Pedestrian Intention Recognition using Latent-dynamic Conditional Random Fields

Andreas Th. Schulz[1] and Rainer Stiefelhagen[2]

Abstract— **We present a novel approach for pedestrian intention recognition for advanced video-based driver assistance systems using a Latent-dynamic Conditional Random Field model. The model integrates pedestrian dynamics and situational awareness using observations from a stereo-video system for pedestrian detection and human head pose estimation. The model is able to capture both intrinsic and extrinsic class dynamics. Evaluation of our method is performed on a public available dataset addressing scenarios of lateral approaching pedestrians that might cross the road, turn into the road or stop at the curbside. During experiments, we demonstrate that the proposed approach leads to better stability and class separation compared to state-of-the-art pedestrian intention recognition approaches.**

I. INTRODUCTION

The field of Advanced Driver Assistance Systems (*ADAS*) gets strong interests in present days. There are several reasons, namely increasing computational power of embedded platforms or emerging technologies in the sector of intelligent sensors like video, radar, laser or ultrasonic. During the past years video-based driver assistance systems build up a strong growing market because of their wide functional applications on low costs. Especially, the static improving performance for video-based pedestrian detection resulted in first commercial active pedestrian protection systems available for a wide range of vehicles, e.g. Mercedes Benz, Volvo and VW. Those kind of systems try to avoid collisions in dangerous situations involving an inattentive driver and pedestrian by triggering an autonomous braking. One of the most challenging task is to interpret situations with lateral approaching pedestrians correctly. Due to the high variability of movement patterns, pedestrians can change their walking direction within a short time period or suddenly start or stop. Therefore, existing systems are designed in a conservative way by decreasing benefit in order to reduce potential false activations that could result in consequential damages involving other traffic participants. To cope with this situation, a reliable pedestrian intention recognition and path prediction states a great value. This work focuses on a reliable intention recognition for pedestrians walking along or towards the road curbside on their way to cross, stop or just moving on in the same direction. Pedestrian candidates are detected by a stereo-video system mounted behind the windshield of an approaching vehicle. Motivated by scientific evidence about pedestrian behavior in daily traffic situations [9], [16], [18], intention decisions will be made by two dominant factors. Firstly, the pedestrian dynamics by means of predicted lateral and longitudinal velocity components as a result of a linear dynamical system. Secondly, the pedestrian's awareness of an oncoming vehicle with the help of observing the human head pose. See Fig. 1 for a rough overview about our system. We propose a latent-dynamic discriminative model for time-series, that is able to integrate all these features, to learn inner connections within a specific type of scenario and external correlations between different types of scenarios. Our approach can be easily integrated into an overall system as an indicator for pedestrian path prediction helping to estimate potential future pedestrian states.

II. RELATED WORK

In this section we list the main recent contributions on pedestrian intention recognition and the closely related field of pedestrian path prediction. Most of them build upon an existing system for video-based pedestrian detection (see [4] for a survey). The main focus here is to address the situation of lateral approaching pedestrians. As accident statistics show ([14]) this covers the main scenario of accidents involving vehicles and pedestrians. [11] for example, proposed two non-linear, higher order Markov models to estimate whether an approaching pedestrian will cross the street or stop at the curbside. First a Probabilistic Hierarchical Trajectory Matching (PHTM) is used to match an actual observation of a pedestrian track with a database of trajectory snippets. Using the information of future locations and pedestrian behavior from the best matching snippets future pedestrian motion is extrapolated. In addition, a Gaussian Process Dynamical Model (GPDM) which models the dense flow for walking and stopping motions is suggested to predict future flow fields. Both suggested models integrate features that capture pedestrian positions and dynamics by means of dense optical flow. For path prediction only [19] analyzed the usability of different linear dynamical systems involving Kalman filters and interacting multiple models to predict future pedestrian positions by propagating pedestrian states for a small time slot of 1 second. All these models only try to access features from pedestrian moving dynamics but do not take the underlying context into account. Initially, [12] presented a Dynamic Bayesian Network (*DBN*) on top of Switching Linear Dynamic System (*SLDS*), where they integrate contextual information using latent information from pedestrian awareness, the pedestrian position w.r.t the curbside and the criticality of the underlying situation. A

[1] Andreas Th. Schulz is a PhD student with the Robert Bosch GmbH, 71229 Leonberg, Germany Andreas.Schulz3@de.bosch.com

[2] Rainer Stiefelhagen is with the Institute of Anthropomatics, Karlsruhe Institute of Technology, Karlsruhe, Germany Rainer.Stiefelhagen@kit.edu

