Intentions of Vulnerable Road Users—Detection and Forecasting by Means of Machine Learning

Michael Goldhammer, Sebastian Köhler, Stefan Zernetsch, Konrad Doll, *Member, IEEE*, Bernhard Sick, and Klaus Dietmayer

***Abstract*— Avoiding collisions with vulnerable road users (VRUs) using sensor-based early recognition of critical situations is one of the manifold opportunities provided by the current development in the field of intelligent vehicles. As, especially, pedestrians and cyclists are very agile and have a variety of movement options, modeling their behavior in traffic scenes becomes a challenging task. In this paper, we propose movement models based on machine learning methods, in particular, artificial neural networks, in order to classify the current motion state and to predict the future trajectory of the VRUs. Both model types are also combined to enable the application of specifically trained motion predictors based on a continuously updated pseudo probabilistic state classification. Furthermore, the architecture is used to evaluate motion-specific physical models for starting and stopping and video-based pedestrian motion classification. A comprehensive dataset consisting of a total of 1068 pedestrian and 494 cyclist scenes acquired at an urban intersection is used for optimization, training, and evaluation of the different models. The results show substantially higher classification rates and the ability, through the machine learning approaches, to earlier recognize motion state changes than by the way of interacting multiple model (IMM) Kalman filtering. The trajectory prediction quality has also been improved for all kinds of test scenes, especially when starting and stopping motions are included. Here, 37% and 41% fewer position errors were achieved on average, respectively.**

***Index Terms*— Road safety, vulnerable road users, movement modeling, intention recognition, motion classification, trajectory prediction, artificial neural networks.**

I. Introduction

According to the World Health Organization's status report on road safety, traffic accidents currently constitute the leading cause of death for young people aged 15 – 29 years. Moreover, they are also one of the most frequent causes among most other age classes [1]. About half of those cases concern vulnerable road users (VRU), i. e., pedestrians, cyclists, and motorcyclists. While the progress of active and passive safety functions in the last decades has steadily improved the protection of car passengers, the protection of VRUs still remains a critical issue. Passive safety concepts such as helmets or energy-absorbing vehicle design play an important role, but often cannot prevent severe injuries, even at relatively low velocities within urban accident scenarios. A unique opportunity to close this gap between vehicle and VRU safety is provided by the current progress in the fields of advanced driver assistant systems (ADAS) and autonomous driving: By predicting critical situations and thus being able to take active countermeasures at an early state, VRU accidents could be avoided, or at least, their consequences could be reduced significantly. A fundamental task in this context is the creation of suitable VRU movement models. In particular for pedestrians and cyclists, behavioral prediction is challenging as they often do not use specific traffic lanes, can abruptly change their motion state (e. g., starting, stopping, turning off) and do not actively communicate their intention through indicators or brake lights. Considering future applications, they also hardly have an option to broadcast their motion state and intention via vehicle-to-everything (V2×) communication as motor vehicles do.

Addressing those issues, we propose and investigate novel VRU movement models based on machine learning techniques in combination with large-scale realistic training data from public traffic scenes. We use an offline learning approach, meaning that we first train the models and then the models are evaluated without further learning input. Our concept includes the detection of motion state changes in an early phase, the prediction of the future trajectory and the combination of both approaches. The models are designed with the goals of having few requirements regarding prior knowledge and a high degree of independence from general conditions such as traffic and surroundings. Thus, only the VRU's trajectory, i. e., past positions, observed in any global coordinate system by an appropriate sensor technique, is sufficient as input data. Additional environment information, e. g., map data, is not required. Based on the same data, also other state-of-the-art models using video features and physical motion parameters are optimized and compared to the proposed approaches.

The remainder of the article is structured as follows: In Section II, an overview of the related work in the areas

Manuscript received July 13, 2017; revised March 8, 2018, August 30, 2018, and March 11, 2019; accepted May 17, 2019. Date of publication June 27, 2019; date of current version June 29, 2020. This work partially results from the project Aktiver Fußgängerschutz (AFUSS), supported in part by the German Federal Ministry of Education and Research (BMBF) under Grant 03FH021I3, and in part by the project Detecting Intention of Vulnerable Road Users based on Collective Intelligence as a Basis for Automated Driving (DeCoInt^2), supported by the German Research Foundation [Deutsche Forschungsgemeinschaft (DFG)] within the priority program SPP 1835: "Kooperativ Interagierende Automobile", under Grant DO 1186/1-1 and Grant SI 674/11-1. Additionally, the work is supported by "Zentrum Digitalisierung Bayern." The Associate Editor for this paper was N. Zheng. *(Corresponding author: Sebastian Köhler.)*

M. Goldhammer, S. Köhler, S. Zernetsch, and K. Doll are with the Faculty of Engineering, University of Applied Sciences Aschaffenburg, 63743 Aschaffenburg, Germany (e-mail: michael.goldhammer@h-ab.de; sebastian.koehler@h-ab.de; stefan.zernetsch@h-ab.de; konrad.doll@h-ab.de).

B. Sick is with the Intelligent Embedded Systems Lab, University of Kassel, 34125 Kassel, Germany (e-mail: bsick@uni-kassel.de).

K. Dietmayer is with the Institute of Measurement Control and Microtechnology, Ulm University, 89081 Ulm, Germany (e-mail: klaus.dietmayer@uni-ulm.de).

Digital Object Identifier 10.1109/TITS.2019.2923319

of pedestrian and cyclist motion modeling and prediction is given. Our approach, broken down into the individual main processing steps, is presented afterwards in Section III. In Section IV, the test site and the VRU datasets used for training and evaluation of the models are described. The performed experiments and results are discussed in Section V, while a final conclusion is drawn and a brief outlook is given in Section VI.

II. Related Work

Modeling pedestrian and cyclist movement has already been a task in different areas of research, e. g., biomechanics, physiotherapy and sports sciences. The main objective of those models is the analysis of basic movement parameters for understanding and improving motion sequences or their individual components for certain groups or activities, e. g., physically restricted persons or athletes in popular and competitive sports. In contrast, the main purpose of VRU movement modeling is to forecast the short-term behavior for a continuous analysis of traffic surroundings. In this sector a certain number of studies concern pedestrians, but hardly any papers on bicyclists were published within the last years.

A frequently used technique for trajectory prediction in many applications is Kalman Filtering (KF), where physical state variables are assumed as constant whenever the state information cannot be updated by observation [2]. Published approaches addressing pedestrians use constant velocity (CV), constant acceleration (CA), and constant turn rate (CT) models, or even combinations of these by way of interacting multiple model (IMM) filters [3]–[6]. Kalman Filtering offers the advantage of few requirements regarding prior knowledge and, therefore, it is suitable in many cases, e. g., small time horizons or walking/cycling at steady state velocity. However, as a matter of principle, it results in larger prediction errors whenever motion state changes occur and, thus, the basic assumption is invalidated, e. g., during starting and stopping. Furthermore, the detection of motion transitions can also be performed with Kalman Filtering [2], which is an important aspect of VRU intention recognition, and can serve as a basis for classifying critical situations (e. g., a person suddenly stepping out on the road) or for choosing suitable movement models. A common approach is the usage of the model probability of an IMM-KF with constant position (CP) and CV model to distinguish standing from walking motion [7]. Another trajectory-based approach is presented by Wakim *et al.*, where the four classes *standing still*, *walking*, *jogging*, and *running* are modeled by Hidden Markov Models (HMM) and updated based on the measured absolute and angular velocity [8].

Besides these exclusively trajectory-based approaches, additional sensor specific information, e. g., gathered from monocular cameras, stereo cameras, or LiDAR systems is used in other publications. Keller and Gavrila compare stopping motion detection by KF approaches to stereo-vision-based methods using dense optical flow [9]. They also perform a trajectory prediction and show that the KF methods are outperformed in this case. Quintero *et al.* determine the joint positions of pedestrians on data acquired by a high-resolution stereo camera and a LiDAR system and characterize the movement via Gaussian Process Dynamical Models (GPDM) [10], [11]. They perform a state classification within the classes *standing*, *starting*, *walking*, and *stopping* as well as a trajectory prediction for a time horizon of one second. The head pose and associated direction of view are detected and tracked in camera images by Schulz and Stiefelhagen [12]. They use this feature, among others, to recognize the intention of pedestrians to cross the road by way of different models, primary IMM filters, and Latent-Dynamic Conditional Random Fields (LDCRF) [13], [14]. Kooij *et al.* use camera-based context information of the scene environment such as the distance of the pedestrian to the curb and his head pose (line of sight) in order to rate the criticality of situations. Using a novel Dynamic Bayesian Network, they perform an early recognition of motion state transitions [15]. Pool *et al.* use road topologies to improve cyclist path prediction [16]. More recently, approaches based on deep learning methods have been proposed by Kim *et al.* [17], Alahi *et al.* [18], Völz *et al.* [19] and Bartoli *et al.* [20]. They use recurrent neural networks with long short-term memory (LSTM) models.

