2014 IEEE Intelligent Vehicles Symposium (IV)
June 8-11, 2014. Dearborn, Michigan, USA

Joint probabilistic pedestrian head and body orientation estimation

Fabian Flohr[1,2], Madalin Dumitru-Guzu[1,3], Julian F. P. Kooij[1,2] and Dariu M. Gavrila[1,2]

***Abstract*— We present an approach for the joint probabilistic estimation of pedestrian head and body orientation in the context of intelligent vehicles. For both, head and body, we convert the output of a set of orientation-specific detectors into a full (continuous) probability density function. The parts are localized with a pictorial structure approach which balances part-based detector output with spatial constraints. Head and body orientation estimates are furthermore coupled probabilistically to account for anatomical constraints. Finally, the coupled single-frame orientation estimates are integrated over time by particle filtering.**

The experiments involve 37 pedestrian tracks obtained from an external stereo vision-based pedestrian detector in realistic traffic settings. We show that the proposed joint probabilistic orientation estimation approach reduces the mean head and body orientation error by 10 degrees and more.

I. INTRODUCTION

Great strides have been made over the last few years in the area of video-based pedestrian detection, leading to the first commercial active pedestrian systems reaching the market (e.g. 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 manouverability; pedestrian can change their walking direction or accelerate/decelerate at a whim. Any auxiliary information that can help to reduce this uncertainty is welcome. The pedestrian body and head orientation is a relevant indicator as to what the pedestrian will do next. In this paper, we address the issue of estimating these orientation quantities using a stereo sensor-setup that is already available on the market. In particular, we describe a method to robustly track the orientation of both head and body, given the output of an external pedestrian tracker (which is outside the scope of this paper). We present a principled probabilistic approach for dealing with faulty detections, continuous orientation estimation, coupling of the body- and head-localization and orientation, and temporal integration. By estimating various parameters of our model from real data, we assure that various anatomical and dynamical pedestrian characteristics are accounted for. Fig. 1 shows an overview of our proposed approach.

II. RELATED WORK

In this section, we focus on techniques for person head and body orientation estimation. For vision-based pedestrian detection, see recent surveys (e.g. [1]).

Approaches for head orientation estimation are largely application dependent (see survey [2]). Applications in Human-Machine interaction (HMI) [3, 4] or entertainment [5] typically consider high resolution images and cooperative subjects. Applications in surveillance [6, 7, 8, 9, 10] and in intelligent vehicle [11] domains need to cope with low resolution images, with complex and dynamic backgrounds, and changing lighting conditions. To cope with these challenges, most approaches use robust lower-level features like SIFT/HOG [12, 6, 7, 13, 14], Haar-like features [11, 15], local receptive field features [14] or distance metrics [12, 4, 8] in combination with different classification schemes (e.g. SVMs [7, 13, 8, 15, 14], NNs [14], Random Regression/Decision Trees or Ferns [12, 6, 4] or Boosting cascades [11]) to perform head/body orientation estimation. In particular, [11] trained a boosting cascade of Haar-like features for eight head orientation classes in an one-versus-all manner. The maximum over all possible hypotheses and the eight discrete orientation classes was selected. [12] used a random fern architecture with a combination of HOG and color based features to infer head orientation. Training was done with eight discrete head orientation classes. While most of the above mentioned methods used manually labeled training data, [6] learned head orientations unsupervised by using the output of a tracking system, supposing that head orientation is dependent on walking direction.

Body orientation estimation can exploit prior knowledge about body shape by matching shape models [16, 17]. Another idea is to use the walking direction as a proxy for body orientation (e.g. [9]), thus assuming that people only move forward. There has also been extensive work done on articulated 3D pose recovery [4, 18]. These typically require

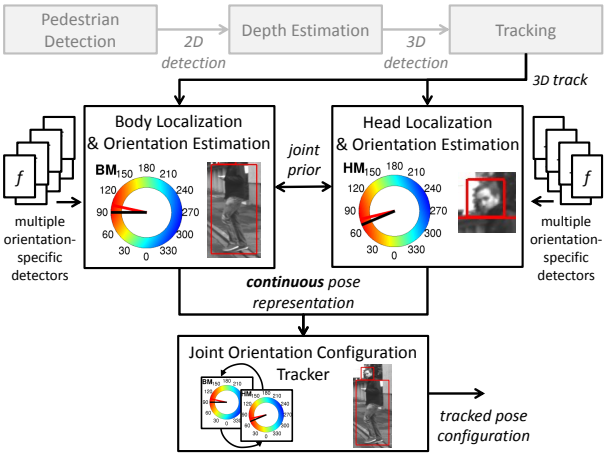

Fig. 1. Proposed joint probabilistic orientation estimation approach

[1]Environment Perception Department, Daimler R&D, Ulm, Germany
[2]Intelligent Systems Laboratory, Univ. of Amsterdam, The Netherlands
[3]Computer Vision Laboratory, Delft Univ. of Tech., The Netherlands

multiple cameras and have issues with robustness.

