Robust Multi-Person Tracking from Moving Platforms

Andreas Ess[1], Konrad Schindler[1], Bastian Leibe[1,2] and Luc van Gool[1,3]

[1]ETH Zürich [2]RWTH Aachen [3]KU Leuven, IBBT
{aess|schindler|leibe|vangool}@vision.ee.ethz.ch

Abstract. In this paper, we address the problem of multi-person tracking in busy pedestrian zones, using a stereo rig mounted on a mobile platform. The complexity of the problem calls for an integrated solution, which extracts as much visual information as possible and combines it through cognitive feedback. We propose such an approach, which jointly estimates camera position, stereo depth, object detection, and tracking. We model the interplay between these components using a graphical model. Since the model has to incorporate object-object interactions, and temporal links to past frames, direct inference is intractable. We therefore propose a two-stage procedure: for each frame we first solve a simplified version of the model (disregarding interactions and temporal continuity) to estimate the scene geometry and an overcomplete set of object detections. Conditioned on these results, we then address object interactions, tracking, and prediction in a second step. The approach is experimentally evaluated on several long and difficult video sequences from busy inner-city locations. Our results show that the proposed integration makes it possible to deliver stable tracking performance in scenes of realistic complexity.

1 Introduction

Recent research successes have fostered the demand for mobile vision systems that can operate in unconstrained scenarios of daily human living. Building such systems is a crucial requirement for many applications in the near future of mobile robotics and smart vehicles. So far the sheer complexity of real-world scenes has however often stymied progress in this direction.

In this paper, we focus on the task of multi-person tracking in busy street scenes as seen from a mobile observer. This could be a mobile robot, an electric wheelchair, or a car passing through a crowded city center. The scenario is extremely challenging due to a variety of factors: motion blur, varying lighting, large numbers of independently moving objects (sometimes covering almost the entire image) frequent partial occlusions between pedestrians, and sub-optimal camera placement dictated by the constraints of moving platforms (cameras are less than 1 meter above ground, so for an object 20 meters away, a localization error of 1 pixel in the image equals about 1 meter in depth).

It has been argued that scene analysis in such complex settings requires the combination of and careful interplay between several different vision modules. However, it is largely unclear how such a combination should be undertaken and which properties are critical for its success. Here, we integrate visual odometry, depth estimation, pedestrian detection, and tracking in a graphical model, and propose a two-step procedure to perform approximate inference in that model. An important component of

Dagstuhl Seminar Proceedings 08091
Logic and Probability for Scene Interpretation
http://drops.dagstuhl.de/opus/volltexte/2008/1617

the proposed integration is the concept of cognitive feedback. The underlying idea is that the higher-level information extracted by each vision module should be fed back to other modules, thereby improving performance. The main contributions of this paper are: (1) We simultaneously estimate scene geometry and track objects in a challenging real-world scenario (from video input), integrating cues from dense stereo, object detection, tracking, visual odometry, and ground plane estimation. (2) We model this integration in a principled fashion using a graphical model that allows depth measurement and object detection in each frame to benefit from each other. With the help of a world coordinate system provided by visual odometry, we then link these single-frame results over time to object tracks. (3) We experimentally validate the proposed method on challenging real-world data and demonstrate that the integrated approach to visual scene understanding improves over the state-of-the-art.

The paper is structured as follows. After discussing related work in the following section, Section 3 presents the system, with a particular focus on improving object detection using a Bayesian network that incorporates ground plane and depth measurements. Next, Section 4 describes a few implementation details for the single-frame detection part. Then, Section 5 presents experimental results on a number of challenging video sequences, and Section 6 concludes the paper with a summary and outlook.

2 Related Work

Pedestrian Detection. Human detection has reached an impressive level [8, 39–41], with many systems also able to estimate the silhouette of the detected pedestrian [26,37, 42]. Still, pedestrian detection remains a difficult task due to large intra-category variability, scale changes, articulation, and frequent partial occlusion. To achieve robustness to adverse imaging conditions, the importance of *context* has been widely recognized. Depending on the authors, the rather loose notion of “context” can refer to different types of complementary information, including motion [9, 40], stereo depth [12, 15], scene geometry [20, 25], temporal continuity [1, 27, 41], or semantics of other image regions [29, 33, 38].

The use of depth to improve object detections suggests itself in systems equipped with camera pairs, *e.g*. [15, 17]. Both of these approaches assume a fixed groundplane and disregard interactions between pedestrians.

Multi-body Tracking. Many approaches are available for multi-object tracking from stationary cameras (*e.g*. [4, 24]). The task is however made considerably harder when the camera itself moves. In such cases, background subtraction is no longer a viable option, and tracking-by-detection appears to be the most promising alternative [2, 15, 18, 25, 32, 41, 44].

Targets are typically followed using classic tracking approaches, such as Extended Kalman Filters (EKF) [16], particle filters [21], or Mean-Shift tracking [6], which rely on a first-order Markov assumption and hence carry the danger of drifting away from the correct target. This danger can be reduced by optimizing data assignment and considering information over several time steps, as in Multi-Hypothesis Tracking (MHT) [7,35] and Joint Probabilistic Data Association Filters (JPDAF) [14]. However, their combinatorial nature limits those approaches to consider either only few time steps [35] or

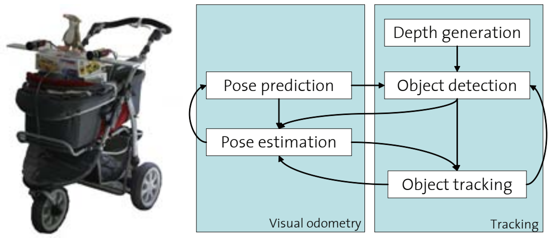


Fig. 1. (Left) Mobile recording system equipped with camera pair. (Right) Components of our mobile vision system and their connections, executed for each frame of a video sequence.

only single trajectories over longer time windows [4, 22]. Recently, [44] suggested a graph-based formulation for multi-target tracking that allows an efficient global solution even in complex situations. The approach operates on the entire video sequence, and requires the detections for all frames as input. This precludes its online application to long sequences. In contrast, our approach works online and simultaneously optimizes detection and trajectory estimation for multiple interacting objects and over long time windows, by operating in a hypothesis selection framework [25, 27].

