Will the Pedestrian Cross? A Study on Pedestrian Path Prediction

Christoph G. Keller and Dariu M. Gavrila

***Abstract*—Future vehicle systems for active pedestrian safety will not only require a high recognition performance but also an accurate analysis of the developing traffic situation. In this paper, we present a study on pedestrian path prediction and action classification at short subsecond time intervals. We consider four representative approaches: two novel approaches (based on Gaussian process dynamical models and probabilistic hierarchical trajectory matching) that use augmented features derived from dense optical flow and two approaches as baseline that use positional information only (a Kalman filter and its extension to interacting multiple models). In experiments using stereo vision data obtained from a vehicle, we investigate the accuracy of path prediction and action classification at various time horizons, the effect of various errors (image localization, vehicle egomotion estimation), and the benefit of the proposed approaches. The scenario of interest is that of a crossing pedestrian, who might stop or continue walking at the road curbside. Results indicate similar performance of the four approaches on walking motion, with near-linear dynamics. During stopping, however, the two newly proposed approaches, with nonlinear and/or higher order models and augmented motion features, achieve a more accurate position prediction of 10–50 cm at a time horizon of 0–0.77 s around the stopping event.**

***Index Terms*—Computer vision, pedestrian safety, prediction methods.**

I. Introduction

PREDICTING the path of a pedestrian is important in several application contexts, such as robot control in human-inhabited environments and driver assistance systems for improved traffic safety. In this paper, we consider the intelligent vehicles context, in which strong gains have been made over the years in improving computer-vision-based pedestrian recognition performance. This has culminated in the first active pedestrian safety systems reaching the market. For example, our company Daimler has recently introduced an innovative stereo-vision-based pedestrian system in its 2013 Mercedes-Benz E- and S-Class models, which incorporates automatic full emergency braking.

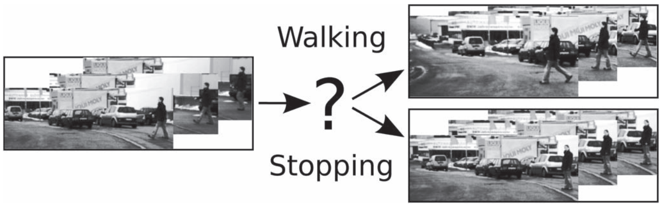


Fig. 1. Pedestrian path prediction and action classification. Where exactly will the pedestrian be in the immediate future? Will the pedestrian cross?

A sophisticated situation assessment requires a precise estimation of the current and future positions of the pedestrian with respect to the moving vehicle. A deviation of, for example, 30 cm in the estimated lateral position of the pedestrian can make all the difference between a "correct" and an "incorrect" maneuver initiation. One major challenge is the highly dynamic behavior of pedestrians, which can change their walking direction in an instance, or start/stop walking abruptly. As a consequence, prediction horizons for active pedestrian systems are typically short; even so, small performance improvements can produce tangible benefits. For example, accident analysis [1] shows that being able to initiate emergency braking 0.16 s (four frames at 25 Hz) earlier, at a time to collision of 0.66 s, reduces the chance of incurring injury requiring a hospital stay from 50% to 35%, given an initial vehicle speed of 50 km/h.

This paper focuses on the task of predicting the position of pedestrians walking toward the road curbside, when viewed from an approaching vehicle. A secondary question is whether the pedestrian will cross or stop. The setting in Fig. 1 is inspired by an earlier human factors study by Schmidt and Färber [2], which had several test participants watch videos of pedestrians walking toward the curbside and decide whether the pedestrians would stop or cross, at various time instants. The study varied the amount of visual information provided to the test participants and examined its effect on their classification performance. In the baseline case, the pedestrian was fully visible, whereas in other cases, parts of the pedestrian's body were masked out. Masking the complete pedestrian, and leaving only positional information (bounding box), decreased human accuracy markedly, showing the importance of augmented visual features for this prediction task.

We address the following questions in this paper.

- At the short prediction horizons typical of the traffic safety context, can nonlinear models outperform linear models,

Manuscript received September 25, 2012; revised June 4, 2013, August 6, 2013, and August 23, 2013; accepted September 2, 2013. Date of publication October 17, 2013; date of current version March 28, 2014.

C. G. Keller was with the Image and Pattern Analysis Group, Department of Mathematics and Computer Science, University of Heidelberg, 69117 Heidelberg, Germany. He is now with the Environment Perception Department, Daimler Research and Development, 71034 Böblingen-Hulb, Germany.

D. M. Gavrila is with the Environment Perception Department, Daimler Research and Development, 89081 Ulm, Germany and also with the Intelligent Autonomous Systems Group, University of Amsterdam, 1098 Amsterdam, The Netherlands.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2013.2280766

or alternatively, can higher order Markov models outperform their first-order counterparts?

- Do augmented visual features (optical flow) improve path prediction and action classification over the use of positional information only?
- How does measurement error (e.g., pedestrian localization error and vehicle egomotion estimation error) affect the results? Can the more complex models still maintain an edge over the simpler ones?

In order to provide answers for the preceding questions, we consider four approaches that differ in their modeling of dynamics and in their use of augmented visual features; together, they cover a broad spectrum of possible approaches. In the category of nonlinear first-order models with augmented visual features, we propose a novel pedestrian path prediction approach, based on Gaussian process dynamical models (GPDMs) [3] and dense optical flow features (see Section III-A). An appealing aspect of this approach is that a low-dimensional latent representation is learned from the data, which takes into account the process dynamics. In the category of nonlinear higher order models with augmented visual features, we propose a novel probabilistic hierarchical trajectory matching (PHTM) approach, based on a low-dimensional motion representation (see Section III-B). Finally, in the category of first-order Markov models using positional information only, and mostly as a baseline, we consider the popular Kalman filter (KF, linear model) and its extension interacting multiple model KF (IMM KF, mixture of linear models) [4] (see Section III-C).

Experimental results on real traffic data are given in Section IV, with pedestrian image location obtained either from ground truth (GT) (optionally corrupted with noise) or obtained by a state-of-the-art pedestrian detection system. Several experimental cases are distinguished (pedestrian stopping versus walking, egovehicle standing versus moving). A discussion of the results, in terms of prediction performance and computational cost, is given in Section V. This paper concludes in Section VI.

II. Previous Work

Here, we discuss pedestrian motion model and path prediction techniques. For an overview of vision-based pedestrian detection, we refer to the surveys of Dollar *et al.* [5] and Enzweiler and Gavrila [6].

One way to perform path prediction relies on closed-form solutions for Bayesian filtering; in the KF [4], the current state of a dynamic system can be propagated to the future by means of the underlying linear dynamical model, without the incorporation of new measurements. The same idea can be applied to KF extensions to either multiple linear dynamical models, e.g., the IMM KF [4], or to nonlinear models, e.g., the extended KF or the unscented KF (see [7] and [8] for applications to pedestrian tracking).

An alternative approach to path prediction involves nonparametric stochastic models. Possible trajectories are generated by Monte Carlo simulations, taking into account the respective dynamical models. For example, Keller *et al.* [9] described an integrated vehicle safety system that combines sensing, situation analysis, decision making, and vehicle control, to automatically brake or evade for pedestrians. Collision detection assumes constant motion from the last estimates of pedestrian position and motion. Abramson and Steux [10] combined a constant motion model with particle filtering. De Nicolao *et al.* [11] distinguished lateral and longitudinal pedestrian velocities and model these independently by a random walk. Wakim *et al.* [12] modeled pedestrian motion by means of four states of a Markov chain, corresponding to standing still, walking, jogging, and running. Each state is associated with probability distributions of magnitude and direction of pedestrian velocity; the state changes are controlled by various transition probabilities. Recently, more complex pedestrian motion models have also accounted for group behavior and spatial layout, e.g., entry/exit points (see Antonini *et al.* [13] for a discussion). These latter approaches, although interesting, are less relevant to the traffic safety domain considered in this paper.

The limited amount of available training data precludes the use of modeling approaches that compute joint probability distributions over time intervals explicitly. Indeed, most pedestrian motion models consist of states that correspond to single time steps and are first-order Markovian. This potentially limits their expressiveness and precision. In contrast, Black and Jepson [14] described an extension of particle filtering to incrementally match trajectory models to input data. It is used for motion classification of 2-D gestures and expression. Sidenbladh *et al.* [15] added an efficient tree search in the context of articulated 3-D human pose recovery. Käfer *et al.* [16] applied this technique to vehicle motion prediction, utilizing the quaternion-based rotationally invariant longest common subsequence (QRLCS) metric for trajectory matching. Keller *et al.* [17] combined positional and optical flow features in the QRLCS matching to perform pedestrian path prediction and action classification (continue walking versus stopping at the curbside) from a vehicle. One of their findings is that humans are still better at this action classification task than the systems considered. Following up on the analysis of pedestrian intention at the curbside, Köhler *et al.* [18] addressed the continue-standing versus starting-to-walk classification task, from a stationary monocular camera. They combined a motion contour image based histogram of oriented gradient (HOG)-like descriptor with a linear support vector machine (SVM). Chen *et al.* [19] proposed a multilevel prediction model, in which the higher levels are long-term predictions based on trajectory clustering matching, whereas the low level uses an autoregressive model to predict the next time step.