To improve orientation estimations, [6, 19, 7, 9, 20, 10] introduced constraints which set head and body orientation into relation of each other. In [7], such constraints were directly used during classifier training, while [20] used body orientation information to differentiate online between opposite head directions. [10] constrained the head location with respect to the body location to obtain physically possible configuration. [6] applied a Conditional Random Field (CRF) for modeling the interaction between the head orientation, walking direction and appearance to recover gaze direction. While [9] constrained the head orientation on the velocity direction, [19] introduced a coupling between body orientation and velocity direction. By tracking, single frame orientation estimations can be smoothed and results can be further improved [3, 21, 22, 13, 15, 10, 23].

Our paper contribution is a principled joint probabilistic head and body orientation estimation approach that handles faulty detections, continuous orientation estimation, coupling of the body- and head-localization and orientation, and temporal integration. We differentiate to [9, 19, 7] in several ways. We consider an intelligent vehicle context. We use stereo vision for localization. We constrain the body orientation also on the previous head orientation (modeling the cases where the pedestrian changes the body orientation in response to what he is observing). Finally, we take into account (external) tracker confidence.

III. JOINT HEAD AND BODY ORIENTATION ESTIMATION

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 $\boldsymbol{x}_t = [x_t, y_t]$, defined in world coordinates on the ground plane, and velocity $\dot{\boldsymbol{x}}_t = [\dot{x}_t, \dot{y}_t]$. The full body tracks are provided as input to our orientation tracker (see Fig. 1), 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 $\boldsymbol{x}$, $\dot{\boldsymbol{x}}$ 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.

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 head or 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 head/body in a certain orientation. The angular domain of $[0°, 360°)$ is discretized into a fixed set of orientation classes, e.g. centered around angles of $0, 45, \cdots, 315$ degrees. Each class then has a detector, e.g. $f_0, f_{45}, \cdots, f_{315}$, for both head and body, such that the detector response $f_o(z)$ is strength for the evidence that image region z contains the head/body 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. An additional non-target or background classifier $f_-(z)$ assigns a likelihood to the case that z does not contain the head/body.

A. 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, $p(\boldsymbol{\omega}_t|\boldsymbol{z}_{1:t}, \dot{\boldsymbol{x}}_{1:t})$, 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,

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

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 the Section III-B.

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

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

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

In our implementation, we use a Particle Filter (PF) [24] to approximate the posterior distribution with a set of particles in the state space. The PF allows us to use a multi-modal dynamic model, and we only need to evaluate a function proportional to the probability density of the observation model. For a new pedestrian track, we initialize a filter by sampling orientations ω_1^H and ω_1^B from a uniform circular distribution, and perform the update step with $\boldsymbol{z}_1$.

B. Dynamic model

The dynamic model for the head and body orientations is

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

These conditional dependencies are also visualized as a directed graphical model in Fig. 2. Similar to [9], we constrain the head orientation at the current time step on the head orientation of the previous time step and on the current body orientation with

$$\begin{aligned} p(\omega_t^H|\omega_{t-1}^H, \omega_t^B) = {} & \alpha_{hh} \mathcal{V}(\omega_t^H; \omega_{t-1}^H, \kappa_{hh}) \\ & + (1 - \alpha_{hh}) \mathcal{V}(\omega_t^H; \omega_t^B, \kappa_{hb}), \end{aligned} \tag{4}$$

where κ_{hh} and κ_{hb} are concentration parameters for the *von Mises* distribution. The *von Mises* $\mathcal{V}(\cdot; \omega, \kappa)$ is an analogue of the normal distribution for the circular domain, with mean angle ω, and 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

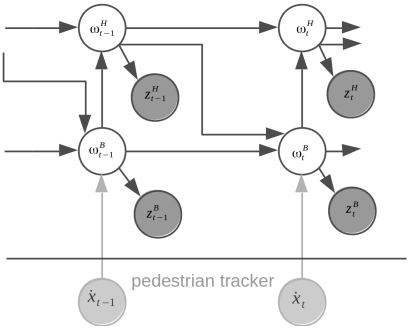

Fig. 2. Dynamic Bayesian network model, showing used constraints between head (ω_t^H) and body orientation (ω_t^B). Inferred hidden state variables are unshaded and observation variables are shaded.