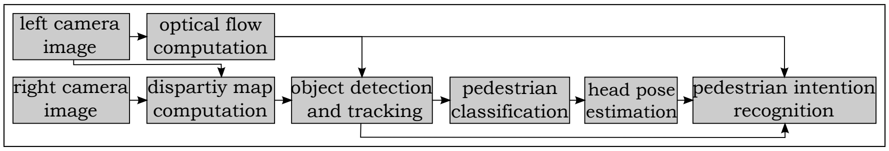
left camera image
optical flow computation
right camera image
dispartiy map computation
object detection and tracking
pedestrian classification
head pose estimation
pedestrian intention recognition

Fig. 1. System Overview. Pedestrians are detected and tracked by a stereo generic obstacle detection approach verified by gray-value based classifier. For each pedestrian the orientation will be estimated. All previous information is input for the intention recognition module.

SLDS uses a top-level discrete Markov chain to select per time step the system dynamics of the underlying LDS. Whereas SLDSs can account for changes in dynamics, a switch in dynamics will only be detected after sufficient observations contradict the currently predominant dynamic model. To forecast pedestrian behavior the model should include possible causes for change. Therefore, they use the expected point of closest approach, presented in [1]. To account for inattentive pedestrians they extract features from the human head pose inspired by the work of [9]. Pedestrian head pose estimation was handled for example by the works of [3], [5] and [21]. At the end an existing system for curb stone detection is used to determine, whether a pedestrian is at the curbside or too far away to state a risk.

In this work we investigate the use of Latent-dynamic Conditional Random Fields (*LDCRF*) for the task of intention recognition in different scenarios. [15] first introduced LDCRF models as an extension of conventional Conditional Random Fields (*CRF*) [13] by adding a layer of hidden latent states. These hidden state variables can model the intrinsic sub-structure of a specific class label and capture extrinsic dynamics between different classes. Furthermore, LDCRFs proved to outperform typical CRF models, the well-known Hidden Markov Models (*HMM*) and conventional machine learning algorithms like Support Vector Machines (*SVMs*) in the field of gesture recognition. See Fig. 2 for a simplified version of a LDCRF. Therefore a benefit of our model is the fact that it can work with time series of arbitrary lengths, where additional confidence can be retrieved over temporal integration. By the nature of our LDCRF model it is also possible to capture dynamical changes for single features, like head turnings or different movement behaviors. Different connections of features can be learned automatically from training data but nevertheless, expert knowledge still can be brought in when designing the structure of the LDCRF. We address a wide range of scenarios including also pedestrians that initially are walking along the sidewalks but then bend in towards the road. The output of our approach can also be easily used for controlling the switching states of the SLDS presented in [12].

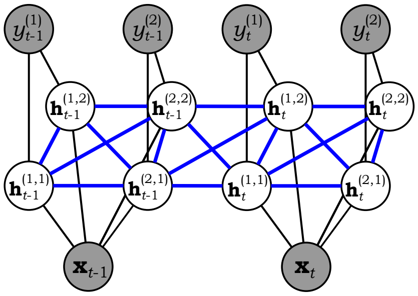

Fig. 2. LDCRF for 2 classes with 2 hidden states per class label. Observables are displayed as shaded nodes. Connections between latent hidden states are blue colored (some connections not visualized for better readability). The hidden states $\mathbf{h}_j^{(i,1)}$ and $\mathbf{h}_j^{(i,2)}$ model the intrinsic structure for class label $y_j^{(i)}$, $i \in \{1,2\}$, while the connections between $\mathbf{h}_j^{(1,k)}$ and $\mathbf{h}_j^{(2,k)}$, $k \in \{1,2\}$, model the extrinsic relations between the class labels $y_j^{(1)}$ and $y_j^{(2)}$.

