A Probabilistic Framework for Joint Pedestrian Head and Body Orientation Estimation

Fabian Flohr, Madalin Dumitru-Guzu, Julian F. P. Kooij, and Dariu M. Gavrila

***Abstract*—We present a probabilistic framework for the joint estimation of pedestrian head and body orientation from a mobile stereo vision platform. For both head and body parts, we convert the responses of a set of orientation-specific detectors into a (continuous) probability density function. The parts are localized by means of a *pictorial structure* approach, which balances part-based detector responses with spatial constraints. Head and body orientations are estimated jointly to account for anatomical constraints. The joint single-frame orientation estimates are integrated over time by particle filtering. The experiments involved data from a vehicle-mounted stereo vision camera in a realistic traffic setting; 65 pedestrian tracks were supplied by a state-of-the-art pedestrian tracker. We show that the proposed joint probabilistic orientation estimation framework reduces the mean absolute head and body orientation error up to 15° compared with simpler methods. This results in a mean absolute head/body orientation error of about 21°/19°, which remains fairly constant up to a distance of 25 m. Our system currently runs in near real time (8–9 Hz).**

***Index Terms*—Active pedestrian safety, computer vision, pose estimation, social robotics.**

I. Introduction

SIGNIFICANT progress has been made over the last few years on video-based pedestrian detection (e.g., [1]). In the area of driver assistance, this has led to the first commercial active pedestrian systems reaching the market. Daimler, for example, has introduced a stereo-vision-based pedestrian system in the 2013–2014 Mercedes-Benz S-, E-, and C-Class models.

A sophisticated situation analysis relies on an accurate path prediction. For pedestrians, the latter is challenging due to their high maneuverability; pedestrians can change their walking direction or accelerate/decelerate on a whim. Any auxiliary information that can help to reduce this uncertainty is welcome. Empirical evidence suggests that the pedestrian body and head orientation is a good indicator as to what the pedestrian will do next. For example, a human factors study by Schmidt and Färber [2] 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. The study showed that head motion was among the most important indicators of future pedestrian action. More recently, Hamaoka *et al.* [3] presented a study on head turning behaviors at pedestrian crosswalks, in order to establish the best point of warning for inattentive pedestrians. They used gyro sensors to record head turning and had test pedestrians press a button when they recognized an approaching vehicle.

In this paper, we address the problem of estimating pedestrian head and body orientation over time from a mobile stereo vision platform, focusing on the application of active pedestrian safety for intelligent vehicles, but the developed techniques could be also applied to mobile robotics, where the aim is that a robot socially interacts with persons in its environment [4]. In this paper, we use a stereo sensor setup (image resolution of 1176×640 pixel, baseline of 22 cm, frame rate of 17 Hz) that is already available in production vehicles on the market. Inputs to our approach are the bounding boxes provided by a state-of-the-art pedestrian tracker developed by the authors (a histogram of oriented gradients (HOG)/linSVM pedestrian detector [5] combined with a Kalman filter, but that particular choice is not material for the paper contributions). We present a principled probabilistic approach for dealing with faulty part detections, continuous orientation estimation, coupling of the body and head localization and orientation, and temporal integration. By estimating parameters of the spatial body part configuration from real data, we assure that anatomical characteristics are accounted for. Fig. 1 shows an overview of our proposed approach.

The outline of this paper is as follows. Section II discusses related work on person pose estimation. Section III covers our probabilistic framework for head and body orientation estimation. The section starts with a discussion of the overall graphical model used (see Section III-B). It is followed by a description of the underlying dynamical and observation models (see Sections III-C and III-D, respectively). Thereafter, the modeling of the spatial prior is discussed (see Section III-E). Section IV provides the experimental results on real traffic data. Section V places the obtained results in context to other approaches and to what a practical application might require. This paper concludes in Section VI.

Manuscript received March 25, 2014; revised September 15, 2014 and October 26, 2014; accepted November 28, 2014. Date of publication January 14, 2015; date of current version July 31, 2015. This research was funded in part by the EC FP7 679 FROG project under Grant 288235. The Associate Editor for this paper was S. Birchfield.

F. Flohr, J. F. P. Kooij, and D. M. Gavrila are with the Environment Perception Department, Daimler Research and Development, 89081 Ulm, Germany, and also with the Intelligent Systems Laboratory, University of Amsterdam, 1098 XH Amsterdam, The Netherlands.

M. Dumitru-Guzu was with the Environment Perception Department, Daimler R&D, Ulm, Germany, and also with the Computer Vision Laboratory, Delft University of Technology, 2628 CD Delft, The Netherlands. He is now with Fotonation, 040205 Bucharest, Romania.

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.2014.2379441

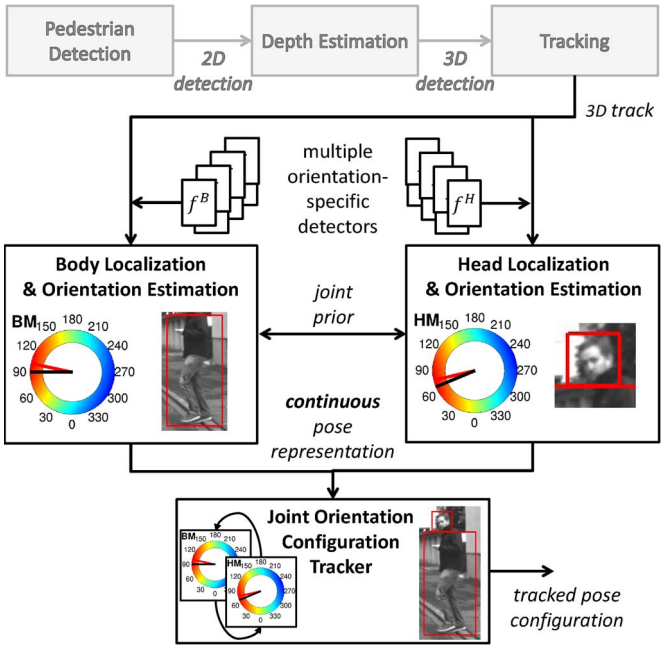

Fig. 1. Proposed joint probabilistic orientation estimation approach (shaded modules are outside the scope of this paper).

II. Related Work

An extensive amount of research, meanwhile, exists on articulated pose estimation. Here, we focus on previous work on head and body orientation estimation.

Approaches in body part orientation estimation are largely application dependent (see surveys [6] and [7] for head orientation estimation). Applications in human–machine interaction [8], [9] or entertainment [10] typically consider high-resolution images and cooperative subjects under controlled backgrounds to estimate the body part orientation. Applications in surveillance [11]–[16] and in intelligent vehicle [17], [18] domains need to cope with low-resolution images, with complex and dynamic backgrounds, and changing lighting conditions.

To cope with these challenges, approaches often use robust lower level image features such as scale invariant feature transform and/or HOG [11], [12], [18]–[20], Haar [17], [21], local receptive fields (LRFs) [18], and distance metrics [9], [13], [19] in combination with different classification schemes (e.g., support vector machines (SVMs) [12], [13], [18], [20], [21], neural networks (NNs) [18], random regression/decision trees or ferns [9], [11], [19], or boosting cascades [17]) to perform orientation estimation.

Such data-driven approaches can be used for both head and body orientation estimation. For example, Schulz *et al.* [17] trained a boosting cascade of Haar-like features for eight head orientation classes in a one-versus-all manner. The maximum classifier response over all possible hypotheses (different scales and locations) and the eight orientation classes was selected as the final estimation result for head position and head orientation. Enzweiler and Gavrila [18] used four orientation-specific classifiers jointly to give the final pedestrian detection output. The same classifiers are reused to infer a continuous orientation estimate for the detected pedestrian. Benfold and Reid [19] used a random fern architecture with a combination of HOG and color-based features to infer head orientation, after the head was found by a HOG-based head detector. Training was done with eight orientation classes. While the majority of aforementioned methods used manually labeled training data, Benfold and Reid [11] learned head orientations unsupervised by using the output of a tracking system [22], supposing that head orientation is dependent on walking direction. The walking direction can be also used as a proxy for body orientation (e.g., [14]), thus assuming that people only move forward.