A common assumption when dealing with human motion is that measurements in a high-dimensional space can be represented in a low-dimensional nonlinear manifold. Nonlinear dimensionality reduction methods allow learning the internal model of the data (see van der Maaten *et al.* [20] for an overview of techniques). It often depends on the data and the task at hand (e.g., visualization and classification) which of the techniques is best suited. Because measurements from human motions are time dependent, it is desirable to consider the dependence of the data over time. The Gaussian process latent variable model [21], which is a generalization of the probabilistic principal component analysis (PCA) [22], can be extended to model the

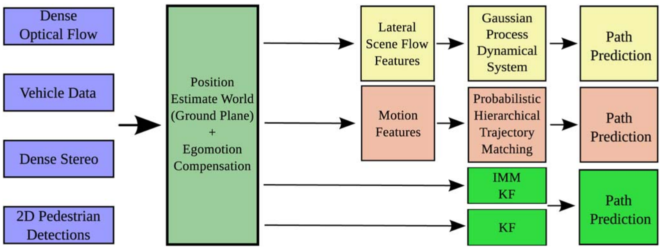

Dense Optical Flow
Vehicle Data
Dense Stereo
2D Pedestrian Detections
Position Estimate World (Ground Plane) + Egomotion Compensation
Lateral Scene Flow Features
Gaussian Process Dynamical System
Path Prediction
Motion Features
Probabilistic Hierarchical Trajectory Matching
Path Prediction
IMM KF
KF
Path Prediction

Fig. 2. Overview of considered approaches for pedestrian path prediction.

dynamics of the data. This GPDM [3] allows for nonlinear mapping from the latent space to the observation space, as well as a smooth prediction of latent points. In particular, the mapping (or prediction) of data points on the latent space makes this technique interesting for tracking application. Urtasun *et al.* [23] used a GPDM to track a small number of 3-D body points that have been derived using an image-based tracker. The system is trained using one gait cycle from six subjects and is able to handle several frames of occlusions. Andriluka *et al.* [24] used a dynamic part based limb detector in combination with a GPDM to allow robust detection and tracking in complex scenes with many persons and long-term occlusions. Raskin *et al.* [25] used a GPDM with an articulated model of the human body in combination with an annealed particle filter for tracking and action classification. Action classification is realized by comparing observed sequences with template sequences in latent space.

The contributions of this paper are as follows. We propose two novel approaches that use augmented features derived from dense optical flow for pedestrian path prediction, one GPDM based (see Section III-A) and the other PHTM based (see Section III-B). Furthermore, we present an experimental study on pedestrian path prediction (including baseline KF approaches), based on real video data from a vehicle and actual sensor processing, as opposed to simulated data. Given the documented benefit of dense stereo for pedestrian sensing [26], we use it as input to our approaches. Section III-B is based on our earlier work [17].

III. General Framework

We compare four different approaches for pedestrian path prediction, involving GPDMs, PHTM, KFs, and IMM KFs. For an overview, see Fig. 2.

To allow meaningful comparisons among the systems, several preprocessing components are set equal. Bounding boxes containing pedestrians are supplied from the same detector module. Dense disparity is computed using the semiglobal matching stereo algorithm [27]. Pedestrian positions on the ground plane are obtained by considering the midpoint of the bounding box and the disparity computed over the part of the bounding box that corresponds to the upper body (assuming typical human proportions). The latter involves clustering disparity values using mean shift [28] and selecting the cluster with the largest weight; the median of the corresponding disparity values provides the desired pedestrian distance.

Vehicle egomotion is compensated by rotation and translation of pedestrian positions to a global reference point using a single-track vehicle model [29] and velocity and yaw rate measurements from onboard sensor data. The two approaches that use augmented visual features (GPDM and PHTM) compute dense optical flow [30] over the bounding boxes provided by the pedestrian detector; this flow is, subsequently, egomotion compensated.

A. GPDM System

The first approach uses scene flow features describing the lateral movement of the pedestrian derived from the dense optical flow field and measured pedestrian distance in the world. Feature dimensionality is reduced by means of a GPDM [3] with a dynamic model in the latent space. To overcome the absence of direct mapping from feature space to latent space, the dynamic model is combined with a particle filter. GPDMs that capture the walking and stopping movements of a pedestrian are separately trained. The learned dynamical models provide optical flow fields at future time instants; future lateral positions can be derived by integration. Longitudinal position is independently estimated by means of a separate KF for each action class (walking versus stopping). Weighting lateral and longitudinal predictions using the probability of each action model results in future pedestrian positions.

1) Feature Extraction: Given the lateral component from dense optical flow and a pedestrian distance derived from dense stereo, the lateral velocity of a pedestrian in the world is computed.

With the pedestrian distance (as disparity $disp$), the horizontal component of the optical flow field V_u, the camera base width b, and the camera cycle time Δt, the lateral speed $v_X(m/s)$ of each pixel is computed using

$$v_X = \frac{V_u \cdot b}{disp \cdot \Delta t}. \tag{1}$$

To obtain only flow values located on the pedestrian body, a mask image is generated from the thresholded disparity image, and velocity values corresponding to the background are set to zero. Applying this distance mask also adds rough pedestrian contour information to the feature. Fig. 3 describes the feature extraction steps.

For further use as a feature, this scene flow image is rescaled to 32 × 16 pixel and concatenated to a feature vector $\mathbf{y_t} \in \mathbb{R}^D$, with $p = 512$. From the scene flow image (*SFlowX*), the lateral velocity of the pedestrian can be directly extracted using the median of velocity values located in the area of the pedestrian upper body (see red box in Fig. 3).

2) Dynamical Model: We are interested in a low-dimensional representation $\mathbf{x}_t \in \mathbb{R}^d$ of features $\mathbf{y}_t \in \mathbb{R}^D$ from a pedestrian sequence with $d < D$. This dimensionality reduction is realized using the GPDM [3], [23], [31], which allows modeling the dynamics of the features over time t in the low-dimensional space. For data in latent space $\mathbf{x}_t$, the relation to the input $\mathbf{y_t}$ can be described using

$$\mathbf{y_t} = g(\mathbf{x_t}; \mathbf{B}) + \mathbf{n}_{y,t} \tag{2}$$

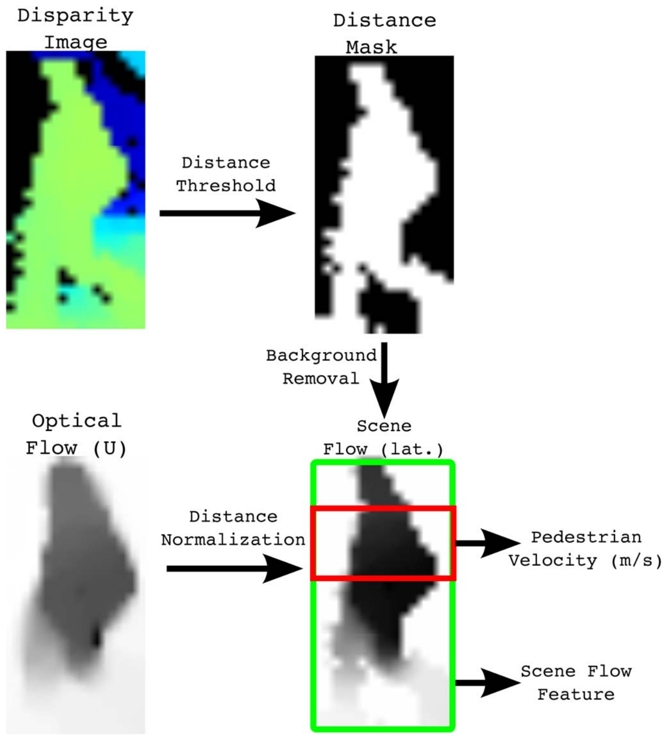


Fig. 3. Feature extraction using dense optical flow and roughly estimated pedestrian contour from dense stereo.

with zero-mean Gaussian noise $\mathbf{n}_{y,t}$ and mapping function g with parameters $\mathbf{B} = [b_1, b_2, \ldots]$. Assuming a first-order Markov model, the dynamics of the data in the latent space $\mathbf{x}_1, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_N$ can be described using

$$\mathbf{x_t} = f(\mathbf{x_{t-1}}; \mathbf{A}) + \mathbf{n}_{x,t} \tag{3}$$

with zero-mean Gaussian noise $\mathbf{n}_{x,t}$ and mapping function f with parameters $\mathbf{A} = [a_1, a_2, \ldots]$.

In a Gaussian process framework, the parameters and basis functions of f and g are marginalized out, and the positions of the latent coordinates are optimized.

a) Latent mapping: In the GPDM framework, the conditional density for the data $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]^T$ given latent positions $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T$ is described using

$$p(\mathbf{Y}|\mathbf{X}, \overline{\beta}, \mathbf{W}) = \frac{|\mathbf{W}|^N}{\sqrt{(2\pi)^{ND}|\mathbf{K}_Y|^D}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{K}_Y^{-1}\mathbf{Y}\mathbf{W}^2\mathbf{Y}^T\right)\right) \tag{4}$$

with kernel matrix $\mathbf{K}_Y$ and kernel hyperparameters $\overline{\beta} = \{\beta_1, \beta_2, \ldots\}$ and $\mathbf{W}$. To equally weight all the feature dimensions, the scale parameter is set to $\mathbf{W} = \mathbf{I}$ and is omitted in the following equations. Entries in the kernel matrix are defined using a kernel function $(\mathbf{K}_Y)_{i,j} = k_Y(\mathbf{x}_i, \mathbf{x}_j)$. For our data, we use a radial basis function (RBF) kernel with an additional noise term, i.e.,

$$k_Y(\mathbf{x}_i, \mathbf{x}_j) = \beta_1 \exp\left(-\frac{\beta_2}{2}\|\mathbf{x}_i - \mathbf{x}_j\|^2\right) + \frac{\delta_{\mathbf{x}_i, \mathbf{x}_j}}{\beta_3}. \tag{5}$$

b) Dynamic mapping: The dynamics of the time series data is incorporated using

$$p(\mathbf{X}|\overline{\alpha}) = \frac{p(\mathbf{x}_1)}{\sqrt{(2\pi)^{(N-1)d}|\mathbf{K}_X|^d}} \times \exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{K}_X^{-1}\mathbf{X}_{2:N}\mathbf{X}_{2:N}^T\right)\right) \tag{6}$$

with $\mathbf{X}_{2:N} = [\mathbf{x}_2, \ldots, \mathbf{x}_N]^T$, the kernel matrix $\mathbf{K}_X$ constructed from $\mathbf{X}_{1:N-1} = [x_1, \ldots, x_{N-1}]^T$ with dimensionality $(N-1) \times (N-1)$ and entries $(\mathbf{K}_X)_{i,j} = k_X(\mathbf{x}_i, \mathbf{x}_j)$. A combination of an RBF and a linear kernel with an additional noise is used for the dynamics

$$k_X(\mathbf{x}_i, \mathbf{x}_j) = \alpha_1 \exp\left(-\frac{\alpha_2}{2}\|\mathbf{x}_i - \mathbf{x}_j\|^2\right) + \alpha_3 \mathbf{x}_i^T \mathbf{x}_j + \alpha_4^{-1}\delta_{\mathbf{x}_i, \mathbf{x}_j} \tag{7}$$

with kernel hyperparameters $\overline{\alpha} = \{\alpha_1, \alpha_2, \ldots\}$.

c) Learning the GPDMs: Combining the latent and dynamics mapping defines the model

$$p(\mathbf{X}, \mathbf{Y}, \overline{\alpha}, \overline{\beta}) = p(\mathbf{Y}|\mathbf{X}, \overline{\beta})p(\mathbf{X}|\overline{\alpha})p(\overline{\alpha})p(\overline{\beta}). \tag{8}$$

Learning a GPDM requires finding the latent positions $\mathbf{X}$ and kernel hyperparameters $\mathcal{H} = \{\overline{\alpha}, \overline{\beta}\}$ with respect to the features $\mathbf{Y}$ by minimizing the negative log posterior $-\ln p(\mathbf{X}, \mathcal{H}|\mathbf{Y})$. Minimization can be done using a scaled conjugated gradient (SCG) method [31]. This requires the inverse of the kernel matrix with a complexity of $O(N^3)$ in each optimization iteration. We select $d = 3$ as the latent space dimensionality.