3 System

Our system is based on a mobile platform equipped with a pair of forward-looking cameras, Fig. 1(left). Under the predominantly occurring forward motion, the stereo setup is a better choice for self-localization than a monocular system, due to the latter's weak geometric configuration [19]. Furthermore, generating depth maps has been well-studied for such setups [36], and dense depth information is of great help for constraining object detection, and thus improving tracking and egomotion estimation. Figure 1(right) gives an overview of the proposed vision system. For each frame, the blocks are executed as indicated: first, a depth map is calculated and the new frame's camera pose is predicted. Then objects are detected, taking advantage of appearance, depth, and trajectory information. The output, along with predictions from the tracker, helps stabilize visual odometry, which updates the pose estimate for the platform and the detections, before running the tracker on these updated detections. The whole system is held entirely causal, *i.e*. at any point in time it only uses information from the past and present. The following subsections describe the three main components of the system, and give details about their robust implementation.

3.1 Object Detection

Fig. 2 shows the Bayesian network we use for inference over object hypotheses o_i, object depth d_i, and the ground plane $\boldsymbol{\pi}$ using evidence from the image $\mathcal{I}$, the depth map $\mathcal{D}$, its occlusion map $\mathcal{O}$, and the ground plane evidence $\boldsymbol{\pi}_{\mathcal{D}}$ in the depth map. Following standard graphical model notation [5], the plate indicates repetition of the

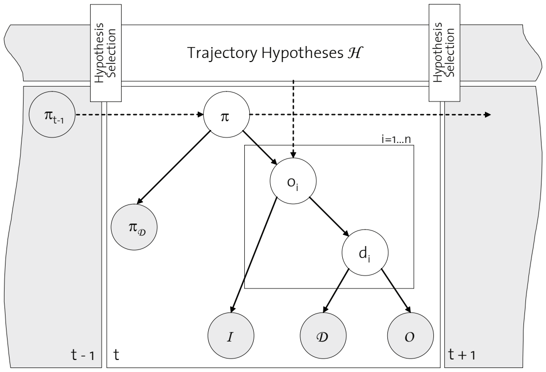


Fig. 2. Graphical model for tracking-by-detection with additional depth information.

contained parts for the number of objects n. Inference in this model is performed as follows:

$$P(\boldsymbol{\pi}, o_i, d_i, \mathcal{E}, \mathcal{F}) \propto P(\boldsymbol{\pi}|\boldsymbol{\pi}_{\mathcal{D}})P(\boldsymbol{\pi}|\boldsymbol{\pi}_{t-1})\cdot \\ \cdot\prod_i \left(P(o_i|\boldsymbol{\pi})P(o_i|d_i)P(d_i)P(o_i|\mathcal{H}_{t0:t-1})P(\mathcal{I}|o_i)P(\mathcal{D}|d_i)P(\mathcal{O}|d_i)\right) \ , \tag{1}$$

where $\mathcal{E} = \{\mathcal{I}, \mathcal{D}, \mathcal{O}, \boldsymbol{\pi}_{\mathcal{D}}\}$ is the evidence observed in the current frame and $\mathcal{F} = \{\boldsymbol{\pi}_{t-1}, \mathcal{H}_{t0:t-1}\}$ is the evidence from previous frames. An object's probability depends both on its geometric world features (distance, size) $P(o_i|\boldsymbol{\pi})$ and its correspondence with the depth map (distance, assumption of uniform depth) $P(o_i|d_i)$. The factor $P(o_i|\mathcal{H}_{t0:t-1})$ incorporates past trajectories $\mathcal{H}$, and $P(\mathcal{I}|o_i)$ is the object probability estimated by the pedestrian detector (the time index t for the current frame was omitted for brevity – all variables without time index refer to time step t).

Finally, we introduce temporal dependencies, indicated by the dashed arrows in Fig. 2. For the ground plane, we propagate the state in the previous frame as a temporal prior $P(\boldsymbol{\pi}|\boldsymbol{\pi}_{t-1}) = (1-\alpha)P(\boldsymbol{\pi})+\alpha P(\boldsymbol{\pi}_{t-1})$ that augments the per-frame information from the depth map $P(\boldsymbol{\pi}|\boldsymbol{\pi}_{\mathcal{D}})$. For the detections, we add a spatial prior for object locations that are supported by tracked candidate trajectories $\mathcal{H}_{t0:t-1}$. As shown in Fig. 2, this dependency is not a first-order Markov chain, but reaches many frames into the past, as a consequence of the tracking framework explained in Section 3.2.

In the following, the components of this Bayesian network are described in detail. All 3D calculations are executed in camera coordinates, *i.e.* the projection matrix is $\mathtt{P} = [\mathtt{K}|\mathbf{0}]$. This not only simplifies calculations and parameterizations, but it also keeps the set of possible ground planes in a range that can be trained in a meaningful way. For the subsequent tracking, the results are easily transferred into world coordinates with the camera orientation provided by visual odometry, Section 3.3.

Ground Plane. As shown in previous publications [15, 20, 25], the ground plane helps substantially to constrain object detection to meaningful locations. It is defined in the current camera frame as $\boldsymbol{\pi} = (\mathbf{n}, \pi_4)$, with the normal vector parameterized by spherical coordinates, $\mathbf{n}(\theta,\phi)=(\cos\theta\sin\phi, \sin\theta\sin\phi, \cos\phi)$. The ground plane parameters $\boldsymbol{\pi}$ are inferred from a combination of a prior, object bounding boxes, and the depth map evidence $\boldsymbol{\pi}_{\mathcal{D}}$, so that the system does not critically depend on any one individual cue. While accurate ground planes can be estimated directly from depth maps, see Section 4, such methods break down in outlier-ridden scenarios. Thus, $\boldsymbol{\pi}_{\mathcal{D}}$ will just act as an additional cue in our Bayesian Network.