Model-driven strategies represent a possible alternative to data-driven approaches. Orientation information of body parts is here often a direct result of an applied shape model. In particular, active shape models (ASMs) and active appearance models (AAMs) introduced by Cootes [23], [24] are often used for inferring head or body orientation estimation. ASMs combine statistical shape models (SSMs) (compact linear-subspace probabilistic representations) with means to match these to images. AAMs [24] extend ASMs by capturing shape and texture information jointly. ASMs and AAMs require feature correspondence, unlike exemplar-based representations. Fitting them to an image can result in suboptimal solutions because of local maxima. Giebel and Gavrila [25] represented shapes by multiple SSMs to account for different shape aspects (pedestrian feet apart versus feet closed) and orientations. The idea of such multiple linear subspaces was also adopted by Lee and Kriegman [26]. They used the method of Hall *et al.* [27], [28] and applied an incremental online update of multiple linear-subspace models, each representing a face orientation. Zhu and Ramanan [29] used a mixture of trees to infer face detection and orientation estimation. The trees share a pool of facial landmarks and use global mixtures (similar to AAM) to capture topological changes due to viewpoint. While there is a limited ability of using accurate shape information for the head with decreasing resolution, body orientation estimation can exploit, even in lower resolution images, prior knowledge about the body shape by matching shape models (e.g., [30]).

To improve orientation estimations, [11], [12], [14], [15], [31], and [32] introduced constraints that set head and body orientation into relation of each other. Chen and Odobez [12] used such constraints directly during classifier training, whereas Zhao *et al.* [32] used body orientation information to differentiate online between opposite head directions. Smith *et al.* [15] constrained the head location with respect to the body location to obtain a physically possible configuration. Benfold and Reid [11] applied a conditional random field for modeling the interaction between the head orientation, walking direction, and appearance to recover gaze direction. While Robertson and Reid [14] constrained the head orientation on the velocity direction, Chen *et al.* [31] introduced a coupling between head and body orientation and between body orientation and velocity direction. The constraints are modeled here by *von Mises* distributions.

By tracking, single-frame orientation estimations can be smoothed, and results can be further improved. One of the simplest approaches here is to choose the most frequent direction over a fixed number of frames [21]. More sophisticated models use, for example, hidden Markov model (e.g., [20])

or particle filter (PF) frameworks (e.g., [8], [14], [15], [31], and [33]) to keep track of a body part orientation distribution over time. Constraints between body parts can be then modeled efficiently within the used dynamic model as in [14] and [31]. Smith *et al.* [15] used a reversible-jump Markov chain Monte Carlo sampling scheme for particle filtering to handle a large state space consisting of interperson (multiperson tracking) and intraperson (localization between head and body) interactions.

Finally, there has been extensive work done on articulated 3-D body pose recovery (e.g., see surveys [34] and [35]). These typically require multiple cameras, are computationally intensive, and still have issues with robustness.

III. Joint Head and Body Orientation Estimation

A. Overview and Contributions

See Fig. 1. Motivated by efficiency and the existence of previous modules, we use a decoupled pedestrian tracker that estimates for each time step t the pedestrian's position $\boldsymbol{x}_t = [x_t, y_t]$ and pedestrian's height h_t, defined in world coordinates on the ground plane, and velocity $\dot{\boldsymbol{x}}_t = [\dot{x}_t, \dot{y}_t]$. The pedestrian tracks are provided as input to our orientation tracker, which, in turn, tracks the head ω_t^H and body orientation ω_t^B jointly as $\boldsymbol{\omega}_t = [\omega_t^H, \omega_t^B]$. We will therefore assume that all $\boldsymbol{x}$, $\dot{\boldsymbol{x}}$, and h are known up to time t and focus in this paper on the estimation of $\boldsymbol{\omega}_t$ only, which we will refer to as the state space. Additional constraints between head and body regarding the orientation are applied within the dynamic model.

Let $\boldsymbol{z}_t = [\boldsymbol{z}_t^H, \boldsymbol{z}_t^B]$ be the observed image data at time t, which can be decomposed into head observations $\boldsymbol{z}_t^H$ and body observations $\boldsymbol{z}_t^B$. Since we only have as input an estimate of the pedestrian's full bounding box in the image, but do not know the exact location of the body parts (head or full body), we have to take multiple image regions into account for both parts. For example, when there are N candidate regions for the head at time t, we can write out the corresponding observation as $\boldsymbol{z}_t^H = [z_t^{H(1)}, z_t^{H(2)}, \cdots, z_t^{H(N)}]$.

We use multiple detectors to evaluate how well an image region corresponds to a specific body part in a certain orientation. The angular domain of [0°, 360°) is discretized into a fixed set of eight orientation classes, centered around $0°, 45°, \ldots, 315°$ (0° and 90° are associated with frontal and left-facing poses, respectively, when viewed from the camera). Each class then has a detector, e.g., $f_0, f_{45}, \ldots, f_{315}$, for both head and body, such that the detector response $f_o(z)$ is strength for the evidence that image region z contains the body part in orientation class o. Note that this gives a tradeoff, as having more classes and detectors requires more training data and computational effort but also yields more precise evidence of the true angle (up to some point). An additional nontarget or background detector $f_-(z)$ assigns a likelihood to the case that z does not contain the body part. The output of all detectors $f_o(z)$ and $f_-(z)$ are then used to determine if and where a body part is present in the image region z, relying on disparity-based image segmentation and a *pictorial structure* (PS) [36] on the head and body configuration as a spatial prior.

The contribution of this paper is a principled joint probabilistic head and body orientation estimation framework that handles faulty part detections, continuous orientation estimation, coupling of the body and head localization and orientation, and temporal integration. We differ to [12], [14], [19], and [31] in several ways. We consider an intelligent vehicle context. While we rely on a small set of detectors for canonical body part orientations, our joint observation model for head and body deals with continuous angles. It is also used to jointly localize the head and body (additionally exploiting disparity information from stereo vision and knowledge of body configuration) and still accounts for the possibility of occluded body parts or false positives. Furthermore, the temporal model has coupled the joint orientation dynamics and enforces temporal consistency. This paper extends our previous work [37].

B. Filtering Orientations

Let $\boldsymbol{z}_{1:t}$ denote all observations up to and including time t and $\dot{\boldsymbol{x}}_{1:t}$ the corresponding pedestrian velocities provided by the position tracker. We use a Bayes' filter to obtain the posterior, i.e., $p(\boldsymbol{\omega}_t|\boldsymbol{z}_{1:t}, \dot{\boldsymbol{x}}_{1:t})$, which represents our belief of the state at time t after observing $\boldsymbol{z}_{1:t}$. For each time instance, the filter performs the following two steps.

First, a prediction is made given all earlier observations, i.e.,

$$p(\boldsymbol{\omega}_t|\boldsymbol{z}_{1:t-1}, \dot{\boldsymbol{x}}_{1:t}) = \int p(\boldsymbol{\omega}_t|\boldsymbol{\omega}_{t-1}, \dot{\boldsymbol{x}}_t) p(\boldsymbol{\omega}_{t-1}|\boldsymbol{z}_{1:t-1}, \dot{\boldsymbol{x}}_{1:t-1}) d\boldsymbol{\omega}_{t-1} \quad (1)$$

where $p(\boldsymbol{\omega}_{t-1}|\boldsymbol{z}_{1:t-1}, \dot{\boldsymbol{x}}_{1:t-1})$ is the posterior for the previous time step. The dynamic model $p(\boldsymbol{\omega}_t|\boldsymbol{\omega}_{t-1}, \dot{\boldsymbol{x}}_t)$ will be discussed in Section III-C.

Second, an update is made to incorporate new evidence $\boldsymbol{z}_t$ in the prediction, i.e.,

$$p(\boldsymbol{\omega}_t|\boldsymbol{z}_{1:t}, \dot{\boldsymbol{x}}_{1:t}) \propto p(\boldsymbol{z}_t|\boldsymbol{\omega}_t) p(\boldsymbol{\omega}_t|\boldsymbol{z}_{1:t-1}, \dot{\boldsymbol{x}}_{1:t}) \quad (2)$$

where $p(\boldsymbol{z}_t|\boldsymbol{\omega}_t)$ is the observation model, which will be discussed in Section III-D.

For the corresponding dynamic Bayesian network (DBN), see Fig. 2. Since exact inference is intractable, we use a PF [38] for approximate inference. The PF represents the posterior distribution by a set of particles in the state space, which facilitates using a nonlinear and multimodal dynamic model.

