Automatic detection and tracking of pedestrians from a moving stereo rig

Konrad Schindler[a], Andreas Ess[b], Bastian Leibe[c], Luc Van Gool[b,d]

[a]*Photogrammetry and Remote Sensing, ETH Zürich, Switzerland*
[b]*Computer Vision Lab, ETH Zürich, Switzerland*
[c]*UMIC research centre, RWTH Aachen, Germany*
[d]*ESAT/PSI–VISICS, IBBT, KU Leuven, Belgium*

Abstract

We report on a stereo system for 3D detection and tracking of pedestrians in urban traffic scenes. The system is built around a probabilistic environment model which fuses evidence from dense 3D reconstruction and image-based pedestrian detection into a consistent interpretation of the observed scene, and a multi-hypothesis tracker to reconstruct the pedestrians' trajectories in 3D coordinates over time. Experiments on real stereo sequences recorded in busy inner-city scenarios are presented, in which the system achieves promising results.

Keywords:

1. Introduction

Automotive safety and autonomous navigation are emerging as important new application areas of close-range photogrammetry. The goal in such applications is to equip a vehicle or robot with cameras, and automatically derive a metric and semantic model of the platform's environment from the recorded image sequences. In road scenes, a particularly important part of such an environment model are the pedestrians. Knowing their locations and motion trajectories is an essential prerequisite for safe navigation, path planning, and collision prevention (Shashua et al., 2004; Gavrila and Munder, 2007; Wedel et al., 2008; Ess et al., 2009a). The topic of this paper is the detection and tracking of people with a stereo camera rig mounted on a moving camera platform.

The described task requires a combination of geometric 3D modelling to obtain a metric environment model, and image understanding to find the people in the observed scene. Furthermore processing must be done online, i.e. at any given

(a) *CharioBot* (b) *CharioBot II* (c) *SmartTer*

Figure 1: Recording platforms used in this work. (a), (b) stereo rig mounted on child strollers. (c) stereo rig mounted on *SmartTer* robotic car. Only synchronised stereo videos serve as measurement data, the further sensors of the *SmartTer* platform were not used.

time the state of the environment must be estimated using only data observed in the past and present. Tracking people in 3D coordinates from a moving vehicle is a challenging combination of several classic problems:

- to establish a 3D reference frame for tracking, the platform's ego-motion needs to be estimated, which amounts to recovering the position and orientation of the stereo rig at each frame in a common coordinate system.
- the people within the cameras' field of view must be detected in the images, and then localised in the 3D reference system.
- the per-frame detections of each individual must be connected over time to form pedestrian trajectories in 3D world coordinates.

In this paper we report on a system for detecting and tracking pedestrians from moving vehicles. The described system uses only stereo vision as input (the recording setup is depicted in Fig. 1), however we stress that the framework is generic: although we use only stereo video in the present study, other sensors like LIDAR, GPS/IMU, conventional odometry, and possibly thermal cameras could be useful for the task. If available, such sensors should be added, and would certainly improve performance. We do however point out that in the automotive sector, and even more in robotics, there is a desire to limit the amount of sensor hardware, and that stereo images are at present the most successful sensor for detecting and localising humans during daytime (e.g. thermal cameras work well for detection at night and to a certain extend during the day, but it is not possible to reliably recover dense 3D depth; LIDAR delivers highly accurate 3D geometry, but in moving platforms is limited to one or a small number of scan-lines, and does not enable robust object recognition).

As building blocks for the presented system, we use several methods of photogrammetry and computer vision, which generate different measurements from the input images: automatic camera orientation is performed to obtain the ego-motion in a 3D reference frame (Sec. 2.1). Automatic image matching is applied to the stereo pair in each frame to obtain dense 3D depth measurements (Sec. 2.2), and robust geometric fitting in the dense 3D point cloud yields observations for the current ground plane (Sec. 2.2). Appearance-based pedestrian detection delivers further observations, which indicate the putative presence and location of people in the field of view (Sec. 2.3).

To fuse all these observations on a per-frame basis, we then introduce a probabilistic model of scene geometry, which combines the measured evidence to obtain a *maximum a posteriori* estimate of the ground plane as well as the 3D locations of pedestrians (Sec. 3). The model allows one to fuse the available evidence in a principled way, while still being simple enough to allow efficient inference.

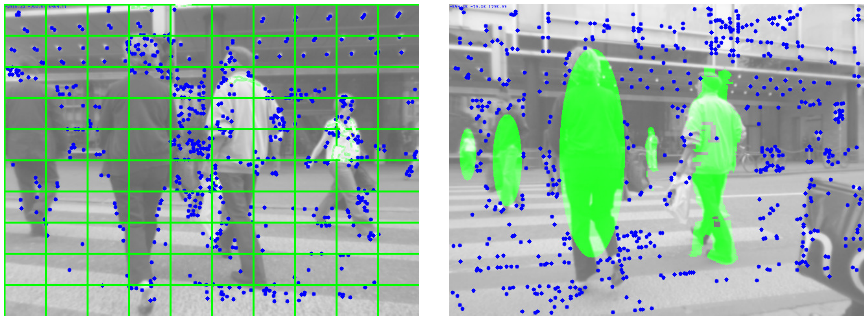
In a second step, the per-frame results are integrated over time to yield an optimal estimate of the platform's environment for the entire observation time up to and including the current frame (Sec. 4). Due to the high number of interacting people in urban traffic scenes, simply tracking each person independently is not sufficient for this step. We therefore include interactions between different people in the representation, which increases its modelling power and substantially improves results in practice.