Inference at this stage works on a per-frame basis using information obtained from $\mathcal{D}$. Specifically, we consider the depth-weighted median residual between $\boldsymbol{\pi}$ and $\mathcal{D}$:

$$r(\boldsymbol{\pi},\mathcal{D}) = \mathrm{med}_{\mathbf{x}\in\mathcal{D}}\ \Delta_z(\boldsymbol{\pi},\mathbf{x}). \tag{2}$$

Here $\mathbf{x}\in\mathcal{D}$ denotes the set of 3D points inferred from $\mathcal{D}$, pruned according to the vehicle's maximally expected tilt angle and restricted to the lower part of the image for increased robustness to outliers. $\Delta_z(\cdot,\cdot)$ is the plane-to-point Mahalanobis distance taking into account the 3D point's uncertainty. The probability that the real ground plane has generated the depth map evidence is modeled as a 1D Gaussian,

$$P(\boldsymbol{\pi}_{\mathcal{D}}|\boldsymbol{\pi}) \propto \mathcal{N}(r(\boldsymbol{\pi},\mathcal{D}); 0, \sigma_{\mathcal{D}}^2)\ , \tag{3}$$

with $\sigma_{\mathcal{D}}$ the standard deviation of a plane-point measurement in $\mathcal{D}$, influenced by the depth map accuracy and the quantization of the parameter space. The prior $P(\boldsymbol{\pi})$ is learned from a training set (see Section 4).

Object Hypotheses. Object hypotheses $o_i = \{v_i, \mathbf{c}_i\}^{i=1\dots n}$ consist of a validity flag $v_i\in\{0,1\}$ and a 2D center point with scale $\mathbf{c}_i=\{x,y,s\}$. The number n is determined by the pedestrian detector (in our case, an ISM [26]) independently for every frame; in our scenes it is usually $\approx 70-90$. Given a specific $\mathbf{c}$ and a standard object size (w,h) at scale $s=1$, a bounding box can be constructed. From the box base point in image coordinates $\mathbf{g}=(x, y+s\frac{h}{2}, 1)$, its counterpart in world coordinates is found as

$$\mathbf{G} = -\frac{\pi_4 \mathtt{K}^{-1}\mathbf{g}}{\mathbf{n}^\top \mathtt{K}^{-1}\mathbf{g}}. \tag{4}$$

The object's depth is thus $z(o_i) = \|\mathbf{G}_i\|$. The box height $\mathbf{G}_i^h$ is obtained in a similar fashion. Because of the large localization uncertainty of appearance-based detection, the detector's output for center and scale are considered estimates, denoted $\tilde{x}_i$, $\tilde{y}_i$, and $\tilde{s}_i$. Taking these directly might yield misaligned bounding boxes, which in turn can result in considerably wrong estimates for distance and size. For each object o_i, we therefore consider a set of possible real centers $\mathbf{c}_i=\{y_i,s_i\}$ (fixing $x_i=\tilde{x}_i$ due to its negligible influence), obtained by sampling around the detection, $y_i=\tilde{y}_i+k\sigma_y\tilde{s}_i$, $s_i=\tilde{s}_i+l\sigma_s\tilde{s}_i$. The number of samples, *i.e.* the range of $\{k,l\}$, is the same for every object. We thus obtain a set of bounding boxes $\mathbf{b}_i^{\{k,l\}}$ for each o_i. This allows the inference to compensate for detection inaccuracies in order to find a better explanation of the scene. In the following, we omit the superscripts for readability. The object term is decomposed as

$$P(o_i|\boldsymbol{\pi}) = P(v_i|\mathbf{c}_i,\boldsymbol{\pi})P(\mathbf{c}_i|\boldsymbol{\pi})\ . \tag{5}$$

By means of Eq. (4), $P(\mathbf{c}_i|\boldsymbol{\pi}) \propto P(\mathbf{G}_i^h)P(z(o_i))$ is expressed as the product of a distance and a size prior for the corresponding real-world object. We formulate the probability for a hypothesis' validity based on this, $P(v_i=1|\mathbf{c}_i,\boldsymbol{\pi}) = \max P(\mathbf{c}_i|\boldsymbol{\pi})$.

Depth Map. The depth map $\mathcal{D}$ is a valuable asset for scene understanding that is readily available in a multi-camera system. However, stereo algorithms frequently fail, especially in untextured regions. For all our calculations, we therefore consider an additional *occlusion map* $\mathcal{O}$, which models the trust in each depth estimate based on a left-right check. Based on this consistency check, we integrate depth into our framework in a robust manner: each object hypothesis is augmented with a depth flag $d_i \in \{0,1\}$, indicating whether the depth map for a bounding box is reliable ($d_i = 1$) or not. As explained above, the depth term is decomposed into two parts:

$$P(o_i|d_i) = P(v_i|\mathbf{c}_i, d_i)P(\mathbf{c}_i|d_i). \tag{6}$$

First, we evaluate the depth measured inside $\mathbf{b}_i$ and its consistency with $z(o_i)$ as an indicator for $P(\mathbf{c}_i|d_i{=}1)$. Second, we test the depth variation inside the box and define $P(v_i = 1|\mathbf{c}_i, d_i = 1)$ to reflect our expectation that the depth is largely uniform when a pedestrian is present. The measurements are defined as follows: the median depth inside a bounding box, $z(\mathcal{D}, \mathbf{b}_i) = \mathrm{med}_{\text{pixel } \mathbf{p}\in\mathbf{b}_i}\mathcal{D}(\mathbf{p})$, yields a robust estimate of the contained object's depth. Assuming additive white noise with covariance $\mathtt{C}_{2\mathrm{D}}$ on pixel measurements, we find the variance $\sigma^2_{(z),i}$ of $z(\mathcal{D}, \mathbf{b}_i)$ using error backpropagation,

$$\mathtt{C}_i = \left(\mathtt{F}_i^{(1)\top}\mathtt{C}_{2\mathrm{D}}^{-1}\mathtt{F}_i^{(1)} + \mathtt{F}_i^{(2)\top}\mathtt{C}_{2\mathrm{D}}^{-1}\mathtt{F}_i^{(2)}\right)^{-1} , \tag{7}$$

where $\mathtt{F}_i^{(j)}$ are the Jacobians of a projection using camera matrix j, thus $\sigma^2_{(z),i} = C_i^{(3,3)}$. This yields

$$P_{(z),i}(x) \propto \mathcal{N}(x; z(\mathcal{D}, \mathbf{b}_i), \sigma^2_{(z),i}) . \tag{8}$$

