state M^P was categorized as *standing* in all scenarios where a pedestrian stops starting from three frames preceding TTE = 0 (see Section V-B for definition of TTE), similarly to AL^P below. The motion state at all other time instants was categorized as *walking*. The motion state transition matrices were then obtained by counting and normalizing the occurrences of the respective transitions. The initial motion states assume the vehicle and pedestrian are driving and walking.

Context: To obtain the parameters for binary context states, we need to establish their ground truth values; we do so in a two-step approach. In the first step, ground truth values were roughly obtained by setting some states to the same values for the entire scenario based on its definition (S^P, S^V, CC), by manual annotation ($AL^P = 1 \iff$ TTE = 0), or by an automatic observable criterion ($AL^V = 1$ for all moments after first deceleration, i.e., pressing the brake pedal). This yields the context likelihood distributions as shown by the histograms in Figure 3. Parametric distributions were fitted by Maximum-Likelihood-Estimation and are shown by line plots. The parametric form of the distributions was chosen heuristically: Gaussian (DL^P, DL^V), Gamma (D^{min}, HO^V) or von-Mises distribution (HO^P).

In a second step, more accurate ground truth values for the context states were obtained on the basis of the obtained context likelihood distributions. For context states AL^V, AL^P and CC, the values were re-assigned based on maximum likelihood criterion (e.g., $CC = 1 \iff D^{min} < 2.6$ m, see Figure 3a). For S^P and S^V, re-assignment was done heuristically. We re-assigned $S^P = 1 \iff HO^P \in [-30, 30]^\circ$ due to the largely overlapping distributions caused by miss-estimation of the head pose estimation algorithm. We re-assigned $S^V = 1 \iff HO^V < 10^\circ$ whenever we use the head orientation and $< 4^\circ$ otherwise for the eye gaze orientation. The transition matrices which represent the transition probabilities conditioned on the input states (i.e., incoming links in the DBN graph) were obtained by counting and normalizing the re-assigned binary context values between adjacent time steps. The transition probabilities of HS^V and HS^P are implemented as a binary OR in order to memorize the last state in accordance with their definition in Section III-B1.

The initial context states values were set conservatively at the beginning of each encounter: driver/pedestrian not looking, vehicle not near crossing location and pedestrian not at curb.

B. Model parameter optimization

We employed the gradient-based method of Pool *et al.* [7] to obtain optimized model parameters. In short, the method performs back-propagation similar to neural networks on the DBN parameters on a differentiable loss function. We maximize the observation log likelihood of the vehicle and pedestrian under their respective predicted Gaussian distributions, see Eq. (4). All intermediate time-steps up to the prediction horizon are incorporated into the loss function to enforce a consistent path. Measurements with time-to-event (TTE) $\in [-2.5\,\mathrm{s}, 3.0\,\mathrm{s}]$ are considered for optimization, to cover periods of typical motion dynamics. Missing intermediate measurements are ignored for optimization. TTE is defined in Section V.

Optimization has been performed while enforcing properties of the DBN variables to keep the state representation interpretable, such as probabilities residing in $[0, 1]$ and process and observation noises remaining positive definite. We also enforce the latter to be diagonal matrices with variability along elements of main direction of travel to reduce degrees of freedom and obtain more stable convergence in the optimization process.

The model parameters chosen for optimization are: process noises (Q) of pedestrian and vehicle, transition probabilities, and context observation distribution parameters. The model was implemented in Python 3 using PyTorch 1.4 and was optimized using Adam [30].

V. Dataset

A. Scenarios

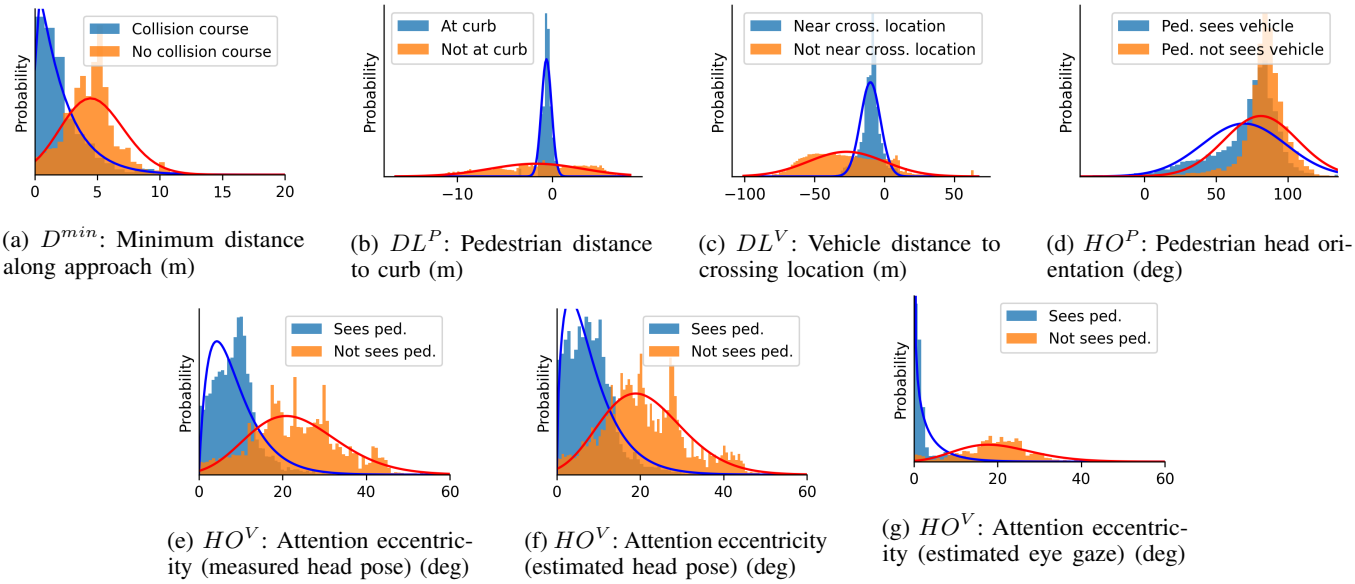
93 vehicle-pedestrian encounters with 4 trained drivers and 4 pedestrians were staged on two empty public roads. Each encounter consisted of a single pedestrian with the intention to cross the street in front of the approaching vehicle. The encounters represented nine disjoint scenarios (8-20 encounters each) with different combinations of situation criticality (collision course/sufficient time to cross), pedestrian behavior (stop at curb/cross), pedestrian awareness of the approaching vehicle (aware/unaware), vehicle behavior (brake/continue) and driver awareness of the approaching pedestrian (aware/unaware). The included scenarios are listed in the left of Table III.