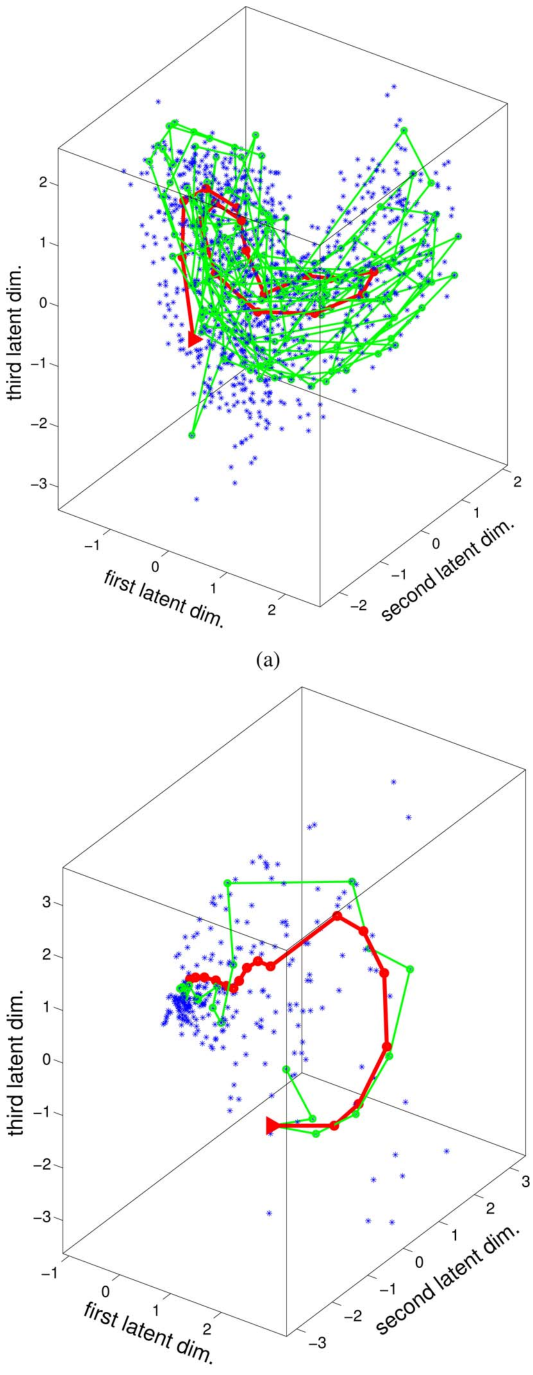
It is difficult to learn a generic GPDM that captures large variations in the data and different motions. Combining trajectory data in which the pedestrian is walking and data in which the pedestrian is stopping results in degenerated models. Urtasun *et al.* [32] introduced additional constraints to prevent the degeneration of models. Selecting the correct constraints for a model that captures the walking and stopping motions of a pedestrian for the used features is difficult, particularly with noisy data. Additionally, the complexity when training the model is increased. To avoid these problems, we train two separate models. The first model is trained using trajectory data segments in which pedestrians are walking. Stopping situations are selected to train the second model. Because the beginning of a stopping action is difficult to define, data from 20 frames (0.91 s) before the stopping of the pedestrians is used. Examples of the two models are plotted in Fig. 4.

d) Mean prediction: With the learned dynamic model, a point $\mathbf{x}$ in the latent space is predicted, and the most likely successor is derived using

$$\mu_{\mathbf{X}}(\mathbf{x}) = \mathbf{X}_{2:N}^T \mathbf{K}_{\mathbf{X}}^{-1} k_{\mathbf{X}}(\mathbf{x}) \tag{9}$$

with the vector $k_{\mathbf{X}}(\mathbf{x})$ containing at the ith entry the results of $k_X(\mathbf{x}, \mathbf{x_i})$ using training sample $\mathbf{x_i}$.

Fig. 4 illustrates this mean prediction of a point for several frames on the low-dimensional space.



(a)

(b)

Fig. 4. Traversal of a (○) training trajectory through the (∗) learned latent space and (○) mean predictions of a (▶) point for 17 frames (0.77 s). Figures depict (a) the walking case and (b) the stopping case. All available training samples are shown.

e) Latent reconstruction: A point $\mathbf{x}$ in the latent space is reconstructed using

$$\mu_{\mathbf{Y}}(\mathbf{x}) = \mathbf{Y}^T \mathbf{K}_{\mathbf{Y}}^{-1} k_{\mathbf{Y}}(\mathbf{x}). \tag{10}$$

An example of the reconstructed scene flow feature is shown in Fig. 5.

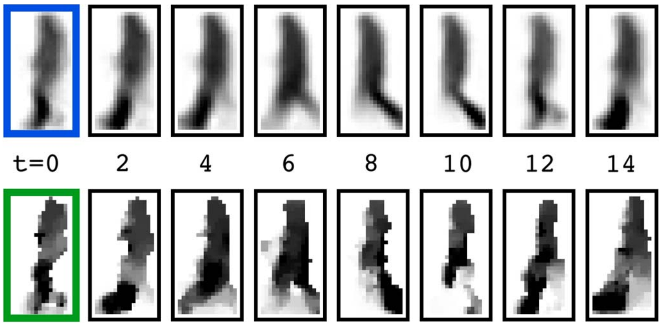


Fig. 5. (Top row) Reconstructed optical flow based on current state ($t = 0$) and state predictions ($t = 2, \ldots, 14$) in low-dimensional latent space. (Bottom row) Optical flow that is (will be) actually measured at the corresponding time steps.

3) Multiple Model Particle Filter: The state of a pedestrian at time t is described using $\phi_t = [\mathbf{x_t}, \mathcal{X}_t]$, where $\mathbf{x_t} \in \mathcal{R}^d$ is a point in the low dimensional space and $\mathcal{X}_t$ the lateral pedestrian position in the world. Given an observed motion feature $\mathbf{y}_t$ and observed lateral position $\mathcal{Y}_t$, the probability of a pedestrian state ϕ_t is computed by

$$p(\phi_t|\mathbf{y_t}, \mathcal{Y}_t) = \eta p(\mathbf{y_t}, \mathcal{Y}_t|\phi_t) \int p(\phi_t|\phi_{t-1}) p(\phi_{t-1}|\mathbf{y_{t-1}}, \mathcal{Y}_{t-1}) d\phi_{t-1} \tag{11}$$

with normalization constant η. The probability $p(\phi_t|\phi_{t-1})$ of observing a future state is computed from the GPDM latent space mean prediction.

This distribution is represented by a set of particles $\{\phi_t^{(s)} : s \in \{1, \ldots, S\}\}$ with corresponding weight $w_t^{(s)}$ that is propagated using a particle filter. Particles are predicted using the learned GPDM model with the predicted state $\hat{\phi}_t^{(s)} = [\hat{\mathbf{x}}_t^{(s)}, \hat{\mathcal{X}}_t^{(s)}]$ with $\hat{\mathbf{x}}_t^{(s)} = \mu_{\mathbf{X}}(\mathbf{x}_{t-1}^{(s)}) + n_x$, and $\hat{\mathcal{X}}_t^{(s)} = \mathcal{X}_{t-1}^{(s)} + s_{\mathcal{X}}(\hat{y}_t^{(s)}, \Delta t)$ which computes the traveled distance from the mean velocity derived from the reconstructed scene flow image $\hat{\mathbf{y}}_t^{(s)} = \mu_{\mathbf{Y}}(\hat{\mathbf{x}}_t^{(s)})$ and camera cycle time Δt. For each particle, the noise term n_x is randomly sampled from $\mathcal{N}(\mathbf{0}, I_d \times \sigma_{n_x}^2)$, with an experimentally derived $\sigma_{n_x} = 0.1$.

Scene flow feature similarity is computed using the Euclidean distance

$$d_f(\mathbf{y}, \hat{\mathbf{y}}) = \|\mathbf{y} - \hat{\mathbf{y}}\|^2. \tag{12}$$

For the lateral position in the world, the distance $d_p(\mathcal{Y}, \mathcal{X})$ is computed with

$$d_p(\mathcal{Y}, \mathcal{X}) = \|\mathcal{Y} - \mathcal{X}\|^2. \tag{13}$$

Using the distances between the observed and predicted data, the observation likelihood $p(y_t, \mathcal{Y}_t|\phi_t^{(s)}) \propto w_t^{(s)}$ is approximated using

$$w_t^{(s)} = \exp\left(-\frac{d_f\left(\mathbf{y}_t, \hat{\mathbf{y}}_t^{(s)}\right)^2}{2\sigma_f^2} - \frac{d_p\left(\mathcal{Y}_t, \hat{\mathcal{X}}_t^{(s)}\right)^2}{2\sigma_p^2}\right) \tag{14}$$

with an empirically estimated $\sigma_f = 7$ for the feature similarity and $\sigma_p = 0.06$ for the deviation of the lateral position. The

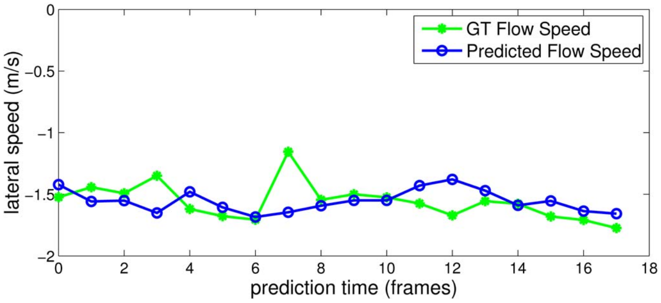


Fig. 6. Predicted speed derived from (○) predicted optical flow and corresponding (*) measured optical flow speed for different prediction horizons.

updated $\phi_t^{(s)}$ is obtained from $\hat{\phi}_t^{(s)}$ by reweighting the particle set.