For reasoning about depth uniformity, we consider the depth variation within $\mathbf{b}_i$, $V = \{\mathcal{D}(\mathbf{p}) - z(\mathcal{D}, \mathbf{b}_i)|\mathbf{p}\in\mathbf{b}_i\}$. To be robust against outliers, the estimate is restricted to the interquartile range $[LQ(V), UQ(V)]$, and depth uniformity is measured by the normalized count of pixels that fall within the confidence interval $\pm\sigma_{(z),i}$,

$$q_i = |\{x \in [LQ, UQ]\big|x^2 < \sigma^2_{(z)i}\}|/(UQ - LQ) . \tag{9}$$

This robust "depth inlier fraction" serves as basis for learning $P(v_i|\mathbf{c}_i, d_i = 1)$, as is shown in Section 4. $P(o_i|d_i = 0)$ is assumed uniform, since an inaccurate depth map gives no information about the object's presence. We infer $P(d_i)$ from the training set based on the data from the occlusion map.

3.2 Tracking, Prediction

After passing the Bayesian network, object detections are placed into a common world coordinate system using camera positions estimated from visual odometry. The actual tracking system follows a multi-hypotheses approach, similar to the one described in [27]. We do not rely on background subtraction, but instead accumulate the detections of the current and past frames in a space-time volume. This volume is analyzed by growing many trajectory hypotheses using independent bi-directional Extended Kalman filters (EKFs). By starting EKFs from detections at different time steps, an overcomplete set of trajectories is obtained, which is then pruned to a minimal consistent explanation using model selection. Overlapping trajectory hypotheses are resolved with a global model selection step, in which trajectories compete for detections and space-time volume. In a nutshell, the pruning step employs quadratic pseudo-boolean optimization to pick the set of trajectories with maximal joint probability, given the observed evidence over the past frames. That probability

- increases as the trajectories explain more detections and as they better fit the detections' 3D location and 2D appearance through the individual contribution of each detection;
- decreases when trajectories are (partially) based on the same object detections, through pairwise corrections to the trajectories' pairwise joint probabilities (these express the constraints that each pedestrian can only follow one trajectory and that two pedestrians cannot be at the same position at the same time);
- decreases with the number of required trajectories through a prior favoring explanations with fewer trajectories – balancing the complexity of the explanation against its goodness-of-fit in order to avoid over-fitting ("Occam's razor").

For the mathematical details, we refer to [27]. The most important features of this method are automatic track initialization (usually, after about 5 detections) and the ability to recover from temporary track loss and occlusion.

The selected trajectories $\mathcal{H}$ are then used in the next frame to provide a spatial prior for the object detections. This prediction has to take place in the world coordinate system, so tracking critically depends on an accurate and smooth egomotion estimate.

3.3 Visual Odometry

To allow reasoning about object trajectories in the world coordinate system, the camera position for each frame is estimated using visual odometry. The employed system builds on previous work by [31], see Fig. 3 for a flow diagram. In short, each incoming image is divided into a grid of 10×10 bins, and an approximately uniform number of points is detected in each bin using a Harris corner detector with locally adaptive thresholds. The binning encourages a feature distribution suitable for stable localization. In the initial frame, stereo matching and triangulation provide a first estimate of the 3D structure. In subsequent frames, we use 3D-2D matching to get correspondences, followed by camera resection (3-point pose) with RANSAC [30]. Bundle adjustment is run on a sliding window of $n_b = 18$ past frames to polish the raw camera estimates. Older frames are discarded, along with points that are only supported by these removed frames.

Important details for reliable performance are the use of 3D-2D matching to bridge temporally short occlusions of feature points and to filter out independently moving objects at an early stage, as well as a Kalman filter to predict the next camera position for feature detection (leading to a feature detection strategy similar to the "active search" paradigm in SLAM, *e.g.* [10]). Scene points are directly associated with a viewpoint-invariant SURF descriptor [3] that is adapted over time. In each frame, the 3D-2D correspondence search is then constrained by the predicted camera position. As mentioned above, only scene points without support in the past n_b frames are discarded. This allows one to bridge temporally short occlusions (*e.g.* from a person passing through the

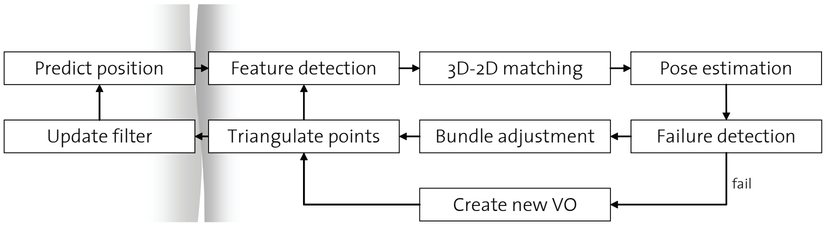


Fig. 3. Flow diagram of the employed visual odometry system. The shaded regions indicate the insertion points for the feedback from object tracking.