All scenarios (except the anomalous scenario 9^a) encode following behaviors:

- An aware pedestrian will yield to the vehicle. Pedestrian awareness is inferred from pedestrian head pose.
- An aware driver brakes for an inattentive pedestrian approaching the curb. Awareness is inferred from driver head or gaze orientation.
- In non-collision-course crossing scenarios, both participants continue walking/driving.
- Unaware participants continue walking/driving.

Scenarios 1 to 4 represent non-collision-course scenarios, meaning the pedestrian has sufficient time to cross. Scenarios 5 to 7 are safe through a change in behavior by either the driver or pedestrian due to awareness of the other participant. Scenario 8 represents a collision where both driver and pedestrian are unaware of each other's presence. Scenario 9^a represents an anomalous scenario: the pedestrian crosses despite being aware of the approaching vehicle. The anomalous scenario is not considered for model parameter estimation.

Pedestrians were instructed to either *"continuously observe the vehicle"* or to *"keep facing forward and don't look at the vehicle"*. Drivers were instructed to either *"keep looking at*



(a) D^{min}: Minimum distance along approach (m)

(b) DL^P: Pedestrian distance to curb (m)

(c) DL^V: Vehicle distance to crossing location (m)

(d) HO^P: Pedestrian head orientation (deg)

(e) HO^V: Attention eccentricity (measured head pose) (deg)

(f) HO^V: Attention eccentricity (estimated head pose) (deg)

(g) HO^V: Attention eccentricity (estimated eye gaze) (deg)

Fig. 3: Original and fitted context likelihood distributions. See Section IV-A for details.

the pedestrian" or to *"avoid looking at the pedestrian"* while approaching the pedestrian.

While scenarios 8 and 9[a] represent collisions, naturally, no actual collision took place during data collection. Instead, the vehicle was brought to a full stop before colliding with the pedestrian. The vehicle's velocity and position data were artificially replaced with a constant velocity model starting just before the onset of braking.

To ensure safety, the road was overseen to halt the experiments when other traffic entered the testing area. A co-driver provided verbal instructions on when to brake. Target driving speed was $20\,\mathrm{km/h}$ and pedestrians adopted their preferred walking pace.

B. Instrumentation, measurements and ground truth

All data were collected with a TU Delft experimental vehicle, whose instrumentation is described in further detail in [31]. Vehicle position, orientation and velocity are obtained from an ego-vehicle localization system which fuses differential GNSS, IMU, steering wheel angle and wheel ticks. We implement this by the Robot Operating System (ROS) `robot_localization` package [32] and gain the transformations from vehicle frame to the world coordinate frame, which is set to identity at the start of the system. The GPS maintains a position accuracy of $4\,\mathrm{cm}$ while drift between GPS updates is limited to 0.8% per unit of distance traveled. The road was observed at $10\,\mathrm{Hz}$ using a forward-facing stereo camera (baseline $22\,\mathrm{cm}$, $1936 \times 1216\,\mathrm{px}$) mounted behind the top-center of the windshield to obtain a dense stereo depth image of the scene in front of the vehicle.

Driver head pose and gaze were recorded with two systems. *Estimated* eye gaze and head pose were recorded with a high-end commercial off-the-shelf eye tracker (Smarteye: 4-camera Smart Eye Pro dx 5.0, software version 8.2, running at $60\,\mathrm{Hz}$ with a gaze accuracy down to $0.5°$). Secondly, *measured* head pose is obtained by a head-worn infrared-reflective marker tracked by an optical marker tracking system (Smarttrack) mounted on the rear seat head rest [17], [33]. Additionally, the driver was observed by a camera mounted above the speedometer for visual verification purposes. All sensor data were spatially calibrated and resampled to a target rate of $20\,\mathrm{Hz}$.

Measured pedestrian positions on the ground plane were obtained in three successive steps: (1) 2D pedestrian bounding boxes were estimated from the forward facing camera by the Single-Shot-Multibox-Detector (SSD) of Braun *et al.* [11]. (2) Distance to camera was found by median stereo disparity [34] of the 2D bounding box. (3) Transformation of this car-relative pedestrian position to ground plane positions in world coordinate frame was performed via ego-vehicle localization. The time between the first pedestrian detection and the pedestrian reaching the curb was (min / max / mean = $1.3\,\mathrm{s}$ / $3.2\,\mathrm{s}$ / $2.9\,\mathrm{s}$) over the various sequences. In that period, the pedestrian detection recall was $83\,\%$.

Similarly to Kooij *et al.* [5], we infer pedestrian's focus-of-attention from pedestrian head orientation. We apply the method of Braun *et al.* [35] to obtain a single yaw angle representing pedestrian head orientation.

In order to temporally compare prediction performance among the various scenarios, a semantically meaningful event was manually annotated for each sequence, as in [5], [15]. For scenarios where the pedestrian crosses, it represents the first frame where a pedestrian's foot crosses the curb. For scenarios where the pedestrian stops, it represents the moment where the last foot is placed on the ground near the curb. This implicitly defines time-to-event (TTE) for each time-step of each sequence (negative TTE: before event).

For each encounter, we obtained ground truth of the pedestrian position in the world coordinate frame. The pedestrian's path of travel is defined in the world coordinate frame as a

straight line and corresponds to the path the participants were instructed to move along. The pedestrian ground plane location is then obtained by the intersection of the annotated path of travel with the vertical plane spanned by the image column of the hip point which we manually annotated in each frame. We employ map information and ego-vehicle localization to estimate the location of the curb side.

VI. Results

To evaluate the incremental benefits of the DBN model components for an intelligent collision warning system, we compare six models with varying access to the used context cues on their joint prediction performance of vehicle- and pedestrian-path and collision risk. We adopt two evaluation metrics: the ability to predict driver and pedestrian location 1.5 s into the future, and collision risk across multiple prediction horizons. Evaluation is performed using 5-fold cross validation.

A. Evaluation metrics

For each time t, each model creates a predictive distribution $\widetilde{P}_{t\rightarrow t+t_p}(X_t)$ for state X_t and prediction horizon t_p. Based on the predictive distributions of both vehicle and pedestrian, we evaluate individual path prediction performance and combined collision risk.