In the present article, the work listed in this paragraph is extended, combined and extensively evaluated. A trajectory prediction method based on two physical models of the pedestrian starting motion is published in [21], while corresponding analyses for stopping motions are given in [22]. A machine learning model using the time series of the pedestrian's ego velocity as input pattern for Multi Layer Perceptrons (MLP) is presented in [23]. This concept is extended with a polynomial representation by least-squares approximation of the input and output time series, leading to a reduction of the feature space dimensionality and an increase of generalizability [24]. In [25], a modification of this model is used for the early recognition of the starting movement intention of pedestrians. A transfer of the approaches with physical and MLP models to cyclists is published in [26]. Video-based methods for motion classification and early recognition of state changes of pedestrians are presented in [27] and [28]. A Motion History Image (MHI) based histogram feature vector (MCHOG) approach is used as input for the classification of monocular images of a static camera [27] and of stereo images from a moving vehicle [28] by Support Vector Machines (SVM).

Besides extending our already published work and putting it into context, the main contributions of this paper are: (1) the combination of motion state recognition and trajectory prediction, (2) an extensive publicly available dataset for pedestrian and cyclist trajectories, and (3) a detailed and extensive evaluation of the proposed approach based on a performance measure taking into account mean detection times and precision/F_1 scores dependent on an intention probability threshold.

III. Methodology

For the recognition of motion states as well as the prediction of the future trajectory we use a method of polynomial approximation of time series in combination with MLPs,

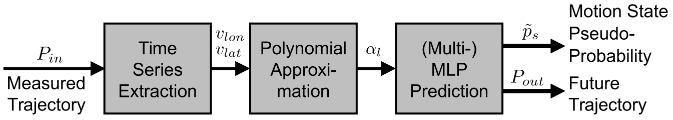

Fig. 1. Overview of the basic processing steps of the method PolyMLP.

which we call "PolyMLP". As input at the current, discrete time step k we use the observed trajectory consisting of N past VRU positions $\mathcal{P}_{in} = \{P_{k-N}, \ldots, P_k\}$ in a world coordinate system. Discretization is done by sampling measurements. As one target output for motion state recognition, a pseudo class posterior probability $\tilde{p}_s$ with $s \in \mathcal{S} = \{Waiting, Starting, Moving, Stopping\}$ of the four considered motion states s is provided. The method is compared to CP/CV-IMM-KF classification as baseline and to the MCHOG/SVM approach as a solely video-based classifier. The past positions are also used for trajectory prediction, which outputs a trajectory of M future positions $\mathcal{P}_{out} = \{P_{k+1}, \ldots P_{k+M}\}$. Each position P_i with $i \in \{k-N, \ldots, k, \ldots, k+M\}$ is represented by 2D world coordinates $P_i = (x_i, y_i)$. The third dimension (height) is not used in our approach. The resulting position accuracy is compared to CV-KF as baseline method and to physical models for starting and stopping optimized on the same training data. An overview of the input/output behavior is given in Fig. 1.

In a final step, we combine the methods of motion state recognition and trajectory prediction to use motion state dependent optimized movement prediction models (e. g. specific trajectory prediction models only trained with starting movements) and evaluate their benefits compared to the single-stage ("monolithic") approach.

Within this section, the successive steps of extracting characteristic time series (Sec. III-A), their representation by polynomial least-squares approximation (Sec. III-B), the prediction by MLP (Sec. III-C) as well as the target outputs of the motion state (Sec. III-D) and the future trajectory (Sec. III-E) are described. Afterwards, in Sec. III-F the motion state specific trajectory prediction model using multiple specifically trained MLPs is presented.

A. Time Series Extraction

The PolyMLP method solely uses the information within the observed VRU trajectory in a static world coordinate system for the prediction of the motion state and/or the future trajectory. Therefore, it is very flexible with regard to the applied sensor system, but also allows for an easy extension of the input space with additional information.

For the acquisition of the trajectories evaluated within this study, an infrastructure-based wide-angle stereo camera system for the pedestrians and, alternatively, an array of multilayer laser scanners for the cyclists is used (see Sec. IV). As an anchor point for measuring trajectories and tracking in 3D world space we use the center of the head detected from video data (see [23]), or the VRUs' centers of gravity (COG) determined from the laser scanner generated point cloud, respectively. As changes of the motion state, especially the beginning of the pedestrian starting motion, are initialized by a slight upper body bending leading to a shift of the COG into the direction of movement [29], the head movement can serve as an early indicator for the intention.

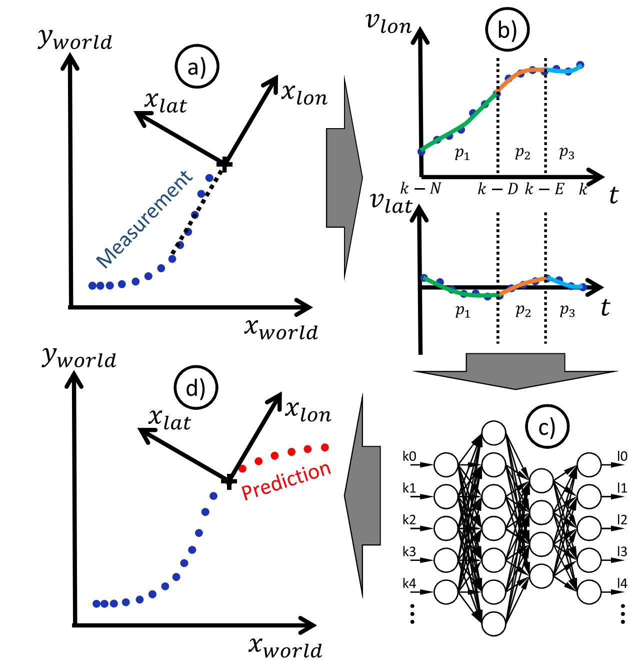

Fig. 2. Overview of the proposed path prediction method.

To extract a time series representing the VRU motion independent from the absolute position and direction, several approaches were implemented and compared, e. g., the two-dimensional trajectory or velocity in the pedestrian's ego coordinate system or the combination of the absolute and angular velocity. The finally applied configuration uses the velocity v_{lon}, v_{lat} as a numerical differentiation of position P_i at the discrete time steps $k, k-1, k-2, \ldots, k-N$ (see Fig. 2 b) in the VRU's ego coordinate system x_{lon}, x_{lat} at current time k (see Fig. 2 a, d). The time series v_{lon} and v_{lat} are additionally processed with a first order exponential smoothing filter.

B. Representation With Approximating Polynomials

The time series extracted within a certain time window, e.g., $\{k-N, \ldots, k\}$, can directly serve as an input pattern of a machine learning predictor [23]. However, we proceed to a further level of abstraction by using the coefficients α_l of an approximating polynomial $\tilde{v}_{lon}(t)$ of v_{lon} based on an orthogonal expansion with polynomials $p_l(t)$:

$$\tilde{v}_{lon}(t) = \sum_{l=0}^{L} \alpha_l \, p_l(t) \tag{1}$$

with l representing the degree of $p_l(t)$ and L the degree of $\tilde{v}_{lon}(t)$. The same approximation is done for v_{lat}. The advantages of this step are a high grade of independence from the input data sampling rate and a significant reduction of the dimensionality of the feature space. Dependent on the polynomial degree and the length of the time window,

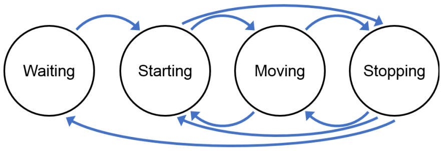

Fig. 3. Modeled motion states and physically possible state transitions.

the approximation also reduces the influence of measurement noise. As the coefficients of the orthogonal expansion are optimal estimates for the average values of the time series' derivatives, they can be interpreted as mean position, velocity, acceleration, etc. during the considered time window. Using the polynomials and update algorithms for sliding window processing of time series presented by Fuchs *et al.* [30], the approximation can be performed very efficiently. In order to allow the machine learning prediction algorithm for weighting different time periods separately as it is possible with direct input of the time series elements, we make use of multiple polynomials per dimension v_{lon}, v_{lat} fitted in sequential temporal sub-windows. Their number, temporal position, length and polynomial degree are model parameters and varied within the optimization. A schematic example with three sub-windows $\{k-N,\dots,k-D\}$, $\{k-D+1,\dots k-E\}$, $\{k-E+1,\dots,k\}$ with $k-N<k-D<k-E<k$ is shown in Fig. 2 b (polynomials p_0, p_1, and p_2).