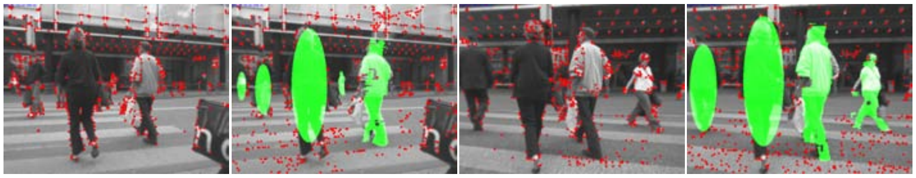


Fig. 4. Object detection and tracking give semantic meaning to the image and can be used to restrict localization efforts to parts that are believed to be static.

image) by re-detecting 3D points that carry information from multiple viewpoints and are therefore already reliably reconstructed. Attaching 2D appearance descriptors to (coarse) 3D geometry is also a recent trend in object recognition [43].

In order to guarantee robust performance, we introduce two measures: first, cognitive feedback from the tracker is used to constrain corner detection for visual odometry: the predictions delivered by the tracker and visual odometry are used to mark parts of the image, which are with a high probability occupied by moving objects (pedestrians), Fig. 4. These parts are then ignored when looking for corners, thereby considerably reducing the number of outlier matches in RANSAC. Second, we introduce an explicit failure detection mechanism, as described in [11]. In case of failure, the Kalman filter estimate is used instead of the measurement, all scene points are cleared, and the visual odometry is restarted from scratch. This allows us to keep the tracker running without resetting it. While such a procedure may introduce a small drift, a locally smooth trajectory is more important for our application. In fact, driftless global localization using only a moving camera rig is inherently impossible (except in retrospect in the case of loop closure). We believe that this capability, if needed, is best achieved by integrating other sensors, such as GPS and INS.

4 Detailed Implementation

The system's parameters have been trained on a sequence (Seq. #1) with 490 frames, containing 1'578 annotations.[1] For learning the ground plane prior, we considered an

[1] We have used data recorded at a resolution of 640×480 pixels (bayered) at 15 FPS, with a camera baseline of 0.4 meters.

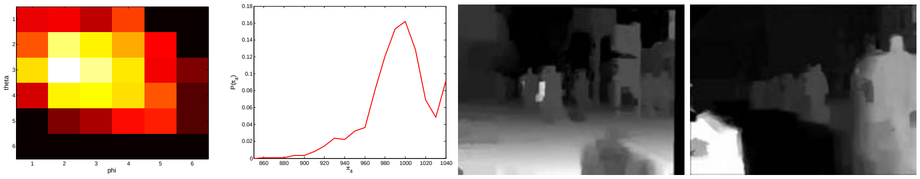


Fig. 5. (First two images:) Learned priors for (θ, ϕ) (left) and π_4 (right), projected onto π_4 and (θ, ϕ), respectively. (Second two images:) Example depth maps. Often, useful cues can be inferred (left), but robust measures have to account for faulty depth maps, *e.g*. missing ground plane (right).

additional 1'600 frames from a few selected environments with hardly any moving objects.

Ground Plane. In imagery with few objects, $\mathcal{D}$ can be used to directly infer the ground plane using Least-Median-of-Squares (LMedS) by means of Eq. (2),

$$\boldsymbol{\pi} = \min_{\boldsymbol{\pi}_i} r(\boldsymbol{\pi}_i, \mathcal{D}). \tag{10}$$

Related but less general methods include *e.g*. the *v*-disparity analysis [23]. All such methods break down if less than 50% of the pixels in $\mathcal{D}$ support $\boldsymbol{\pi}$. For training, we use the estimate from Eq. (10), with bad estimates discarded manually.

For tractability, (θ, ϕ, π_4) are discretized into a $6 \times 6 \times 20$ grid, with bounds inferred from the training sequences. The discretization is chosen such that quantization errors are below 0.01 for θ and ϕ, resulting in component-wise abberations of maximally $5 \cdot 10^{-7}$ from the original $\mathbf{n}$. In our tests, the errors ensuing from the discretization of $\boldsymbol{\pi}$ were below 0.05 meters in depth for a pedestrian 15 meters away. Note that other choices of spherical coordinates for the normal vector would be better suited to the variability of the tilt angle. However, the described parameterization is sufficient, and alternative choices for discretizing turn out to be more cumbersome because of switches from $-180°$ to $180°$. The training sequences also serve to construct the prior distribution $P(\boldsymbol{\pi})$. Fig. 5 visualizes $P(\boldsymbol{\pi})$ in two projections onto π_4 and (θ, ϕ).

Object Hypotheses. Object hypotheses are detected using a single-category ISM detector [26], trained on a mixed set of frontal and side views of pedestrians. The detector is run without the final global optimization stage of [26] to retain the necessary flexibility – in the context of the additional evidence we are using, final decisions based only on appearance would be premature. The range of detected scales corresponds to pedestrian heights of 60–400 pixels. Other detectors could also be included in our system, as long as they provide confidence maps.

As the original detection centers $\tilde{x}, \tilde{y}, \tilde{s}$, and hence the bounding boxes, may not always be sufficiently accurate for reliable depth estimation, we model the variance between real and detected object centers by Gaussians, see [12] for details.

The object size distribution is chosen as in [20], $P(\mathbf{B}^h|h) \sim \mathcal{N}(1.7, 0.085^2)$ [m], though we consider different standard deviations σ_h in a first systematic experiment in Section 5. This is mainly to account for children and for the remaining discretization

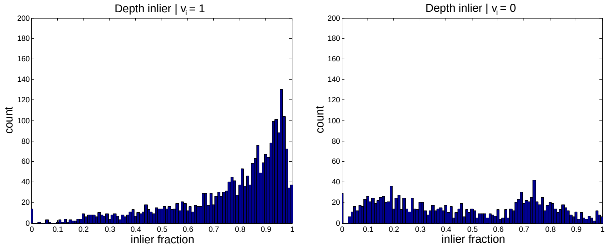


Fig. 6. Distribution of depth inliers for correct (left) and incorrect (right) detections, learned from 1,578 annotations and 1,478 negative examples. A sigmoid is used to represent this.

errors due to the sampling of $\mathbf{c}_i$. The depth distribution $P(z(o_i))$ is assumed uniform in the system's operating range of 0.5–30 [m].