C. Dynamic Model

The dynamic model for the head and body orientations is

$$p(\boldsymbol{\omega}_t|\boldsymbol{\omega}_{t-1}, \dot{\boldsymbol{x}}_t) = p\left(\omega_t^H|\omega_{t-1}^H, \omega_t^B\right) p\left(\omega_t^B|\omega_{t-1}^B, \omega_{t-1}^H, \dot{\boldsymbol{x}}_t\right) \quad (3)$$

(see Fig. 2). Similar to [14], we constrain the head orientation at the current time step on the head orientation of the previous

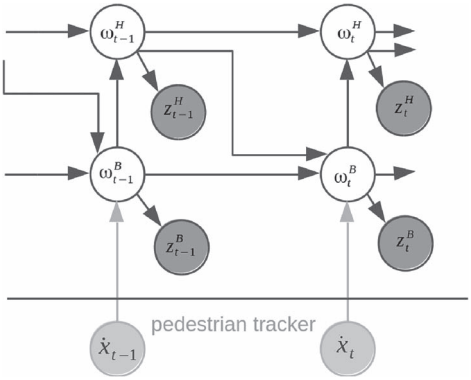

Fig. 2. DBN showing the dependencies between latent variables (head and body orientation, ω^H and ω^B) and observed variables (head and body measurements, $\boldsymbol{z}^H$ and $\boldsymbol{z}^B$). Latent variables are unshaded, and observed variables are shaded. Pedestrian velocity ($\dot{\boldsymbol{x}}$) is estimated by an external tracker.

time step and on the current body orientation with

$$p\left(\omega_t^H|\omega_{t-1}^H,\omega_t^B\right)=\alpha_{hh}\mathcal{V}\left(\omega_t^H;\omega_{t-1}^H,\kappa_{hh}\right)+(1-\alpha_{hh})\mathcal{V}\left(\omega_t^H;\omega_t^B,\kappa_{hb}\right) \quad (4)$$

where κ_{hh} and κ_{hb} are concentration parameters for the von Mises distribution. The von Mises $\mathcal{V}(\cdot;\omega,\kappa)$ is an analog of the normal distribution for the circular domain, with mean angle ω and concentration κ. It reduces to a circular uniform distribution when $\kappa=0$. The balance between temporal consistency and the assumption that the head orientation is around the body orientation is given by the weight α_{hh}. The first term in (4) models the case that the current head orientation is distributed around the previous head orientation. The second term covers the (possibly alternative) case where the head has moved to a similar orientation as the body.

We condition the body orientation on the body and head orientation of the previous time step and on the current pedestrian velocity, i.e.,

$$\begin{aligned}p(\omega_t^B|\omega_{t-1}^B,\omega_{t-1}^H,\dot{\boldsymbol{x}}_t)&=\alpha_{bb}\mathcal{V}(\omega_t^B;\omega_{t-1}^B,\kappa_{bb})+\alpha_{bh}\mathcal{V}(\omega_t^B;\omega_{t-1}^H,\kappa_{bh}).\\&+(1-\alpha_{bb}-\alpha_{bh})\mathcal{V}(\omega_t^B;ang(\dot{\boldsymbol{x}}_t),\kappa_{bv}).\end{aligned} \quad (5)$$

With $ang()$, we denote the angle of the velocity vector and with $\alpha_{bb,bh}\in[0,1]$ (with $\alpha_{bb}+\alpha_{bh}\leq 1$) the weighting factors for the terms. The first term in (5) expresses that the body orientation is typically centered around its previous orientation. Furthermore, there are cases when the body orientation changes to where the pedestrian is looking, which are captured by the second term. The last term expresses that the body orientation is typically aligned with the direction of motion. κ_{bb}, κ_{bh}, and κ_{bv} denote concentration parameters. Concentration κ_{bv} depends on the velocity magnitude $\|\dot{\boldsymbol{x}}_t\|$ and on the pedestrian tracker confidence ξ_t (between 0 and 1). We found that this dependence can be well represented by a logistic growth model [39] given parameters $\boldsymbol{\theta}=[\theta_1,\theta_2,\theta_3]$, i.e.,

$$\kappa_{bv}(\dot{\boldsymbol{x}}_t,\xi_t)=\frac{\theta_1\xi_t}{1+\exp\left(-\theta_2\left(\|\dot{\boldsymbol{x}}_t\|-\theta_3\right)\right)} \quad (6)$$

where θ_1 denotes the upper concentration asymptote, θ_2 denotes the growth rate, and θ_3 denotes the velocity magnitude of maximum growth. For a new confirmed pedestrian track, we initialize the orientation tracker by sampling orientation ω_1^B and ω_1^H based on the pedestrians' walking direction given concentrations κ_{bv} and κ_{hb}. Then, new evidence is incorporated by the given measurement $\boldsymbol{z}_1=[\boldsymbol{z}_1^H,\boldsymbol{z}_1^B]$ at this timestep.

D. Observation Model

We assume conditional independence between the head and body observations and obtain two terms, i.e.,

$$p(\boldsymbol{z}_t|\boldsymbol{\omega}_t)=p\left(\boldsymbol{z}_t^H|\omega_t^H\right)p\left(\boldsymbol{z}_t^B|\omega_t^B\right). \quad (7)$$

The superscripts refer again to H for head and B for body. Since both terms are computed in the same way, we will drop this superscript when referring to either term and, likewise, drop the time index t for simplicity, e.g., we write $p(\boldsymbol{z}|\omega)$ when referring to both $p(\boldsymbol{z}_t^H|\omega_t^H)$ and $p(\boldsymbol{z}_t^B|\omega_t^B)$.

1) From Continuous to Discrete Orientations: The orientation $\omega\in\mathbb{R}$ is a continuous value in the domain [0°, 360°), but for the observation likelihoods, we will use the detectors for the discretized orientation classes. We therefore define the likelihood in terms of the class Ω of the current $\boldsymbol{z}$, i.e.,

$$p(\boldsymbol{z}|\omega)=\sum_{\Omega}p(\boldsymbol{z}|\Omega)p(\Omega|\omega). \quad (8)$$

The probability $p(\Omega|\omega)$ expresses the probabilistic relationship between the continuous orientation angle ω and discrete class Ω, which is found by Bayes' rule, i.e.,

$$p(\Omega=o|\omega)=\frac{p(\omega|\Omega=o)p(\Omega=o)}{\sum_{k\in\Omega}p(\omega|\Omega=k)p(\Omega=k)}. \quad (9)$$

Here, $p(\Omega)$ is a prior on the discrete class, and for each class o, $p(\omega|\Omega=o)$ is a von Mises distribution, i.e.,

$$p(\omega|\Omega=o)=\mathcal{V}(\omega;c_o,\kappa_o) \quad (10)$$

where c_o and κ_o are the mean and concentration of the distribution for orientation class o, respectively. We now need to define the term $p(\boldsymbol{z}|\Omega)$, which is the observation likelihood given an orientation class instead of a continuous angle.

2) Likelihood With Auxiliary Variables: We introduce two auxiliary variables, namely, R and V, and express first the likelihood $p(z|\Omega,R,V)$. In Section III-D3, we will then define $p(z|\Omega)$ in terms of this extended likelihood. The indicator variable $R=r, r\in\{1\ldots N\}$ specifies which region $z^{(r)}$ of the possible regions in $\boldsymbol{z}$ fits the sought head/body (and as a consequence, also specifies that all other regions do not fit the head/body). Additionally, the Boolean variable $V=v$, with $v\in\{0,1\}$, indicates whether there exists a head/body in any of the N regions at all $(V=1)$ or whether none of the regions contain it $(V=0)$.

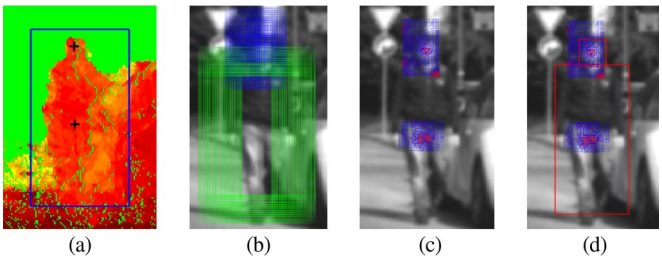

Fig. 3. (a) We use the disparity map to estimate (black crosses) head and body center $(\boldsymbol{h}_c, \boldsymbol{b}_c)$, which serves in (b) as prior information over (blue) R^H and (green) R^B in addition to the detector responses. (c) Region probabilities after including the detector responses. Ambiguities here can be efficiently resolved by combined head and body localization with the PS model. (d) Selected joint maximum.

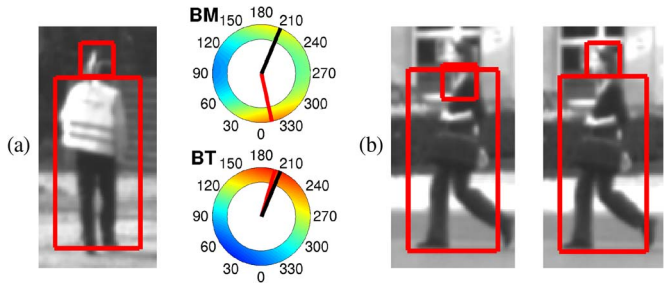

Fig. 4. (a) Accurate localization of (left) head and body can still lead to multimodal likelihood estimations, here shown for the (top right) body (BM). In the tracked (bottom right) body posterior (BT), this ambiguity is resolved. We show the (red line) maximum likelihood/posterior estimate and (black line) GT orientation. (b) Integrating the (right) PS constraint results in a better localization of head and body.