C. Multilayer Perceptron for Prediction

We apply an MLP to predict the motion state and/or the future trajectory of the VRU based on the extracted patterns (see Fig. 2 c). The MLP provides high flexibility and efficiency for the given application as it allows predicting multiple output dimensions at the same time. The applied artificial neural network consists of neurons with sigmoid activation functions, whereas other configurations (identity, Gaussian) were evaluated, too. After normalizing the input patterns using a z-transformation, a training is performed with the Resilient Backpropagation (RPROP) algorithm [31]. The number and sizes of the hidden layers are variable and also part of the optimization process.

D. Recognition of Motion State

For the recognition of the motion state, a model of the four states *Waiting*, *Starting*, *Moving*, and *Stopping* is used, where *Waiting* is defined for all time steps where the pedestrian stands at a fixed position while upper body movements are possible, *Starting* is defined to be between the last *Waiting* time step and the time step when he reaches his constant walking velocity, *Stopping* is defined to be during the deceleration from constant walking velocity to standstill, and *Moving* is defined to be between the last *Starting* time step and the first *Stopping* time step. This basically allows the 8 (out of a possible basic quantity of 12) state transitions depicted in Fig. 3, as the transitions between *Waiting* and *Moving* are always separated by one of the states *Starting* or *Stopping*. For the online classification, an MLP output layer with 4 neurons, each representing one pseudo class posterior probability $\tilde{p}_s$ for each state, is configured. The MLP input layer consists of the polynomial coefficients of the input trajectory which are extracted in the same manner as for the trajectory prediction MLP (see Section III-B). The trajectories used for training are labeled with a distinct class label for every time step, generating output training patterns with a one (1) at the element representing the current ground truth class and zeros (0) at all others. The detection of state transitions results from changes in the output probabilities $\tilde{p}_s$ over time and, therefore, no further measures are taken to define physically meaningful transitions or transition probabilities explicitly, as the neural network should be given the possibility to learn these aspects from scratch without further conditions. With this approach, the trained classifier provides a pseudo-probabilistic rating for each state. As the four output neurons have no direct interconnection, the cumulative probability does not explicitly add up to 100%. Thus, an additional scaling step has to be applied if this output information is required by the subsequent application.

E. Prediction of Trajectory

The predicted trajectory is represented by polynomial coefficients, in the same manner as the past input trajectory. The coefficients are generated by the neural network. To train the predictor, again characteristic time series are extracted from the ground truth trajectory within a certain time window (cf. Section III-A) and the time series are approximated using orthogonal basis polynomials to get a low-dimensional representation independent of the sampling rate f (cf. Section III-B), which in our case is 50 Hz for the cameras and 12.5 Hz for the laser scanners. The choice of the MLP output pattern is completely independent of the one of the input pattern, which allows for different temporal sub-windows and polynomial degrees. Each polynomial coefficient is estimated by a separate MLP output neuron, which also uses a z-transformation for normalization. During online application, the process is performed vice versa: The time series is evaluated by sampling the output polynomials at the requested discrete points in time. It is then transformed to an estimation of the future trajectory in the original global coordinate system using the current VRU position and movement direction (see Fig. 2 d).

F. State-Specific Trajectory Prediction

After training, the weights of the MLP computed in Sec. III-E implicitly contain the knowledge of the trained movements (starting, stopping, etc.) within a single model. Besides this monolithic approach, a second approach combining state classification with several specifically trained models for trajectory prediction is implemented and evaluated. The architecture of the prediction system using the PolyMLP method is shown in Fig. 4.

The polynomial coefficients representing the measured trajectory are fed into a classifier module that predicts the current motion state as described in Sec. III-D. In parallel, the trajectory is fed into several (here: four) PolyMLP modules for trajectory prediction. Each predictor is specifically trained with scenes containing movements according to one output

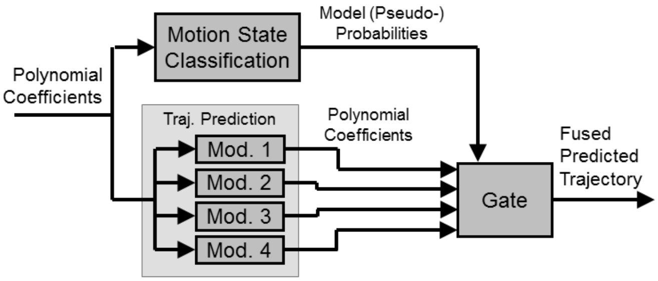

Fig. 4. Motion-state-specific trajectory prediction.

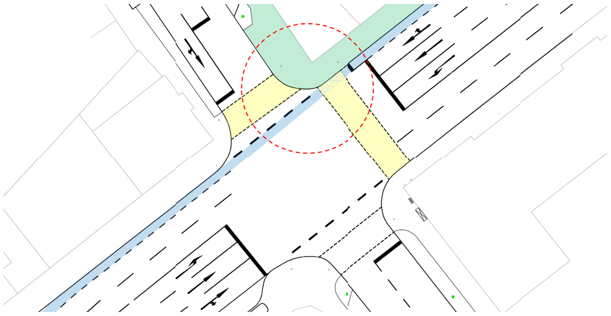

Fig. 5. Map of the public intersection used to acquire VRU behavior data. The highlighted areas represent the observed sidewalk, the crosswalks, and the bicycle lane. The area of overlapping fields of view of the high resolution cameras allowing high-precision 3D position measurements is marked by the circle.

of the classifier model, i. e. motion state. As the prediction modules can use the same configuration for preprocessing the trajectory and polynomial approximation, only the output of the specifically trained MLPs has to be processed separately. The predicted polynomial coefficients $\alpha_{l,s}$ with $s \in \mathcal{S}$ are fed into a gating module and added as a weighted sum based on the corresponding class posterior probabilities $\tilde{p}_s$:

$$\alpha_{l,fused} = \sum_{s \in \mathcal{S}} \alpha_{l,s} \cdot \tilde{p}_s \qquad (2)$$

Finally, the resulting fused coefficient set is used to generate the final trajectory prediction. This method of explicit separation of classification and trajectory prediction has the advantage that the classification as well as the prediction modules can be replaced by other approaches. For our evaluation we are thus able to use the MCHOG/SVM approach within the classification module or physical starting and stopping models in the prediction module as comparison techniques.

IV. Test Sites and Data Sets

For training, optimization, and evaluation of the models, appropriate datasets containing realistic VRU data play an essential role. We use an urban traffic intersection in order to gain measurements of realistic, unaffected pedestrian and cyclist behavior under real conditions. The observed area includes two sidewalks, two crosswalks, and a bicycle lane which is separated from the motor vehicle lanes by road markings (see Fig. 5). The intersection is equipped with eight low-resolution (640 × 480 px) and two high-resolution (1920 × 1080px) grayscale cameras, as well as 14 eight-layer laser scanners. All sensors are mounted on infrastructure elements, e. g., street lamps and traffic light posts, at in heights of between 4 and 11 meters in order to reduce the risk of occlusions. The sensor system is described in detail in [32]. The high-resolution cameras are pointed towards a part of the intersection where two crosswalks meet and, therefore, a high number of pedestrians is expected (Fig. 5, dashed circle). They form a wide-angle stereo system allowing precise 3D measurements with an accuracy better than 3 cm for corresponding image points. The laser scanners cover the central intersection and the area of three approaching roads up to 100 m. They provide object point clouds and are used to track cyclists in a larger area beyond the stereo range of the camera system. The low-resolution cameras are currently only used for manual scene observation and the labeling of movement phases.