Depth Cues. The depth map $\mathcal{D}$ for each frame is obtained with a publicly available, belief-propagation-based disparity estimation software [13]. See Fig. 5 for two example depth maps. The true distribution of $P(\mathbf{c}_i|d_i = 1)$ given the object's depth $z(o_i)$ and the depth map estimate $z(\mathcal{D}, \mathbf{b}_i)$ is very intricate to find. It involves many factors: first, the uncertainty of the object's center propagated to its distance. Due to the sampling of $\mathbf{c}_i$, we can neglect this factor. Second, it depends on $P_{(z),i}$ as defined in Eq. (8). Finally, using a fixed set of disparities introduces a quantization error, which is only to some extent covered by $P_{(z),i}$.

In Section 5, we compare two ways for modeling $P(\mathbf{c}_i|d_i = 1)$. The first option uses a non-parametric distribution $P(v_i|z(o_i) - z(\mathcal{D}, \mathbf{b}_i))$, learned from the training sequence. The second option models it using the dominating factor $P_{(z),i}(z(o_i))$ only.

For learning $P(v_i|\mathbf{c}_i, d_i = 1)$, we find the percentage q_i of pixels that can be considered uniform in depth for correct and incorrect bounding boxes using Eq. (9). As can be seen in Fig. 6, q_i is a good indicator of an object's presence. Using logistic regression, we fit a sigmoid to arrive at $P(v_i|\mathbf{c}_i, d_i = 1)$. In Section 5, we also test the use of $P(v_i = 1|\mathbf{c}_i, d_i = 1) = \max P(\mathbf{c}_i|d_i = 1)$. With the same training set as above, we found $P(d_i = 1) \approx 0.96$.

Belief Propagation. The network of Fig. 2 is constructed in Matlab using the BayesNet toolbox [28], with all variables modeled as discrete entities and their conditional probability tables defined as described above. Inference is conducted using Pearl's Belief Propagation [34]. For efficiency reasons, the set of possible ground planes is pruned to the 20% most promising ones (according to prior and depth information).

Resolving Interactions. The Belief Propagation framework does not easily lend itself to expressing the notion of exclusion. When two pedestrians detections overlap in the image (which often happens in our application), their supporting pixels may be overcounted. We therefore employ the global optimization procedure from [12] to let hypotheses compete for pixels in a model selection framework. This procedure is able to resolve interactions between overlapping detection hypotheses and select the most consistent subset.

5 Experiments

In order to evaluate our vision system, we applied it to four additional sequences, showing strolls through busy pedestrian zones. In total, we have 4’000 frames. All sequences were acquired with similar mobile platforms and consist of two synchronized video streams recorded at 13–14 fps.[2] The first test sequence (“Seq. #2”) extends over 1’208 frames. We manually annotated all visible pedestrians in every fourth frame, resulting in 1’894 annotations. The second sequence (“Seq. #3”), with 5’193 annotations in 999 frames, has considerably worse contrast. Finally, as a demonstration of the breaking point of our system, we show two other sequences with fast turns (“Seq. #4”) and an extreme number of moving pedestrians (“Seq. #5”). For testing, all system parameters are kept the same throughout all sequences. We measure performance by comparing generated and annotated bounding boxes and plotting recall over false positives per image.

5.1 Systematic Experiments

The experiments in this section are performed on the training sequence and are used to determine the remaining parameters for the test sequences. First, we consider the sampling steps $\{k, l\}$ for $\mathbf{c}_i$, along with the standard deviation σ_h of the size prior. We consider no sampling, 3×3 ($k, l \in \{-1, 0, 1\}$), and 5×5 ($k, l \in \{-1, -0.5, 0, 0.5, 1\}$) sampling. Fig. 7 (left) shows the resulting detection performance. As expected, a higher σ_h yields better precision at first, but recall grows too slowly. Due to the increased number of choices in Belief Propagation, the use of 5×5 sampling steps has also a negative effect on the performance. By just fixing the object center, recall is limited, as the algorithm cannot compensate for misaligned bounding boxes. A 3×3 sampling with $\sigma_h = 0.12$ thus seems a good compromise.

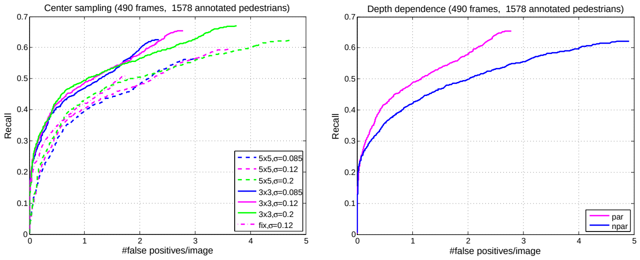


Fig. 7. Left: Influence of center/scale sampling and σ_h on performance. In all future experiments, we use 3×3 sampling and $\sigma_h = 0.12$. Right: Influence of depth term choice on performance, a parametric distribution performs better.

Second, we experimentally establish how to integrate the depth cues into our system. For $P(\mathbf{c}_i|d_i = 1)$, we consider either the learned, non-parametric distribution

[2] Data, including annotations, is available from http://vision.ee.ethz.ch/~aess/.

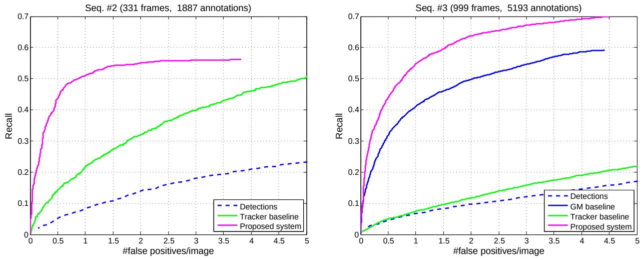
$P(v_i|z(o_i) - z(\mathcal{D}, \mathbf{b}_i))$ ("npar") or a normal distribution inferred from Eq. (7) ("par"). As can be seen from the result plot (Fig. 7, right), the non-parametric distribution for $P(\mathbf{c}_i|d_i=1)$ performs worse. This is mostly due to a relatively small number of samples (especially at larger depths) for creating the necessary tables, as well as to a bias introduced by annotations and the training ground plane.