Eq. (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:

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

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 Eq. (5) expresses that the body orientation is typically 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 might also be aligned with the direction of motion. κ_{bb}, κ_{bh} and κ_{bv} denote concentration parameters. Concentration κ_{bv} however depends, similar to [19], on the velocity magnitude $||\dot{\boldsymbol{x}}_t||$, but also on the track state (T_S), with $T_S \in \{0$ (initialized), 1 (preliminary), 2 (confirmed)$\}$, and on the track probability (T_P):

$$\kappa_{bv} = \begin{cases} \kappa_v \cdot (||\dot{\boldsymbol{x}}_t|| - t_v)^2 T_P T_S & \text{if } ||\dot{\boldsymbol{x}}_t|| > t_v \ \&\ T_P > t_p, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Here t_v denotes a threshold for the velocity magnitude and t_p is a threshold for the track probability. κ_v is an initial concentration parameter.

C. Observation model

We assume conditional independence between the head and body observation, and obtain two terms:

$$p(\boldsymbol{z}_t|\boldsymbol{\omega}_t) = p(\boldsymbol{z}_t^H|\omega_t^H)p(\boldsymbol{z}_t^B|\omega_t^B) \tag{7}$$

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 $\boldsymbol{z}$,

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

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,

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

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

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

with c_o and κ_o the mean and concentration of the distribution for orientation class o. 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, R and V, and express first the likelihood $p(z|\Omega, R, V)$. In Section III-C.3 we will then define $p(z|\Omega)$ in terms of this extended likelihood. The indicator variable $R = r, r \in \{1...N\}$ then 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 ($V = 0$).

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

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

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

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

Note that if we belief that the head/body is not present in any region, it then follows from Eq. (11) and (12) that the data likelihood is independent of the orientation Ω and selected region R,

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

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

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

With $p(R^H, R^B|\Omega^B, \boldsymbol{D})$ we introduce prior knowledge about the joint region configuration of head and body, in form of a *Pictorial Structure* (PS) model [25] dependent on the body orientation. Additional knowledge about the regions is modeled based on disparity data $\boldsymbol{D}$, as will be explained in the experiments section. With $p(\Omega)$ we can introduce additional prior information on the orientation classes at this region selection step.

Finally, let $p(V)$ express our prior belief that a head/body is present in any of the observed regions. We integrate out the variable V and obtain,

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

By expanding the terms with Eq. (11) and (12), we see that we can efficiently evaluate Eq. (15) up to a constant factor,

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

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

We also see from Eq. (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 Eq. (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.

IV. EXPERIMENTS

A. Setup

1) Datasets:

a) Training: We use 9300 manually contour labeled pedestrian samples from 6389 images with a minimum / maximum / mean height of 69 / 344 / 122 pixels to train our orientation-specific body and head 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. No single pedestrian occurring in the training set also occurs in the test set.

b) Test: Ground truth consists of 37 manually labeled pedestrian tracks (bounding box labels per frame). As test set we use track estimates on these sequences based on a state-of-the-art HOG/linSVM pedestrian detector [26] and a Kalman Filter. In each frame an estimated track is associated with a ground truth label when the distance between them is smaller than a threshold. The threshold is set according to a percentage of the Euclidean distance of the ground truth label to the camera. We select a different percentage of 8 % and 12 % for lateral and longitudinal direction, since uncertainty in lateral direction is in general smaller. We evaluate all confirmed estimated tracks which have an overlap of more than 80 % to any ground truth track. All other estimated tracks are regarded as false positives and are not used for evaluation. Furthermore, we only include samples with a maximum lateral / longitudinal distance of $5\ m$ / $35\ m$ to the camera. We get 37 valid estimated tracks with 1840 samples for evaluation and ignore 12 false positive tracks.