III. LATENT-DYNAMIC CONDITIONAL RANDOM FIELDS FOR PEDESTRIAN INTENTION RECOGNITION

The task of action-/intention recognition is to find a sequence of labels $\mathbf{y} = \{y_1, y_2, \ldots, y_T\}$ that best explains the sequence of observations $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ for a total of T time steps. LDCRF takes root in Conditional Random Fields [13], which is one of famous activity recognition models that can capture extrinsic dynamics between the class labels. As Fig. 2 shows, a LDCRF as an undirected graph consisting of sequential variable pairs of state variables $\mathbf{x}_t$ and class labels y_t plus an additional layer of hidden, non-observable variables $H = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_T\}$, on every time step. Using a larger set of observations as training data, the LDCRF is able to learn an intention recognition model and given a new observation sequence $X = \{\mathbf{x}_t\}_{t=1,\ldots,T}$ to infer the intention labels $\mathbf{y} = \{y_t\}_{t=1,\ldots,T}$. Each $\mathbf{h}_t$ is member of a set $\mathcal{H}_{y_t}$ of possible hidden states for the intention label $y_t \in \{1, 2, \ldots, L\}$, where L is the number of intention labels. For different labels, we restrict the model to have disjoint sets of hidden states $\mathcal{H}_{y_t}$. We define $\mathcal{H}$ to be the union of all $\mathcal{H}_{y_t}$ sets, i.e. $\mathcal{H} = \bigcup_{y_t} \mathcal{H}_{y_t}$. So all possible hidden states are contained in $\mathcal{H}$. The LDCRF defines a latent conditional model as

$$P(\mathbf{y}|X; \boldsymbol{\theta}) = \sum_{H \in \mathcal{H}} P(\mathbf{y}, H|X; \boldsymbol{\theta}), \tag{1}$$

resulting in

$$P(\mathbf{y}|X; \boldsymbol{\theta}) = \sum_{H \in \mathcal{H}} P(\mathbf{y}|H, X; \boldsymbol{\theta}) P(H|X; \boldsymbol{\theta}), \tag{2}$$

where $\boldsymbol{\theta}$ contains the parameters of the model. By definition, for sequences having any $\mathbf{h}_t \notin \mathcal{H}_{y_t}$ it holds $P(\mathbf{y}|H, X; \boldsymbol{\theta}) = 0$ and the model Eq. 2 can be simplified to

$$P(\mathbf{y}|X; \boldsymbol{\theta}) = \sum_{H : \forall \mathbf{h}_j \in \mathcal{H}_{y_j}} P(H|X; \boldsymbol{\theta}), \tag{3}$$

Identical to the usual CRF formulation, $P(H|X; \boldsymbol{\theta})$ can be defined as

$$P(H|X; \boldsymbol{\theta}) = \frac{1}{\mathcal{Z}(X, \boldsymbol{\theta})} \exp\left(\sum_k F_k(H, X)\right), \tag{4}$$

with the partition function

$$\mathcal{Z}(X,\boldsymbol{\theta}) = \sum_{H} \exp\left(\sum_{k} F_k(H,X)\right). \tag{5}$$

The feature functions $F_k(H,X)$ can be written as a linear combination of state functions $s_l(\mathbf{h}_t, X, t)$ and transition functions $t_m(\mathbf{h}_t, \mathbf{h}_{t-1}, X, t)$

$$F_k(H,X) = \sum_{t=1}^{T} \left\{ \sum_{l} \lambda_l s_l(\mathbf{h}_t, X, t) + \sum_{m} \mu_m t_m(\mathbf{h}_t, \mathbf{h}_{t-1}, X, t) \right\}, \tag{6}$$

with

$$\boldsymbol{\theta} = \{\theta_k\}_k = \{\lambda_1, \ldots, \lambda_l\} \cup \{\mu_1, \ldots, \mu_m\}. \tag{7}$$

State functions s_l depend on a single hidden variable and observations in the model while transition functions t_m depend on pairs of hidden variables. The number of state functions, s_l, will be equal to the dimension of the feature vector d times the number of possible hidden states. With the L intention labels for our model and assuming M hidden states per label, the total number of state functions, s_l, and total number of associated weights λ_l will be $d \times L \times M$. For each hidden state pair $(\mathbf{h}', \mathbf{h}'')$, the transition function t_m is defined as

$$t_m(\mathbf{h}_{t-1}, \mathbf{h}_t, \mathbf{x}, t) = \begin{cases} 1, & \text{if } \mathbf{h}_{t-1}\text{=}\mathbf{h}' \text{ and } \mathbf{h}_t\text{=}\mathbf{h}'' \\ 0, & \text{otherwise} \end{cases}. \tag{8}$$

The weights μ_m associated with the transition functions model both the intrinsic and extrinsic dynamics. Weights associated with a transition function for hidden states that are in the same subset $\mathcal{H}_{y_t}$ will model the substructure patterns, while weights associated with the transition functions for hidden states from different subsets will model the external dynamic between intention labels.