5.2 Experimental Validation

Fig. 8 shows performance plots for Seqs. #2 and #3. Besides raw detector output ("Detector"), we consider an additional baseline: we emulate the system of [27] by an offline step of running VO, fitting ground planes through wheel contact points, and then running our tracker without depth-map information ("Tracker baseline"). For the first sequence, we also show the output of the original Bayesian network from [12]. Even though our proposed system needs a few frames before initializing a track (losing recall) and even though it reports currently occluded hypotheses (increasing false positives), the single-frame baseline is outperformed.

An interesting observation is the bad performance of the baseline tracker on Seq. #3. Here, the detector yields multiple hypotheses at different scales for many pedestrians. Due to the low camera placement, these cannot be disambiguated by the ground plane alone. Thus, misplaced detections generate wrong trajectories that in turn encourage bad detections, resulting in a very unstable system. Our system breaks this vicious circle by using depth information.

We manually evaluated tracking performance in 450 frames of Seq. #2 using similar criteria as described in [41] (Tab. 8). We consider the number of pedestrians, the number of trajectories (if a pedestrian is occluded for >10 frames, we count a new trajectory), the number of mostly hit trajectories ($> 80\%$ covered), mostly missed trajectories ($<$ 20% covered), the number of false alarms, and the number of ID switches (meaning the tracker drifts from one person to another). On average, 75% of a trajectory are covered by the tracker. The missed trajectories belong mostly to pedestrians at smaller scales and to two children that do not fit the size prior. Example tracking results for Seq. #2 are shown in the first two rows of Fig. 9. Our system's ability to track through occlusion is demonstrated in the top row: please note how the woman entering from the left has temporarily occluded almost every part of the image. Still, the tracker manages to pick up the trajectory of the woman on the right again (in red). Fig. 9 also shows additional tracking results for Seqs.#3, #4, and #5. Again, our system manages to produce long and stable tracks in complex scenarios with a considerable degree of occlusion. In the third row, a pedestrian gets successfully tracked on his way around a few standing people and two pedestrians are detected at far distances. The middle row again demonstrates tracking through major occlusion. Finally, the bottom row shows an example scenario from Seq. #5 with many pedestrians blocking the camera's field-of-view. As mentioned above, scenes of this complexity are at the limit of what is currently possible with our system. Further work is required to address typical failures, such as false positives on trees or reflections and missing detections at too large or small scales.



# persons	30
# trajectories	45
mostly hit	36
mostly missed	6
false alarms	4
ID switches	2

Fig. 8. (Left,Middle) Single-frame detection performance on Seq. #2 and #3. (Right) Quantitative tracking results for part of Seq. #2– see text.

6 Conclusion

In this paper, we have presented an integrated system for multi-person tracking from a mobile platform. The different modules (here, appearance-based object detection, depth estimation, tracking, and visual odometry) were integrated using a set of feedback channels. This proved to be a key factor in improving system performance. We showed that special care has to be taken to prevent system instabilities caused by erroneous feedback. Therefore, a set of failure prevention, detection, and recovery mechanisms was proposed. The resulting system can handle very challenging scenes. Still, there is some way to go before it becomes deployable in a real-world application. The individual components still need to be optimized further, both with respect to speed and performance. For instance, very close pedestrians, with only parts of their torso visible, are often missed by the current detector. A graceful degradation in form of image-based tracking might be a possibility to prevent system breakdown in such cases. Further combinations with other modules, such as world knowledge inferred *e.g.* from map services, provide other exciting feedback possibilities that we plan to investigate in the future.

Acknowledgments. This project has been funded in parts by Toyota Motor Corporation/Toyota Motor Europe and the EU projects DIRAC (IST-027787) and HERMES (IST-027110).

References

1. M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *CVPR*, 2008.
2. S. Avidan. Ensemble tracking. In *CVPR*, 2005.
3. H. Bay, A. Ess, T. Tuytelaars, and L. van Gool. Speeded-up robust features (SURF). *CVIU*, 110(3):346–359, 2008.
4. J. Berclaz, F. Fleuret, and P. Fua. Robust people tracking with global trajectory optimization. In *CVPR*, 2006.
5. C. M. Bishop. *Pattern recognition and machine learning*. Springer Verlag, 2006.
6. D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE TPAMI*, 25(5):564–575, 2003.
7. I. J. Cox. A review of statistical data association techniques for motion correspondence. *IJCV*, 10(1):53–66, 1993.