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. As suggested in [11], we do not use the direct neighbor classes for the positive class. For the background detector $f_-(z)$ we use all background samples versus all orientation specific samples. For all detectors we use multi-layer *Neural Network* architectures (NN/LRF) [27] with a 5x5 *Local Receptive Field*. We extracted the head by a fixed aspect ratio of 15 % of the whole body from top of the contour labeled shape. All head samples were scaled to 16 x 16 pixels for training and testing, while the body samples were scaled to 48 x 96 pixels. For the body detectors we use only the lower 85 % of the whole body to make sure that the head part is completely ignored and not affecting the body orientation estimate. A border of 2 pixels was added to head and body samples to avoid border effects. We allow a maximum of 8 jittered additional samples per sample.

3) Priors and parameters: 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 position. We factor the prior from Eq. (14) into

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

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 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 Semi Global Matching (SGM) [28] to calculate the disparity map $\boldsymbol{D}$. The likelihood in case of the head region is then modeled with

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

$\boldsymbol{\mu}(r^H)$ 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. Due to efficiency reasons, we generate possible head regions only around the estimated head

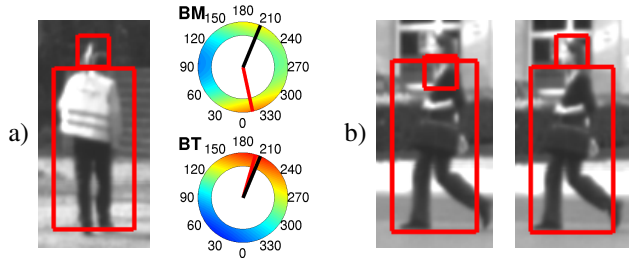

Fig. 3. a) Example of a multi-modal likelihood estimation of the body (BM), where in the tracked posterior (BT) this ambiguity is resolved. We show also the maximum likelihood/posterior estimate (red line) and ground truth orientation (black line). b) Integrating the PS constraint (right) results in a better localization of head and body.

position $\boldsymbol{h}_c$. Since we are already given a track bounding box, the number of necessary body hypotheses is much less than it is for the head. The likelihood $p(\boldsymbol{b}_c(\boldsymbol{D})|R^B = r^B)$ in case of the body region is modeled similar. While the size of the head/body region is set according to the estimated head height, the step size between regions is set dependent on the pedestrian distance.

b) Joint region prior: To model the joint spatial prior $p(R^H, R^B|\Omega^B)$ we use a *Pictorial Structure* (PS) model [25], which is dependent on the body orientation:

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

$\boldsymbol{l}^D(r^H, r^B)$ denotes the distance of head and body region center relative to the width of the body region. We learned the parameters $\boldsymbol{\mu}^D_{o^B}$ and $\boldsymbol{C}^D_{o^B}$ for the PS between head and body region for each discrete orientation from training data.

c) Other settings: The priors $p(\Omega)$ and $p(V)$ are modeled with an uniform distributions for head and body orientation. Furthermore we set the parameters $\alpha_{hh} = \alpha_{bb} = 0.7$, $\alpha_{bh} = 0.2$, $\kappa_{hh} = \kappa_{bb} = 4$, $\kappa_{hb} = \kappa_{bh} = 1.0$, $\kappa_v = 2$ heuristically and learned from training data an average of $c_{o^H} = 0.78$, $c_{o^B} = 0.68$ (see Eq. (10)) for each class o.

B. Results

We show in Fig. 3 a) a sample that gives a multi-modal likelihood estimate, caused by confusing opposite directions. This ambiguity can be successful corrected by our joint tracking approach. The effect of integrating the PS spatial constraint is shown in Fig. 3 b) where localization of head and body is improved with the spatial PS constraint (right image). In Fig. 4 we show disparity and gray image of every sixth frame of an estimated track with continuous estimation results of our proposed approach.[1] Disparity data is used as prior information for region generation and localization. As can be seen the joint tracking delivers good localization and a robust continuous orientation estimate of head and body.

We perform a quantitative evaluation on the complete test set using all 37 valid, estimated tracks. In Fig. 5 we show the absolute angular mean error for head and body

[1]Follow the links from *http://www.gavrila.net* for a video animation.