A. Learning the model parameters

Our training set consists of n labeled sequences $(X_i, \mathbf{y}_i)$, $i = 1 \ldots n$. Following [13] and [15], we use the objective function

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log P(\mathbf{y}_i | X_i, \boldsymbol{\theta}) - \frac{1}{2\sigma^2} \|\boldsymbol{\theta}\|^2 \tag{9}$$

to learn the optimal parameter set $\boldsymbol{\theta}^*$. Eq. 9 combines the conditional log-likelihood of the training data with the log of a Gaussian prior with variance σ^2, i.e., $P(\boldsymbol{\theta}) \sim \exp(\frac{1}{2\sigma^2} \|\boldsymbol{\theta}\|^2)$. The optimal parameter values under the criterion $\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$ can be found by gradient ascent using the belief propagation technique. To save space we refer to [15] for further details.

B. Inference

To test a previously unseen sequence X, the most probable label sequence $\mathbf{y}^*$ will be estimated that maximizes the trained model using the optimal parameter values $\boldsymbol{\theta}^*$:

$$\mathbf{y}^* = \arg\max_{\mathbf{y}} P(\mathbf{y}|X, \boldsymbol{\theta}^*), \tag{10}$$

Applying Eq. 3 once again, we get

$$\mathbf{y}^* = \arg\max_{\mathbf{y}} \sum_{H:\forall \mathbf{h}_t \in \mathcal{H}_{y_t}} P(H|X, \boldsymbol{\theta}^*). \tag{11}$$

To predict the label y_t^* of frame t, the marginal probabilities $P(\mathbf{h}_t = \mathbf{h}|X, \boldsymbol{\theta}^*)$ are calculated for all possible hidden states $\mathbf{h} \in \mathcal{H}$. Single marginal probabilities are summed according to the disjoint sets of hidden states $\mathcal{H}_{y_t}$. Finally, the label associated with the optimal set is chosen.

C. Feature Computation

Here, we present the set of features we are using for pedestrian intention recognition. [9], [16] and [18] describe dominant features to recognize persons' intention during road crossing events. We try to derive most of these features using the measurements from an on-board stereo-video based object detection and head pose estimation system. Features are extracted for each frame an concatenated into a time-series of T frames as an input for our LDCRF model.

1) Pedestrian dynamics (Pos + Vel): Detected pedestrians will be tracked in lateral and longitudinal direction using a Kalman filter [2]. We assume a simple CV model for pedestrians. See [19] for details. As measurements we take the pedestrian image bounding box plus an additional median disparity value calculated over the upper pedestrian body. Vehicle dynamics are incorporated into the dynamical model for ego-motion compensation similar to [10] and [19]. As a result, we get the pedestrians filtered relative world positions wrt. the ego vehicle (x, z) and the absolute velocities (v_x, v_z) in lateral and longitudinal direction, for each frame.

2) Pedestrian head pose (Hp): To capture the pedestrians' awareness of an oncoming vehicle the human head pose gives a dominant cue [9], [16]. We build upon a system presented in [20], that tries to estimate pedestrian head poses in monocular gray value images. The basic idea is to train multiple classifiers for different head pose classes related to a specified pan angle range. Furthermore, the single head pose estimation results are filtered for usage in video sequences by implementing a particle filter [21]. To more robustly guide the particles around future head regions, depth information within a detected pedestrian bounding box as well as the estimated human movement direction is incorporated (see [5]). Hence, for a given pedestrian track the continuous head pan angle $\omega \in [-180, 180)$ will be extracted for each frame, where $\omega = 0°$ relates to a frontal face.

3) Formation of a time-series: For a video sequence the above mentioned features will be integrated into a time-series of observations. A time-series X over T frames can be then defined as

$$X = \{(x^{(1)}, z^{(1)}, v_x^{(1)}, v_z^{(1)}, \omega^{(1)}), \dots, \\ (x^{(T)}, z^{(T)}, v_x^{(T)}, v_z^{(T)}, \omega^{(T)})\}. \quad (12)$$

IV. Experimental Results

A. Evaluation Dataset

[19] presented a new dataset containing labeled stereo-video images. The images recorded at 16 fps show different situations of persons approaching the curb. Ground truth is provided by means of labeled pedestrian bounding boxes, distance measurements estimated from calculated disparity maps and ego-motion data from on-board inertial sensors. For pedestrian path prediction the so called "time-to-event" (*TTE*, in frames) is labeled, identifying, when a person is crossing, starting to cross, bending in or stopping at the curb. In total 36 training and 32 test scenarios were recorded. Figure 3 shows a one of the evaluated images. We also address the scenario of a walking pedestrian on the sidewalk, which we call "straight". "straight"-samples are extracted out of "bending-in"-scenarios with an adequate time distance to the turning event. Without loss of generality, we restrict our system to lateral approaching pedestrians from to right side. There is one sample in the training set and testing set respectively, where a pedestrian is crossing from the left side. We overcome this problem and convert both samples into right-to-left-crossings by inverting the extracted features for lateral position, lateral velocity and head pose. Additionally, the contained samples for "starting"-scenarios are ignored for further evaluation. If we take all data together into an overall set we get the distribution over addressed scenarios displayed in TABLE I.