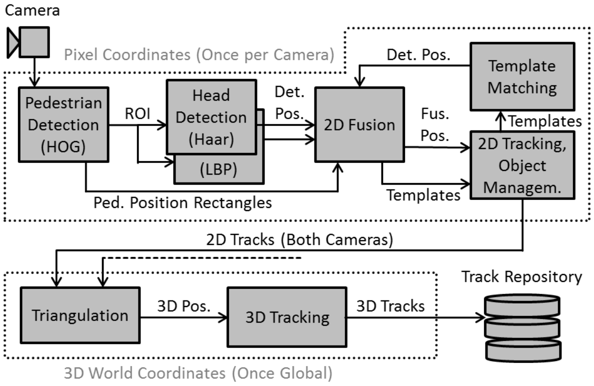

Fig. 6. Framework for extraction of trajectories.

Fig. 6 shows the framework for the extraction of pedestrian and cyclist head trajectories from stereo video data. At the first stage, both synchronous camera frames are individually scanned by a sliding-window histogram of oriented gradients (HOG) pedestrian detector [33]. The upper half of the detection rectangles serve as region of interest (ROI) for the head detectors, which are based on Haar [34] and local binary pattern (LBP) features [35]. In the next step, the most likely head position in pixel coordinates is determined fusing HOG, Haar, and LBP detections. If the head position is already tracked, template matching is used as a fourth measurement value that is fed into the fusion module. The tracked head positions from both cameras are merged into 3D world coordinate positions via triangulation. In a final step, valid positions are connected to 3D trajectories and stored.

A. Full Pedestrian Dataset

The *Full Pedestrian Dataset* contains 1068 scenes of pedestrians of 4 to 10 seconds length recorded by the high-resolution stereo system. The scenes are categorized into one of the following scene labels (the number in brackets indicates the number of scenes):

- The category *"Moving scenes"* (288) contains pedestrians crossing the observed area without stopping or significant deceleration, including scenes of walking along the sidewalk, crossing one of the roads, walking straight ahead or making a turn.

TABLE I

Training- and Test Data in the Full Pedestrian Dataset. The Table Shows the Number of Scenes Used and of the According Extracted Patterns in Brackets

	Waiting	**Starting**	**Moving**	**Stopping**
Train	177 (54.6 k)	239 (66.5 k)	201 (45.2 k)	130 (42.6 k)
Test	82 (25.3 k)	97 (25.1 k)	87 (20.1 k)	55 (17.7 k)
Total	259 (79.8 k)	336 (91.6 k)	288 (65.3 k)	185 (60.3 k)

- The category *"Waiting scenes"* (259) contains pedestrians standing on the sidewalk, mostly waiting for a green pedestrian light signal (head velocity lower than 0.3 m/s).
- The category *"Starting scenes"* (336) contains pedestrians accelerating from a standing position. These scenes also contain up to 3 seconds before and after the actual acceleration phase if available within the data.
- The category *"Stopping scenes"* (185) contains scenes where moving pedestrians decelerate to standstill. These scenes also contain up to 3 seconds before and after the actual deceleration phase.

The *Starting* scenes include time labels of the start, and if existent within the scene, also of the end of the motions state *Starting*. The beginning is determined as the time where the ground truth head velocity exceeds 0.2 m/s, its end is labeled at the first local maximum after the velocity exceeds 80% of their steady state value. The labeling of *Stopping* is done vice versa. Table I gives an overview of the training and test scenes and extracted patterns (combination of predictor in- and output, generated for each time step with complete in- and output time window around).

B. Detailed Pedestrian Dataset

The *Detailed Pedestrian Dataset* is a subset of 136 *Starting*, 107 *Stopping*, and 69 *Moving* scenes from the full set extended by finer differentiated ground truth motion states. It contains manually determined labels from video data observation, e. g., the state, the timestamps of heel-off, the first and second heel-down, the entering of the roadway and partial occlusions.

C. Cyclist Dataset

The *Cyclist Dataset* contains 494 trajectories of bicyclists including 86 *Moving*, 133 *Waiting*, 197 *Starting* and 78 *Stopping* scenes. As the cyclists' phases of acceleration and deceleration typically take more time and distance than those of pedestrians, the laser scanner system was used to also capture tracks leaving the more limited common field of view of the high resolution stereo cameras. In return, a lower spatial and temporal resolution was accepted for those tracks.

Both the pedestrian and the cyclist dataset with labeled trajectories are available to the scientific community [36].

V. Experiments and Results

A. Training and Optimization

The training and optimization of the pedestrian models is performed using the training split (747 scenes) of the *Full Pedestrian Dataset*. The training scenes are split again into 70% for the MLP training with RPROP and 30% as validation set for the optimization of the meta parameters (pre-filtering parameters, sliding window positions and sizes, polynomial representation). The target function of the MLP training is the mean squared error (MSE) of the z-normalized output, while the meta parameters are optimized with a view to a maximum accuracy (state classification) and a minimum average specific average Euclidean error (ASAEE, see Sec. V-D). For the selection of the input time series representation (Sec. III-A) and network architecture a five-fold cross validation within the training set is applied. The best validation score was achieved using a network with 2 hidden layers of 16 and 12 neurons respectively. The results produced using the test set of the full pedestrian dataset and the ASAEE, show that the time series of the 2D velocity in the pedestrian's ego system perform best as a basis for the input patterns of both state classification (see Sec. III-D) and trajectory prediction (see Sec. III-E), while only minor differences to other tested representations (see Sec. III-A) were observed (0% – 3% for state clas., 3% – 9% for traj. pred.). The representation with polynomials shows slight advantages compared to the direct input of each time series (0% – 5%). One parameter to consider is the length of the used historical trajectory. We assume that 1.0 s time is available in most cases where VRUs are tracked in practical traffic applications, but also we investigated the influence of varying this parameter. A decrease from 1.0 to 0.2 s, for example, results in a relatively slight increase of 8% of the resulting ASAEE. This fact shows that a majority of the information used by the ANN can already be extracted from few historic trajectory elements. In contrast, a further increase of the input trajectory length only shows a very slight influence on the quality increase (<0.5% improve up to 1.4 s), and even a decrease for larger values. In order to optimize the input configuration, the number, temporal positions, and degrees of the input polynomials are varied. As a large number of possible combinations exists, a manual selection of 18 configurations with varying polynomial degrees (0 to 8) for each are trained and tested. The final outcome yields two consecutive time windows of 800 ms and 200 ms length and polynomial degrees of 3. The results show no further improvement when increasing the number of windows, and even a slightly decreasing quality for higher polynomial degrees.

For the shown evaluation, an overall prediction time horizon of 2.5 s has been chosen, although the presented methods allow a variation of this parameter. The required horizon depends on the accident prevention strategy used in a concrete practical application. A driver warning may thereby potentially need a larger time horizon than an autonomous maneuver, as the human reaction time of 1–1.5 s [37] has to be considered in addition. A study using driving simulators show appropriate driver reactions when the warning is realized 2–3 s before a potential collision [38]. For the output pattern of the trajectory prediction, the time series of the 2D position in the ego system with five consecutive sliding windows (500 ms each, polynomial degree 2) is used, leading to the chosen overall prediction time horizon of 2.5 s.

B. Motion State Classification

In the first experiment, the quality of the proposed method for motion state classification is evaluated. For comparison,

TABLE II

ACCURACY (ACC) AND F_1 SCORE OF THE CLASSIFIERS FOR THE 40 STARTING AND 35 STOPPING TEST SCENES OF THE DETAILED PEDESTRIAN DATASET

	Start		Stop	
	ACC	**F_1**	**ACC**	**F_1**
IMM-KF	0.9803	0.9565	0.9248	0.9548
PolyMLP (4 Cl.)	0.9819	0.9601	0.9381	0.9644
PolyMLP (2 Cl.)	0.9824	0.9612	0.9363	0.9636
MCHOG	0.9844	0.9659	0.9141	0.9503

an IMM-KF using CP/CV models and the directly video-based method MCHOG/SVM [28] are applied to the same scenes. As the evaluation requires the additional labels of the *Detailed Pedestrian Dataset* (i.e. the time of heel-off to distinguish between standing and starting phase), the following tests are performed using these patterns. In the experiment, the overall quality of the different motion state classifiers is evaluated by means of the accuracy (ACC)

$$ACC = \frac{TP + TN}{P + N} \tag{3}$$

and the F_1 score

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{4}$$

with the number of positives P, negatives N, true positives TP, true negatives TN and false positives FP.