Finally, we give an extensive experimental evaluation on several long and challenging real-world stereo sequences, in order to assess performance both quantitatively and qualitatively (Sec. 5). The paper ends with a discussion and outlook (Sec. 6). Some rather lengthy mathematical details have been collected in an appendix.

2. Pre-processing

2.1. Camera Orientation

In order to model and track pedestrians in 3D, a common reference frame must be established for the video data collected along the vehicle's path. This amounts to solving for the six parameters of the stereo rig's absolute orientation



(a) (b)

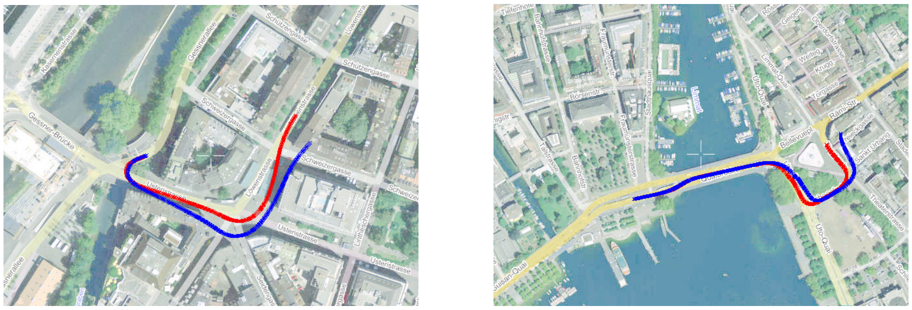
Figure 2: Camera resection. (a) feature binning ensures that the point distribution is suitable for localisation. (b) tracked pedestrians are masked out, since they move w.r.t. the background scene.

in every frame.[1] An obvious way of determining the absolute orientation is to equip the platform with a GPS/IMU unit and measure position and orientation directly (“direct geo-referencing”), possibly also including odometer readings.

A different approach is classical photogrammetric triangulation: in applications where video needs to be recorded anyway (e.g. robotics mapping) it is becoming more and more popular to determine the camera orientation from observed scene points by resectioning. This can nowadays be performed robustly in real-time (“visual odometry”, [e.g. Davison, 2003; Nistér et al., 2004; Ess et al., 2008; Mei et al., 2009). For simplicity, the latter method is used in the experiments reported here: ego-motion estimation is purely visual. This proved to be sufficiently accurate for pedestrian tracking, although it would obviously be beneficial to also include GPS, IMU and/or odometry.

The employed processing pipeline is straightforward: in each frame, the incoming images are divided into a grid of 10×10 bins, see Fig. 2. Image regions corresponding to tracked people are masked out, since they violate the assumption of a static scene (c.f. Ess et al., 2008). In the unmasked part of the image, feature points are detected with the Förstner corner detector (Förstner and Gülch, 1987) with locally adaptive thresholds, such that the number of points per bin is approximately constant. This binning improves the feature distribution in the presence of uneven contrast. The local structure around the corner points is then described by robust SURF descriptors (Bay et al., 2008).

[1] In the general case also the interior and relative orientations may need to be determined. For our stereo rig we have confirmed that the calibration is stable.



(a) (b)

Figure 3: Camera trajectories for Seq. LOEWENPLATZ and Seq. BELLEVUE, obtained by terrestrial camera triangulation. Red: with bundle adjustment and using double precision. Blue: without bundle adjustment and using single precision on the GPU (computation time < 20 ms per frame, applicable under hard real-time constraints).

In the first frame initial 3D points are reconstructed by matching the SURF descriptors and triangulating the corresponding image points. The SURF vectors are stored as appearance descriptors for the triangulated 3D points. In each subsequent frame the image corners are matched directly to the 3D structure points, using a Kalman filter to predict the camera position and constrain point matching accordingly, similar to the "active search" paradigm in robotic SLAM (e.g. Davison, 2003).

With the 2D-3D correspondences, the new camera orientation is found by robust resection (RANSAC estimation of 3-point pose), and the SURF descriptors of the 3D points are updated. Bundle adjustment is run on a sliding window of 18 past frames to polish the camera parameters and scene points. The camera parameters of older frames are discarded, as are the 3D points only supported by the removed frames. Importantly, points are remembered until they have not been matched over 18 consecutive frames, so that short occlusions (e.g. by a person) can be bridged. The robustness of SURF against viewpoint changes makes it possible to re-detect points after several frames.

The system is implemented largely on the graphics card, taking advantage of both GPU-SURF (Cornelis and Van Gool, 2008) for feature description and the parallel nature of RANSAC to simultaneously generate and test multiple hypotheses for the camera pose.

In our specific application, where the aim is not precise 3D scene reconstruction, but a reference frame for people detection and tracking, gradual drift of the camera path does not hurt. Hence it is even possible to limit least-squares adjustment only to the newly estimated orientation parameters, if computation time is an issue.