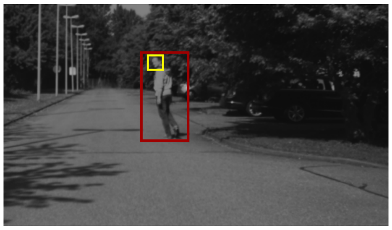

Fig. 3. Bending In scenario, Daimler dataset [19]

TABLE I
DISTRIBUTION OF SEQUENCES PER SCENARIO

scenario	bending in	stopping	crossing	straight
number of sequences	23	17	18	20

B. Setup

We reprocessed the sequences and calculated the features mentioned in Sec. III-C. Similar to [11] we add a uniform noise of up to 10% of the original height of the labeled bounding boxes to their height and center to simulate real-world performance. To capture the pedestrians awareness of the underlying scene, continuous head pose angels are calculated using the algorithms of [20] and [21]. For each of 8 discrete head pose classes we trained boosting cascades including MCT-Features [6] on a large set of manually labeled pedestrian head images (≈2300 per class). For a rough comparison, pedestrian dynamics are captured using a similar approach to PHTM [11], where motion histogram features are extracted from dense optical flow. We use a public available version of the TV-L1 flow [22],[17]. Vehicle ego-motion compensation is achieved by applying an efficient and highly accurate algorithm for visual odometry [7]. In [11] only flow vectors inside a pedestrian depth mask contribute to the histogram calculation. Therefore, we compute disparity maps over whole frames using the method of [8]. Facing the problem of a very low number of samples we perform leave–one–out cross–validation (*LOO*). We train One-Vs-One-Classifiers to differentiate between "stopping" and "crossing" (SC) or "bending-in" and "straight" (BS). The idea for LDCRFs is now to replace one abstract label by a specified number of latent variables in order to model the extrinsic and intrinsic class dependencies. The number of feature functions/parameter can be controlled by setting a time window for temporal feature dependencies to be learned. During experiments, we evaluated different window sizes $(0, \dots, 5)$ and different number of hidden states per class label $(1, \dots, 4)$ for training of our LDCRF models. Only the best performing models are visualized. The training algorithm is adapted to not take data of future frames into account for actual observations. This will prevent latency of our system. Observations taken from a stopping scenario related to TTE values larger than 5 are members of the crossing class. Indeed, modifying this threshold has strong impact on the course of the performance plots displayed later. For evaluation, a sliding window is shifted over a whole pedestrian trajectory to collect frame-based system responses. We want to analyze the system's capability to early detect critical situations with a high reliability. Therefore, we show the system's output, namely the event probability for a pedestrian to stop at the curbside or turning in. We are interested in the behavior of the system's output within a short time range around the actual event for both types of scenarios. As proposed in literature [11], [12], we focus on TTE interval of $[-5, 20]$. Plotted are mean and standard deviation of event probabilities over all tested sequences. Positive TTE-values relate to frames prior to the event, while for negative TTE-values the event already occurred. A system for pedestrian action/intention-recognition to be used for collision warning or even avoidance, the desired event-probabilities should have the following course. In an early stage (20 to 15 frames prior to the event) the event probabilities should still have considerable low values. Towards the occurring event, the probabilities should increase rapidly showing a strong gradient. This behavior will lead to systems less sensitive to false activations. We consider this fact during interpretation of our results.