For efficiency, an estimated state ϕ_T representing the pedestrian state in the future $T = t + \Delta T$ is derived from the weighted mean $\mathbf{x}^*$ of the particle set $\{\phi_t^{(s)}\}$ and iteratively predicted using $\mu_{\mathbf{X}}(\mathbf{x}^*)$. From the reconstructed predicted scene flow data (see Fig. 5), the pedestrian velocity in the future is computed (see Fig. 6). Integrating over the velocity predictions results in the predicted pedestrian position.

As mentioned in Section III-A2c, we trained two models for the different pedestrian motions. Models are combined using an interacting multiple model particle filter (*MM-PF*) similar to [33]. For each model a fixed number of particles ($S = 200$) is used to represent the state. From the set of particles M_i in model i the model probability is derived using

$$\gamma_i(t) = \frac{\sum_{\phi_t^{(s)} \in M_i} w^{(s)}}{\sum_{\phi_t^{(s)} \in M_1} w^{(s)} + \sum_{\phi_t^{(s)} \in M_2} w^{(s)}}. \tag{15}$$

Model probabilities are updated similar to the *IMM-KF* scheme, described in Section III-C. The conditional probability γ_{ij} of a transition from model i to j is computed using

$$\gamma_{ij}(t) = \frac{\Psi_{ij} \cdot \gamma_i(t)}{\sum_{k=1}^{2} \Psi_{kj} \cdot \gamma_k(t)} \tag{16}$$

with the state transition matrix Ψ.

We assume the lateral and longitudinal pedestrian dynamics to be weakly dependent. Longitudinal state estimation is decoupled, and to each of the lateral models (GPDM), a KF with a corresponding constant velocity (CV) or constant position (CP) model is assigned to track the position. Longitudinal positions $s_{\mathcal{Z}}^i(\Delta t)$ are linearly predicted with the estimated velocity of each filter. Mixing the lateral model predictions $s_{\mathcal{X}}^i(\Delta t)$ and the longitudinal KF prediction $s_{\mathcal{Z}}^i(\Delta t)$ with the state transition probabilities at t results in the pedestrian position

$$s_{\mathcal{X},\mathcal{Z}}(\Delta t) = \sum_{i=1}^{2} \gamma_i \cdot \begin{pmatrix} s_{\mathcal{X}}^i(\Delta t) \\ s_{\mathcal{Z}}^i(\Delta t) \end{pmatrix}. \tag{17}$$

In the following, the approach using the GPDMs in combination with scene flow features is abbreviated with *SFlowX/GPDM*.

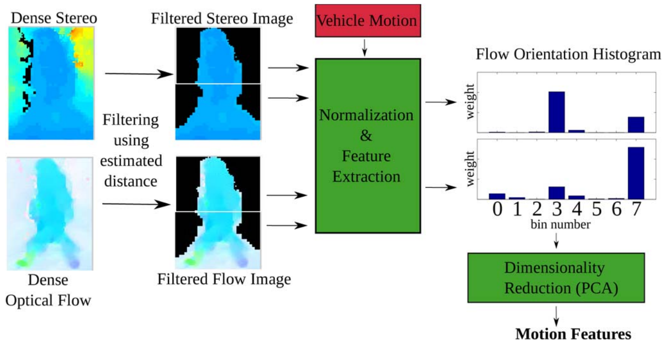


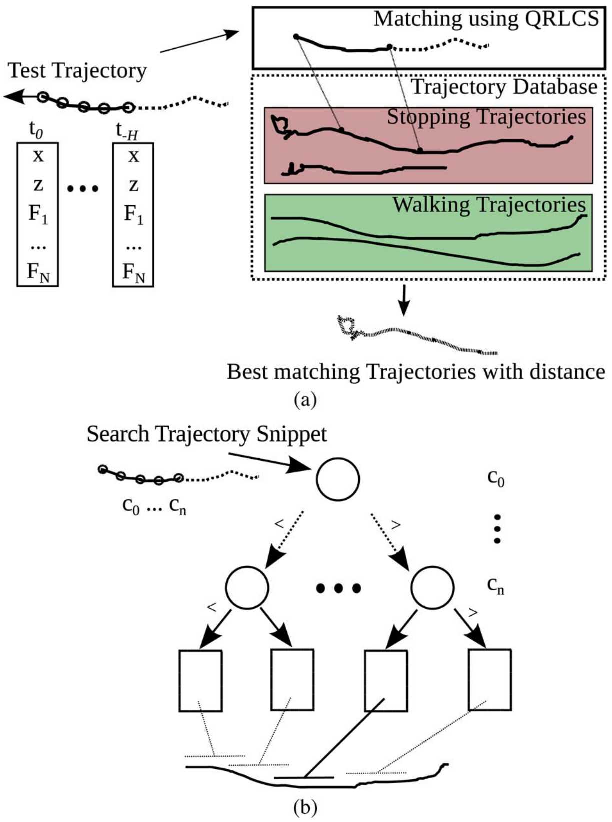
Fig. 7. Motion feature extraction in the PHTM-based system.

B. PHTM System

The second approach uses motion features involving a low-dimensional histogram representation of optical flow. Measured pedestrian positions and motion features are subsequently used in a trajectory matching and filtering framework. From the filter state, a future pedestrian position is derived by looking ahead on matched trajectories of the training set.

1) Motion Features: The low-dimensional feature captures flow variations on the pedestrian legs and upper body. In order to operate from a moving vehicle, additional invariance to pedestrian distance and vehicle motion is important. Features are designed to allow bounding box localization errors from a pedestrian detection system. Fig. 7 illustrates the feature extraction steps. Flow vectors are normalized with the camera cycle time to account for asynchronous capture and frame drops. Flow vectors are further normalized with measurements from dense stereo for invariance to different pedestrian distances. The resulting normalized motion field is used to extract features given a bounding box detection and distance estimation z_{ped} from a pedestrian detection system. To ensure that the pedestrian is located in the box for all possible limb extensions and slight localization errors, a bounding box aspect ratio of 4 : 3 is used. Motion vectors not belonging to the pedestrian body are suppressed by using only values at a depth similar to the estimated pedestrian distance. Remaining values in the motion field are used to compute the median object motion and extract orientation histograms. To capture motion differences between torso and legs, the bounding box is split into upper and lower subboxes. For each subbox, the median motion is removed to compensate the pedestrian egomotion. Resulting orientation vectors $v = [v_x, v_y]^T$ are assigned to bins $b \in [0, 7]$ using their 360° orientation $\theta = \operatorname{atan2}(v_y, v_x)$ and bin index $b = \lfloor \theta / \pi / 4 \rfloor$. Bin contributions are weighted by their magnitude, and resulting histograms are normalized with the number of contributions. A feature vector is formed by concatenating the histogram values and the median flow for the lower and upper boxes. Dimensionality reduction of the feature vector is achieved by applying PCA. The first three PCA dimensions with the largest eigenvalue are used as final histogram of orientation motion (*HoM*) features.

2) Trajectory Matching: A pedestrian trajectory Ω is represented using the ordered tuples $\Omega = ((\omega_1, t_1), \ldots, (\omega_N, t_N))$. For every time stamp t_i, the pedestrian state ω_i consists of the lateral and longitudinal positions of the pedestrian and



(a)

(b)

Fig. 8. (a) Test trajectory with a history of length H containing position and feature information for every entry is matched to the training database. Resulting matching position and similarity distance to trajectories in the training database describe a possible trajectory course and class label. (b) Tree representation of the trajectory training database. Leaf nodes represent trajectory snippets of fixed length. Similar trajectories are searched by traversing the tree using the trajectory descriptors for every level.

additional features extracted from optical flow [see Fig. 8(a)]. For path prediction, it is possible to compare an observed test trajectory with a history of H pedestrian states to each trajectory in a training database using a similarity measure. With the QRLCS metric [16], the optimal translation and rotation parameters to superimpose two trajectories are derived. The distance $\text{dist}_{\text{QRLCS}}(\Omega_i, \Omega_j) \in [0, 1]$ between two trajectories is given by the number of possible assignments determined by an ε area around each pedestrian state, normalized by the number of pedestrian states. Fig. 8(a) illustrates this comparison process.

We replace this exhaustive search by a probabilistic search framework [15], [16]. A set of overlapping subtrajectories (snippets, e.g., [34]) with a fixed history of pedestrian states is created from a training database. Information of the snippet position in the origin trajectory and successor snippets is kept for later use. By piling the features for each state in a snippet into a description vector and applying the PCA method to these vectors, their principal dimensions can be ordered according to the largest eigenvalue. The resulting transformed description vector $\mathbf{c}$ is used to build a binary search tree. For each level l, the snippet is assigned to the left or right subtree depending on the sign of c_l. Given N training snippets, the depth of the search tree, i.e., n, is $O(\log(N))$. Fig. 8(b) illustrates this search tree.

Given a trajectory $\Omega_{1:t}$, the probability of the state ϕ_t is computed by

$$p(\phi_t|\Omega_{1:t}) = \eta p(\Omega_{1:t}|\phi_t) \int p(\phi_t|\phi_{t-1}) p(\phi_{t-1}|\Omega_{1:t-1}) d\phi_{t-1} \tag{18}$$

with a normalization constant η. The distribution $p(\phi_t|\Omega_{1:t})$ is represented by a set of samples or particles $\{\phi_t^{(s)}\}$, which are propagated in time using a particle filter [14]. Each particle $\phi_t^{(s)}$ represents a snippet describing a pedestrian state with a history and an assigned likelihood. Our transition model $p(\phi_t|\phi_{t-1})$ is determined by a probabilistic search in the binary tree. Particle prediction is performed by a probabilistic search in the constructed binary tree and a lookup for the successor snippet in the training database. The distribution $p(\Omega_{1:t}|\phi_t)$ represents the likelihood that the measurement trajectory $\Omega_{1:t}$ can be observed given the current state. In the context of particle filters, this value corresponds to the weight of a particle and is approximated using $w^{(s)} = 1 - \text{dist}_{\text{QRCLS}}$ for each particle $\phi_t^{(s)}$.