Sample camera trajectories for the *SmartTer* platform are shown in Fig. 3, both with bundle adjustment over 18 frames, and with adjustment of only the last frame. The average uncertainty of the camera position is $\sigma_{\mathbf{x}} = \pm\, 1.4\,\mathrm{cm}$ with adjustment over 18 frames, respectively $\sigma_{\mathbf{x}} = \pm\, 2.0\,\mathrm{cm}$ when only adjusting the newly added viewpoint. The standard deviations of the viewing direction are $\sigma_{\psi} = \pm\, 0.49°$, respectively $\sigma_{\psi} = \pm\, 0.64°$. Note, the standard deviations attest only to the local smoothness of the camera paths, whereas the lack of tie points between distant frames leads to considerable drift over time, which as expected is a lot stronger if only adjusting a single new viewpoint.

2.2. Dense Depth

Since we are aiming for a 3D environment model, the scene depth w.r.t. the stereo rig must be measured. Again there are two main alternatives, namely direct range sensing, or dense image matching followed by stereo triangulation.

While direct range measurement with LIDAR may seem the obvious choice, it has some important disadvantages: first of all it has significantly higher weight and power consumption than passive sensors, which can be important on moving platforms; second, and more importantly, practical LIDAR systems measure range by sequentially scanning the field of view, which means that covering the relevant solid angle at an appropriate resolution takes a significant amount of time (typically several seconds). Hence, depth maps are not available at an adequate frame-rate, and when recorded from a fast-moving platform are also distorted by the ego-motion. Additionally, thin objects are not well modelled because of the limited angular resolution: the resolution of a typical high-speed laser scanner is $0.5°$ (0.17 m sampling distance at a range of 20 m); in comparison, the radial resolution of our *SmartTer* setup is $0.07°$. We hence prefer to recover depth from stereo images, in spite of the lower range accuracy. Still, sensor fusion is an important option to consider in future work.

Another option for 3D localisation of people detected in an image is not to measure depth, but instead project the foot point of a person from the image to the ground plane (Gavrila and Munder, 2007; Hoiem et al., 2006; Leibe et al., 2008; Havlena et al., 2009). While this method is also applicable with monocular video, it is considerably less accurate: on the one hand, 2D detection accuracy is rather low (typically about $\pm\, 5$ pixels), and localisation errors in the image are greatly amplified, because the corresponding rays intersect the terrain at grazing angles; on the other hand, the ground surface itself cannot be reconstructed accurately

Figure 4: Stereo depth maps for an example image pair from Seq. LOEWENPLATZ. middle: local smoothing, right: global optimisation. Parts that are believed to be inaccurate (by a left-right check) are painted black. Advanced algorithms give visually better results, but take more time and are often not necessary.

with the recording geometry of realistic vehicles (see Sec. 2.2). We thus believe that measuring depth is currently inevitable for 3D environment modelling.

For a calibrated stereo pair, estimating depth is equivalent to estimating image disparity: w.l.o.g. the two images can be assumed to be in standard configuration, i.e. their epipolar lines are horizontal and corresponding lines have the same y-coordinate. Hence, disparity is inversely proportional to depth, and its estimation amounts to a 1D search for the best-matching pixel. Due to the nonlinear relationship between disparity and depth, it is important to properly account for the uncertainty in all subsequent computations, see Appendix A.

Nowadays, a plethora of stereo algorithms is available. For an overview and taxonomy see Scharstein and Szeliski (2002), or for a more recent update the associated *Middlebury Stereo Evaluation Page*.[2] The main requirements for an algorithm in our application are speed and the ability to handle lack of texture. We present two representative methods from different extremes of the spectrum. Example outputs on a typical street scene are shown in Fig. 4. The fastest breed of stereo matchers at present are methods which alternate between depth estimation and smoothing of the disparity field. All operations are local and can be carried out in parallel. This allows for GPU implementations which take less than 20 ms per VGA image, e.g. Cornelis and Van Gool (2005). On the other end of the spectrum, the best results under difficult conditions are achieved by methods based on global optimisation of an appropriately designed energy function. An excellent recent example is the method of Zach et al. (2009). The downside is that even when implemented on modern GPUs, computation times per image pair exceed 1 s.

In the context of our system, where robust methods are used to derive higher-level cues from raw depth, we observe that top-of-the-line stereo methods bring

[2] `http://vision.middlebury.edu/stereo/`

little improvement at the system level, in spite of visually superior depth maps – see experimental results in Sec. 5.

Confidence map. Disparity estimation will not be accurate everywhere, due to problems such as occlusions, specularities, untextured areas and over-smoothing. Usually, algorithms simply ignore these problems and return incorrect results. To prevent such measurement errors from propagating, we try to label bad pixels according to the following two rules:

- *Appearance.* If the sum of absolute intensity differences between the neighbourhoods of two matched pixels exceeds a threshold, the pixel is labelled as occluded. This identifies most mistakes due to occlusion.
- *Disparity.* In untextured areas depth is filled in by assuming smoothness of the scene. If that assumption is not justified, smoothing will give different results depending on the viewpoint. Therefore, the disparity w.r.t. the left image will differ from the one w.r.t. the right image for such pixels. The further condition that the two disparities must be the same identifies most incorrect labels in untextured regions.