C. System Performance

Evaluation is done on single features only and on their combined versions. Fig. 4 and Fig. 5 show the evaluation results for our SC-models and BS-models respectively. Compared to the single-feature-based models, the combined SC-

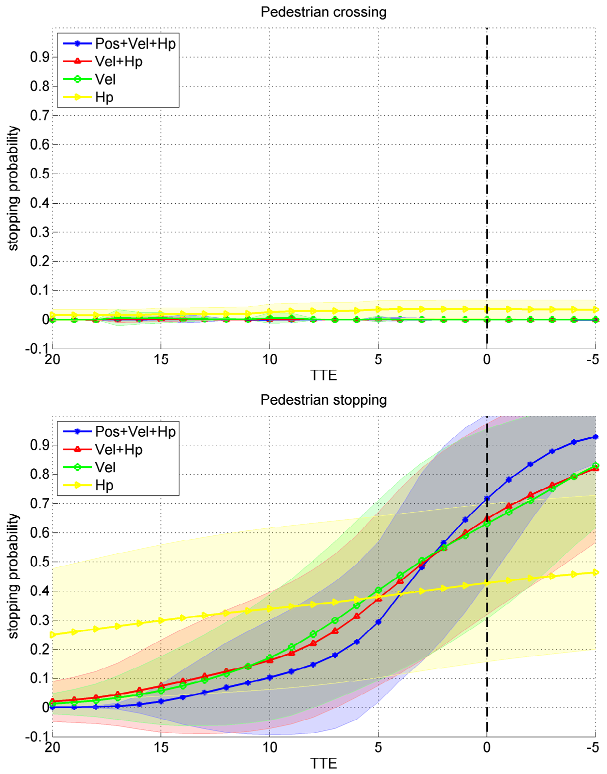

Fig. 4. Stopping probabilities for our LDCRF model. Road-crossing scenarios (upper) and stopping scenarios (lower). Visualized are mean and standard deviation (shaded area) over all tested sequences.

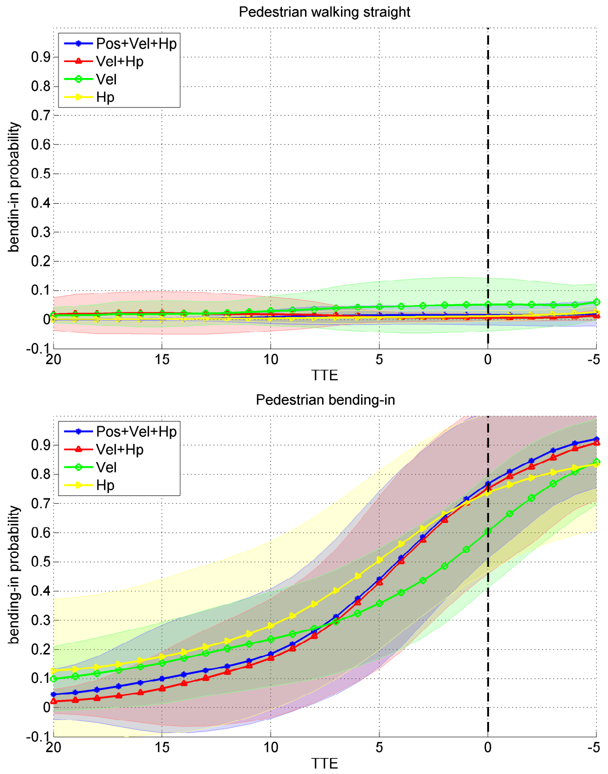

Fig. 5. Bending-in probabilities for our LDCRF model. Straight-walking scenarios (upper) and bending-in scenarios (lower).

model (Pos+Vel+Hp) is able to reliably recognize crossing situations, i.e. the stopping probability is approximately zero within the considered time period for most of test sequences. For stopping scenarios the probability increases continuously towards the stopping event (TTE=0). While at an early stage (TTE=20) for most of the stopping sequences the system still tends to predict a crossing scenario (low average stopping probability smaller than 0.025 with a standard deviation near to 0), this behavior rapidly changes for getting closer the actual stopping event (TTE$\in [0, 5]$). The combined BS-model shows similar behavior. The velocity components do not seem to have the impact on an accurate intention recognition for BS-scenarios compared to CS-scenarios. We explain this by the fact, that absolute velocity values do not change that significant for a bending in rather than for a stopping scenario where the values tend to zero. Another interesting observation is related to the power of the head pose feature. While for the CS-model, the head pose cannot contribute significantly to a performance gain, for the BS-model the opposite is the case. Here, the model only trained on the head pose feature outperforms the one trained on velocity inputs. We explain this by the presence of a dominant pedestrian head turning towards the oncoming vehicle in most of the sequences. This also reflects pedestrian behavior in real-world scenarios. The motion histogram features introduced by [11] were used in two ways. Firstly, by applying a PHTM-like version and secondly by integrating the features into a LDCRF model, see Fig. 6. Compared to out LDCRF model, the PHTM approach (green) results in more unstable intention estimates over the whole dataset especially for crossing situations. Nevertheless, the motion histogram features show high potential for a robust intention recognition in combination with a LDCRF model (red). We also trained standard machine learning approaches – here, SVMs and Random Forests for SC – to test their suitability compared to LDCRF. Therefore single observations were integrated over time window of 20 frames. Results are given in Fig. 7. Compared to LDCRF, SVM (green) and RF models (red) result in unstable estimates for the underlying scenarios. At an earlier stage there is still a comparably high confusion between ”stopping”- and ”crossing”-scenarios for both SVM and RF models, whereas the LDCRF results in more stable estimates.