We then express the region likelihood, given the auxiliary variables, in terms of the detector responses as

$$p\left(z^{(s)}|\Omega=o,R=r,V\right)=\begin{cases}f_o\left(z^{(s)}\right) & \text{if } s=r\wedge V=1\\ f_-(z^{(s)}) & \text{otherwise.}\end{cases} \tag{11}$$

Since we assume that the all candidate regions are conditionally independent, the complete data likelihood is only

$$p(\boldsymbol{z}|\Omega,R,V)=\prod_{z^{(s)}\in\boldsymbol{z}}p(z^{(s)}|\Omega,R,V). \tag{12}$$

Intuitively, one would expect that all orientation classes are equally likely when the head/body is not contained in any region and therefore unobserved. This property indeed follows from (11) and (12) since the observations are independent of the orientation Ω and selected region R when $V=0$, i.e.,

$$p(\boldsymbol{z}|\Omega,R,V=0)=p(\boldsymbol{z}|V=0)=\prod_{z^{(s)}\in\boldsymbol{z}}f_-(z^{(s)}). \tag{13}$$

3) Removing the Auxiliary Variables: We first use the region likelihood to select an optimal value $\hat{r}$ for the region indicator R. Assuming that there is a head $(V^H=1)$ and a body $(V^B=1)$ in one of the head and body regions, we select the most probable head and body region configuration $\hat{r}=[\hat{r}^H,\hat{r}^B]$ by

$$\hat{r}=\underset{R^H,R^B}{\operatorname{argmax}}\left[\sum_{\Omega^H}p(\boldsymbol{z}^H|\Omega^H,V^H=1,R^H)p(\Omega^H)\right.$$
$$\left.\times\sum_{\Omega^B}p(\boldsymbol{z}^B|\Omega^B,V^B=1,R^B)p(R^H,R^B|\Omega^B,\boldsymbol{D})p(\Omega^B)\right]. \tag{14}$$

At this region selection step, we utilize several priors (i.e., prior to observing $\boldsymbol{z}$). With $p(R^H,R^B|\Omega^B,\boldsymbol{D})$, which will be described in detail in Section III-E, we introduce prior knowledge about the joint region configuration of head and body from a PS model [36] dependent on the body orientation and disparity data $\boldsymbol{D}$. The fixed Bernoulli distribution $p(V)$ and categorical distribution $p(\Omega)$ can be used to incorporate prior knowledge on the occurrences of false positives (e.g., set high probability for $p(V=0)$ if false positives are common) and orientation classes or just set to uniform to rely on the observation likelihoods only.

The term $p(\boldsymbol{z}|\Omega)$ without auxiliary variables can now be obtained by fixing R to $\hat{r}$ and integrating out the variable V, i.e.,

$$p(\boldsymbol{z}|\Omega)=\sum_{v\in\{0,1\}}p(\boldsymbol{z}|\Omega,V=v,R=\hat{r})p(V=v)$$
$$=p(\boldsymbol{z}|\Omega,V=1,\hat{r})p(V=1)+p(\boldsymbol{z}|V=0)p(V=0). \tag{15}$$

Using (11) and (12) to expand (15) further, we see that the term can be efficiently evaluated up to a constant factor, i.e.,

$$p(\boldsymbol{z}|\Omega=o)\propto f_o(z^{(\hat{r})})p(V=1)+f_-(z^{(\hat{r})})p(V=0). \tag{16}$$

It follows that the same is true for $p(\boldsymbol{z}|\omega)$ in (8). This constant can be ignored, since it does not affect the posterior distribution of (2) after normalization.

We also see from (16) that the stronger the background detector response f_- is (relative to the orientation detectors f_o), the higher the weight of the second term, and therefore the smaller the relative differences between the likelihoods of the different orientation classes. This means that, in the extreme case where only f_- gives a strong response, the likelihood term is the same for all orientations. The posterior of (2) would then reduce to just the prior distribution from the prediction step, i.e., no information on the true orientation was gained at this time step.

E. Spatial Prior Over Body Part Regions

Let $\boldsymbol{h}_c(\boldsymbol{D})$ and $\boldsymbol{b}_c(\boldsymbol{D})$ be functions on the disparity $\boldsymbol{D}$ that give us an estimate of head and body positions. We factor the prior from (14) into

$$p(R^H,R^B|\Omega^B,\boldsymbol{D})\propto$$
$$p\left(\boldsymbol{h}_c(\boldsymbol{D})|R^H\right)p(\boldsymbol{b}_c(\boldsymbol{D})|R^B)p(R^H,R^B|\Omega^B) \tag{17}$$

a) Disparity-based region priors: $\boldsymbol{h}_c$ and $\boldsymbol{b}_c$ return the mean pixel location of head and body based on disparity values $\tilde{\boldsymbol{D}}$ in the range $\boldsymbol{D}<\tilde{d}-\epsilon$ and $\boldsymbol{D}>\tilde{d}+\epsilon$. Here, $\tilde{d}$ denotes the

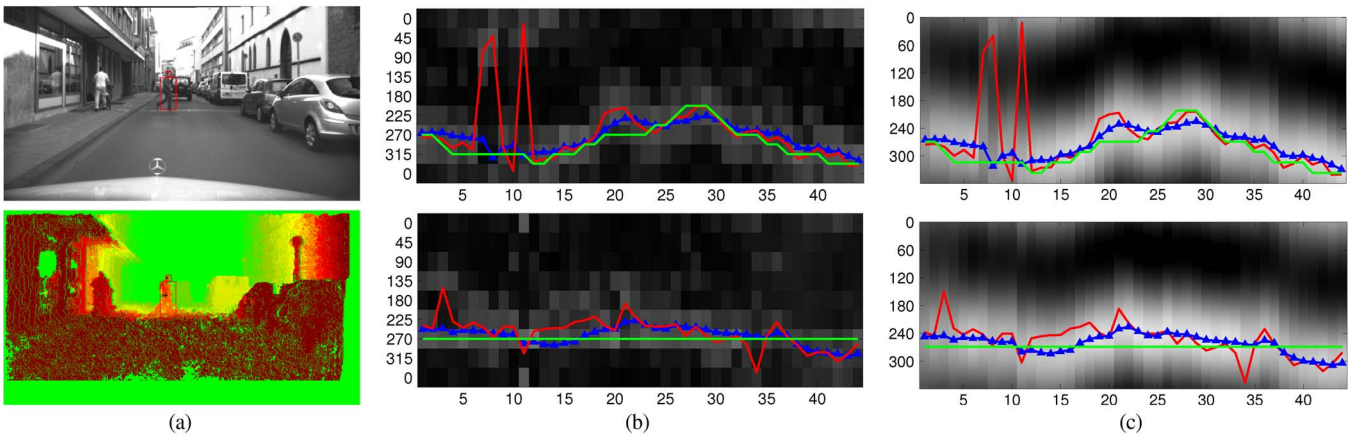

Fig. 5. Sample tracks. (a) Rectified (top) gray value image and (bottom) disparity image. The red boxes show the selected head and body region. (b) Orientation-specific output of (top) head detectors and (bottom) body detectors over the track time (frames). Brighter values indicate higher detector confidences. (c) Continuous posterior distributions estimated by joint tracking of (top) head and (bottom) body over the track time (frames). In (b) and (c), we show (green line) labeled GT orientation, (red line) single-frame estimation with PS, and (blue line) joint tracking result.

median value over all disparity values calculated over the given pedestrian track bounding box. The learned parameter $\epsilon = 1.5$ accounts for disparity estimation errors. We use semiglobal matching [40] to calculate the disparity map $\boldsymbol{D}$. The likelihood of the head region is then modeled with

$$p(\boldsymbol{h}_c(\boldsymbol{D})|R^H = r^H) = \mathcal{N}\left(\boldsymbol{h}_c(\boldsymbol{D}); \boldsymbol{\mu}(r^H), \boldsymbol{C}^H\right). \quad (18)$$

$\boldsymbol{\mu}(r^H)$ denotes the center (u- and v- coordinate) of a given head region r^H in image coordinates, and $\boldsymbol{C}^H$ denotes the corresponding covariance. The likelihood $p(\boldsymbol{b}_c(\boldsymbol{D})|R^B = r^B)$ of the body region is modeled similarly.