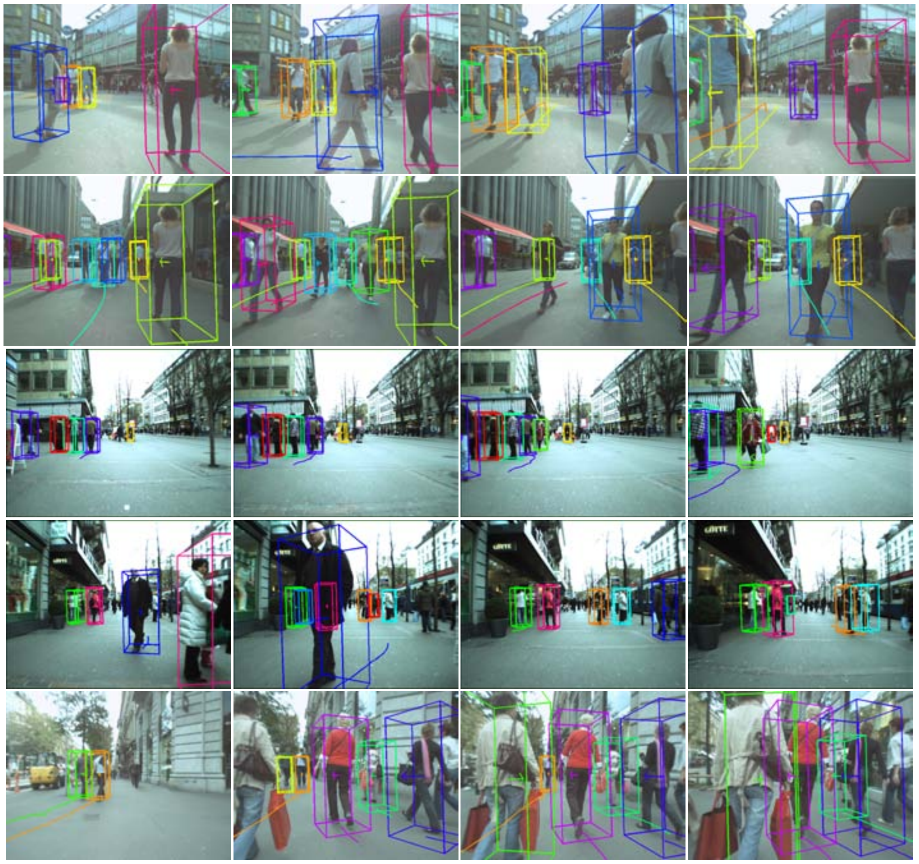


Fig. 9. Top two rows: exemplary subsequences from Seq.#2. Note the long trajectories and ability of the tracker to handle temporary occlusions in complex scenarios. Other rows: selected tracking results for Seqs. #3, #4, and #5.

8. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR'05*.
9. N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV*, 2006.
10. A. J. Davison. Real-time simultaneous localization and mapping with a single camera. In *ICCV*, 2003.
11. A. Ess, B. Leibe, K. Schindler, and L. van Gool. A mobile vision system for robust multi-person tracking. In *CVPR*, 2008.
12. A. Ess, B. Leibe, and L. van Gool. Depth and appearance for mobile scene analysis. In *ICCV*, 2007.
13. P. F. Felzenszwalb and D. P. Huttenlocher. Efficient belief propagation for early vision. *IJCV*, 70:41–54, 2006. Available from http://people.cs.uchicago.edu/~pff/bp/.
14. T.E. Fortmann, Y. Bar Shalom, and M. Scheffe. Sonar tracking of multiple targets using joint probabilistic data association. *IEEE J. Oceanic Engineering*, 8(3):173–184, 1983.

15. D. M. Gavrila and S. Munder. Multi-cue pedestrian detection and tracking from a moving vehicle. *IJCV*, 73:41–59, 2007.
16. A. Gelb. *Applied Optimal Estimation*. MIT Press, 1996.
17. J. Giebel, D.M. Gavrila, and C. Schnörr. A bayesian framework for multi-cue 3d object tracking. In *ECCV*, 2004.
18. H. Grabner and H. Bischof. On-line boosting and vision. In *CVPR'06*, 2006.
19. R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.
20. D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. In *CVPR*, 2006.
21. M. Isard and A. Blake. CONDENSATION–conditional density propagation for visual tracking. *IJCV*, 29(1), 1998.
22. R. Kaucic, A. G. Perera, G. Brooksby, J. Kaufhold, and A. Hoogs. A unified framework for tracking through occlusions and across sensor gaps. In *CVPR*, 2005.
23. R. Labayrade, D. Aubert, and J.-P. Tarel. Real time obstacle detection on non flat road geometry through 'v-disparity' representation. In *IVS*, 2002.
24. O. Lanz. Approximate bayesian multibody tracking. *IEEE TPAMI*, 28(9):1436–1449, 2006.
25. B. Leibe, N. Cornelis, K. Cornelis, and L. van Gool. Dynamic 3d scene analysis from a moving vehicle. In *CVPR*, 2007.
26. B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *IJCV*, 77(1-3):259–289, 2008.
27. B. Leibe, K. Schindler, , and L. van Gool. Coupled detection and trajectory estimation for multi-object tracking. In *ICCV*, 2007.
28. K. Murphy. The bayes net toolbox for Matlab. In *Computing Science and Statistics*, 2001.
29. K. Murphy, A. Torralba, and W. T. Freeman. Using the forest to see the trees: A graphical model relating features, objects, and scenes. In *NIPS*, 2003.
30. D. Nistér. A minimal solution to the generalised 3-point pose problem. In *CVPR*, 2004.
31. D. Nistér, O. Naroditsky, and J. R. Bergen. Visual odometry. In *CVPR*, 2004.
32. K. Okuma, A. Taleghani, N. de Freitas, J. Little, and D. Lowe. A boosted particle filter: Multitarget detection and tracking. In *ECCV*, 2004.
33. B. Ommer and J. M. Buhmann. Object categorization by compositional graphical models. In *EMMCVPR*, 2005.
34. J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann Inc., 1988.
35. D. B. Reid. An algorithm for tracking multiple targets. *IEEE T. Automatic Control*, 24(6):843–854, 1979.
36. D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47:7–42, 2002.
37. V. Sharma and J. Davis. Integrating appearance and motion cues for simultaneous detection and segmentation of pedestrians. In *ICCV*, 2007.
38. E. B. Sudderth, A. B. Torralba, W. T. Freeman, and A. S. Willsky. Learning hierarchical models of scenes, objects, and parts. In *ICCV*, 2005.
39. O. Tuzel, F. Porikli, and P. Meer. Human detection via classification on riemannian manifolds. In *CVPR*, 2007.
40. P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *ICCV*, 2003.
41. B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet part detectors. *IJCV*, 75(2):247–266, 2007.
42. B. Wu and R. Nevatia. Simultaneous object detection and segmentation by boosting local shape feature based classifier. In *CVPR*, 2007.
43. P. Yan, S. M. Khan, and M. Shah. 3d model based object class detection in an arbitrary view. In *ICCV*, 2007.
44. L. Zhang, Y. Li, and R. Nevatia. Global data association for multi-object tracking using network flows. In *CVPR*, 2008.