This binary labelling will be captured in a confidence map $\mathcal{C}$, with $\mathcal{C}(\mathbf{p}) = 1$ indicating a valid pixel $\mathbf{p}$, and $\mathcal{C}(\mathbf{p}) = 0$ an invalid one, for which no reliable disparity could be estimated (black pixels in Fig. 4). As can be seen, the simplistic smoothing of the GPU-based estimator results in far more invalid pixels. These pixels will be ignored in subsequent steps.

Ground plane. An important part of the environment model for navigation is the terrain on which both the moving platform and the people move. It substantially helps pedestrian detection through the twin constraints that people should stand on the ground and that their height should be that of a human (Hoiem et al., 2006; Ess et al., 2007; Gavrila and Munder, 2007; Leibe et al., 2008). The low viewpoint and limited resolution of vehicle-mounted cameras do not allow one to reliably recover the DTM, therefore we opt for a local approximation: the terrain is modelled as a plane, which is robustly fitted to the 3D points in front of the platform, and dynamically updated in every video frame, to adapt terrain undulations and vehicle tilt due to the suspension.

The plane is parametrised in normal form in the camera coordinate system as $\boldsymbol{\pi} = (\mathbf{n}, \pi^{(4)})$, with the normal vector given in spherical coordinates: $\mathbf{n}(\theta, \phi) = (\cos\theta \sin\phi, \sin\theta \sin\phi, \cos\phi)$.

The ground plane is not determined from the depth map directly, which is unreliable in scenarios like ours, where it is not easy to decide which depth points

Figure 5: Calculation of ground plane evidence is distributed over several stripes of decreasing size in order to alleviate the effect of uneven sampling.

really belong to the terrain. Instead, it is inferred jointly with the pedestrians, using the depth map as uncertain measurement – see Sec. 3. To this end a distribution $P(\boldsymbol{\pi}|\mathcal{D}) \sim P(\mathcal{D}|\boldsymbol{\pi})P(\boldsymbol{\pi})$ over the ground plane parameters must be defined, which measures the probability of a certain parameter vector $\boldsymbol{\pi}$, given the observed depth map $\mathcal{D}$. To measure the goodness-of-fit and define $P(\mathcal{D}|\boldsymbol{\pi})$, we consider the depth-weighted median residual between $\boldsymbol{\pi}$ and the depth map $\mathcal{D}$, averaged over three horizontal stripes $\mathcal{S}_i$ (to account for unequal sampling):

$$r_i(\boldsymbol{\pi}, \mathcal{D})^2 = \underset{\{\mathbf{p}\in\mathcal{S}_i|\mathcal{C}(\mathbf{p})=1\}}{\mathrm{med}} \left(\frac{1}{\sigma_{\mathcal{D}}^2} (\mathbf{n}^\top \mathcal{D}(\mathbf{p}) - \pi^{(4)})^2 \right) , \tag{1}$$

$$r(\boldsymbol{\pi}, \mathcal{D})^2 = \frac{1}{3} \Big(\sum_{i=1}^{3} r_i(\boldsymbol{\pi}, \mathcal{D}_i)^2 \Big) \quad . \tag{2}$$

Here $\mathbf{p} \in \mathcal{S}_i$ denotes the pixels from a vertical stripe of $\mathcal{D}$, deemed valid by the confidence map ($\mathcal{C}(\mathbf{p}) = 1$). To account for the decreasing number of points at greater distances, the height $h_y(i)$ of the stripes $\mathcal{S}_i$ increases towards the lower image border (we use the progression $h_y(i) = \frac{h}{2(i+1)} = \{120, 80, 40\}$, with h the total image height; see Fig. 5). $\sigma_{\mathcal{D}}$ accounts for the uncertainty of the plane-to-point distance. Given this robust estimate, we set

$$P(\mathcal{D}|\boldsymbol{\pi}) \sim e^{-r(\boldsymbol{\pi},\mathcal{D})^2} \quad . \tag{3}$$

In the scene model, this distribution is complemented with an empirically learnt ground plane prior $P(\boldsymbol{\pi})$ and combined with evidence from pedestrian detection to fit the most likely plane; see Appendix B.

2.3. Pedestrian Observations

Evidence for the presence of people is generated by running a state-of-the-art pedestrian detector. Methods for recognising and localising people in images can be broadly grouped into two types: those which generate hypotheses by evidence aggregation (e.g. Leibe et al., 2005; Felzenszwalb et al., 2008), often using part-based human body models; and sliding-window methods, which exhaustively scan all positions and scales of the input image and for each window return a detection score, i.e. a pseudo-likelihood that the window contains a pedestrian. So far, the sliding-window approach has proved more successful in practice, despite its conceptual simplicity.

Since the pioneering works of Papageorgiou and Poggio (2000) and Viola et al. (2003), many improvements of the basic sliding-window method have been proposed. The most common features are variants of the HOG framework, i.e. local histograms of gradients (Dalal and Triggs, 2005; Felzenszwalb et al., 2008; Wang et al., 2009), and different flavours of generalised Haar wavelets, e.g. (Viola et al., 2003; Dollar et al., 2009). Classifiers are mostly standard methods from statistical learning, predominantly support vector machines (Shashua et al., 2004; Dalal and Triggs, 2005; Sabzmeydani and Mori, 2007; Lin and Davis, 2008) and variants of boosting (Viola et al., 2003; Zhu et al., 2006; Wu and Nevatia, 2007; Wojek et al., 2009).