b) Joint region prior: To model the joint spatial prior $p(R^H, R^B|\Omega^B)$, we use a PS model [36], which is dependent on the body orientation

$$p(R^H = r^H, R^B = r^B|\Omega^B = o^B) = \mathcal{N}\left(\boldsymbol{l}^D(r^H, r^B); \boldsymbol{\mu}^D_{o^B}, \boldsymbol{C}^D_{o^B}\right). \quad (19)$$

$\boldsymbol{l}^D(r^H, r^B)$ denotes the difference of head and body region center divided by the width of the body region.

c) Region generation: Due to efficiency reasons, we generate possible head and body regions based on the tracked pedestrian height (h_t) and the estimated horizontal gravity line (g^x_c) inside the tracked bounding box. To calculate g^x_c, we find the dominant peak in the histogram of the horizontally projected image locations of all disparity values $\tilde{\boldsymbol{D}}$. Since the tracked bounding box should already give an appropriate estimation of the body location, the number of used body hypotheses can be much less than it is for the head. The size of the generated head and body regions is set according to the estimated pedestrian height (h_t). The step size between regions is set dependent on the pedestrian distance. Fig. 3 shows an example of the region generation process and the calculated region probabilities based on detector responses and the described prior information.

IV. Experiments

A. Setup

1) Data Sets: For our training set, we extract head and body bounding boxes from 9300 manually labeled pedestrian samples from 6389 images. Pedestrian samples have a minimum/maximum/mean height of 69/344/122 pixels and are combined with background samples to train our detectors. Half of the background samples were sampled from false positive pedestrian detections in the area of the sought head/body. The other half was sampled around the head/body of a true positive pedestrian detection with a maximum overlap of 25% to the true head/body bounding box. No pedestrian occurring in the training set also occurs in the test set.

Our validation/test set consists of 32/60 image sequences. They contain pedestrians against various traffic backdrops in mostly benevolent illumination conditions (i.e., no strong backlighting) and with limited occlusion (up to 20% of pedestrian). Pedestrians predominantly cross laterally or walk longitudinally with respect to the vehicle. The vehicle mostly drives straight (i.e., no vehicle turns at intersection) at speeds up to 37 km/h. Ground truth (GT) was obtained by manual labeling of the orientation and location of the body part bounding box. Input to our framework were the bounding boxes of a state-of-the-art HOG/linSVM pedestrian detector [5] and Kalman filter, implemented by the authors (cf., the shaded modules in Fig. 1). In each frame, a pedestrian location estimate is associated with a GT label when the distance between them is smaller than a threshold. This threshold is set according to a percentage of the Euclidean distance of the GT label to the camera. We select a different percentage of 8% and 12% for lateral and longitudinal directions, since uncertainty in lateral direction is, in general, smaller.

For our evaluation of the orientation estimation performance, we only consider estimated tracks in the test set that follow more than 80% of their duration a particular GT track (several estimated tracks can correspond to a single GT track). All other estimated tracks are regarded as false positives and were set aside.

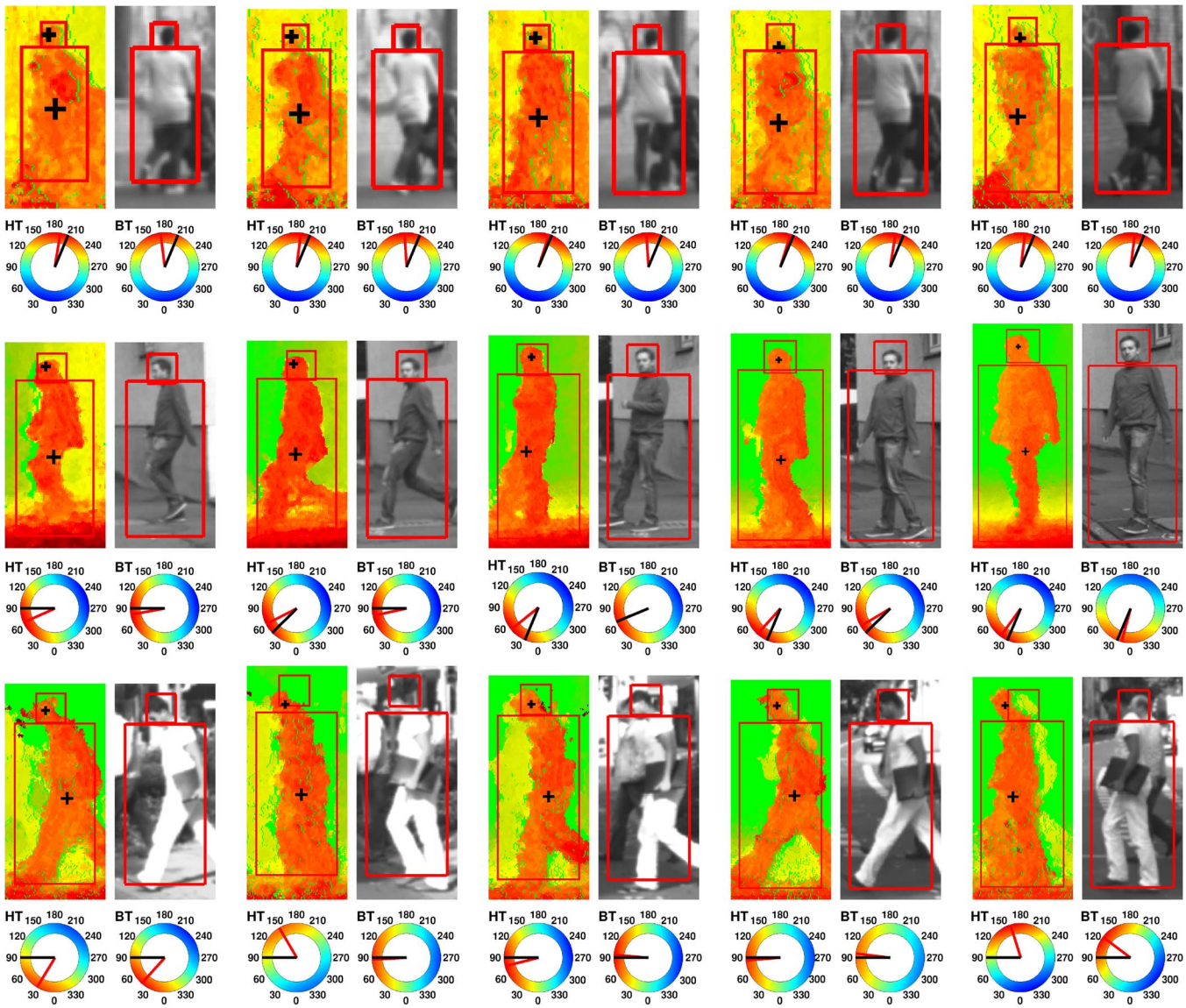

Fig. 6. (Left) Disparity and (right) gray image of every sixth frame of three estimated tracks. The red boxes show again the selected head and body region. Below the images, we show the posterior distributions of our approach for the head (HT) and body orientation (BT), (red line) maximum posterior estimate, and (black line) GT orientation. Black crosses in disparity images denote estimated head and body center. The last track (last row) shows bad estimation results due to inaccurate localization of head and body.

Furthermore, we only include track samples with a maximum lateral/longitudinal distance of 5 m/35 m to the camera. Doing so, we obtain 65 "valid" estimated tracks on the test set (i.e., five GT tracks were split by our pedestrian tracker) with 3167 samples (0.02 false positive track samples per image were set aside.)

2) Detectors: We train eight orientation-specific detectors $f_o(z)$ with class centers $o \in \{0°, 45°, \cdots, 315°\}$ in a modified one-versus-all manner, including the background class. Our originally labeled 16 orientation bins were reduced to 8 GT orientation bins, by merging three neighboring labeled orientation bins (i.e., labeled bins 337.5°, 0°, and 22.5° are merged to GT bin 0°, labeled bins 22.5°, 45°, and 67.5° are merged to GT bin 45°, etc.). Preliminary experiments showed that this gives a better performance when combining the detectors to estimate a continuous orientation.

For the background detector $f_{-}(z)$, we use all background samples versus all orientation specific samples. For all detectors, we use multilayer NN architectures (NN/LRF) [41] with a 5×5 LRF. We extracted the head by a fixed aspect ratio of 15% of the whole body centered below the highest labeled pedestrian contour point (contour labels were already available to us from an earlier pedestrian segmentation study). For the body detectors, we use the lower 85% part of the pedestrian bounding box. All head samples were scaled to 16×16 pixels for training and testing, whereas the body samples were scaled to 18×36 pixels. A border of 2/4 pixels was added to head/body samples to avoid border effects. We also generate eight additional samples from each original sample by shifting the corresponding bounding box by 1–2 pixels.