Path prediction performance: Two performance metrics are used to evaluate path prediction performance [5] [15]: (a) Euclidean distance error between predicted expected position and future ground truth position GT_{t+t_p}:

$$\text{error}\,(t_p|t) = \left|\mathbb{E}\left[\widetilde{P}_{t\rightarrow t+t_p}\,(X_t)\right] - \mathrm{GT}_{t+t_p}\right| \qquad (3)$$

and (b) the log likelihood of the future ground truth position GT_{t+t_p} under the predictive distribution:

$$\text{loglik}\,(t_p|t) = \log\left[\widetilde{P}_{t\rightarrow t+t_p}\left(\mathrm{GT}_{t+t_p}\right)\right] \qquad (4)$$

loglik encapsulates both the spatial error and certainty about the position observation. Larger *loglik* values denote better prediction performance.

Collision risk: We determine the probability for a collision by taking the integral of the predictive distributions over a collision area, which is defined by all possible intersections between vehicle and pedestrian locations. Let $\widetilde{P}_{t\rightarrow t+t_p}(X_t) = \mathcal{N}(\mu_{t\rightarrow t+t_p}, \sigma^2_{t\rightarrow t+t_p})$ be a single Gaussian predictive position of either pedestrian P or vehicle V. The *combined* predictive position is then defined as $\widetilde{P}^{\phi}_{t\rightarrow t+t_p}(X^{\mathrm{P}}_t, X^{\mathrm{V}}_t) = \mathcal{N}(\mu^{P}_{t\rightarrow t+t_p} - \mu^{V}_{t\rightarrow t+t_p}, (\sigma^{P}_{t\rightarrow t+t_p})^2 + (\sigma^{V}_{t\rightarrow t+t_p})^2)$. The collision risk predicted from t for $t+t_p$ is given by:

$$\text{CR}\,(t_p|t) = \int_{A^{\phi}} \widetilde{P}^{\phi}_{t\rightarrow t+t_p}(X^{\mathrm{P}}_t, X^{\mathrm{V}}_t) dX^{\mathrm{P}}_t dX^{\mathrm{V}}_t \qquad (5)$$

with A^{ϕ} being the combined spatial extent of vehicle and pedestrian. If the predictive distributions for the vehicle and the pedestrian are represented as Gaussian Mixtures (SLDS and DBN variants), the overall collision risk is given by the weighted pairwise collision risk between the Gaussian Mixture components. This extends the collision risk estimation method of Braeuchle *et al*. [25].

Context cue	LDS	SLDS	DBN.p [6]	DBN.pv	DBN.pvh	DBN.pvg
Pedestrian at-curb	-	-	x	x	x	x
Pedestrian awareness	-	-	x	x	x	x
Collision course	-	-	x	x	x	x
Vehicle near-crossing	-	-	-	x	x	x
Driver awareness	-	-	-	-	head pose	eye gaze
# Ped. motion models	1	2	2	2	2	2
# Veh. motion models	1	2	2	2	2	2

TABLE II: Context cues and number of motion models per road user used in the models. DBN suffixes denote used context: p: pedestrian [6]; v: vehicle (AL^V); h: driver head pose; g: driver gaze. E.g., *DBN.pvg* uses pedestrian, vehicle and driver eye gaze awareness context.

For the application of collision risk warning, collision probability has to be classified into collision or no collision, and classification performance requires a ground truth for collision outcome. We define collision ground truth as true for any time instance where the vehicle and pedestrian ground truth overlap given their position and spatial extent. In order to assess the collision risk prediction performance at various prediction horizons, we select a fixed false positive rate (FPR) and find the attainable true positive rate (TPR) for each prediction horizon t_p.

B. Model variants

We evaluate four context-aware models, including the method of Kooij *et al*. [6], which differ in their access to pedestrian and vehicle context, and compare them to two context-agnostic models. An overview of the used context cues of the models is given in Table II. All models were optimized individually as described in Section IV.

Context-agnostic LDS*:* Both linear dynamical systems for pedestrian and vehicle path prediction are instantiated by constant velocity motion models.

Context-agnostic SLDS*:* Vehicle and pedestrian motion are both modeled by context-agnostic SLDSes with the same underlying motion models as the context-aware models (driving/braking, walking/standing) described below.

Context-aware models with varying pedestrian- and vehicle-context: We analyze four variants of the model presented in Figure 2 which take different amounts of context into account: *DBN.p* represents the context-based pedestrian path prediction method of Kooij *et al*. [6]. The method is driver-agnostic and models the vehicle dynamics as a context-agnostic SLDS. *DBN.pv* is vehicle-aware and extends *DBN.p* with vehicle static environment cues but remains driver-agnostic. It includes proximity of the vehicle to the crossing location of the pedestrian (AL^V). *DBN.pvh* additionally uses driver head pose as an awareness cue (S^V). *DBN.pvg* uses driver eye gaze instead of driver head pose.

C. Path prediction

Table III depicts average path prediction performance over various encounters of a certain scenario in terms of *loglik* and Euclidean distance error of both pedestrian and vehicle

for a prediction horizon $t_p = 1.5\,\mathrm{s}$ averaged over periods where typical changes in dynamics occur (pedestrian: TTE $\in [-0.5\,\mathrm{s}, 2.0\,\mathrm{s}]$, vehicle: TTE $\in [-0.5\,\mathrm{s}, 3.0\,\mathrm{s}]$; TTE ranges define times where predictions are made for). Let us consider three scenario types.

1) Normal scenarios with no motion change

We first consider the normal scenarios where no motion change occurs for a certain road user (i.e. scenarios 1-4 and 7-8 for the pedestrian, and scenarios 1-6 and 8 for the vehicle; the respective average performances are listed in two separate rows of Table III).

We see that the *LDS* for that road user has a comparatively poor *loglik* overall (-4.4 and -13.9, resp.), as the uncertainty region of its single-Gaussian state representation is large to account for possible motion changes. On the other hand, its maximum likelihood estimate is comparatively accurate: the Euclidean distance error is smaller than that of other models (65 cm and 52 cm, for pedestrian and vehicle resp.); this is to be expected as its linear model precisely fits the actual motion.

We also observe that context-aware models are at least on-par-with their context-agnostic (multi-motion) counterparts; cases of outperformance suggest that the context in the former provides more selective guidance when a motion change is probable. Specifically, models that incorporate pedestrian context (all DBN variants) are on-par-with (outperform) *SLDS* in terms of the *loglik* (Euclidean distance error) metric for the pedestrian. Models that incorporate vehicle context (*DBN.pv*, *DBN.pvh* and *DBN.pvg*) are on-par-with *SLDS* in terms of the *loglik* and Euclidean distance error metric for the vehicle.