As MCHOG and IMM-KF only support binary classification, the output states of the PolyMLP are combined to correspond to the classes of the comparison methods.

To detect the transition between *Waiting* and *Starting*, the classes are divided into *Waiting* and *Not Waiting* $=$ $\{$*Starting*, *Moving*, *Stopping*$\}$. Therefore, the ground truth label *Waiting* is assigned to time steps before the manually determined heel-off within the 40 *Starting* scenes. The label *Not Waiting* is assigned to the remaining time steps. Two PolyMLP predictors are trained: The first uses a four-class output and a binarization after the prediction step by thresholding the sum probability $\tilde{p}_{Sum}$ of the three estimates for *Starting*, *Moving* and *Stopping*: $\tilde{p}_{Sum} = \tilde{p}_{Starting} + \tilde{p}_{Moving} + \tilde{p}_{Stopping}$. The second is directly trained with this two-class split and uses a single output neuron to predict $\tilde{p}_{Sum}$ while the class *Waiting* is defined by the complement probability $1 - \tilde{p}_{Sum}$.

To detect the transition between *Moving* and *Stopping* the classes are divided into $\{$*Starting*, *Moving*$\}$ and $\{$*Stopping*, *Waiting*$\}$. The 35 *Stopping* scenes are also tested with a four-class model, where a threshold on $\tilde{p}_{Sum} = \tilde{p}_{Stopping} + \tilde{p}_{Waiting}$ is used for binarization. The two-class model is again trained directly with this split and predicts the output with a single neuron. As the classifier quality measures depend on the binarization threshold, the optimal value is chosen by maximization of the accuracy of the training data (an optimization of the F_1 score leads to the same operating point). The resulting quality measures of both test scene types using optimal bias values are shown in Table II. For the regarded *Starting* scenes, the quality measures show only minor differences with slight advantages for MCHOG and slight disadvantages for the IMM-KF compared to PolyMLP. Clearer differences can be observed for the *Stopping* scenes: Here, PolyMLP performs best, while IMM-KF outperforms MCHOG. The optimization thereby results in a relatively high threshold for the CV model probability P_{CV} of the IMM-KF of 98.9%, which means that stopping can be recognized at an early state.

TABLE III

CONFUSION MATRIX FOR THE CLASSIFICATION OF THE MOTION STATE. THE ROWS CONTAIN THE SPECIFIC GROUND TRUTH LABEL. THE COLUMNS CONTAIN THE PERCENTAGE OF DETECTIONS WITH REGARD TO THE TOTAL NUMBER OF PATTERNS OF THE GROUND TRUTH LABEL (SUM OF EVERY ROW EQUALS 100%)

	Waiting	**Starting**	**Walking**	**Stopping**
Waiting	98.6%	0.7%	0.0%	0.7%
Starting	11.8%	77.1%	8.8%	2.3%
Walking	2.0%	4.8%	88.1%	5.0%
Stopping	2.1%	2.2%	34.8%	60.9%

TABLE IV

CONFUSION MATRIX FOR THE CLASSIFICATION OF THE MOTION STATE. THE ROWS CONTAIN THE SPECIFIC GROUND TRUTH LABEL. THE COLUMNS CONTAIN THE TOTAL NUMBER OF PATTERNS

	Waiting	**Starting**	**Walking**	**Stopping**	**Recall**
Waiting	24916	177	0	177	98.6%
Starting	2960	19341	2208	577	77.1%
Walking	402	965	17704	1005	88.2%
Stopping	373	390	6176	10807	60.9%
Precision	86.7%	92.7%	67.9%	86.0%	

As the PolyMLP method performs a multi-class prediction of all four motion states, it is additionally evaluated using the 321 test scenes of the *Full Pedestrian Dataset*. The result shows an accuracy of 88.6%, which is lower than that of the *Detailed Pedestrian Dataset* as the classifier now also has to distinguish between *Starting*, *Walking*, and *Stopping*. More detailed results are set out in the confusion matrices in Table III and Table IV.

The state *Waiting* can be separated best from the three others, leading to a correct classification rate of 98.6%. Most challenging is the distinction between *Walking* and *Stopping*: here, especially the transition from *Walking* to *Stopping* is difficult to distinguish from velocity variations during walking, crossing one of the roads, or turning off.

C. Early Recognition of State Transitions

For an evaluation of the prediction model ability to recognize state transitions at an early stage, the temporal development of the classification quality depending on the binarization threshold of the pseudo class probability is analyzed. Here, the trade-off between the classifier's sensitivity leading to short reaction times in movement transitions on the one hand, and low false alarm rates on the other hand should be examined. For this purpose, the *Detailed Pedestrian Dataset* is evaluated scene-wise on false positives under variation of the chosen threshold value: a scene is already considered as false positive if the classifier produces one false positive output at any single time step. Under that condition, the quality measures of the precision and the F_1 score can be evaluated depending on the threshold. The additional determination of the accuracy does not make sense here as the number of true negatives is zero

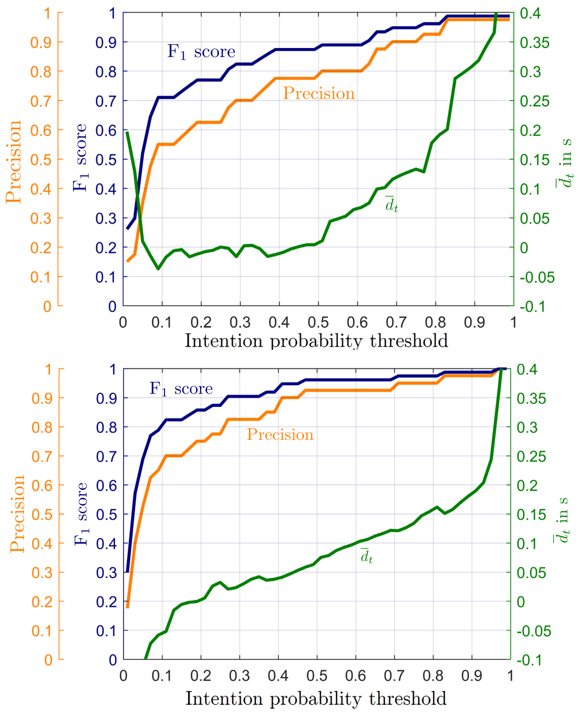

Fig. 7. Evaluation of the recognition of starting movements using 40 scenes of the *Detailed Pedestrian Dataset* for PolyMLP (upper plot) and MCHOG (lower plot). The plots show the precision, the F$_1$ score, and the mean time $\overline{d}_t$ it takes to detect the starting motion relative to the manually labeled time of the heel-off. The chosen threshold on the pseudo-probabilistic classifier output is drawn on the horizontal axis.

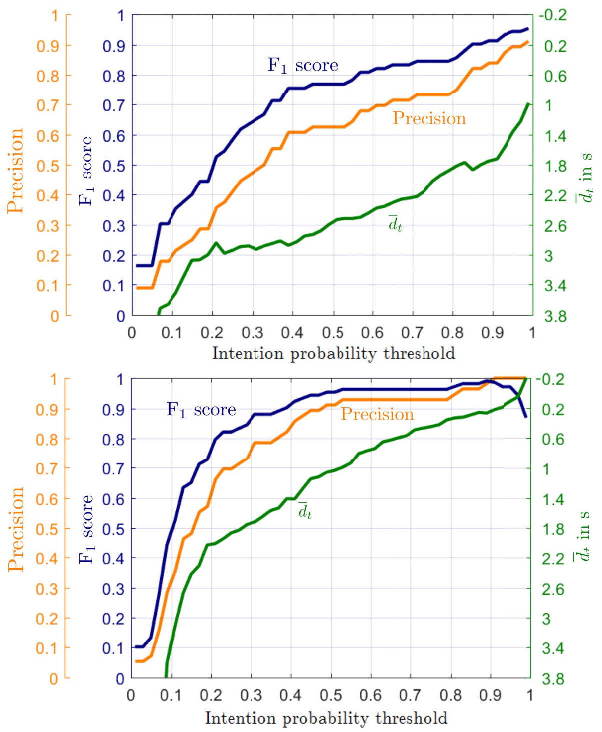

Fig. 8. Evaluation of the recognition of stopping movements using 35 test scenes of the *Detailed Pedestrian Dataset* for PolyMLP (upper plot) and MCHOG (lower plot). The axes and plot configurations correspond to Fig. 7.

for most of the possible values (state transition is detected earlier or later in any case.).