V. CONCLUSION

We presented a method to estimate the intention of lateral approaching pedestrians in the domain of intelligent vehicles. Multiple features capturing the pedestrian dynamics and the awareness of the nearby traffic situation were used to learn a LDCRF model. The proposed model has the advantage to automatically learn intrinsic structure and feature dependencies as well as temporal dynamics between different actions. Evaluation of the model showed stable intention estimates for different scenarios compared to other machine learning approaches. The model provides evidence for potential risky

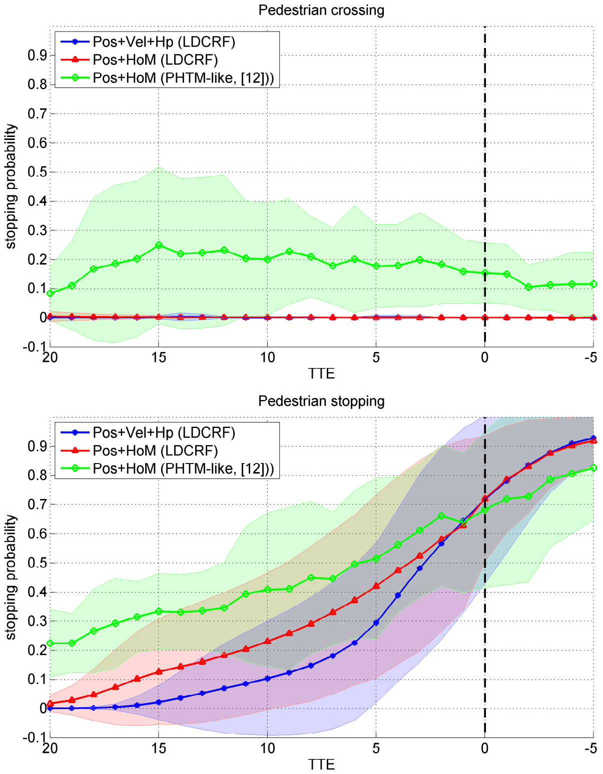

Fig. 6. Stopping probabilities for our best LDCRF (blue), a LDCRF learned on ped. position and motion histograms (red) and a PHTM-like approach [11] (green).

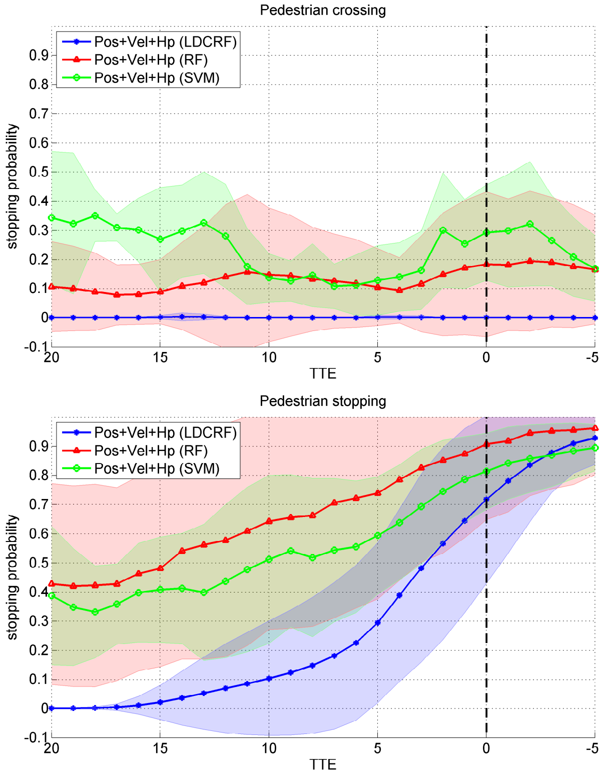

Fig. 7. Stopping probabilities for our best LDCRF model compared to other machine learning approaches. Crossing scenarios (upper) and stopping scenarios (lower).

situations and therefore can serve for better pedestrian path prediction or be directly integrated into a system implementing a pedestrian warning or emergency braking function for reduction of false alarms.