2) Normal scenarios with motion change

Let us now consider the normal scenarios where motion change occurs for a certain road user (i.e. scenarios 5-6 for the pedestrian, and scenario 7 for the vehicle; the respective average performances are listed in two separate rows of Table III).

We see that the context-aware models for a road user mostly outperform their context-agnostic counterparts (*LDS* and *SLDS*) in terms of *loglik* and Euclidean distance error for that road user. We observe that having the full context of a road user does not necessarily improve performance for that road user as opposed to using only partial context (e.g. for the vehicle, *DBN.pvh* and *DBN.pvg* underperform *DBN.pv* on Euclidean distance error.)

We also observe that adding context related to the other user does not improve performance for the original road user (e.g. adding vehicle context *DBN.pvh* and *DBN.pvg* does not outperform pedestrian prediction performance by *DBN.p*). An outperformance might have been expected, as a motion change indicates an interaction between the road users, where such other road user context could be helpful. Apparently, the motion coupling by means of the *CC* state variable in the DBN is (too) weak, and is possibly overshadowed by data issues (e.g. measurement noise, insufficient data).

Figure 4 shows a temporal analysis of vehicle path prediction performance for sequences where the vehicle stops (scenario 7). While the vehicle approaches the pedestrian with constant velocity ($TTE < -0.2\,\mathrm{s}$), the three compared models (*LDS*, *SLDS*, *DBN.pvg*) show similar performance. As the vehicle slows down, both *LDS* and *SLDS* increase in spread over various sequences (shown by the standard deviations) and gradually decrease in vehicle *loglik*. The *SLDS* model adapts more quickly to the change of dynamics (switch from driving to braking) compared to the *LDS*. The *DBN.pvg* model variant anticipates the change in motion dynamics resulting in a higher *loglik* and less uncertainty than the context-agnostic models, therefore resulting in a better path prediction performance for the vehicle.

3) Anomalous scenario

Finally, let us consider the anomalous scenario 9^{a}. It is anomalous as the pedestrian crosses despite seeing the vehicle. We observe in Table III a lower prediction performance of the context-aware models (all *DBN* variants) regarding the pedestrian compared to the context-agnostic models (*SLDS* and *LDS*). This is no surprise, as the context-aware models were trained to expect stopping behaviour. Despite this, performance degrades gracefully, since the measurements of the walking pedestrian allow the context-aware models to infer decent motion state estimates.

Figure 5 shows a comparison between driver gaze (*DBN.pvg*) and driver head pose (*DBN.pvh*) as contextual cue for S^V (sees-pedestrian). For $S^V = 1$, driver gaze provides higher classification confidence in HS^V (has-seen-pedestrian) compared to head pose. For $S^V = 0$, both models incorrectly believe that the driver has seen the pedestrian for a similar fraction of sequences. However, this classification accuracy did not yield a better vehicle path prediction performance when comparing *DBN.pvg* to *DBN.pvh* in Table III. We attribute this to the memorizing effect of HS^V.

Measured driver head pose (Smarttrack) provided virtually identical results to estimated head pose (Smarteye) on all scenarios, and was therefor excluded from analysis.

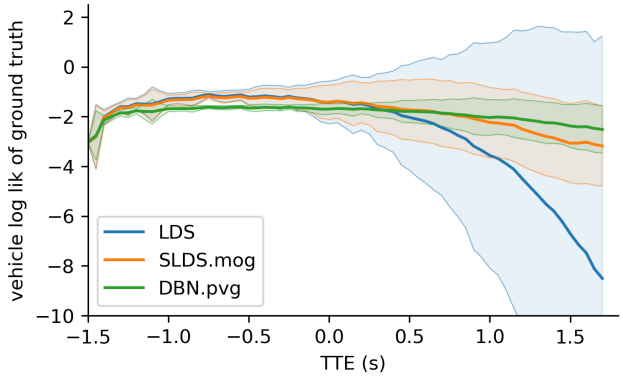


Fig. 4: *Loglik* and standard deviation over time for a braking vehicle (scenario 7) for a prediction horizon $t_p = 1.5\,\mathrm{s}$, and drawn at the moment for which the prediction was created (i.e., the values shown at TTE = $0.0\,\mathrm{s}$ were predicted from measurements of TTE = $-1.5\,\mathrm{s}$). The vehicle initiates braking for the crossing pedestrian between $-1.8\,\mathrm{s}$ and $0.6\,\mathrm{s}$, with most vehicles braking from $0.0\,\mathrm{s}$ onward.

TABLE III: Scenario decomposition (left), mean path prediction performance in terms of *loglik* (center) and Euclidean distance error (right) of various models for a prediction horizon of $t_p = 1.5\,\mathrm{s}$. The top and lower halves of the table capture the prediction performances of pedestrian and vehicle along the dimension of main travel (i.e. lateral and longitudinal vs. vehicle main axis). See Section VI-B for model definitions. Higher *loglik* and lower Euclidean distance error denote better prediction performance. Bold numbers denote best-performing model per scenario. Grey rows denote scenarios with a change in dynamics of the respective road user.