Furthermore, the mean time $\overline{d}_t$ for the correct classification of the movement transition relative to the corresponding manually labeled point in time is determined. For this detection time evaluation, only scenes without false positives for the regarded threshold value are considered. Fig. 7 shows the four quality measures evaluated in the *Starting* scenes dependent on the chosen threshold for the PolyMLP and the MCHOG/SVM classifier. The results show that an F$_1$ score of 95% and a precision of 90% are reached 130 ms after the labeled heel-off. The MCHOG classifier is about 3 frames faster at the same operating point, and thus able to detect the starting motion about 70 ms earlier. In contrast, the IMM-EKF takes 160 ms (plot not shown).

Fig. 8 shows the plots for the same kind of evaluation regarding the recognition of stopping motion. Here, the detection time is measured in relation to the heel-down of the last step. In this case, the PolyMLP method performs best, reaching an F$_1$ score of 95% and a precision of 90% already 1.4 s before the heel-down (on average). Both other classifiers, MCHOG/SVM and IMM-KF, reach this quality level only 400 ms later, which is a large time span in the field of active accident prevention.

Actually, we want to reach 100% F$_1$ score and precision, but there is a tradeoff between F$_1$ score/precision and mean detection time and sometimes we are not able to achieve 100% F$_1$ score/precision. Therefore, we choose the 95% F$_1$ score and 90% precision levels for comparison of the PolyMLP and the MCHOG/SVM classifier on *Starting* motions (Fig. 7) and *Stopping* motions (Fig. 8). When analyzing accidents it was shown that an initiation of an emergency break 160 ms earlier at a Time-to-Collision of 660 ms and a vehicle speed of 50 km/h reduces the probability of an injury resulting in a hospital stay from 50% to 35% [7]. Fig. 7 shows that for an F$_1$ score/precision of 100% some starting motions can only be detected 0.4 sec or later after heel-off. This indicates that there is still research to be done.

D. Trajectory Prediction on Full Pedestrian Dataset

In this section, the monolithic and state specific PolyMLP approaches are evaluated and compared to CV Kalman Filtering as baseline using the *Full Pedestrian Dataset*. As performance measure, we use the ASAEE (Eq. 6) with the AEE (Eq. 5), being the average Euclidean position error between all H predicted positions $(\hat{x}, \hat{y})$ and the ground truth (x_{gt}, y_{gt}), for a specific time span $t_{pred}(H) = H/f$, and the total number M of discrete time spans predicted into the future.

$$\mathrm{AEE}(H) = \frac{1}{H}\sum_{i=1}^{H}\sqrt{(\hat{x}_i - x_{gt,i})^2 + (\hat{y}_i - y_{gt,i})^2} \quad (5)$$

$$\mathrm{ASAEE} = \frac{1}{M}\sum_{H=1}^{M}\frac{\mathrm{AEE}(t_{pred}(H))}{t_{pred}(H)} \quad (6)$$

In our case, $M = 125$ steps with $t_{pred} = 0.02$ s, 0.04s, ..., 2.50s are used. The ASAEE is evaluated separately for the methods and the four scene types within the *Full Pedestrian Dataset*, leading to the results shown in Table V, rows CV-KF and PolyMLP.

TABLE V

ASAEE (IN cm/s) OF PREDICTION METHODS FOR THE FOUR SCENE CLASSES. THE TABLE SHOWS THE RESULTS FOR CV KALMAN FILTERING (CV-KF), THE SINGLE STAGE (MONOLITHIC) MLP MODEL (POLYMLP), THE TWO-STAGE MLP MODEL (POLYMLP (2 ST.)) AND THE COMBINATION OF THE GROUND TRUTH CLASS LABELS WITH THE SPECIFICALLY TRAINED POLYMLP MODELS (GT+POLYMLP)

	Waiting	Starting	Walking	Stopping	Mean
CV-KF	7.8	44.2	27.6	33.5	**28.3**
PolyMLP	6.9	33.6	25.5	22.7	**22.2**
PolyMLP (2 St.)	4.7	34.4	23.4	25.2	**21.9**
GT+PolyMLP	3.9	32.3	22.8	22.5	**20.4**

The results of the Kalman Filter already show the varying degrees of difficulty of the single scene types: As expected, *Waiting* generates the lowest errors (7.8 cm/s). The ASAEE rises with increasing presence of velocity changes resulting in the further order *Walking*, *Stopping*, and *Starting*. *Starting* includes the most abrupt velocity changes in combination with the absence of a defined direction regarding the trajectory at the beginning of the movement. For *Waiting* and *Walking*, the monolithic PolyMLP shows moderate improvements of 12% (from 7.8 cm/s to 6.9 cm/s) and 8% (from 27.6 cm/s to 25.5 cm/s) to the Kalman Filter. The improvement for scenes including velocity changes is substantially larger: *Starting* scenes improve by 24% (from 44.2 cm/s to 33.6 cm/s), *Stopping* scenes even by 32% (from 33.5 cm/s to 22.7 cm/s). Considering only predictions within the labeled *Starting* and *Stopping* phases, an improvement of 42% and 43% can be observed, respectively.

In comparison to the monolithic PolyMLP, the state specific approach shows a slight further improvement of $1-2$ cm/s on *Waiting* and *Walking* scenes while the error for *Starting* and *Stopping* increases by approximately the same values. This means that the additional classification error outweighs the improvement of the specific trajectory prediction models for the transition scenes. Overall, the state specific model leads to slightly lower errors, whereas the difference is only 1% (from 22.2 cm/s to 21.9 cm/s).

In order to examine the maximum potential of the state-specific approach, the classification stage is replaced by the ground truth class labels. Thus, the trajectory prediction stage is always able to choose the optimal model. The experimental results show a potential limit at a 7% (from 22.2 cm/s to 20.4 cm/s) lower ASAEE. Considering the separate scene types, especially the *Starting* and *Stopping* prediction of the monolithic model already performs very close to the optimum.

E. Trajectory Prediction on Detailed Pedestrian Dataset

In this test series different approaches for motion state classification, i. e., MCHOG, PolyMLP, and Ground Truth, within state-specific trajectory prediction are evaluated and compared to each other using the *Detailed Pedestrian Dataset* for *Starting* and *Stopping* scenes. We evaluate the predictions only during the labeled movement transition, as the dataset contains longer phases of *Waiting* or *Moving* before the *Starting* or *Stopping* movement, respectively. The third scene class *Moving* is included in order to evaluate the sensitivity for false alarm *Stopping* classifications.