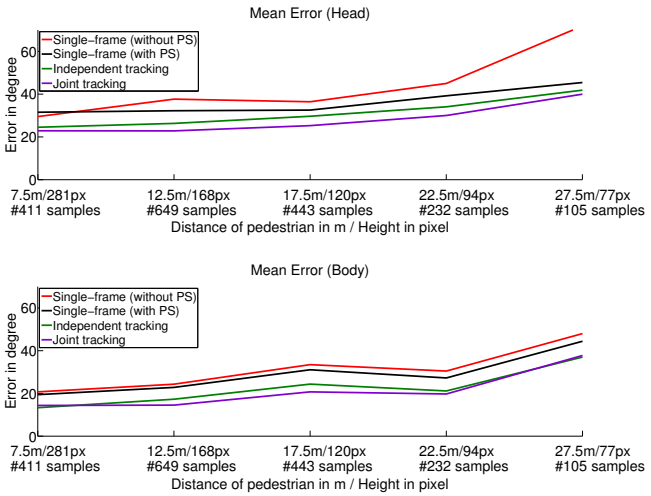

Fig. 5. Absolute angular mean error over increasing distance for head (top) and body (bottom) orientation, using joint (purple), independent tracking (green) and single-frame estimation with (black) and without (red) PS.

orientation estimation. We compare our proposed joint tracking to the results of independent tracking and single-frame orientation estimation with and without PS (see Eq. (18)). Independent tracking refers to tracking of head and body without an orientation coupling as defined in Sec. III-B. 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 error for head/body orientation in total by $13°$ / $10°$ compared to single-frame estimation without PS. This benefit is mainly caused by the removal of outliers compared to single-frame estimation (e.g. confusion between opposite body directions, which visually can look very similar). Furthermore in comparison to independent tracking, we decrease the error by $3°$ / $2°$ for head/body orientation. Anatomical and movement constraints within tracking as defined in Sec. III-B help here to reject impossible configurations between head and body orientation. In Fig. 6 we show an additional boxplot to get a better impression of the estimation uncertainty and the error distribution. It can be seen that joint tracking reduces the uncertainty and outliers. Our unoptimized implementation, running on a 3.33 GHz i7-CPU processor, needs on average one second per image. We expect to reach real-time performance within few months.

V. Conclusion

In this paper, we presented a method for continuous estimation of pedestrian head and body orientation by applying multiple orientation-specific detectors. We evaluated our method on gray-value images recorded from a camera on board a moving vehicle. Quantitative experiments showed that joint tracking of head and body decreases the angular mean error for head/body orientation by $13°$ / $10°$ compared to single frame estimation and further by $3°$ / $2°$ compared to independent tracking. Future work involves the incorporation of head and body orientation for situation analysis, and the

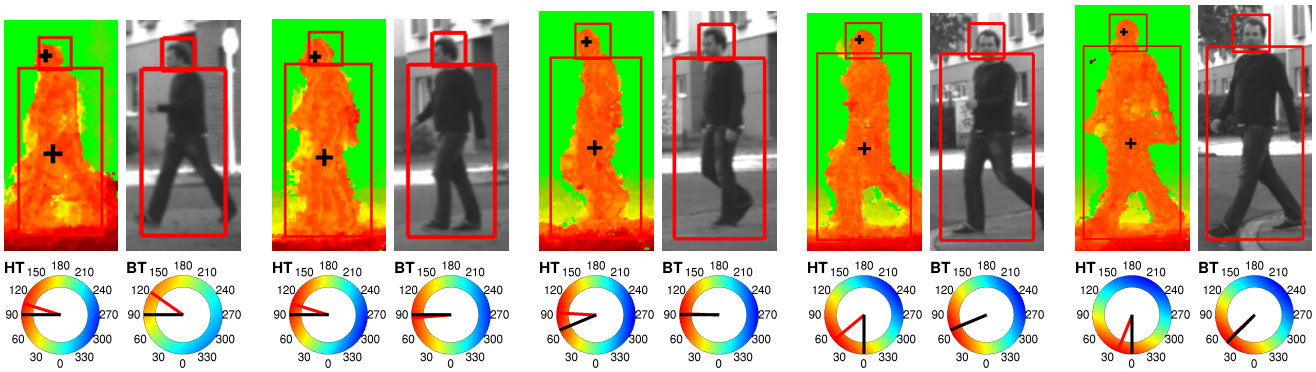

Fig. 4. Disparity (left) and gray image (right) of every sixth frame of an estimated track. The red boxes show the selected head and body region. Below the images we show the posterior distributions of our approach for the head (HT) and body orientation (BT), maximum posterior estimate (red line) and ground truth orientation (black line). Black crosses in disparity images denote estimated head and body center, which are used as prior information for localization and region generation.