Scen.	CC	Ped. stops	Ped. sees	Veh. stops	Driver sees	LDS	SLDS	DBN p [6]	DBN pv	DBN pvh	DBN pvg	LDS	SLDS	DBN p [6]	DBN pv	DBN pvh	DBN pvg
						Pedestrian 1.5 s *loglik*						Pedestrian 1.5 s Euclidean error (cm)					
1	0	0	0	0	0	−3.3	−2.2	**−2.1**	**−2.1**	−2.2	−2.2	64	99	**48**	51	52	51
2	0	0	0	0	1	−2.8	−2.7	−2.5	−2.5	**−2.4**	**−2.4**	**83**	140	112	110	110	111
3	0	0	1	0	0	−9.2	−3.5	**−3.1**	**−3.1**	−3.6	−3.7	77	133	**68**	71	77	73
4	0	0	1	0	1	−9.0	−2.3	−2.3	**−2.2**	−2.3	−2.3	54	73	55	50	**46**	49
5	1	1	1	0	1	−4.0	−2.4	**−1.8**	**−1.8**	−2.2	−2.2	122	131	**84**	86	91	91
6	1	1	1	0	0	−4.2	−2.5	**−1.7**	**−1.7**	−1.8	−1.8	114	131	**83**	87	87	87
7	1	0	0	1	1	**−1.1**	−1.5	−1.9	−1.8	−1.7	−1.7	**58**	90	71	70	70	70
8	1	0	0	0	0	**−1.0**	−1.3	−2.0	−1.9	−1.9	−1.9	**52**	74	63	61	63	63
9[a]	1	0	1	0	0	**−1.5**	−1.8	−2.1	−2.0	−2.0	−2.0	**63**	100	79	77	73	73
non-anomalous, motion change (5-6)						−4.1	−2.5	**−1.8**	**−1.8**	−2.0	−2.0	118	131	**84**	87	89	89
non-anomalous, no motion change (1-4, 7-8)						−4.4	**−2.3**	**−2.3**	**−2.3**	−2.4	−2.4	**65**	102	70	69	70	70
						Vehicle 1.5 s *loglik*						Vehicle 1.5 s Euclidean error (cm)					
1	0	0	0	0	0	−6.2	**−2.2**	−2.8	−2.8	−2.8	−2.8	54	53	**46**	52	55	55
2	0	0	0	0	1	−38.0	−7.4	−8.8	**−6.0**	−6.1	−6.1	60	62	**49**	53	55	55
3	0	0	1	0	0	−31.2	**−6.1**	−7.9	−7.9	−7.0	−7.0	48	52	**39**	44	51	50
4	0	0	1	0	1	−12.9	**−2.8**	−3.7	−3.6	−3.7	−3.8	63	66	**55**	56	58	58
5	1	1	1	0	1	−4.5	**−1.5**	−2.4	−2.1	−2.0	−2.0	**48**	54	**48**	117	69	69
6	1	1	1	0	0	−3.4	**−1.4**	−2.0	−2.0	−1.8	−1.8	43	52	**40**	103	61	61
7	1	0	0	1	1	−7.8	−2.7	−2.6	**−2.1**	−2.2	−2.2	245	189	195	**149**	175	175
8	1	0	0	0	0	**−1.0**	**−1.0**	−1.6	−1.7	−1.6	−1.6	46	47	**39**	81	45	45
9[a]	1	0	1	0	0	**−1.1**	**−1.1**	−1.6	−1.8	−1.7	−1.7	38	47	**34**	78	45	45
non-anomalous, motion change (7)						−7.8	−2.7	−2.6	**−2.1**	−2.2	−2.2	245	189	195	**149**	175	175
non-anomalous, no motion change (1-6, 8)						−13.9	**−3.2**	−4.2	−3.7	−3.6	−3.6	52	55	**45**	72	56	56

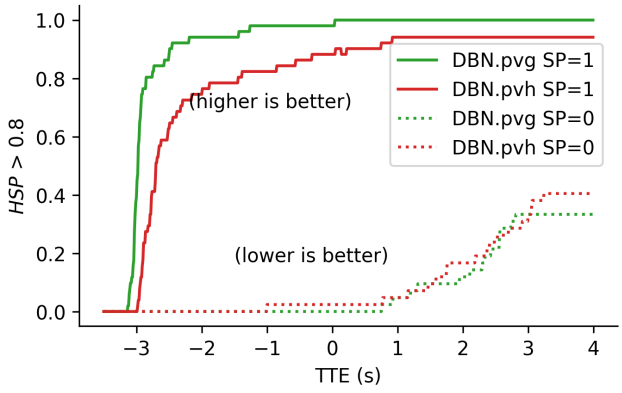


Fig. 5: Classification performance of *DBN.pvg* and *DBN.pvh* on the hidden HS^V state on sequences where driver is instructed to be attentive ($S^V = 1$) and inattentive ($S^V = 0$).

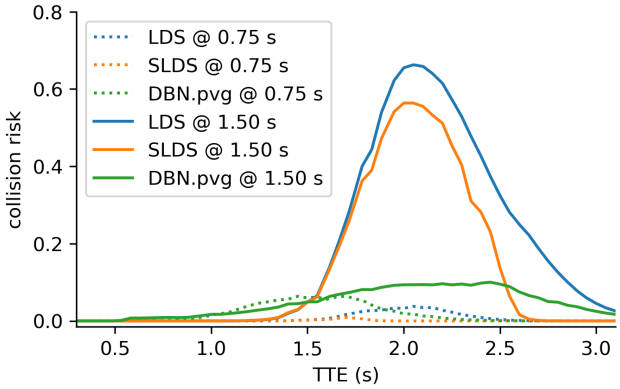
D. Collision risk estimation

We first compare how collision risk estimates evolve over time for the *LDS*, *SLDS* and *DBN.pvg* models on two exemplary sequences with changing vehicle dynamics (scenario 7) and collision (scenario 8), followed by an assessment of overall collision risk prediction performance as function of prediction horizon.

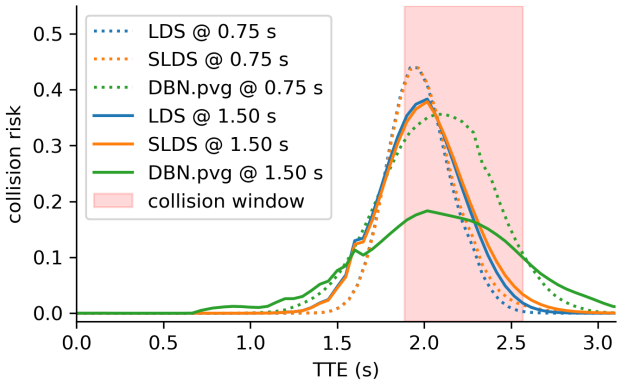
Scenario-based collision risk: Figure 6a shows collision risk prediction for a sequence from scenario 7, where the vehicle brakes due to an aware driver. Thus, a low predicted collision risk is expected. For a prediction horizon $t_p = 0.75\,\mathrm{s}$, all models predict a negligible collision risk (dashed lines). Predicting $t_p = 1.5\,\mathrm{s}$ into future, the *LDS* and *SLDS* models anticipate a collision risk of 66% and 56% respectively while the *DBN.pvg* model keeps a collision risk below 10% throughout the sequence.

Figure 6b shows collision risk over time for one sequence from the collision scenario (scenario 8), where both the vehicle and the pedestrian continue their respective motion, being unaware of each other. The *collision window* depicts all time instances defined as a collision in accordance with Section VI-A, i.e., where the geometries of vehicle and pedestrian overlap. Predicting 0.75 s into the future, all compared models (*LDS*, *SLDS*, *DBN.pvg*) depict similar maxima of collision risk within the collision window. With increasing prediction horizon, each model becomes less certain, resulting in a lower predicted collision risk value.