3) Parameter Setting: Since we use discrete GT orientation steps, we manually annotated time slots where there is a

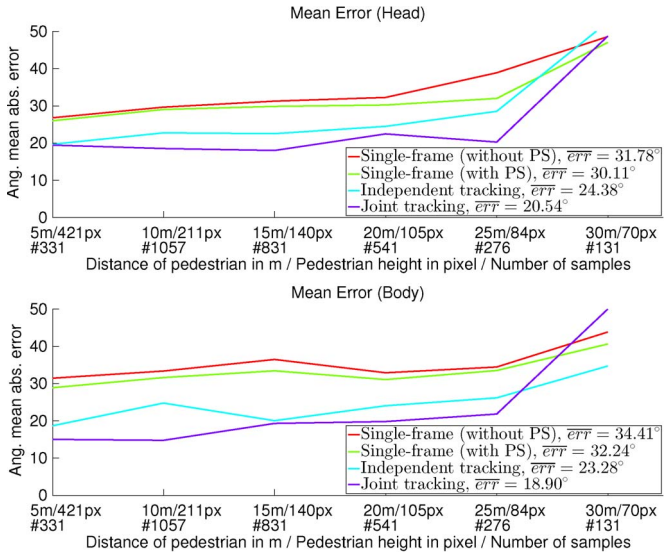

Fig. 7. Angular mean absolute error over increasing distance for (top) head and (bottom) body orientation estimation, using (purple) joint tracking, (cyan) independent tracking, and single-frame orientation estimation (green) with and (red) without PS. Legend is also showing mean error over all samples ($\overline{err}$).

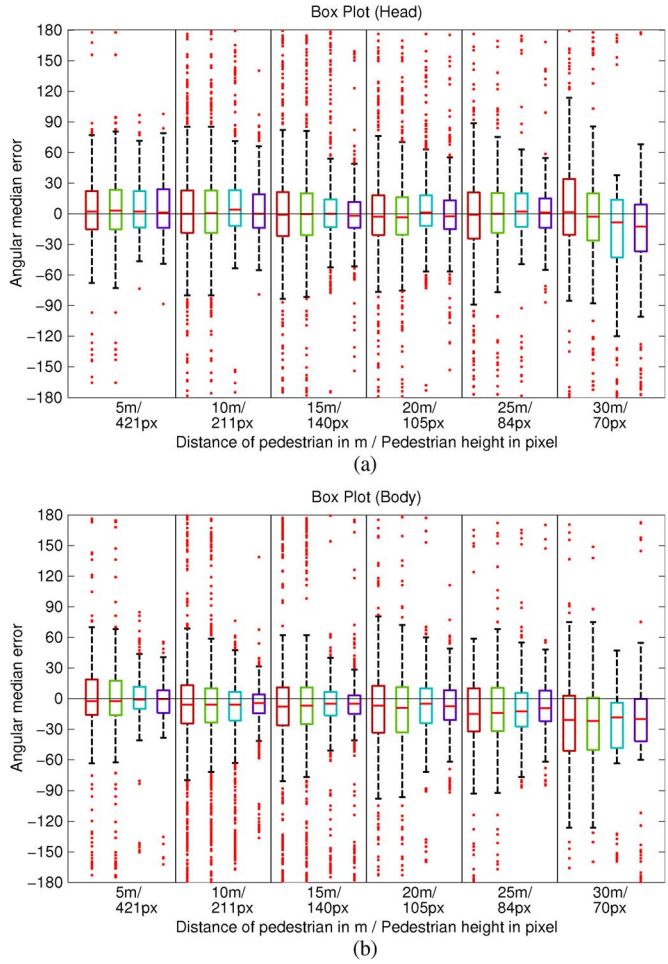

Fig. 8. Boxplots showing (red line) median error and (red crosses) outliers for (a) head and (b) body orientation estimation in case of single-frame estimation (red box) without PS and (green box) with PS, (cyan box) independent tracking, and (purple box) joint tracking. Boxes contain 50% of samples. Used whiskers define 99.3% data coverage. By joint tracking, we get a more robust estimation compared with single-frame estimations and independent tracking.

head or body movement in the validation set. In these time slots, we estimated temporal constraints for head [κ_{hh}, (4)] and body [κ_{bb}, (5)] from differences of GT orientations between adjacent time steps. Anatomical constraints $(\kappa_{hb}, \kappa_{bh})$ are estimated on displacement cases (difference between head and body GT orientation greater than 45°) on the complete validation set. For a set of pedestrian velocities $k \in \boldsymbol{V}$ (where $k = \{0, 0.1, \cdots 2\}(m/s)$), we estimate concentration κ_{bv}^{k} on differences between inferred velocity direction and GT body direction on the complete validation set. Based on these concentration/velocity tuples, we estimate parameters $\boldsymbol{\theta}$ of the logistic growth model [see (6)] by keeping $\xi_t = 1$ constant. Mixture weights $(\alpha_{hh}, \alpha_{bb}, \alpha_{hb})$ are tuned on the validation set. Concentration parameters [κ_{o^H}, κ_{o^B}, (10)] are estimated on the training set using the 16 GT orientations. We consider the interval between the canonical orientations of left and right neighbored detectors as the domain where we estimate these detector concentrations. More specifically, we consider the distribution of the average belief of a detector when presented with samples with GT orientation from this domain. Disparity-based region prior parameters for head/body [$\boldsymbol{C}^H, \boldsymbol{C}^B$, (18)] are estimated on image distances between estimated and GT head/body centers, normalized by GT pedestrian height. Joint region prior parameters [$\boldsymbol{\mu}^D, \boldsymbol{C}^D$, (19)] are estimated on image distances between head and body GT bounding boxes, normalized by GT pedestrian height.

Priors $p(\Omega)$ and $p(V)$ are modeled with uniform distributions for head and body orientation.

B. Results

We show in Fig. 4(a) a sample that gives a multimodal likelihood estimate, caused by confusing opposite directions. This ambiguity can be successfully corrected by our joint tracking approach. The effect of integrating the PS spatial constraint is shown in Fig. 4(b), where localization of head and body is improved with the spatial PS constraint (right image).

In Fig. 5, we show detector outputs and filtered orientation distributions. The tracker is successfully smoothing over outliers and is able to react also to small changes in the orientation. In Fig. 6, we show disparity and gray image of every sixth frame of three estimated tracks with continuous estimation results of our proposed approach. As can be seen, the joint tracking delivers good localization and a robust continuous orientation estimate of head and body. Even in cases with limited stereo support for the head (e.g., first row, fourth and fifth images), the head is localized correctly due to the detector outputs. In the last track, we show various problem cases of our current approach causing a wrong localization of head and body and therefore a bad orientation estimation. Reasons for that are stronger deviations of mean head position and rotation (second image), pedestrian groups (fifth image), or contrast and lighting issues.

We perform a quantitative evaluation on the complete test set using all 65 valid estimated tracks.

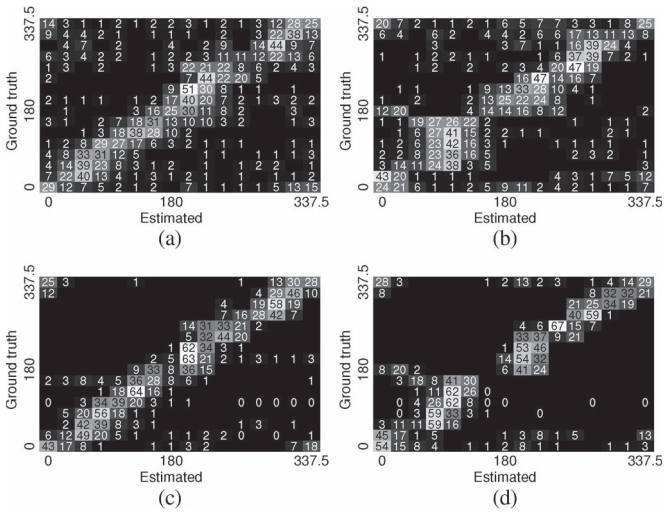

Fig. 9. Confusion matrices for the single-frame case without PS for (a) head and (b) body orientation, respectively. Idem for joint probabilistic tracking, for (c) head and (d) body orientation, respectively. Occurrences with brighter colors are more frequent. Numbers define the occurrences in percent.