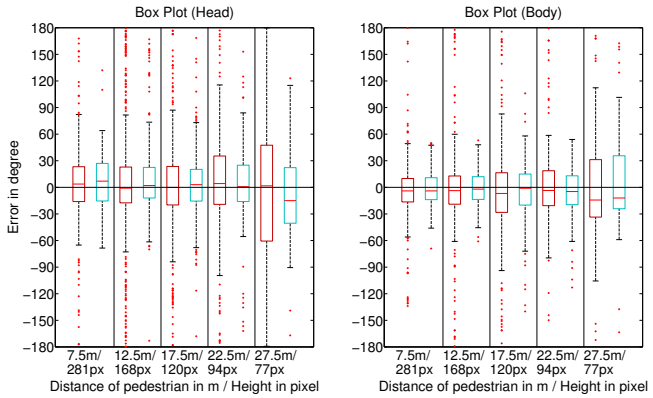

Fig. 6. Boxplots showing median error (red line) and outliers (red crosses) for a) head and b) body orientation estimation in case of single-frame estimation without PS (red box) and joint tracking (blue box). Boxes contain 50 % of samples. Used whiskers define 99.3 % data coverage. By joint tracking we get a more robust estimation and smaller whisker lengths.

real-time integration in a test vehicle.

REFERENCES

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

[2] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE PAMI*, vol. 31, no. 4, pp. 607–626, 2009.

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

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

[5] 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.

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

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

[8] J. Orozco, S. Gong, and T. Xiang, "Head pose classification in crowded scenes." in *Proc. BMVC*, vol. 1, 2009, p. 3.

[9] N. Robertson and I. Reid, "Estimating gaze direction from low-resolution faces in video," in *Proc. ECCV*, 2006, pp. 402–415.

[10] 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 PAMI*, vol. 30, no. 7, pp. 1212–1229, 2008.

[11] 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*, 2011, pp. 51–60.

[12] B. Benfold and I. Reid, "Guiding visual surveillance by tracking human attention." in *Proc. BMVC*, 2009, pp. 1–11.

[13] T. Gandhi and M. M. Trivedi, "Image based estimation of pedestrian orientation for improving path prediction," in *IEEE Intell. Veh*, 2008, pp. 506–511.

[14] M. Enzweiler and D. M. Gavrila, "Integrated pedestrian classification and orientation estimation," in *Proc. CVPR*, 2010, pp. 982–989.

[15] H. Shimizu and T. Poggio, "Direction estimation of pedestrian from multiple still images," in *IEEE Intell. Veh*, 2004, pp. 596–600.

[16] F. Flohr and D. M. Gavrila, "Pedcut: an iterative framework for pedestrian segmentation combining shape models and multiple data cues," in *Proc. BMVC*, 2013.

[17] D. M. Gavrila and S. Munder, "Multi-cue pedestrian detection and tracking from a moving vehicle," *IJCV*, vol. 73, no. 1, pp. 41–59, 2007.

[18] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *CVIU*, vol. 104, no. 2, pp. 90–126, 2006.

[19] 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.

[20] G. Zhao, M. Takafumi, K. Shoji, and M. Kenji, "Video based estimation of pedestrian walking direction for pedestrian protection system," *Journal of Electronics (China)*, vol. 29, no. 1-2, pp. 72–81, 2012.

[21] S. O. Ba and J.-M. Odobez, "Probabilistic head pose tracking evaluation in single and multiple camera setups," in *Proc. CLEAR Workshop*, 2008, pp. 276–286.

[22] R. Cucchiara, C. Grana, A. Prati, and R. Vezzani, "Probabilistic posture classification for human-behavior analysis," *IEEE Trans. on Sys., Man, and Cyb.*, vol. 35, no. 1, pp. 42–54, 2005.

[23] R. Yang and Z. Zhang, "Model-based head pose tracking with stereovision," in *Proc. FG*, 2002, pp. 255–260.

[24] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Journal of Statisitcs and Computing*, vol. 10, no. 3, pp. 197–208, 2000.

[25] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *IJCV*, vol. 61, no. 1, pp. 55–79, 2005.

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

[27] C. Wöhler and J. K. Anlauf, "A time delay neural network algorithm for estimating image-pattern shape and motion," *IVC*, vol. 17, no. 3, pp. 281–294, 1999.

[28] H. Hirschmüller, "Stereo processing by semiglobal matching and mutual information," *IEEE PAMI*, vol. 30, no. 2, pp. 328–341, 2008.