The maxima are above 18% within the collision window for the exemplaricly depicted sequence. Figure 6 further shows that only for *DBN.pvg*, there exists a range of collision risk thresholds (10%–18%) for which a collision warning is triggered in the collision sequence (Figure 6b) but not in the non-collision sequence (Figure 6a).



(a) Sequence from scenario 7. Lower collision risk denotes better performance.



(b) Sequence from collision scenario 8. Higher collision risk denotes better performance. The collision window CW is shaded in red.

Fig. 6: Collision risk estimates obtained from different models for a braking vehicle (top) and collision (bottom) sequence. TTE indicates the time for which the predictions were made. Values are shown for prediction horizons t_p of $0.75\,\mathrm{s}$ and $1.5\,\mathrm{s}$.

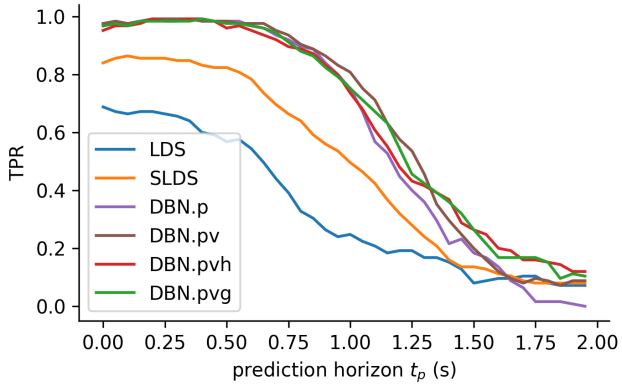


Fig. 7: Collision risk TPR of different models obtained under a 1% FPR for various prediction horizons. Higher values denote better performance.

Overall collision risk prediction: To examine how collision risk prediction performance changes with prediction horizon t_p, we select a FPR of 1% and evaluate the attainable TPR as a function of t_p, see Figure 7. One observes that the context-agnostic models (*LDS* and *SLDS*) significantly under-perform the context-aware models (*DBN* variants). For a prediction horizon up to $0.75\,\mathrm{s}$, all DBN variants achieve a TPR close to 1.0. They continue to perform similarly until a prediction horizon of about $1.3\,\mathrm{s}$, after which point the driver aware models *DBN.pvh* and *DBN.pvg* obtain a small edge. Towards a horizon of $2.0\,\mathrm{s}$, the TPR of the models drops towards 10%.

VII. Discussion

We evaluated path prediction performance in three scenario types within a time interval of a few seconds around a potential motion change: in normal scenarios with no motion change, in normal scenarios with motion change and in an anomalous scenario. We did so as reporting aggregate performance would not have been very insightful. This is because in reality, the time steps in which "normal" scenarios apply with no motion changes vastly outnumber the two other scenario types. Just considering aggregate performance would strongly favor simple models like the *LDS* (or a parameter setting of a more complex model that essentially implements such a simple model). However, the time instants involving motion changes should arguably carry more weight, as they might strongly induce changes in collision risk. Listing separate performance values for various scenario types allows to side-step this weighting issue for now.

For normal scenarios with no motion change, the single-motion model *LDS* performs best in terms of Euclidean distance error, albeit with by far the worst *loglik* performance of all models. Context-aware models (*DBN.p*, *DBN.pv*, *DBN.pvh*, *DBN.pvg)* were at least on-par-with their context-agnostic (multi-motion) versions (*SLDS*). They remained competitive with the *LDS* on Euclidean distance error. The normal scenarios with motion changes are those settings where the context-aware models can potentially shine. Indeed, we found the context-aware models to mostly outperform their context-agnostic counterparts (*LDS* and *SLDS*). Anomalous situations which defy the anticipated motions, but still occur in real-world traffic, provide a challenge to a context-aware model. They might contradict the expert knowledge encoded in the DBN structure or will not adhere to the parameters estimated on a training set. Fortunately, the probabilistic modeling allows for softer decisions: the switch of motion dynamics not only depends on the pre-conditioning context, but also on the current positional observations. Indeed, the performances of context-aware models were shown to remain competitive with that of context-agnostic counterparts.

Overall, one observes that the models using both pedestrian and vehicle context (*DBN.pv*, *DBN.pvg*, *DBN.pvh*) performed best over the three time scenario types. Full context was not shown to improve path prediction performance (i.e. *DBN.pvg* and *DBN.pvh* not outperforming *DBN.pv*). While *DBN.pv*, *DBN.pvh* and *DBN.pvg* encode typical vehicle braking locations, variation in braking behavior seems to limit the predictive value of the driver awareness cue. Contrary to our expectations, measuring driver gaze (*DBN.pvg*) yielded similar path prediction and collision risk estimation performance compared

to measuring driver head pose (*DBN.pvh*), i.e. see Figure 7. However, when multiple road users or driving distractions are introduced, it is likely that driver awareness will be disambiguated more accurately from gaze compared to head-pose. Other fixation-related metrics may provide further insights in driver awareness, such as number of fixations, total fixation duration and angle of first saccade landing within 2° of the pedestrian [36], though such evaluations would require natural as opposed to instructed viewing behavior, and other spatial regions competing for attention.

In this paper, we chose to model mutual awareness and interaction between vehicle and pedestrian loosely, by means of the shared context state CC (collision course) of the respective DBN sub-graphs. This has the advantage that we could easily scale-up to multiple road users, as their DBN sub-graphs can be designed and optimized individually, and the number of dependencies grow linearly. On the other hand, some limitations result from this loose motion coupling. The driver-aware models (*DBN.pvh, DBN.pvg*) encode the following: if one road user *A* is aware of the other *B*, this influences the motion of *A* which affects the shared collision course latent state CC, which in turn influences the motion of *B*. Not modeling the dependency between awareness of *A* and motion of *B* directly might lead to decreased performance. Consider the path prediction performance of the vehicle in scenarios 5 and 7. In both scenarios, the driver sees the pedestrian, however, only in scenario 7 the vehicle stops (due to the unaware pedestrian). The fact that the vehicle motion in the driver-aware models is not directly influenced by the pedestrian's awareness might contribute to why *DBN.pvh* and *DBN.pvg* are not the best performing models for scenario 7.

DBNs provide a versatile structure to model expert knowledge. Dependencies amongst pairs of road users could be added, but limiting them to close spatial proximity, to remain scalable with increasing number of road users. Additional cues could be integrated, such as "exchanged" awareness [37], i.e. modeling the driver's belief about the pedestrian's awareness in addition to the driver's awareness of the pedestrian's presence.