For automotive applications, two recent surveys (Dollar et al., 2009; Enzweiler and Gavrila, 2009) conduct extensive experiments over several hours of urban driving to assess the performance of current detection algorithms. In short, it turns out that for large and medium-sized pedestrians ($>$ 50 pixels) the HOG (histogram of oriented gradients) feature of Dalal and Triggs (2005) performs very well even with a linear SVM classifier. Another advantage of HOG is that it is highly parallelisable – GPU implementations exceed 10 frames per second on VGA-size images (Wojek et al., 2008).

In the present work we have used the standard HOG approach. In a nutshell, HOG collects 3D histograms over the (x, y)-location and gradient orientation within the sliding window. Each pixel's contributes to the histogram is weighted with the local gradient magnitude, and the histogram entries are normalised over larger regions of (2×2) bins. All histogram bins are then concatenated to a feature vector and classified with a linear SVM. For details we refer to the original publication (Dalal and Triggs, 2005).

Following the original work, we scan down to a minimum window height of 48 pixels. This corresponds to a maximum distance of about 19 m for the child strollers (*CharioBot*, *CharioBot II*) and 30 m for the *SmartTer* platform, both assuming a pedestrian height of 1.8 m. In future work, we plan to also include optic flow between consecutive frames, which has been shown to consistently improve detection in a dynamic environment (Wojek et al., 2009; Walk et al., 2010). We emphasise that the output of people detection is not regarded as final result, but rather as one more type of image measurement to be considered during inference. The detector is set to a low threshold to generates *hypotheses*, such that it may produce false alarms, but misses as few actual people as possible.

3. Single-frame inference

In real images of urban environments, the automatically generated measurements described in the previous section will not always be correct. Appearance-based pedestrian detection tends to become unreliable in low-contrast regions, in the far field, and in the presence of (partial) occlusion, which frequently occur between different people in the scene. Stereo matching returns inaccurate and even grossly wrong depths in homogeneous image areas and around specular reflections. The accuracy of ground plane fitting depends both on the quality of the underlying depth estimates and on an unobstructed view of the ground, much of which is at times occluded by people, vehicles, and street furniture.

We therefore treat the observations made by image processing and computer vision algorithms not as final results, but as noisy observations, from which a consolidated, consistent environment model shall be derived. In the following section we describe a probabilistic way to jointly exploit the observations. For the moment, we will restrict the discussion to a *single* stereo pair. Using input from pedestrian detection and dense stereo, we want to find the correct ground plane, identify the true people among the detector responses, and localise them in the 3D reference frame.

By mapping the problem to a *Bayesian network*, inference can be conducted such that an optimal solution is found based on all input observations (Ess et al., 2009b). A good example to illustrate how clean probabilistic modelling allows for more reliable estimates is the ground plane: if it covers a large part of the image, it can be robustly estimated from depth, and strongly constrains pedestrians, by penalising people not standing on the ground; conversely, for scenes crowded with people, independent ground plane estimation is bound to fail because too little of the ground is visible – but the people themselves will constrain the ground plane, since a consensus is required such that all pedestrians stand on the same plane. In the Bayesian network both cases are naturally accounted for in a single

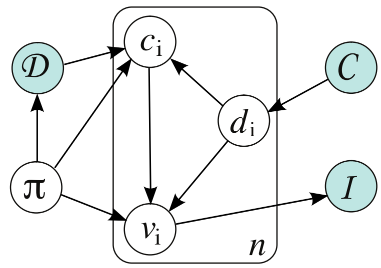


Figure 6: Probabilistic scene model for single-frame inference. For a given stereo pair, the observed evidence consists on the one hand of the pedestrian detection scores $\mathcal{I}$ in the two images, and on the other hand of the depth map $\mathcal{D}$ and the associated confidence map $\mathcal{C}$. The unknown quantities that need to be inferred are the ground plane parameters $\boldsymbol{\pi}$, the presence or absence v_i of a potential pedestrian in the most likely model, and the locations $\mathbf{c}_i$ of all present pedestrians. The auxiliary variable d_i indicates whether the depth is reliable for the bounding box of a potential pedestrian.

model. The network is shown in Fig. 6. Following standard graphical model notation (Bishop, 2006), the plate denotes n-fold repetition of the contained parts (corresponding to the n potential pedestrians).

The input of the model are a set of potential pedestrian detections $o_i = \{\mathbf{c}_i, v_i\}$ found by analysing the two images $\mathcal{I}$ of the stereo pair, the depth map $\mathcal{D}$ of the stereo pair, and the associated confidence map $\mathcal{C}$.[3] The unknown variables to be determined are the three parameters $\boldsymbol{\pi}$ of the ground plane, a binary flag v_i for each bounding box declaring it valid or invalid, and the locations $\mathbf{c}_i = (x_i, y_i)$ of all valid boxes.