TABLE VI

ASAEE (IN cm/s) OF THE DIFFERENT OPTIMIZED TRAJECTORY PREDICTION METHODS FOR THE THREE SCENE CLASSES IN COMPARISON

	Starting	Stopping	Moving
KF	49.94	36.04	23.35
PolyMLP	31.71	21.37	22.41
PolyMLP+PhysMod	35.23	21.75	41.97
PolyMLP+PolyMLP	31.54	22.84	23.10
MCHOG+PolyMLP	31.85	22.20	22.72
GT+PolyMLP	30.17	19.01	21.43

The results presented in Table VI show the superiority of the machine learning models compared to the prediction of the (generally optimized) CV Kalman Filter. The application of the monolithic PolyMLP model leads to a reduction of the prediction error by 36.5% (from 49.94 cm/s to 31.71 cm/s) for *Starting* and 40.7% (from 36.04 cm/s to 21.37 cm/s) for *Stopping* scenes. As shown in the last row of the table by combining the ground truth classification with the specifically trained PolyMLP prediction, the potential for improvement by state specific modeling is only 4.9% (from 31.71 cm/s to 30.17 cm/s) for *Starting*, but 11.0% (from 21.37 cm/s to 19.01 cm/s) for *Stopping* scenes. As classification stage we evaluated the four-class PolyMLP classifier already tested on the *Full Pedestrian Dataset* (PolyMLP) and the image-based two-class MCHOG/SVM classifiers (MCHOG). For trajectory prediction, specifically trained PolyMLP models and physically based models for starting and stopping (PhysMod) are evaluated. In the case of *Starting*, the specific PolyMLP models outperform the physical model by 9% (from 35.23 cm/s to 31.71 cm/s) while they perform almost equally well (from 21.75 cm/s to 21.37 cm/s) on *Stopping* motions. A distinct disadvantage of the physical *Stopping* model thereby is its sensitivity towards false alarms of the classifier stage, leading to a particularly high prediction error for *Moving* scenes. Considering the machine learning models, it is remarkable that no state specific-approach is able to outperform the single-stage PolyMLP. While with *Starting* scenes no substantial differences between the single-stage PolyMLP and the state-specific models with PolyMLP and MCHOG classification are measured, the monolithic approach slightly outperforms both at *Stopping* by 6.4% (from 22.84 cm/s to 21.37 cm/s) resp. 3.7% (from 22.20 cm/s to 21.37 cm/s).

F. Cyclists

For the *Cyclist Dataset*, the applicability of the PolyMLP concept to this VRU type could also be shown. As typical cyclist velocities at the test intersection are $2-3$ times higher than those of pedestrians, their reach within the prediction horizon is much larger. Thus, the position prediction errors rise here by 51% in *Starting* and by 79% in *Stopping* scenes compared to the *Full Pedestrian Dataset*. The fundamental insights in addition to those gained for pedestrians: Especially regarding *Starting* scenes, the PolyMLP model could outperform an optimized CV-KF by 46%. Here, also physical modeling already leads to an improvement of 27% compared to CV-KF. A significant advantage (34%) of

PolyMLP is also observed for *Stopping* scenes, while the Kalman Filter prediction did not improve for *Waiting* (±0%) and *Moving* scenes (−11%). For a more detailed evaluation of the cyclist dataset, see [26].

G. Processing Time

Another important aspect is the processing time for training and for the online application of the predictor. As the dimensionality of the in- and output features is relatively low due to the polynomial representation, the training time of the MLP also remains short. Using a current desktop PC (Intel Core i7-3770,4 × 3.4 GHz, 8 GB RAM), it takes 12 min to train the single stage PolyMLP model, while 97% of the final prediction quality is already reached after approx. 1 min.

For the state-specific approach, the preprocessing and the polynomial approximation steps have only to be performed once as long as the same polynomial configuration is used for all models included. With regard to the one-stage approach, the computation time rises by only 35% to 70 μs, compared to 52 μs. Altogether, the algorithms perform very efficiently with computational times several orders of magnitude lower than the sampling rates of commonly used sensors, which in our case is 20 ms and 80 ms. The used algorithms of video-based head detection and tracking, e. g., are processed within less than 40 ms in the test system. Therefore, a HOG/SVM pedestrian detector processing full-frames on the GPU in order to deliver ROIs for head detection is the limiting factor.

VI. Conclusion and Future Work

In this article we proposed VRU movement models based on machine learning methods. The presented approach uses the measured VRU trajectory to predict the current motion state and the future trajectory. It is compared to CV and IMM Kalman Filtering, physically-based models for starting and stopping and video-based motion classification with a MCHOG descriptor and SVMs. The results show that PolyMLP clearly outperforms Kalman Filtering for classification and trajectory prediction, in particular in the cases of starting and stopping motions. Physical models also perform better than KF, but are outperformed by PolyMLP on the other side, especially considering the handling of false positive classifications. MCHOG/SVM uses additional image-based information as input and thus shows slightly more reliable early detection of starting motion (70 ms faster on average), but is approx. 400 ms slower for the classification of *Stopping*. The two-stage approach of motion state classifications and trajectory prediction did not lead to a significant further improvement of the prediction quality in our case, but reveals the potential to include other classification of prediction models, perhaps basing on complementary sensors. Altogether, PolyMLP represents a promising method for intention recognition of VRUs with state-of-the-art prediction quality and a high degree of flexibility with regard to VRU type, used sensor system, predicted time horizon and the concrete field of ADAS application.

Our future work comprises the further evaluation of the techniques using on-board sensors in a moving vehicle under various traffic conditions. Here, it is especially the handling of measurement noise and the vehicle's ego motion that represent additional challenges. Other promising approaches are the inclusion of additional information, e. g. from high-precision maps to include road topology into the forecasting process, other on-board sensors, such as RADAR to include velocity measurements as additional input, infrastructure knots or vehicles via V2I or V2V communication to enhance the input data quality by cooperation, and the consideration of Bayesian networks or deep learning algorithms, e. g. recurrent neural networks with long short-term memory models.

References

[1] *Global Status Report on Road Safety*. Geneva, Switzerland: World Health Organization, 2015.

[2] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation With Applications to Tracking and Navigation: Theory Algorithms and Software*. Hoboken, NJ, USA: Wiley, 2001.

[3] M. Bertozzi, A. Broggi, A. Fascioli, A. Tibaldi, R. Chapuis, and F. Chausse, "Pedestrian localization and tracking system with Kalman filtering," in *Proc. IEEE Intell. Veh. Symp.*, Jun. 2004, pp. 584–589.

[4] E. Binelli *et al.*, "A modular tracking system for far infrared pedestrian recognition," in *Proc. IEEE Intell. Veh. Symp.*, Jun. 2005, pp. 759–764.

[5] M. Meuter, U. Iurgel, S.-B. Park, and A. Kummert, "The unscented Kalman filter for pedestrian tracking from a moving host," in *Proc. IEEE Intell. Veh. Symp.*, Jun. 2008, pp. 37–42.

[6] N. Schneider and D. M. Gavrila, "Pedestrian path prediction with recursive Bayesian filters: A comparative study," in *Proc. German Conf. Pattern Recognit.* Berlin, Germany: Springer, Sep. 2013, pp. 174–183.

[7] C. G. Keller, C. Hermes, and D. M. Gavrila, "Will the pedestrian cross probabilistic path prediction based on learned motion features," in *Proc. Joint Pattern Recognit. Symp.*, Springer, 2011, pp. 386–395.

[8] C. F. Wakim, S. Capperon, and J. Oksman, "A Markovian model of pedestrian behavior," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, vol. 4, Oct. 2004, pp. 4028–4033.

[9] C. G. Keller and D. M. Gavrila, "Will the pedestrian cross a study on pedestrian path prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 2, pp. 494–506, Apr. 2014.

[10] R. Quintero, J. Almeida, D. F. Llorca, and M. A. Sotelo, "Pedestrian path prediction using body language traits," in *Proc. IEEE Intell. Veh. Symp.*, Jun. 2014, pp. 317–323.

[11] R. Quintero, I. Parra, D. F. Florca, and M. A. Sotelo, "Pedestrian intention and pose prediction through dynamical models and behaviour classification," in *Proc. IEEE Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2015, pp. 83–88.

[12] A. T. Schulz and R. Stiefelhagen, "Video-based pedestrian head pose estimation for risk assessment," in *Proc. IEEE Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2012, pp. 1771–1776.

[13] A. T. Schulz and R. Stiefelhagen, "Pedestrian intention recognition using latent-dynamic conditional random fields," in *Proc. IEEE Intell. Veh. Symp. (IV)*, Jun. 2015, pp. 622–627.

[14] A. T. Schulz and R. Stiefelhagen, "A controlled interactive multiple model filter for combined pedestrian intention recognition and path prediction," in *Proc. IEEE 18th Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2015, pp. 173–178.

[15] J. F. P. Kooij, N. Schneider, F. Flohr, and D. M. Gavrila, "Context-based pedestrian path prediction," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 618–633.

[16] E. A. I. Pool, J. F. P. Kooij, and D. M. Gavrila, "Using road topology to improve cyclist path prediction," in *Proc. IEEE Intell. Vehicles Symp. IV*, Los Angeles, CA, USA, Jun. 2017, pp. 289–296.