One of the main insights of this paper is that context cues can help. However, simply using more complex motion models with additional context cues does not necessarily help prediction performance, if those context cues are not sufficiently informative or they cannot be reliably inferred from sensor measurements. Differences in path prediction performance between context cues can be very subtle and might also not materialize due to small data sample effects and due to errors in the estimation of ground truth.

VIII. Conclusion

We presented a novel method for vehicle-pedestrian path prediction that takes into account the awareness of the driver and the pedestrian towards each other. The method jointly modeled the paths of a vehicle and a pedestrian within a single Dynamic Bayesian Network (*DBN*). Subsequently, collision risk was estimated by a probabilistic intersection operation. Overall, this work demonstrated an integrated system from on-board sensing up to collision warning.

We evaluated the incremental benefits of pedestrian- and vehicle-context in six models with varying access to the used context cues, namely Linear Dynamical System (*LDS*, one motion model), Switching Linear Dynamical System (*SLDS*, two motion models), *DBN.p* (pedestrian aware), *DBN.pv* (vehicle-aware and driver-agnostic), *DBN.pvg* (driver-gaze as awareness cue) and *DBN.pvh* (driver head pose as awareness cue).

For normal scenarios with no motion change, the single-motion model *LDS* performed best in terms of Euclidean distance error, albeit with the worst *loglik* performance by far of all models. Context-aware models (*DBN.p*, *DBN.pv*, *DBN.pvh*, *DBN.pvg)* were at least on-par-with their context-agnostic (multi-motion) versions (*SLDS*). They remained competitive with the *LDS* on Euclidean distance error. On the normal scenarios with motion changes we found the context-aware models to mostly outperform their context-agnostic counterparts (*LDS* and *SLDS*). Even in an anomalous scenario, the performances of context-aware models were shown to remain competitive with that of context-agnostic counterparts. Overall, models using both pedestrian and vehicle context (*DBN.pv*, *DBN.pvg*, *DBN.pvh*) performed best on path prediction. This was also reflected in collision risk estimation performance. For example, the collision risk warning true positive rate (TPR) was raised from 18% (pedestrian-aware model *DBN.p* of Kooij *et al*. [6]) to 27% for *DBN.pvg* for a prediction horizon of 1.5 s and a false positive rate (FPR) of 1% over the dataset.

Future work could involve improved pedestrian localization (e.g. sensor data fusion), additional and more realistic motion models within the SLDS, and more sophisticated context modeling (e.g. driver awareness by fixation cues). Tests are needed on large naturalistic datasets, consisting of a rich set of traffic scenarios with possibly multiple road users.

Acknowledgment

We thank Ewoud Pool for sharing his code and expertise on DBN. We also thank Markus Braun for providing pedestrian detections and head pose on our dataset. This project was partially funded by the NWO-TTW Foundation, the Netherlands, within the IAVTRM project (#13712).

References

[1] WHO, "Global Status Report on Road Safety," *World Health Organization*, pp. 1–20, 2018.
[2] European Commission, "Pedestrians and Cyclists," *European Commission, Directorate General for Transport*, pp. 1–44, February 2018.
[3] C. G. Keller, T. Dang, H. Fritz, A. Joos, C. Rabe, and D. M. Gavrila, "Active pedestrian safety by automatic braking and evasive steering," *IEEE Trans. on Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1292—-1304., 2011.
[4] H. Winner, "Fundamentals of Collision Protection Systems," in *Handbook of Driver Assistance Systems*. Springer, 2016, pp. 1149–1176.
[5] J. F. P. Kooij, N. Schneider, F. Flohr, and D. M. Gavrila, "Context-based Pedestrian Path Prediction," in *European Conf. on Computer Vision (ECCV)*, 2014, pp. 618–633.
[6] J. F. P. Kooij, F. Flohr, E. A. I. Pool, and D. M. Gavrila, "Context-Based Path Prediction for Targets with Switching Dynamics," *Int. Journal of Computer Vision (IJCV)*, vol. 127, no. 3, pp. 239–262, 2019.
[7] E. A. I. Pool, J. F. P. Kooij, and D. M. Gavrila, "Crafted vs. Learned Representations in Predictive Models - A Case Study on Cyclist Path Prediction," *IEEE Trans. on Intelligent Vehicles*, DOI:10.1109/TIV.2021.3064253, 2021.

[8] S. Lefèvre, D. Vasquez, and C. Laugier, "A survey on motion prediction and risk assessment for intelligent vehicles," *ROBOMECH Journal*, vol. 1, no. 1, p. 1, 2014.

[9] D. Ridel, E. Rehder, M. Lauer, C. Stiller, and D. Wolf, "A Literature Review on the Prediction of Pedestrian Behavior in Urban Scenarios," in *IEEE Int. Conf. on Intelligent Transportation Systems (ITSC)*, 2018, pp. 3105–3112.

[10] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras, "Human motion trajectory prediction: a survey," *The Int. Journal of Robotics Research*, vol. 39, no. 8, pp. 895–935, 2020.

[11] M. Braun, S. Krebs, F. Flohr, and D. Gavrila, "EuroCity Persons: A Novel Benchmark for Person Detection in Traffic Scenes," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1844–1861, 2019.

[12] A. Palffy, J. Dong, J. F. P. Kooij, and D. M. Gavrila, "CNN based road user detection using the 3d radar cube," *IEEE Robotics and Automation Letters (RAL)*, vol. 5, no. 2, pp. 1263–1270, 2020.

[13] J. R. van der Sluis, E. A. I. Pool, and D. M. Gavrila, "An Experimental Study on 3D Person Localization in Traffic Scenes," in *IEEE Intelligent Vehicles Symposium (IV)*, 2020.

[14] M. Roth, D. Jargot, and D. M. Gavrila, "Deep End-to-end 3D Person Detection from Camera and Lidar," in *IEEE Int. Conf. on Intelligent Transportation Systems (ITSC)*, 2019, pp. 521–527.

[15] C. G. Keller and D. M. Gavrila, "Will the Pedestrian Cross? A Study on Pedestrian Path Prediction," *IEEE Trans. on Intelligent Transportation Systems*, vol. 15, no. 2, pp. 494–506, 2014.

[16] R. Quintero, I. Parra, D. F. Llorca, and M. A. Sotelo, "Pedestrian Intention and Pose Prediction through Dynamical Models and Behaviour Classification," in *IEEE Int. Conf. on Intelligent Transportation Systems (ITSC)*, 2015, pp. 83–88.