For each potential person, back-projection of the bounding box onto the ground plane yields a 3D location $\mathbf{x}$ and height h. Its distance to the camera should then coincide with the dominant stereo depth inside the bounding box (within the uncertainty bounds). The height h should correspond to the expected height of humans, represented by a Gaussian distribution. Furthermore, the bounding box is more likely to correspond to a person if its detection score is higher, and if the depth of most pixels inside the bounding box is constant within the measurement accuracy. Finally, the ground plane should match the observed scene depths, while at the same time passing through the foot points of the valid people.

[3] Note, a simplification is made by considering the detection scores, the depth map, and the confidence map as independent, although they are ultimately all derived from the same image intensities.

MAP Estimation. Inference in the model is performed according to the factorisation

$$
\begin{aligned}
&P(\boldsymbol{\pi}, \mathbf{c}_i, v_i, d_i, \mathcal{I}, \mathcal{C}, \mathcal{D}) \sim \\
&\quad \sim P(\boldsymbol{\pi})P(\mathcal{D}|\boldsymbol{\pi})\prod_i P(\mathbf{c}_i|\boldsymbol{\pi}, \mathcal{D}, d_i)P(v_i|\mathbf{c}_i, \boldsymbol{\pi})P(v_i|d_i)P(d_i|\mathcal{C})P(\mathcal{I}|v_i)\ .
\end{aligned} \tag{4}
$$

The probability for a certain person location $\mathbf{c}_i$ depends on the geometric consistency of depth map and ground plane localisation, $P(\mathbf{c}_i|\boldsymbol{\pi}, \mathcal{D}, d_i)$. The validity flag v_i, which indicates whether at a certain position a pedestrian is present or absent, depends both on the person's geometric location and size $P(v_i|\mathbf{c}_i, \boldsymbol{\pi})$, and on the depth distribution in the bounding box $P(v_i|d_i)$. The detection likelihood $P(\mathcal{I}|v_i)$ is derived from the detector score of hypothesis o_i. $P(d_i|\mathcal{C})$ encodes the reliability of the depth map. The variables, along with their domains, are also summarised in Tab. B.1. Detailed definitions for the single terms in Eq. (4) are given in Appendix B.

All 3D calculations are done in camera-centric coordinates, i.e. the camera orientation is $\mathbf{P} = (\mathtt{I}, \mathbf{0})$. This not only simplifies calculations, but also keeps the ground plane parameters in a limited range that can be meaningfully trained. For the subsequent tracking stage, the results are transformed into world coordinates by applying the known absolute orientations.

The graph of Fig. 6 is constructed for each frame of the video sequence. Once all probabilities have been defined, joint inference over all variables is performed by maximising the posterior, which can be done efficiently with Belief Propagation (BP, Pearl, 1988): after discretising all variables and filling in their conditional probability tables (CPTs) as described in the appendix, sum-product BP yields the posterior marginals of the variables. Due to the loopy nature of our model, BP is not guaranteed to find a global optimum, but in practice it nevertheless works very well, a finding also confirmed by other researchers (e.g. Murphy et al., 1999). The results of single-frame inference form the input for the subsequent tracking step.

4. Object Tracking

Given the output of single-frame inference, tracking amounts to fitting a set of trajectories to the detected people in 3D world coordinates, such that these trajectories together explain the observations over time well, i.e. they have a high posterior probability.

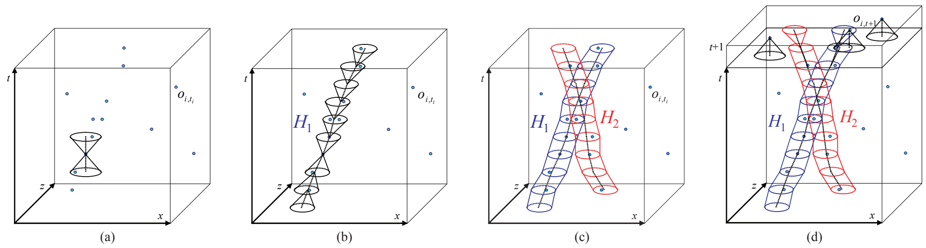


Figure 7: Generating candidate trajectories. (a) Starting from an object detection, detections in nearby frames are found which are within reach according to the dynamic model. (b), (c) Based on the new detections, the trajectory is adapted. Adding new detections and updating the trajectory are iterated forward and backwards in time. (d) For efficiency reasons, trajectories are grown incrementally.

Since standard 1st-order Markov tracking frequently fails in multi-target scenarios, we employ a hypothesise-and-verify strategy to find the set of trajectories that best explains the evidence from past and present frames. The *hypothesise* step samples a large, over-complete set of candidate trajectories with standard methods, and the *verify* step selects an optimal subset and discards the remaining candidates.

The basic units of the tracker are *candidates* for possible object trajectories. A candidate trajectory is defined as $H_j = [\mathcal{S}_j, \mathcal{M}_j, \mathcal{A}_j]$, with $\mathcal{S}_j$ the supporting detections, $\mathcal{M}_j$ its dynamic model, and $\mathcal{A}_j$ its appearance model. At each time step, an exhaustive set of plausible candidates is instantiated, and pruned to a minimal consistent subset.