An estimated state $\phi_T^{(s)}$ representing the pedestrian state in the future $T = t + \Delta T$ can be derived by looking ahead on the associated origin trajectory for the current state $\phi_t^{(s)}$. This results in many hypotheses, which are compensated using a weighted mean shift algorithm [28] with a Gaussian kernel and weights $w^{(s)} \sim p(\phi_T^{(s)}|\Omega_{1:t})$. At the final predicted state ϕ_T^*, the cluster center with the highest accumulated weight is selected.

The trajectory database contains two classes of trajectory snippets: the class $\mathcal{C}_s$, in which the pedestrian is stopping, and the class $\mathcal{C}_w$, in which the pedestrian continues walking. For the predicted object state ϕ_T^* derived using cluster members $L = \{\phi_t^{(l)}\}$ and the corresponding weight $w^{(l)}$, the stopping probability can be approximated using

$$p(\mathcal{C}_s|L) \approx \frac{\sum_{\phi_t^{(l)} \in \mathcal{C}_s} w^{(l)}}{\sum_{\phi_t^{(l)} \in \mathcal{C}_s} w^{(l)} + \sum_{\phi_t^{(l)} \in \mathcal{C}_w} w^{(l)}}. \tag{19}$$

In the following, the PHTM approach using HoM features is abbreviated with *HoM/Traj*.

C. KF-Based Systems

1) KF: As a third approach, a linear KF [4] is used. The state $\hat{\mathbf{X}}$ of the filter is modeled as

$$\hat{\mathbf{X}} = [\, z \quad x \quad v_z \quad v_x \,]^T$$

with z/x being the longitudinal/lateral position of the pedestrian to the vehicle and v_z/v_x being its absolute longitudinal/lateral velocity in the world. Pedestrian positions are pseudomeasurements provided by the stereo pedestrian detection component, as described at the beginning of this section. A CV model is assumed as a pedestrian motion model. Using this model means that all deviations from a constant pedestrian motion have to be captured as process noise. With the assumption that a pedestrian moving at 1.8 m/s can stop in 1 s, we select a process noise parameter $q = 1.8$ for the filter.

2) IMM-KF: The fourth approach extends the previous KF with an additional CP model; this way, the *IMM-KF* [4] is realized. The basic idea is to maintain a KF for each possible motion model with state $\hat{\mathbf{x}}_j(t)$ and model probability $\gamma_j(t)$. This means that a steady walking pace is represented using a filter with the CV model with process noise parameter q_{CV}. For nonmoving pedestrians, the CP model with q_{CP} applies. Each iteration consists of three steps: interaction, filtering, and mixing. The interaction step computes the mixing probability γ_{ij} from the current model probability γ_j and the state transition probability Ψ_{ij} [see (16)]. From the mixing probability, the mixed state mean $\hat{\mathbf{x}}_{0j}(t)$ and the covariance matrix $\hat{\mathbf{P}}_{0j}(t)$ are computed as initial input for each filter in the filtering step using

$$\hat{\mathbf{x}}_{0j}(t) = \sum_{i=1}^{2} \hat{\mathbf{x}}_i(t)\gamma_{ij}(t). \quad (20)$$

For the computation of $\hat{\mathbf{P}}_{0j}(t)$, see [4]. In the filtering step, a *KF* predict/update step is done using the mixed state mean $\hat{\mathbf{x}}_{0j}(t)$ and covariance matrix $\hat{\mathbf{P}}_{0j}(t)$ derived in the interaction step. Given the likelihood function $\Lambda_j(t+1) = \mathcal{N}(r_j(t+1), \mathbf{S_j}(t+1))$ with residuum $r_j(t+1)$ and residual covariance $\mathbf{S}_j(t+1)$, the updated probabilities $\gamma_j(t+1)$ are computed using

$$\gamma_j(t+1) = \frac{1}{c}\Lambda_j(t+1)\sum_{i=1}^{2}\Psi_{ij}\gamma_i(t) \quad (21)$$

with normalization factor c. An approximation of the resulting mixture model is then computed in the mixing step using

$$\hat{\mathbf{x}}(t+1) = \sum_{i=1}^{2} \hat{\mathbf{x}}_i(t+1)\gamma_i(t+1). \quad (22)$$

For the following evaluation, $q_{\text{CV}} = 0.21$ and $q_{\text{CP}} = 0.41$ have been derived from the set of training trajectories, with respect to the positions minimum root-mean-square error (RMSE). The matrix Ψ describing the transition probabilities between the CV and CP models has been experimentally derived from the available training data $\Psi = [0.999, 0.001, 0.001, 0.999]$. Choosing larger values for the model transitions results in more frequent undesired switches, particularly with noisy measurements. The *IMM-KF* is said to be nonsensitive to improperly selected transition probabilities [35].

IV. Experiments

Video data of two scenarios (see Fig. 1) were recorded using a stereo camera system (baseline 30 cm, 22 fps) mounted behind the windshield of a vehicle. The first scenario features the stopping of a pedestrian at the curbstone. In the second scenario, the pedestrian crosses the street. In both scenarios, the pedestrian was not occluded. In some test runs, the vehicle is stationary, whereas in others, the vehicle is moving at speeds of 20–30 km/h. The data set involved four different pedestrians in three different locations at a distance range of 5–34 m to the vehicle. Table I provides some further statistics on the data set. Fig. 9 illustrates some test images.

The GT locations of the pedestrians in the world were obtained by manual labeling the pedestrian shapes in the images. The median disparity value on the pedestrian upper body and the center foot point of the shape is used to obtain the longitudinal and lateral positions on the ground plane. In terms of alignment along the time axis, for each trajectory in which the pedestrian is stopping, the moment of the last placement of the foot is labeled as the stopping moment. The time-to-stop (TTS) value counts the number of frames until this event; frames earlier to the stopping event have positive TTS values; frames after the stopping event have negative TTS values. In sequences in which the pedestrian continues walking, the closest point to the curbstone (with closed legs) is labeled. Analogous to the TTS definition, the latter is called the time-to-curb (TTC) value.

TABLE I
Number of Sequences With Different Pedestrian and Vehicle Actions in Our Data Set

Sequences	**vehicle standing**	**vehicle moving**	**vehicle standing+moving**
ped. stopping	11	5	16
ped. walking	9	4	13

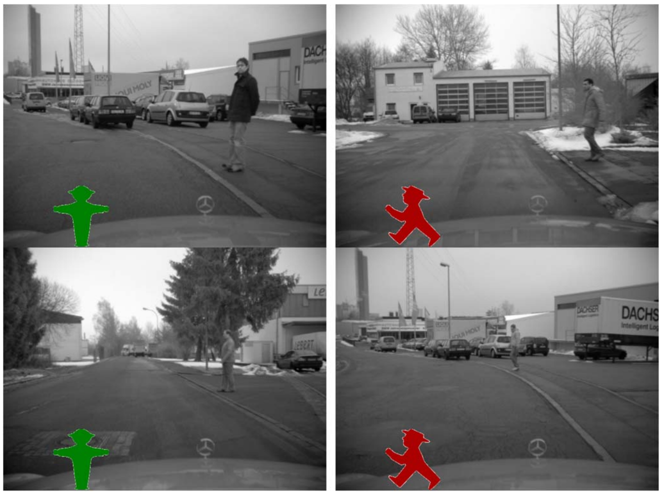


Fig. 9. Example images from the data set showing the pedestrian action. Images show the labeled (left) stopping or (right) walking moment.

Performance evaluation is done using input data with different noise characteristics with regard to the image bounding box positions. Two-dimensional bounding boxes derived from manually labeled pedestrian shapes (termed *label box*) are used as the most accurate input data for feature extraction and localization; it reflects the case of an "ideal" pedestrian detector. We further consider the case in which these ideal 2-D bounding boxes are perturbed by uniform noise; we add up to 10% of the original height of the bounding boxes to their height and center (the resulting bounding boxes are termed *jittered*). Finally, we consider 2-D bounding boxes provided by a state-of-the-art HOG/linSVM pedestrian detector [36] (termed *system detections*). Hereby, detection "gaps" are filled in by means of a standard correlation tracker. Considering data with artificial noise allows abstracting away from the noise bias of a particular pedestrian detector. As we will see shortly, the overall noise level artificially added is realistic, in the sense that it is similar to that of a state-of-the-art detector.

The lateral and longitudinal position errors on the ground plane for different input data are summarized in Table II. In these experiments, we compared with a smoothed version of

TABLE II
MEAN DEVIATION (IN METERS) OF THE PEDESTRIAN POSITION ON THE GROUND PLANE (LATERAL AND LONGITUDINAL) COMPARED WITH THE SMOOTHED GT DATA

	veh. standing+moving	
	lat.	long.
GT	0.03	0.10
label box	0.05	0.22
jittered box	0.13	0.68
sys. detections	0.06	0.64

the GT ground plane positions. GT positions from walking trajectories, in which we are certain that the pedestrian is moving with an approximately constant velocity ($-40 \leq \text{TTC} \leq 40$), were fitted with a curvilinear model, minimizing pedestrian velocity and yaw changes by nonlinear least squares. For the stopping trajectories, smoothing was only applied to the cases in which the pedestrian is standing ($\text{TTS} \leq 0$), by simple averaging. Note that smoothed GT was only used for the purpose of Table II. In the path prediction experiments, comparisons involved the nonsmoothed GT. Following observations can be made from Table II. As expected, the longitudinal error is larger than the lateral error due to stereo vision characteristics. Adding aforementioned uniform noise on 2-D bounding boxes results in degradation of positional accuracy of about 10 and 50 cm in lateral and longitudinal directions, respectively. Positional errors are similar for the artificial noise and real detector case.

A. Parameter Settings and Evaluation Setup

We compare the four approaches using equal parameter settings, whenever possible. Lateral and longitudinal noise parameters for the *KF* and *IMM-KF* and longitudinal noise parameters for tracking the distance of the *SFlowX/GPDM* are selected from Table II. Process noise parameters and state transition matrix were heuristically derived (see Section III-C). The same state transition matrix is used for the *MM-PF* of the *SFlowX/GPDM* system and the *IMM-KF*.