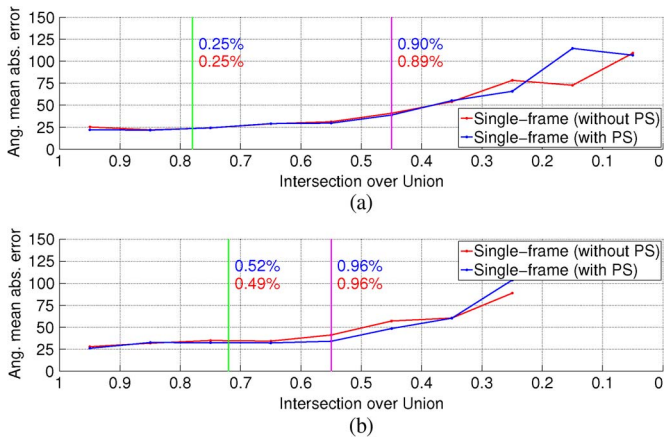

Fig. 10. Angular mean absolute error over decreasing image localization accuracy (IoU) in intervals of 0.1 for (a) head and (b) body. We show the (green line) theoretical maximum localization error compared with the (magenta line) true (observed) maximum localization error. Percentages at the borders indicate the amount of test data having smaller localization error for single-frame measurements (red) without and (blue) with PS.

In Fig. 7, we show the angular mean absolute error for head and body orientation estimation with increasing distance. The averaged error over all distances is shown in the legend. We compare our proposed joint tracking to the results of independent tracking and single-frame orientation estimation with and without PS [see (19)]. Independent tracking refers to tracking of head and body without an orientation coupling, as defined in Section III-C. For both independent and joint tracking, we use the spatial PS constraint. We see that the mean error can be significantly reduced by tracking. Joint tracking decreases the head/body orientation error over all samples by 11°/15° compared with single-frame estimation without PS. This benefit is mainly caused by the removal of outliers compared with single-frame estimation (e.g., confusion between opposite body directions, which visually can look very similar). Furthermore, in comparison with independent tracking, we decrease the error by 4°/4° for head/body orientation. Anatomical and movement constraints within tracking as defined in Section III-C help here to reject impossible configurations between head and body orientation.

By evaluating only displacement cases (difference between head and body orientation greater than 45°), the angular mean error turns out to increase by 7°/1° for head/body in the joint tracking case. Above 25 m, the mean error and the outlier rate increase significantly due to weak detector responses at initial tracking states. Separate experiments (not included here) indicated that a further improvement of the pedestrian track quality (up to the use of actual GT tracks) did not lead to an appreciable improvement of orientation estimation. We take this as a strength of our stereo-based head and body localization procedure, which can already handle the displaced pedestrian bounding boxes at this quality level.

In Fig. 8, we show an additional boxplot to get a better impression of the orientation error distribution. It can be seen that adding more constraints reduces the uncertainty and outliers.

Fig. 9 shows confusion matrices for the single-frame case without PS and for the joint tracking case, for head and body orientation, respectively. Adding the spatial and orientation constraints between the body parts and the temporal filtering results in a clear more diagonal structure of the confusion matrix [i.e., comparing Fig. 9(a) with Fig. 9(c) and Fig. 9(b) with Fig. 9(d)]. Some heightened confusion can still be observed in Fig. 9(d) for the body around 0° and 180°.

Fig. 10 shows how the angular mean absolute error is affected by image localization performance on our test set. We compute the intersection over union (IoU) measure between GT and estimated bounding box such that a value of 1 corresponds to a perfect overlap and a value of 0 corresponds to no overlap. In Fig. 10(a), we show that 90% (89% without PS) of the head samples have a localization performance better than 0.45 (magenta line), while still getting acceptable orientation estimates. The green line shows the computed localization performance threshold with regard to a possible shift of 1 pixel for a 16×16 image in each direction, as done in the training to increase the amount of training samples. In practice, we can accept a lower localization performance (as showed by the magenta line). There are only a few samples with an IoU measure less than 0.2, resulting in the mean absolute angular error to be noisy. Fig. 10(b) shows the same for the body. Since relying only on the measurements with regard to head and body locations is suboptimal, in future work, we will filter these over time.

Our implementation, running on a 3.33-GHz i7 central processing unit processor, needs on average less than 120 ms per image (this is down from 1 s per image in [37] due to multithreading within each module). Fig. 11 shows how this time is distributed among the different components used in the system. Starting with a state-of-the-art HOG/linSVM pedestrian detector (Ped. Detection), we perform a 3-D world estimation using stereo vision (Ped. 3-D Processing) afterward. The pedestrian is then tracked by a standard Kalman filter (Ped. 3-D Tracking).

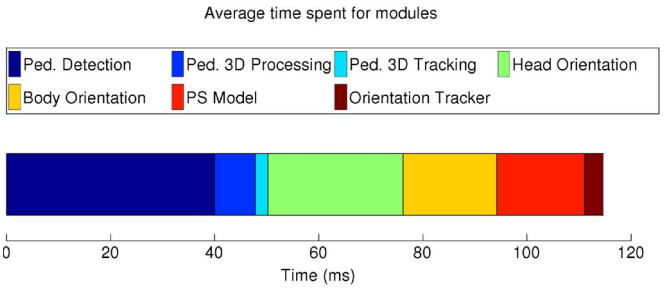

Fig. 11. Different modules and their running time. Altogether, the modules need on average approximately 120 ms per image.

As described in this paper, orientation configuration of head and body is estimated (Head Orientation/Body Orientation) using the PS model and tracked by a PF (Orientation Tracker) using various constraints. Note that the head orientation estimation needs more time, since more region hypotheses are generated on average, than for the body orientation estimation. In the near future, we expect an additional time benefit by using feature sharing between the detectors and by a further parallelization of modules.

V. Discussion

Given input pedestrian tracks of some decent quality (cf., Section IV-A1), our proposed approach results in a mean absolute orientation error on pedestrian head and body orientation, which is fairly constant up to a distance of 25 m, namely, about 21° and 19°, respectively (cf., joint tracking in Fig. 7). As this is the first work on pedestrian head and body orientation estimation from a mobile stereo vision platform to our knowledge, we cannot make direct comparisons with other work. We note that Chen and Odobez listed pedestrian head/body orientation errors of 36.0°/35.6°, 30.0°/29.4°, 23.6°/23.6°, 18.4°/17.4° in their four surveillance scenarios with static monocular camera (cf., [12], Table 1).

In our experiments, we used a 0.75-megapixel camera, which is already integrated into production vehicles on the market. In the foreseeable future, when the image resolution is doubled, the distance range for which we obtain the current performance can be extended from 25 to 35 m, without any algorithm modification. This is a sensible range for a practical application, considering typical urban vehicle speeds and prediction horizons that are not likely to exceed 2 s, given the high pedestrian maneuverability.

Apart from using higher resolution images, we expect performance benefits from a larger training set and from a more accurate head and body orientation localization method, i.e., accurate pedestrian segmentation in images combining prior model knowledge (shape, texture) with data-driven cues (e.g., [42]).

Very recently, we have integrated the currently proposed head and body orientation estimation approach into a system for context-based pedestrian prediction [43]. Preliminary vehicle experiments suggest that we can expedite driver warning and vehicle braking significantly using these body pose cues, without introducing false alarms.

VI. Conclusion

We have presented a probabilistic framework for the joint estimation of pedestrian head and body orientation in the context of stereo-vision-based active pedestrian safety. The framework involved a principled way to deal with faulty part detections, continuous orientation estimation, coupling of the body and head localization and orientation, and temporal integration.

Experiments showed that the proposed joint tracking of head and body orientations decreases the mean absolute head/body orientation error by 11°/15° compared with single-frame estimation and further by 4°/4° compared with independent tracking. In absolute terms, this comes down to mean absolute head/body orientation error of about 21°/19°, which remains fairly constant up to a distance of 25 m. Further work involves improving the underlying pedestrian tracker (vehicle turning, pedestrian groups), obtaining a more accurate segmentation, and dealing specifically with occlusions and adverse lighting conditions. Based on the current results, we feel confident that pedestrian head/body orientation estimation will play an important role in the next-generation intelligent driver warning and vehicle control strategies.

References

[1] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pp. 2179–2195, Dec. 2009.

[2] S. Schmidt and B. Färber, "Pedestrians at the kerb–Recognising the action intentions of humans," *Transp. Res. Part F, Traffic Psychol. Behaviour*, vol. 12, no. 4, pp. 300–310, Jul. 2009.

[3] 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 *Proc. IEEE Intell. Veh. Symp.*, 2013, pp. 106–110.