[17] M. Roth, F. Flohr, and D. M. Gavrila, "Driver and Pedestrian Awareness-based Collision Risk Analysis," in *IEEE Intelligent Vehicles Symposium (IV)*, 2016, pp. 454–459.

[18] E. A. I. Pool, J. F. P. Kooij, and D. M. Gavrila, "Using Road Topology to Improve Cyclist Path Prediction," in *IEEE Intelligent Vehicles Symposium (IV)*, 2017, pp. 289–296.

[19] S. Neogi, M. Hoy, K. Dang, H. Yu, and J. Dauwels, "Context Model for Pedestrian Intention Prediction Using Factored Latent-Dynamic Conditional Random Fields," *IEEE Trans. on Intelligent Transportation Systems*, pp. 1–12, 2020.

[20] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *IEEE Int. Conf. on Computer Vision (ICCV)*, 2009, pp. 261–268.

[21] G. Welch and G. Bishop, "An Introduction to the Kalman Filter," *In Practice*, vol. 7, no. 1, pp. 1–16, 2006.

[22] N. Schneider and D. M. Gavrila, "Pedestrian Path Prediction with Recursive Bayesian Filters: A Comparative Study," in *German Conf. on Pattern Recognition (DAGM GCPR)*, 2013, pp. 174–183.

[23] Y. Li, X.-Y. Lu, J. Wang, and K. Li, "Pedestrian Trajectory Prediction Combining Probabilistic Reasoning and Sequence Learning," *IEEE Trans. on Intelligent Vehicles*, pp. 1–1, 2020.

[24] D. Helbing and P. Molnár, "Social force model for pedestrian dynamics," *Physical Review E*, vol. 51, no. 5, pp. 4282–4286, 1995.

[25] C. Braeuchle, J. Ruenz, F. Flehmig, W. Rosenstiel, and T. Kropf, "Situation analysis and decision making for active pedestrian protection using Bayesian networks," in *6. Tagung Fahrerassistenzsysteme, TÜV SÜD*, 2013.

[26] S. Gupta, M. Vasardani, and S. Winter, "Negotiation between Vehicles and Pedestrians for the Right of Way at Intersections," *IEEE Trans. on Intelligent Transportation Systems*, vol. 20, no. 3, pp. 888–899, 2019.

[27] T. P. Minka, "A family of algorithms for approximate Bayesian inference," Ph.D. dissertation, Massachusetts Institute of Technology, 2001.

[28] H. Saptoadi, "Suitable Deceleration Rates for Environmental Friendly City Driving," *Int. Journal of Research in Chemical, Metallurgical and Civil Engineering*, vol. 4, no. 1, pp. 2–5, 2017.

[29] M. I. Nazir, K. M. A. Al Razi, Q. S. Hossain, and S. K. Adhikary, "Pedestrian Flow Characteristics At Walkways In Rajshahi Metropolitan City Of Bangladesh," in *Int. Conf. on Civil Engineering for Sustainable Development*, 2014, pp. 978–984.

[30] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *International Conference on Learning Representations*, 2015.

[31] L. Ferranti, B. Brito, E. Pool, Y. Zheng, R. M. Ensing, R. Happee, B. Shyrokau, J. F. P. Kooij, J. Alonso-Mora, and D. M. Gavrila, "SafeVRU: A Research Platform for the Interaction of Self-Driving Vehicles with Vulnerable Road Users," in *IEEE Intelligent Vehicles Symposium (IV)*, 2019, pp. 1660–1666.

[32] T. Moore and D. Stouch, "A Generalized Extended Kalman Filter Implementation for the Robot Operating System," in *Int. Conf. on Intelligent Autonomous Systems*, 2016, pp. 335–348.

[33] M. Roth and D. M. Gavrila, "DD-Pose - A large-scale Driver Head Pose Benchmark," in *IEEE Intelligent Vehicles Symposium (IV)*, 2019, pp. 927–934.

[34] H. Hirschmüller, "Stereo Processing by Semi-Global Matching and Mutual Information," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 328–341, 2008.

[35] M. Braun, Q. Rao, Y. Wang, and F. Flohr, "Pose-RCNN: Joint Object Detection and Pose Estimation Using 3D Object Proposals," in *IEEE Int. Conf. on Intelligent Transportation Systems (ITSC)*, 2016, pp. 1546–1551.

[36] J. Stapel, M. E. Hassnaoui, and R. Happee, "Measuring Driver Perception: Combining Eye-Tracking and Automated Road Scene Perception," *Human Factors*, 2020.

[37] Y. Wang, Y. Ren, S. Elliott, and W. Zhang, "Enabling Courteous Vehicle Interactions through Game-based and Dynamics-aware Intent Inference," *IEEE Trans. on Intelligent Vehicles*, vol. 5, no. 2, pp. 217–228, 2020.

Markus Roth received the Diploma degree in computer science from Karlsruhe Institute of Technology, Germany, in 2014. Since then he is working toward the Ph.D. degree at TU Delft, The Netherlands. He is also currently with Mercedes-Benz R&D in the Environment Perception department, Stuttgart, Germany. His research interests include machine learning and video analysis for driver analysis, with a focus on driver head pose estimation and joint awareness between driver and pedestrians.

Jork Stapel received the MSc degree in control and simulation at the faculty of Aerospace Engineering at the TU Delft, The Netherlands, in 2015. He received the Ph.D. degree investigating on-road assessment of driver state and driver behavior in automated driving in 2021 at the same university.

Riender Happee received the Ph.D. degree from TU Delft, The Netherlands, in 1986 and 1992, respectively. He investigated road safety and introduced biomechanical human models for impact and comfort at TNO Automotive (1992-2007). Currently, he investigates the human interaction with automated vehicles focussing on safety, comfort and acceptance at the Delft University of Technology, the Netherlands, where he is an Associate Professor.

Dariu M. Gavrila received the Ph.D. degree in computer science from Univ. of Maryland at College Park, USA, in 1996. From 1997 until 2016, he was with Daimler R&D, Ulm, Germany, where he became a Distinguished Scientist. In 2016, he moved to TU Delft, where he since heads the Intelligent Vehicles group as a Full Professor. His current research deals with sensor-based detection of humans and analysis of behavior in the context of self-driving vehicles. He was awarded the Outstanding Application Award 2014 and the Outstanding Researcher Award 2019, both from the IEEE Intelligent Transportation Systems Society.