The analysis of walking trajectories showed an average gait cycle of 10–14 frames for different pedestrians. The trajectory database for the *HoM/Traj* contains subtrajectories, generated in a sliding window fashion, with a fixed length of ten frames. For test trajectories, a history of 14 frames is used to capture gait cycle variations. Approximating the current probability density is done with $S = 400$ particles and a tree search deviation parameter of $\beta = 0.05$. The mean shift position procedure operates with a kernel width value $h = 0.1$.

Training and testing data have been processed using *leave-one-out* cross-validation. This means that one sequence is used for testing and the remaining training sequences are used to learn the GPDM models *SFlowX/GPDM* or for search tree generation (*HoM/Traj*).

B. Pedestrian Path Prediction

We are interested in the ability of each system to predict future pedestrian positions accurately. Tables III and IV list ground plane localization accuracy (i.e., longitudinal and lateral dimensions combined) at different prediction horizons, for each system. Localization accuracy is measured in terms of the mean and standard deviation of the per-sequence RMSE. Per-sequence RMSE is determined by comparing system predictions at various time horizons with the GT, when the pedestrian is inside the frame range [20, -10], when frame 0 denotes the manually labeled TTS/TTC moment. This corresponds to an evaluation time range of [0.91, -0.45] s around the TTS/TTC event. Pedestrian positions are predicted up to 17 frames (0.77 s) into the future. Tables III and IV list the results for walking and stopping trajectories, respectively. Results are further differentiated based on whether the own vehicle is standing or moving or whether all data are used (cf., Table I).

On walking scenarios (see Table III), all approaches show a similar prediction performance when pedestrian bounding boxes are set precise (*label box*). In the more realistic case of inaccurate image localization (*jittered box*) and moving vehicle, we see that *HoM/Traj*, unlike the other approaches, shows no performance degradation and thus gains a slight edge. We attribute this to the robustness of trajectory matching to outliers in the longitudinal dimension.

On stopping scenarios (see Table IV), in which the constant velocity assumption is violated, incorporation of motion features leads to a path prediction performance advantage of up to a factor of 2 for *HoM/Traj* and *SFlowX/GPDM* compared with the *KF*-based variants (e.g., *jittered* data and vehicle standing and moving cases). As before, the trajectory matching of *HoM/Traj* shows added robustness to noise caused by bounding box position errors and vehicle egomotion.

In the intelligent vehicle pedestrian safety context, the *lateral* component of the localization error is particularly relevant; it determines whether the pedestrian enters the vehicle driving corridor and a collision potentially occurs. Fig. 10 lists the mean lateral localization error at various time offsets to the labeled TTS/TTC moment (*jittered* data and vehicle standing and moving cases). Separate plots are shown depending on the prediction horizon (0 or 17 frames) and whether the pedestrian is walking or stopping. We observe no significant performance difference between walking trajectories [see Fig. 10(a) and (c)]. For stopping scenarios [see Fig. 10(b) and (d)], the advantage of the additional motion model of the *IMM-KF* versus the *KF* becomes visible (in Table IV, this advantage was averaged away over the frame range [20, -10], due to the inclusion of time instants still involving walking). Stopping of the pedestrian leads to a switch to the CP model and a lower localization error compared with the *KF* with CV model. Fig. 10 also shows that *HoM/Traj* and *SFlowX/GPDM* are more quickly able to adjust to the change in the pedestrian motion, resulting in a lower lateral localization error than the KF-based approaches. Fig. 11 illustrates the distribution of the lateral prediction error *difference* between *IMM-KF* and *HoM/Traj* for stopping trajectories. Performance differences are clearly visible close to the stopping event $\text{TTS} = 0$.

Results using tracked detections from a state-of-the-art HOG/linSVM pedestrian detector [36] are listed in Table V. For this experiment, we used a subset of 7 walking and 13 stopping trajectories (five trajectories with a moving vehicle) in which the pedestrian detector had a decent performance in the first place (detection gaps no longer than ten frames consecutively).

TABLE III
MEAN COMBINED LONGITUDINAL AND LATERAL RMSE (IN METERS) FOR *WALKING TRAJECTORIES* AND DIFFERENT PREDICTION HORIZONS (FRAMES)

		veh. standing						veh. standing + moving				veh. moving			
		label box				jittered box		label box		jittered box		label box		jittered box	
		0	**5**	**11**	**17**	**0**	**17**	**0**	**17**	**0**	**17**	**0**	**17**	**0**	**17**
KF	Mean	0.15	0.19	0.22	0.28	0.24	0.35	0.17	0.38	0.24	0.45	0.21	0.62	0.24	0.69
	± *Std*	*0.07*	*0.07*	*0.09*	*0.12*	*0.2*	*0.2*	*0.07*	*0.25*	*0.17*	*0.28*	*0.05*	0.29	*0.06*	*0.29*
IMM-KF	Mean	0.14	0.17	0.2	0.25	0.23	0.34	0.19	0.41	0.26	0.5	0.29	0.77	0.33	0.86
	± *Std*	*0.07*	*0.08*	*0.1*	*0.12*	*0.21*	*0.22*	*0.1*	*0.3*	*0.19*	*0.34*	*0.09*	*0.26*	*0.09*	*0.27*
HoM/Traj	Mean	0.13	0.17	0.23	0.29	0.14	0.30	0.14	0.33	0.15	0.35	0.15	0.43	0.15	0.44
	± *Std*	*0.03*	*0.03*	*0.05*	*0.07*	*0.03*	*0.07*	*0.03*	*0.13*	*0.03*	*0.11*	*0.02*	*0.17*	*0.02*	*0.12*
SFlowX/GPDM	Mean	0.15	0.2	0.26	0.34	0.17	0.5	0.17	0.41	0.19	0.52	0.21	0.62	0.25	0.69
	± *Std*	*0.04*	*0.08*	*0.13*	*0.18*	*0.06*	*0.32*	*0.05*	*0.24*	*0.08*	*0.31*	*0.06*	*0.25*	*0.08*	*0.22*

TABLE IV
MEAN COMBINED LONGITUDINAL AND LATERAL RMSE (IN METERS) FOR *STOPPING TRAJECTORIES* AND DIFFERENT PREDICTION HORIZONS (FRAMES)

		veh. standing						veh. standing + moving				veh. moving			
		label box				jittered box		label box		jittered box		label box		jittered box	
		0	**5**	**11**	**17**	**0**	**17**	**0**	**17**	**0**	**17**	**0**	**17**	**0**	**17**
KF	Mean	0.20	0.36	0.61	0.93	0.27	0.81	0.24	1.04	0.33	1.14	0.32	1.25	0.44	1.39
	± *Std*	*0.04*	*0.06*	*0.1*	*0.15*	*0.14*	*0.23*	*0.08*	*0.26*	*0.15*	*0.37*	*0.1*	*0.33*	*0.15*	*0.51*
IMM-KF	Mean	0.18	0.31	0.55	0.87	0.27	0.77	0.22	0.98	0.32	1.08	0.32	1.19	0.47	1.26
	± *Std*	*0.03*	*0.04*	*0.07*	*0.12*	*0.13*	*0.24*	*0.08*	*0.19*	*0.16*	*0.2*	*0.1*	*0.17*	*0.17*	*0.15*
HoM/Traj	Mean	0.12	0.19	0.34	0.58	0.12	0.56	0.13	0.63	0.12	0.62	0.15	0.74	0.15	0.76
	± *Std*	*0.02*	*0.04*	*0.11*	*0.17*	*0.02*	*0.10*	*0.02*	*0.21*	*0.02*	*0.18*	*0.02*	*0.23*	*0.03*	*0.19*
SFlowX/GPDM	Mean	0.23	0.27	0.35	0.51	0.29	0.62	0.16	0.53	0.23	0.64	0.21	0.66	0.41	0.89
	± *Std*	*0.07*	*0.07*	*0.06*	*0.07*	*0.08*	*0.22*	*0.05*	*0.23*	*0.26*	*0.29*	*0.05*	*0.32*	*0.42*	*0.36*

We observe that using actual system detections, rather than simulated detections, does not change the performance ranking of the approaches considered (compare Table V with the entries in Tables III and IV, in which the vehicle is standing and moving). In fact, performance with actual system detections is similar to that obtained with noise-perturbed GT *jittered* data; this is not surprising given similar per-frame localization measurement error (cf., Table II).

C. Pedestrian Action Classification

We also tested the ability of various systems to classify pedestrian actions, i.e., whether the pedestrian will cross or not. Fig. 12 illustrates the performance of each system on stopping and walking test trajectories; depicted is the estimated probability of stopping, as a function of TTS or TTC. For the *SFlowX/GPDM* and *HoM/Traj* systems, this was achieved by means of (15) and (19), respectively. For *IMM-KF*, stopping was estimated by means of the probability of the CP model, following (21).