[4] V. Evers *et al.*, "The development and real-world deployment of FROG, the fun robotic outdoor guide," in *Proc. ACM/IEEE Human-Robot Interaction*, 2014, pp. 100–100.

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

[6] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 607–626, Apr. 2009.

[7] E. Hjelmås and B. K. Low, "Face detection: A survey," *Comput. Vis. Image Understanding*, vol. 83, no. 3, pp. 236–274, Sep. 2001.

[8] S. O. Ba and J.-M. Odobez, "A Rao-Blackwellized mixed state particle filter for head pose tracking," in *Proc. ACM-ICMI Workshop MMMP*, 2005, pp. 9–16.

[9] G. Fanelli, J. Gall, and L.Van Gool, "Real time head pose estimation with random regression forests," in *Proc. IEEE Conf. CVPR*, 2011, pp. 617–624.

[10] M. J. Marín-Jiménez, A. Zisserman, and V. Ferrari, ""Here's looking at you, kid". Detecting people looking at each other in videos." in *Proc. BMVC*, 2011, pp. 1–12.

[11] B. Benfold and I. Reid, "Unsupervised learning of a scene-specific coarse gaze estimator," in *Proc. ICCV*, 2011, pp. 2344–2351.

[12] C. Chen and J. Odobez, "We are not contortionists: Coupled adaptive learning for head and body orientation estimation in surveillance video," in *Proc. IEEE CVPR*, 2012, pp. 1544–1551.

[13] J. Orozco, S. Gong, and T. Xiang, "Head pose classification in crowded scenes." in *Proc. BMVC*, 2009, pp. 120.1–120.11.

[14] N. Robertson and I. Reid, " *Estimating Gaze Direction From Lowresolution Faces in Video*," in Lecture Notes in Computer Science. New York, NY, USA: Springer-Verlag, 2006, vol. 3952, pp. 402–415.

[15] K. Smith, S. O. Ba, J.-M. Odobez, and D. Gatica-Perez, "Tracking the visual focus of attention for a varying number of wandering people," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 7, pp. 1212–1229, Jul. 2008.

[16] M. C. Liem and D. M. Gavrila, "Coupled person orientation estimation and appearance modeling using spherical harmonics," *Image Vis. Comput.*, vol. 32, 10, pp. 728–738, Oct. 2014.
[17] A. Schulz, N. Damer, M. Fischer, and R. Stiefelhagen, "Combined head localization and head pose estimation for video-based advanced driver assistance systems," in *Proc. DAGM Symp. Pattern Recog.*, 2011, pp. 51–60.
[18] M. Enzweiler and D. M. Gavrila, "Integrated pedestrian classification and orientation estimation," in *Proc. IEEE Conf. CVPR*, 2010, pp. 982–989.
[19] B. Benfold and I. Reid, "Guiding visual surveillance by tracking human attention." in *Proc. BMVC*, 2009, pp. 1–11.
[20] T. Gandhi and M. M. Trivedi, "Image based estimation of pedestrian orientation for improving path prediction," in *Proc. IEEE Intell. Veh. Symp.*, 2008, pp. 506–511.
[21] H. Shimizu and T. Poggio, "Direction estimation of pedestrian from multiple still images," in *Proc. IEEE Intell. Veh. Symp.*, 2004, pp. 596–600.
[22] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *Proc. CVPR*, 2011, pp. 3457–3464.
[23] T. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active shape models–their training and application," *Comput. Vis. Image Understanding*, vol. 61, no. 1, pp. 38–59, Jan. 1995.
[24] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.
[25] J. Giebel and D. M. Gavrila, "Multimodal shape tracking with point distribution models," *Pattern Recog.*, vol. 2449, pp. 1–8, 2002.
[26] K.-C. Lee and D. Kriegman, "Online learning of probabilistic appearance manifolds for video-based recognition and tracking," in *Proc. IEEE CVPR*, vol. 1, 2005, pp. 852–859.
[27] P. Hall, D. Marshall, and R. Martin, "Merging and splitting eigenspace models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 9, pp. 1042–1049, Sep. 2000.
[28] P. M. Hall, A. D. Marshall, and R. R. Martin, "Incremental eigenanalysis for classification." in *Proc. BMVC*, vol. 98, 1998, pp. 286–295.
[29] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Conf. CVPR*, 2012, pp. 2879–2886.
[30] D. M. Gavrila and S. Munder, "Multi-cue pedestrian detection and tracking from a moving vehicle," *Int. J. Comput. Vis.*, vol. 73, no. 1, pp. 41–59, Jun. 2007.
[31] C. Chen, A. Heili, and J.-M. Odobez, "A joint estimation of head and body orientation cues in surveillance video," in *Proc. ICCV Workshops*, 2011, pp. 860–867.
[32] G. Zhao, M. Takafumi, K. Shoji, and M. Kenji, "Video based estimation of pedestrian walking direction for pedestrian protection system," *J. Electron. (China)*, vol. 29, no. 1/2, pp. 72–81, Mar. 2012.
[33] S. O. Ba and J.-M. Odobez, "Probabilistic head pose tracking evaluation in single and multiple camera setups," in *Proc. Workshop Classification Events, Activities Relationships (Multimodal Technologies for Perception of Humans)*. Springer, 2008, pp. 276–286.
[34] D. M. Gavrila, "The visual analysis of human movement: A survey," *Comput. Vis. Image Understanding*, vol. 73, no. 1, pp. 82–98, 1999.
[35] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Comput. Vis. Image Understanding*, vol. 104, no. 2, pp. 90–126, 2006.
[36] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *Int. J. Comput. Vis.*, vol. 61, no. 1, pp. 55–79, 2005.
[37] F. Flohr, M. Dumitru-Guzu, J. F. P. Kooij, and D. M. Gavrila, "Joint probabilistic pedestrian head and body orientation estimation," in *Proc. IEEE Intell. Veh. Symp.*, 2014, pp. 617–622.
[38] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *J. Statist. Comput.*, vol. 10, no. 3, pp. 197–208, 2000.
[39] G. A. F. Seber and C. J. Wild, *Nonlinear Regression*. Hoboken, NJ, USA: Wiley, 1989.
[40] 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.
[41] C. Wöhler and J. K. Anlauf, "A time delay neural network algorithm for estimating image-pattern shape and motion," *Image Vis. Comput.*, vol. 17, no. 3, pp. 281–294, Mar. 1999.
[42] F. Flohr and D. M. Gavrila, "PedCut: An iterative framework for pedestrian segmentation combining shape models and multiple data cues," in *Proc. BMVC*, 2013, pp. 66.1–66.11.
[43] J. F. P. Kooij, N. Schneider, F. Flohr, and D. M. Gavrila, "Context-based pedestrian path prediction," in *Proc. Comput. Vis.*, 2014, vol. 8694, in Lecture Notes in Computer Science, pp. 618–633, Springer-Verlag.

Fabian Flohr received the M.Sc. degree in computer science from Karlsruhe Institute of Technology, Karlsruhe, Germany, in 2012. He is currently working toward the Ph.D. degree at University of Amsterdam, Amsterdam, The Netherlands.

He is also currently with Daimler Research and Development, Ulm, Germany. His research interests include machine learning and video analysis for intelligent vehicles, with a focus on pedestrian detection, segmentation, and tracking. His personal web site is www.fabian-flohr.de.

Madalin Dumitru-Guzu received the B.Sc. degree in computer science and information technology from University Politehnica of Bucharest, Bucharest, Romania, in 2012 and the M.Sc. degree from Delft University of Technology, Delft, The Netherlands, in 2014. His M.Sc. thesis research was performed while being employed at Daimler Research and Development, Ulm, Germany.

He is currently with Fotonation, Bucharest. His research interests include computer vision and multimedia signal processing.

Julian F. P. Kooij received the M.Sc. degree in artificial intelligence in 2008 from University of Amsterdam, Amsterdam, The Netherlands, where he is currently working toward the Ph.D. degree.

He is also currently with Daimler Research and Development, Ulm, Germany. His research interests include machine learning and Bayesian data analysis, with a focus on modeling trajectory dynamics and unsupervised discovery of motion patterns.

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

Since 1997 he has been with Daimler Research and Development, Ulm, Germany, where he is currently a Principal Scientist. He has led the multiyear pedestrian detection research effort at Daimler, which was incorporated in the Mercedes-Benz S-, E-, and C-Class models (2013–2014). In 2003 he also became a part-time Professor with 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.

Prof. Gavrila received the I/O Award 2007 from the Dutch Science Foundation (NWO) and the Outstanding Application Award 2014 from the IEEE Intelligent Transportation Systems Society. His personal Web site is www.gavrila.net.