Dynamic model. As dynamic model for candidate generation, we assume a constant velocity vector in 2D ground plane coordinates. Only few dynamic models are in common use: when tracking in 3D, the constant velocity assumption is the standard choice (e.g. Gavrila and Munder, 2007). When tracking in the image plane, 3D position is replaced by 2D position and object scale (Wu and Nevatia, 2007; Zhang et al., 2008), usually again with a 1st-order dynamic model. Few authors have investigated higher-order models for erratic motions such as in sports (e.g. Okuma et al., 2004).

In our implementation, we employ a standard Extended Kalman Filter (EKF, Gelb, 1996) to describe an individual object's motion pattern. Specifically, we use an extension of linear Kalman filtering with a uni-modal Gaussian distribution of the current state, and 1st-order (constant velocity) motion. The model is specified by defining the transition function $f^{\mathcal{M}}(\cdot)$ and the measurement function $f^{\mathcal{X}}(\cdot)$

(the observed location), and their respective Jacobians. The state space is $\mathbf{s}_t = [x_t, y_t, \theta_t, v_t]^\top$, with (x_t, y_t) the 2D position, θ_t the person's orientation, and v_t their speed. The latter two are initialised to 0, since for the first detection the speed and orientation are unknown. The transition function is

$$f^{\mathcal{M}}(\mathbf{s}_{t-1}, w_{t-1}) = \begin{pmatrix} x_{t-1} + v_{t-1}\cos(\theta_{t-1})\Delta t \\ y_{t-1} + v_{t-1}\sin(\theta_{t-1})\Delta t \\ \theta_{k-1} \\ v_{t-1} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ w_\theta \\ w_v \end{pmatrix} , \tag{5}$$

where w_θ and w_v are additive random noise in the orientation and velocity, respectively. Given a current position $\mathbf{x}_t^s$, the likelihood of an object o_i located at $\mathbf{x}_i$ under the motion model is

$$p(o_i|\mathcal{M}_j) \sim e^{-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_t^s)^\top (\mathtt{C}_t + \mathtt{C}_{x_i})^{-1}(\mathbf{x}_i - \mathbf{x}_t^s)} \quad . \tag{6}$$

Here, $\mathtt{C}_t$ is the covariance matrix specifying the uncertainty in the system, and $\mathtt{C}_{x_i}$ is the localisation uncertainty of the detection, estimated from the stereo geometry (Appendix A). The latter is especially important to handle far away objects correctly, for which the depth uncertainty is high. Correct uncertainty modelling is crucial to achieve good tracking results across a large depth range.

Observation model. We follow the *tracking-by-detection* approach and use the output of the HOG detector, together with a colour histogram in HSV space, as observation. The observation model for visual tracking has evolved a lot over the years. Early approaches often employed background subtraction (Stauffer and Grimson, 1999; Toyama et al., 1999), which is not applicable for moving cameras. Many also rely on low-level image cues such as edges (Isard and Blake, 1998) or local regions (Bibby and Reid, 2008) as observations, which are notoriously unstable. The most successful approach in recent years has been *tracking-by-detection*, which regards the output of an object detector as observation (Okuma et al., 2004; Avidan, 2005; Gavrila and Munder, 2007; Wu and Nevatia, 2007; Zhang et al., 2008; Leibe et al., 2008). For a richer description, the observation is often augmented with local image statistics, mostly colour histograms (e.g. Nummiaro et al., 2003; Okuma et al., 2004; Wu and Nevatia, 2007).

The basis for tracking-by-detection are the pedestrian detections $o_i^{t_i} = [\mathbf{x}_i, \mathtt{C}_i, t_i, a_i]$, where $\mathbf{x}_i$, $\mathtt{C}_i$ are the 2D position on the ground plane and its uncertainty, t_i is the frame index, and a_i the colour histogram describing the appearance. For a given frame t_i, we denote by $P(o_i^{t_i}|\mathcal{I}^{t_i})$ the probability of a person being present given the image evidence (in the following, the superscript t_i in omitted whenever it is clear from the context). Detections are accumulated in a space-time volume $\mathcal{O}$ that spans all past frames up to and including the current one. In practice, only the last few hundred time steps are considered, starting at some frame t_0. The purpose of tracking hence is to fit smooth trajectories H_j to the locations $[\mathbf{x}_i, t_i]^\top$ within $\mathcal{O}$.

While $\mathbf{x}_i$ and $\mathtt{C}_i$ are determined during single-frame inference, the colour model still needs to be defined. In our implementation a trajectory's appearance $\mathcal{A}_j$ is represented with a $(8\times8\times8)$-bin colour histogram in HSV space. For each observation o_i, we compute the colour histogram a_i in an elliptic region inside the bounding box, with Gaussian weighting to put more emphasis on pixels close to the centre. To improve robustness, colour values are distributed over neighbouring histogram bins with trilinear interpolation. The similarity between a detection and a trajectory is then defined as the Bhattacharyya distance between their histograms

$$p(o_i|\mathcal{A}_j) \sim \sum_{q,r,s} \sqrt{a_i(q,r,s)\mathcal{A}_j(q,r,s)}\,, \tag{7}$$

with (q, r, s) indices over the three histogram dimensions.

Every time a new observation o_i is added to a trajectory, its appearance model $\mathcal{A}_j$ is updated with an Infinite Impulse Response (IIR) filter,

$$\mathcal{A}_j(q) = w\mathcal{A}_j(q) + (1-w)a_i(q) \quad . \tag{8}$$

The appearance model contributes to the association probability, but it is not propagated through the EKF, which would prohibitively increase the dimension of the state vector.