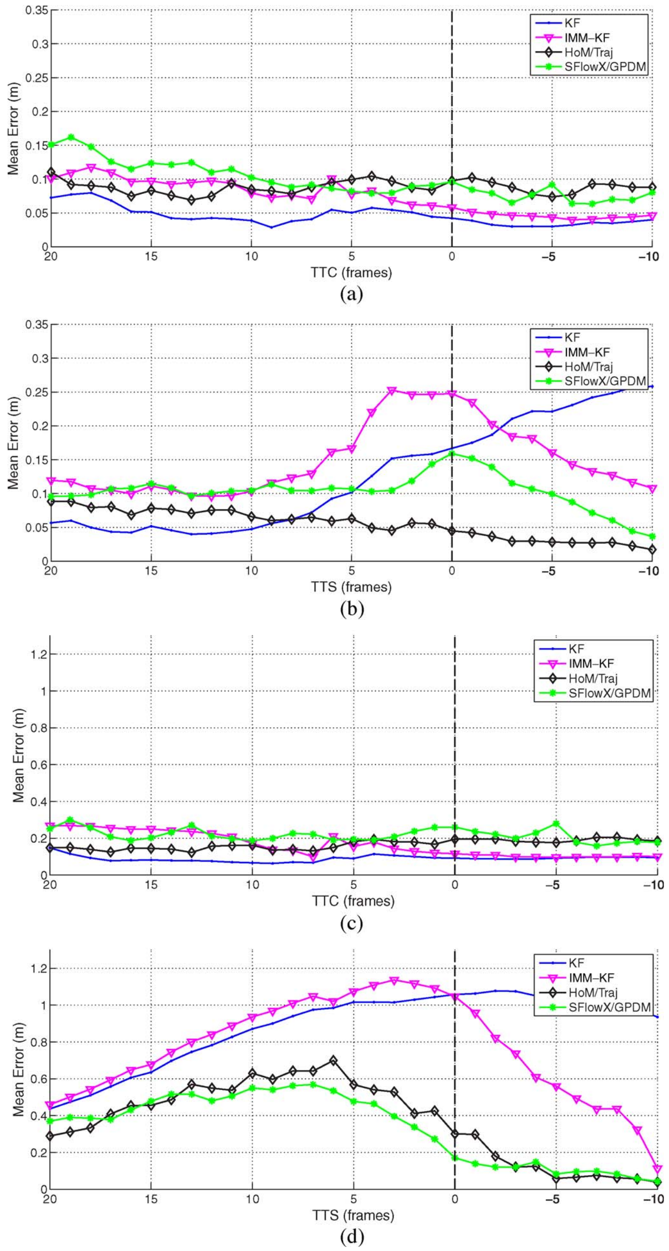
To put the performance of the systems in context, we also evaluated human performance. Video data were presented to several test subjects using graphical user interfaces, where playback was automatically stopped at five different TTC or TTS moments (20, 11, 8, 5, and 3). For each run, the test subjects had to decide whether the pedestrian will stop at the curbstone or cross the street and provide a probability (i.e., confidence) using a slider ranging from 0 to 1. Sequence and playback stopping point were randomly selected before being presented to the test subjects to avoid the effect of reidentification.

In Fig. 12, on walking trajectories, all systems show a low and relatively constant stopping probability. On stopping trajectories, all systems initially start with a low stopping probability, since stopping is preceded by walking. However, within a dozen frames before the stopping event, the stopping probability increases more markedly.

Class membership is determined at each time instant of an input trajectory assigned by thresholding the estimated stopping probability (cf., Fig. 12). Based on the training set, we selected for each system and for the human group a threshold that minimizes its classification error (i.e., stopping classified as walking and vice versa) over all sequences and time instants. Fig. 13 illustrates the resulting classification accuracy over time using these "optimal" thresholds. As can be seen, the humans outperform the various automatic systems at this action classification task. The humans reach accuracy of 0.8 in classifying the correct pedestrian action about 570 ms before the event. This accuracy is only reached about 230 ms before the event by the newly developed *SFlowX/GPDM* and *HoM/Traj* systems, which use augmented visual features. The baseline *IMM-KF* system does worst, reaching the corresponding accuracy only about 90 ms before the event.

V. DISCUSSION

Table V indicates that the proposed more advanced methods for pedestrian path prediction (*SFlowX/GPDM* or *HoM/Traj*) can achieve more accurate path prediction than basic approaches (linear KF or IMM extension thereof). The associated benefit, in terms of reduction of the combined lateral and longitudinal position error, is 10–50 cm at a time horizon of 0–17 frames (up to 0.77 s) around the stopping event. Fig. 10(d) indicates that a 50-cm improvement in lateral position estimation is reached at several time instants. Tables III and IV also suggest that the vehicle egomotion compensation is done reasonably but not perfectly. Further benefits can be obtained when localizing the pedestrian more accurately and improving upon the vehicle egomotion compensation. Comparing the columns



(a)

(b)

(c)

(d)

Fig. 10. Mean lateral localization error at each time step for *jittered* data and vehicle standing and moving (walking versus stopping trajectories, prediction horizon 0 versus 17 frames). (a) Pedestrian walking, prediction time 0 frames. (b) Pedestrian stopping, prediction time 0 frames. (c) Pedestrian walking, prediction time 17 frames. (d) Pedestrian stopping, prediction time 17 frames.

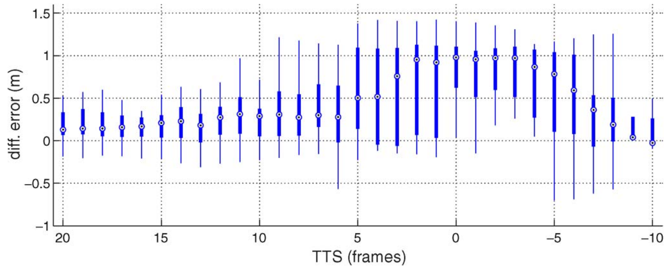
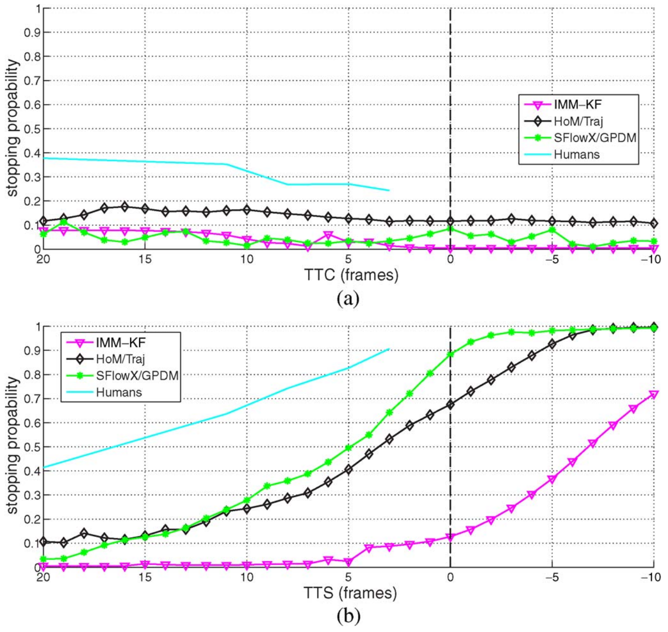


Fig. 11. Distribution of the lateral prediction error difference (*IMM-KF-HoM/Traj*). Results for the *jittered* data, prediction horizon of 17 frames and stopping trajectories.

TABLE V

MEAN COMBINED LONGITUDINAL AND LATERAL RMSE (IN METERS) FOR *STOPPING AND WALKING TRAJECTORIES* USING SYSTEM DETECTIONS WITH DIFFERENT PREDICTION HORIZONS (FRAMES)

		system detections			
		walking		**stopping**	
		0	**17**	**0**	**17**
KF	Mean	0.2	0.55	0.27	1.08
	± *Std*	0.05	0.28	0.08	0.29
IMM-KF	Mean	0.21	0.55	0.29	1.04
	± *Std*	0.06	0.3	0.14	0.25
HoM/Traj	Mean	0.14	0.39	0.14	0.63
	± *Std*	0.03	0.12	0.04	0.22
SFlowX/GPDM	Mean	0.15	0.43	0.21	0.52
	± *Std*	0.06	0.27	0.06	0.19



(a)

(b)

Fig. 12. Estimated probability of stopping over time for (a) walking and (b) stopping test trajectories (averaged over all respective sequences). (a) Pedestrian walking. (b) Pedestrian stopping.

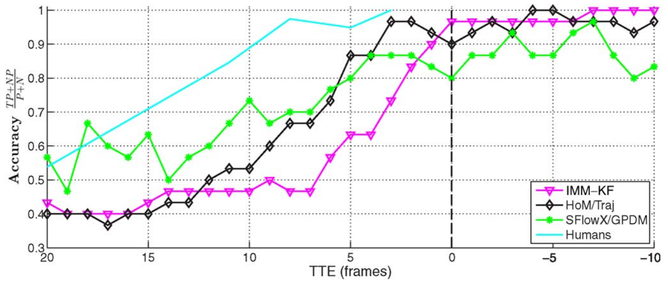


Fig. 13. Classification accuracy of the different systems over time. Results for the *jittered* data.

"vehicle standing, label box, 17" (ideal situation) and "vehicle moving, jittered box, 17" (currently achievable situation) shows that position prediction can be improved by approximately 15–81 cm for the various systems.

These findings are encouraging in terms of the expected benefits that can be achieved, when integrating more sophisticated path planning in pedestrian safety systems that perform emergency vehicle maneuvers (braking, steering).

We now turn to computational cost issues. The popularity of the simple linear KF can be explained due to its relative effectiveness and its low computational requirements. Although the computational cost doubles for the two-process model *IMM-KF*, it remains moderate compared with the *HoM/Traj* and *SFlowX/GPDM* approaches. For the latter, the cost of motion feature extraction needs to be accounted for first. Furthermore, for the *HoM/Traj* approach, a prediction step requires traversing the search tree for each particle and looking up the successor snippet. Computational costs to predict a snippet is linear in the depth of the search tree. To incorporate new measurements, the *QRCLS* distance to each particle has to be computed to update the particle weights. Looking ahead pedestrian position requires applying the mean shift procedure to the predicted particle positions to find the main mode. Main computational costs of the *SFlowX/GPDM* can be subdivided into the costs of predicting a GPDM latent space position and reconstructing the feature to apply the particle weight update. To predict a single particle, the mean prediction on the latent space has to be applied [see (9)]. Because the first part of the formula $(\mathbf{X}_{2:N}^T \mathbf{K}_{\mathbf{X}}^{-1})$ can be precomputed, the online costs for a latent space prediction result from evaluating the kernel function $k_{\mathbf{X}}(\mathbf{x})$ between the particle latent position $\mathbf{x}$ and all inducing variables. Similarly, reconstructing the feature requires evaluation of (10) with a precomputed $\mathbf{Y}^T \mathbf{K}_{\mathbf{Y}}^{-1}$ and evaluation of the kernel function $k_{\mathbf{Y}}(\mathbf{x})$. Costs for an update and prediction of a particle are limited by the number of inducing variables.

Using an unoptimized MATLAB implementation on a 2.53-GHz central processing unit, the path prediction 17 frames into the future requires, on average, 0.003 s for *KF* and 0.017 s for *IMM-KF*. The MATLAB version of the *HoM/Traj* approach with an optimized version of the trajectory matching and mean shift procedure in *C* requires 0.6 s. Without code optimization, the *SFlowX/GPDM* approach requires, on average, 5.4 s for the prediction. Processing times for both *HoM/Traj* and *SFlowX/GPDM* can be much improved by special hardware (i.e., graphics processing unit, digital signal processor, and field-programmable gate array) by parallelizing the particle computation.