[17] B. Kim, C. M. Kang, J. Kim, S.-H. Lee, C. C. Chung, and J. W. Choi, "Probabilistic vehicle trajectory prediction over occupancy grid map via recurrent neural network," in *Proc. IEEE Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2017, pp. 399–404.

[18] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: Human trajectory prediction in crowded spaces," in *Proc. CVPR*, Jun. 2016, pp. 961–971.

[19] B. Vólz, K. Behrendt, H. Mielenz, I. Gilitschenski, R. Siegwart, and J. Nieto, "A data-driven approach for pedestrian intention estimation," in *Proc. ITSC*, Nov. 2016, pp. 2607–2612.

[20] F. Bartoli, G. Lisanti, L. Ballan, and A. D. Bimbo, "Context-aware trajectory prediction," 2017, *arXiv:1705.02503*. [Online]. Available: http://arxiv.org/abs/1705.02503

[21] M. Goldhammer, M. Gerhard, S. Zernetsch, K. Doll, and U. Brunsmann, "Early prediction of a pedestrian's trajectory at intersections," in *Proc. IEEE 16th Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2013, pp. 237–242.

[22] M. Goldhammer *et al.*, "Analysis on termination of pedestrians' gait at urban intersections," in *Proc. 17th IEEE Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2014, pp. 1758–1763.

[23] M. Goldhammer, K. Doll, U. Brunsmann, A. Gensler, and B. Sick, "Pedestrian's trajectory forecast in public traffic with artificial neural networks," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2014, pp. 4110–4115.

[24] M. Goldhammer, S. Köhler, K. Doll, and B. Sick, "Camera based pedestrian path prediction by means of polynomial least-squares approximation and multilayer perceptron neural networks," in *Proc. SAI Intell. Syst. Conf. (IntelliSys)*, Nov. 2015, pp. 390–399.

[25] M. Goldhammer, S. Köhler, K. Doll, and B. Sick, "Track-based forecasting pedestrian behavior by polynomial approximation multilayer perceptrons," in *Proc. Intell. Syst. Appl.*, 2016, pp. 259–279. doi: 10.1007/978-3-319-33386-1_13.

[26] S. Zernetsch, S. Kohnen, M. Goldhammer, K. Doll, and B. Sick, "Trajectory prediction of cyclists using a physical model and an artificial neural network," in *Proc. IEEE Intell. Veh. Symp.*, Jun. 2016, pp. 833–838.

[27] S. Köhler *et al.*, "Stationary detection of the pedestrian's intention at intersections," *IEEE Intell. Transp. Syst. Mag.*, vol. 5, no. 4, pp. 87–99, Oct. 2013.

[28] S. Köhler, M. Goldhammer, K. Zindler, K. Doll, and K. Dietmayer, "Stereo-vision-based pedestrian's intention detection in a moving vehicle," in *Proc. IEEE 18th Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2015, pp. 2317–2322.

[29] N. Shiozawa, S. Arima, and M. Makikawa, "Virtual walkway system and prediction of gait mode transition for the control of the gait simulator," in *Proc. Int. 26th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Sep. 2004, pp. 2699–2702.

[30] E. Fuchs, T. Gruber, J. Nitschke, and B. Sick, "Online segmentation of time series based on polynomial least-squares approximations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 12, pp. 2232–2245, Dec. 2010.

[31] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: The RPROP algorithm," in *Proc. IEEE Int. Conf. Neural Netw.*, vol. 1, Mar. 1993, pp. 586–591.

[32] M. Goldhammer, E. Strigel, D. Meissner, U. Brunsmann, K. Doll, and K. Dietmayer, "Cooperative multi sensor network for traffic safety applications at intersections," in *Proc. 15th IEEE Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2012, pp. 1178–1183.

[33] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.

[34] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Dec. 2001, p. 1.

[35] S. Liao, X. Zhu, Z. Lei, L. Zhang, and S. Li, "Learning multi-scale block local binary patterns for face recognition," in *Proc. Int. Conf. Biometrics*, Aug. 2007, pp. 828–837.

[36] (Feb. 1, 2018). *VRU Trajectory Dataset*. [Online]. Available: https://www.h-ab.de/vru-trajectory-dataset

[37] M. Green, "'How long does it take to stop' methodological analysis of driver perception-brake times," *Transp. Hum. Factors*, vol. 2, no. 3, pp. 195–216, Jun. 2000.

[38] F. Naujoks, H. Grattenthaler, and A. Neukum, "Zeitliche Gestaltung effektiver Fahrerinformationen zur Kollisionsvermeidung auf der Basis kooperativer Perzeption," in *Proc. 8th Workshop FAS*, 2012, pp. 1–11.

Michael Goldhammer received the M.Eng. degree in electrical engineering and information technology from the University of Applied Sciences Aschaffenburg, Germany, in 2009, and the Dr. rer. nat. degree in computer sciences from the University of Kassel, Germany, in 2016. His Ph.D. thesis focuses on self-learning algorithms for video-based intention detection of pedestrians. His research interests include machine vision and self-learning methods for sensor data processing and automotive safety purposes.

Sebastian Köhler received the Dipl.Ing. (FH) degree in mechatronics and the M.Eng. degree in electrical engineering and information technology from the University of Applied Sciences Aschaffenburg, Germany, in 2010 and 2011, respectively. He is currently pursuing the Ph.D. degree in cooperation with the Institute of Measurement, Control, and Microtechnology, University of Ulm, Germany, focusing on intention detection of pedestrians in urban traffic. His main research interests include stereo vision, sensor and information fusion, road user detection, and short-term behavior recognition for ADAS.

Stefan Zernetsch received the B.Eng. and M.Eng. degrees in electrical engineering and information technology from the University of Applied Sciences Aschaffenburg, Germany, in 2012 and 2014, respectively. He is currently pursuing the Ph.D. degree in cooperation with the Faculty of Electrical Engineering and Computer Science, University of Kassel, Germany. His research interests include cooperative sensor networks, data fusion, multiple view geometry, pattern recognition, and behavior recognition of traffic participants.

Konrad Doll (M'00) received the Diploma (Dipl.Ing.) degree and the Dr. Ing. degree (equivalent to a Ph.D. degree) in electrical engineering and information technology from the Technical University of Munich, Germany, in 1989 and 1994, respectively. In 1994, he joined the Semiconductor Products Sector of Motorola, Inc., (currently Freescale Semiconductor, Inc.) In 1997, he was a Professor at the University of Applied Sciences Aschaffenburg in the field of computer science and digital systems design. His research interests include intelligent systems, their real-time implementations on platforms like CPU, GPUs, and FPGAs, and their applications in advanced driver assistance systems.

Bernhard Sick received the Diploma, Ph.D., and Habilitation degrees from the University of Passau, Germany, in 1992, 1999, and 2004, respectively, all in computer science.

He is currently a Full Professor for intelligent embedded systems with the Faculty for Electrical Engineering and Computer Science, University of Kassel, Germany, where he conducts research in the areas autonomic and organic computing and technical data analytics with applications in biometrics, intrusion detection, energy management, and automotive engineering. He has authored more than 90 peer-reviewed publications in these areas. He holds one patent and received several thesis, best paper, teaching, and inventor awards. He is a member of the IEEE Systems, Man, and Cybernetics Society, Computer Society, the IEEE Computational Intelligence Society, and GI Gesellschaft für Informatik. He is an Associate Editor of the IEEE TRANSACTIONS ON CYBERNETICS.

Klaus Dietmayer was born in Celle, Germany, in 1962. He received the Diploma degree (equivalent to a M.Sc. degree) in electrical engineering from the Braunschweig University of Technology, Braunschweig, Germany, in 1989, and the Dr. Ing. degree (equivalent to a Ph.D. degree) from Helmut Schmidt University, Hamburg, Germany, in 1994.

In 1994, he joined the Philips Semiconductors Systems Laboratory, Hamburg, as a Research Engineer. Since 1996, he has been a Manager in the field of networks and sensors for automotive applications. In 2000, he was appointed to a professorship at Ulm University, Ulm, Germany, in the field of measurement and control. He is currently a Full Professor and the Director of the Institute of Measurement, Control and Microtechnology, School of Engineering and Computer Science, Ulm University. His research interests include information fusion, multiobject tracking, environment perception for advanced automotive driver assistance, and e-mobility. He is a member of the German Society of Engineers VDI/VDE.