REFERENCES

[1] G. Antonini, S.V. Martinez, M. Bierlaire and J.P. Thiran, "Behavioral priors for detection and tracking of pedestrians in video sequences." In *IJCV*, 69(2), 2006

[2] Y. Bar-Shalom, T. Kirubarajan and X.-R. Li. "Estimation with Applications to Tracking and Navigation." John Wiley & Sons, Inc., New York, USA, 2002.

[3] B. Benfold and I. Reid, "Guiding visual surveillance by tracking human attention." In *British Machine Vision Conference (BMVC)*, Vol. 20, 2009.

[4] P. Dollár, C. Wojek, B. Schiele, P. Perona, "Pedestrian detection: An evaluation of the state of the art." In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 34(4), 2012

[5] F. Flohr, M. Dumitru-Guzu, J. F. P. Kooij and D. M. Gavrila, "Joint probabilistic pedestrian head and body orientation estimation" In *Intelligent Vehicles Symposium (IV)*, 2014.

[6] B. Fröba and A. Ernst. "Face detection with the modified census transform". In *Face and Gesture Recognition (FG)*, Vol. 6, 2004.

[7] A. Geiger, J. Ziegler and C. Stiller, "StereoScan: Dense 3d Reconstruction in Real-time". In *Intelligent Vehicles Symposium (IV)*, 2011.

[8] A. Geiger M. Roser and Raquel Urtasun, "Efficient Large-Scale Stereo Matching", In *Asian Conference on Computer Vision (ACCV)*, 2010.

[9] H. Hamaoka, T. Hagiwara, M. Tada, and K. Munehiro. "A study on the behavior of pedestrians when confirming approach of right/left-turning vehicle while crossing a crosswalk." In *Intelligent Vehicles Symposium (IV)*, 2013.

[10] C. Keller, M. Enzweiler, and D. M. Gavrila, "A New Benchmark for Stereo-based Pedestrian Detection." In *Intelligent Vehicles Symposium (IV)*, 2011.

[11] C. G. Keller and D. M. Gavrila, "Will the pedestrian cross? A study on pedestrian path prediction." In *Intelligent Transportation Systems (ITS)*, 15(2), 2014.

[12] J. F. P. Kooij, N. Schneider, F. Flohr and D. M. Gavrila, "Context-Based Pedestrian Path Prediction" In *European Conference on Computer Vision (ECCV)*, 2014.

[13] J. Lafferty, A. Mccallum, and F. F. C. N. Pereira, "Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data." In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, 2001.

[14] P. Marchal, D. M. Gavrila, L. Letellier, M.-M. Meinecke, R. Morris and M. Töns, "SAVE-U: An innovative sensor platform for Vulnerable Road User protection". In *Intelligent Transportation Systems (ITS)*, 2003.

[15] L. Morency, A. Quattoni and T. Darrell, "Latent-dynamic discriminative models for continuous gesture recognition." In *Computer Vision and Pattern Recognition (CVPR)*, 2007.

[16] J. Oxley B. Fildes, E. Ihsen, R. Day and J. Charlton, "An Investigation of Road Crossing Behaviour of older pedestrians." In *Monach University-Accident Research Centre*, Report No.81, 1995.

[17] J. Sanchez, E. Meinhardt-Llopis and G. Facciolo. "TV-L1 Optical Flow Estimation". In *Image Processing Online*, http://www.ipol.im, 2012.

[18] S. Schmidt and B. Färber, "Pedestrians at the kerbRecognizing the action intentions of humans." In *Transp. Res. F, Traffic Psychol. Behav.*, vol. 12, no. 4, 2009.

[19] N. Schneider and D. M. Gavrila, "Pedestrian Path Prediction with Recursive Bayesian Filters: A Comparative Study." In *German Conference on Pattern Recognition (GCPR)*, 2013.

[20] A. Schulz, N. Damer, M. Fischer, and R. Stiefelhagen, "Combined head localization and head pose estimation for videobased advanced driver assistance systems." In *Proceedings of Pattern Recognition (DAGM)*, 2011.

[21] A. Schulz and R. Stiefelhagen, "Video-based Pedestrian Head Pose Estimation for Risk Assessment". In *Intelligent Transportation Systems (ITS)*, 2012.

[22] C. Zach, T. Pock and H. Bischof, "A Duality Based Approach for Realtime TV-L1 Optical Flow", In *Proceedings of Pattern Recognition (DAGM)*, 2007.