In terms of scalability, learning a GPDM quickly becomes unfeasible for larger data sets (for example, ≥ 1000 samples) without an approximation method. The fully independent training conditional (FITC) [37] method reduces the complexity from $O(N^3)$ for the SCG method [31] (cf., Section III-A3) to $O(k^2 N)$, when k is the number of data points that remain in the computation of the covariance matrix. Our full data set contains approximately 1700 training samples, and we set $k = 100$. When using the FITC approximation with a fixed number of inducing variables k, the online computational costs do not increase when extending the size of the training set. Without an approximation method, kernel evaluations between all samples in the training set have to be applied. Regarding scalability with the number of pedestrian motion patterns considered, training a single model containing different motion patterns leads to degenerated models on our data set. Degenerated models showed an insufficient latent space prediction performance. Although methods exist to prevent model degeneration [32] when using sequences with a large variety of motion patterns, the computational complexity during training increases. Extending the *SFlowX/GPDM* system with additional motion patterns requires training separate GPDMs for each motion pattern. In the online case, the computational costs linearly increase in the number of models.

Since the *HoM/Traj* system is an instance-based learning approach using a probabilistic search tree, different motion patterns can be added to the training set without complication. Adding additional snippets to the training set leads to an increase in the depth of the binary search tree. Online costs to predict the state of the particle filter are thus sublinear (logarithmic) in the number of training samples.

VI. Conclusion

We have considered four approaches (*SFlowX/GPDM, HoM/Traj, KF*, and *IMM-KF*) for stereo-vision-based pedestrian path prediction from a vehicle. Two scenarios were considered: in one, the pedestrian walking toward the curbside, lateral to the vehicle driving direction, would stop; whereas in the other, the pedestrian would continue walking.

Experiments indicated similar path prediction performance of the four approaches on walking motion, with near-linear dynamics. During stopping, however, the newly proposed approaches (*SFlowX/GPDM* or *HoM/Traj*), with nonlinear and/or higher order models and augmented motion features, achieved a more accurate (longitudinal and lateral) position prediction of 10–50 cm at a time horizon of 0–0.77 s around the stopping event. During stopping, a 50-cm improvement in lateral position prediction was reached at several time instants. Further benefits are possible when localizing the pedestrian more accurately and improving upon the vehicle egomotion compensation: We obtained improvements in lateral position prediction of 15–81 cm for the various systems.

These are encouraging results, indicating that more advanced pedestrian path prediction approaches can make a real difference, when integrated in the next-generation active pedestrian safety systems that perform emergency vehicle maneuvers (braking, steering). However, more work is necessary on improving pedestrian localization, enlarging the set of pedestrian motion patterns considered and increasing the size of the data set, before these benefits can materialize.

Acknowledgment

The authors would like to thank C. Schnörr (University of Heidelberg) and M. Enzweiler (Daimler Research and Development) for valuable discussions and feedback. They would also like to thank N. Lawrence (University of Sheffield) for making his GPDM software publicly available.

References

[1] M.-M. Meinecke, M. Obojski, D. M. Gavrila, E. Marc, R. Morris, M. Töns, and L. Lettelier, “Strategies in terms of vulnerable road user protection,” EU Project SAVE-U, Deliverable D6, 2003.

[2] S. Schmidt and B. Färber, “Pedestrians at the kerb—Recognizing the action intentions of humans,” *Transp. Res. F, Traffic Psychol. Behav.*, vol. 12, no. 4, pp. 300–310, Jul. 2009.

[3] J. M. Wang, D. J. Fleet, and A. Hertzmann, "Gaussian process dynamical models for human motion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 283–298, Feb. 2008.

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

[5] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Proc. IEEE Conf. CVPR*, 2009, pp. 304–311.

[6] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2179–2195, Dec. 2009.

[7] 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.*, 2008, pp. 37–42.

[8] J. Tao and R. Klette, "Tracking of 2D or 3D irregular movement by a family of unscented Kalman filters," *J. Inf. Convergence Commun. Eng.*, vol. 10, no. 3, pp. 307–314, Jul. 2012.

[9] C. Keller, T. Dang, A. Joos, C. Rabe, H. Fritz, and D. M. Gavrila, "Active pedestrian safety by automatic braking and evasive steering," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1292–1304, Dec. 2011.

[10] Y. Abramson and B. Steux, "Hardware-friendly pedestrian detection and impact prediction," in *Proc. IEEE Intell. Veh. Symp.*, 2004, pp. 590–595.

[11] G. De Nicolao, A. Ferrara, and L. Giacomini, "A collision risk assessment approach as a basis for the on-board warning generation in cars," in *Proc. IEEE Intell. Veh. Symp.*, 2002, pp. 436–441.

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

[13] G. Antonini, S. V. Martinez, M. Bierlaire, and J. Thiran, "Behavioral priors for detection and tracking of pedestrians in video sequences," *Int. J. Comput. Vis.*, vol. 69, no. 2, pp. 159–180, Aug. 2006.

[14] M. J. Black and A. D. Jepson, "A probabilistic framework for matching temporal trajectories," in *Proc. ECCV*, 1998, pp. 909–924.

[15] H. Sidenbladh, M. J. Black, and L. Sigal, "Implicit probabilistic models of human motion for synthesis and tracking," in *Proc. ECCV*, 2002, vol. 2350, pp. 784–800.

[16] E. Käfer, C. Hermes, C. Wöhler, H. Ritter, and F. Kummert, "Recognition of situation classes at road intersections," in *Proc. IEEE ICRA*, 2010, pp. 3960–3965.

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

[18] S. Köhler, M. Goldhammer, S. Bauer, K. Doll, U. Brunsmann, and K. Dietmayer, "Early detection of the pedestrian's intention to cross the street," in *Proc. IEEE Intell. Transp. Syst. Conf.*, 2012, pp. 1759–1764.

[19] Z. Chen, D. Ngai, and N. Yung, "Pedestrian behavior prediction based on motion patterns for vehicle-to-pedestrian collision avoidance," in *Proc. IEEE ITSC*, 2008, pp. 316–321.

[20] L. van der Maaten, E. Postma, and H. van den Herik, "Dimensionality reduction: A comparative review," Tilburg Univ., Tilburg, The Netherlands, Tech. Rep. TiCC-TR 2009-005, 2009.

[21] N. Lawrence, "Gaussian process latent variable models for visualization of high dimensional data," in *Proc. Adv. NIPS*, 2004, vol. 16, pp. 329–336.

[22] M. Tipping and C. Bishop, "Probabilistic principal component analysis," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 61, no. 3, pp. 611–622, 1999.

[23] R. Urtasun, D. J. Fleet, and P. Fua, "3D people tracking with Gaussian process dynamical models," in *Proc. IEEE Conf. CVPR*, 2006, pp. 238–245.

[24] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *Proc. IEEE Conf. CVPR*, 2008, pp. 1–8.

[25] L. Raskin, M. Rudzsky, and E. Rivlin, "Dimensionality reduction using a Gaussian process annealed particle filter for tracking and classification of articulated body motions," *Comput. Vis. Image Understanding*, vol. 115, no. 4, pp. 503–519, Apr. 2011.

[26] C. Keller, M. Enzweiler, C. Schnörr, M. Rohrbach, D.-F. Llorca, and D. M. Gavrila, "The benefits of dense stereo for pedestrian detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1096–1106, Dec. 2011.

[27] H. Hirschmüller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, Feb. 2008.

[28] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.

[29] P. Riekert and T. E. Schunck, "Zur fahrmechanik des gummibereiften kraftfahrzeugs," *Arch. Appl. Mech.*, vol. 11, no. 3, pp. 210–224, Jun. 1940.

[30] A. Wedel, T. Pock, J. Braun, U. Franke, and D. Cremers, "Duality TV-L1 flow with fundamental matrix prior," in *Proc. Image Vis. Comput. New Zealand*, 2008, pp. 1–6.

[31] J. M. Wang, D. J. Fleet, and A. Hertzmann, "Gaussian process dynamical models," in *Proc. Adv. NIPS*, 2006, pp. 1441–1448.

[32] R. Urtasun, D. Fleet, and N. Lawrence, "Modeling human locomotion with topologically constrained latent variable models," in *Proc. Human Motion–Understanding, Modeling, Capture Animation*, 2007, pp. 104–118.

[33] Y. Boers and J. Driessen, "Interacting multiple model particle filter," *IET Radar, Sonar Navig.*, vol. 150, no. 5, pp. 344–349, Oct. 2003.

[34] N. Howe, M. Leventon, and W. Freeman, "Bayesian reconstruction of 3D human motion from single-camera video," in *Proc. NIPS*, 2000, pp. 820–826.

[35] S. Blackman and R. Popoli, *Design and Analysis of Modern Tracking Systems*. Norwood, MA, USA: Artech House, 1999.

[36] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. CVPR*, 2005, pp. 886–893.

[37] N. Lawrence, "Learning for larger datasets with the Gaussian process latent variable model," in *Proc. Workshop Artif. Intell. Stat.*, 2007, pp. 21–24.

Christoph G. Keller received the M.Sc. degree in computer science from the University of Freiburg, Freiburg, Germany, in 2007. He is currently working toward the Ph.D. degree with the University of Heidelberg, Heidelberg, Germany.

In 2006 and 2007, he was a visiting Student Researcher with Siemens Corporate Research, Princeton, NJ, USA. Since 2012, he has been with Daimler Research and Development, Böblingen-Hulb, Germany. His research interests include video analysis for intelligent vehicles, particularly pedestrian tracking and vehicle localization.

Dariu M. Gavrila received the Ph.D. degree in computer science from the University of Maryland at College Park, MD, USA, in 1996.

In 1996, he was a Visiting Researcher with the Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. Since 1997, he has been a Senior Research Scientist with Daimler Research and Development, Ulm, Germany. Since 2003, he has additionally been a part-time Professor with the University of Amsterdam, Amsterdam, The Netherlands, in the area of intelligent perception systems. Over the past 15 years, he has focused on visual systems for detecting human presence and activity, with application to intelligent vehicles, smart surveillance, and social robotics. He led the multiyear pedestrian detection research effort at Daimler, which materialized in the Mercedes-Benz E- and S-Class models (2013).

Prof. Gavrila was the recipient of the I/O Award 2007 from the Dutch Science Foundation (NWO). His personal Web site is www.gavrila.net.