4.1. Trajectory candidates

The set of putative candidate trajectories is generated by running bi-directional Extended Kalman Filters (EKFs) starting from each detection in the past and present (for computational efficiency, only candidates starting from new detections are generated from scratch, whereas candidates from previous frames are cached and extended). Each filter generates a candidate trajectory which obeys the dynamic model and bridges short gaps due to occlusion or detector failure – see Fig. 7. The important difference to conventional 1$^{\text{st}}$-order Markov tracking is that candidates do *not* originate only from the previous frame.

Data association between trajectory candidates and detections amounts to checking how well an observed O_I fits the candidate's dynamic model $\mathcal{M}_j$ and appearance model $\mathcal{A}_j$:

$$P(o_i|H_j) = P(o_i|\mathcal{A}_j) \cdot P(o_i|\mathcal{M}_j)\ . \tag{9}$$

The association probability $P(o_i|H_j)$ is computed for all detections at a given time step, and the one with the highest probability is used to update H_j ("winner takes all"). To prevent gross association errors $P(o_i|H_j)$ is gated to exclude overly unlikely associations.

4.2. Trajectory selection

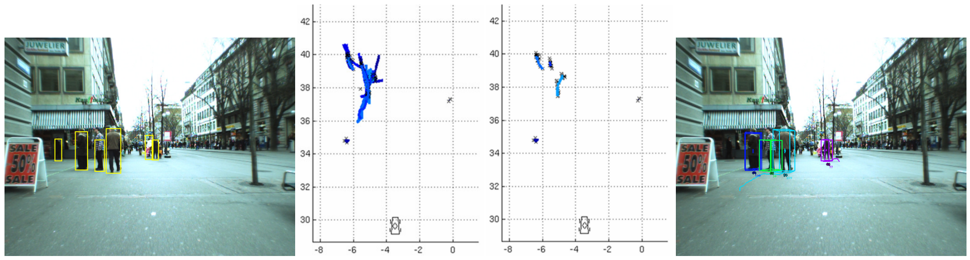
At this point the set of candidates is highly redundant. The different candidates are not independent because of the constraint that two pedestrians cannot be at the same location at the same time, and because each detection may only be assigned to one trajectory so as to avoid over-counting the evidence. Selecting the most likely subset of trajectories amounts to a binary labelling, where each candidate is declared either a member or a non-member of the optimal set, such that the set is as small as possible and conflict-free, while at the same time explaining as much as possible of the evidence observed up to the present frame.

The example in Fig. 8 visualises candidate generation and trajectory selection. People are standing close together, which leads to candidates that contain detections from several different persons. Note for example the long curve going to the left: selecting such a candidate is suboptimal in spite of its high individual score, because the exclusion constraints rule out all other candidates that are based on the same data points, leaving many detections unexplained. Hence, a globally better solution is reached by selecting multiple candidates which each explain less data, but are mutually consistent.

To select the jointly optimal subset of trajectories, we compute a support $\mathcal{U}$ for each candidate H_j, which is based on the strength of the associated detections $\{o_i\}$, weighted by their association probability according to the dynamic model $\mathcal{M}$ and the appearance model $\mathcal{A}$:

$$\begin{aligned}\mathcal{U}(H_j|\mathcal{I}^{t_0:t}) &= \sum_i \mathcal{U}(o_i|H_j, \mathcal{I}^{t_i}) = \\ &= \sum_i P(o_i|\mathcal{I}^{t_i}) \cdot P(o_i|\mathcal{A}_j) \cdot P(o_i|\mathcal{M}_j)\ .\end{aligned} \tag{10}$$

Choosing the best subset $\{H_j\}$ from the list of all candidates is a model selec-



(a) (b) (c) (d)

Figure 8: Tracking by means of a hypothesise-and-test framework: given object detections from the current and past frames (a), we construct an exhaustive, over-complete set of trajectory hypotheses (b) and prune it back to an optimal subset with model selection (c), yielding the final trajectories (d).

tion problem. If we restrict ourselves to interactions between pairs of candidates[4] the optimum is given by the quadratic binary expression

$$\max_{\mathbf{m}} \left[\mathcal{D}(\mathbf{m})\right] = \max_{\mathbf{m}} \left[\mathbf{m}^{\top}\mathbf{Q}\mathbf{m}\right] \quad , \quad \mathbf{m} \in \{0,1\}^N \ . \tag{11}$$

The Boolean vector $\mathbf{m}$ indicates whether a candidate shall be selected ($m_i = 1$) or discarded ($m_i = 0$). The diagonal entries q_{ii} are the individual utilities of the candidates, reduced by a constant "model penalty", which expresses the preference for solutions with fewer trajectories. The off-diagonal entries $q_{ij} \leq 0$ encode the interaction cost between candidates i and j. They are composed of a penalty proportional to the overlap of the two trajectories' footprints on the ground plane, and a correction term for the over-counting of detections consistent with both candidates, that would occur if both are selected.

$$\begin{aligned} q_{ii} &= -\,\epsilon_1 + \sum_{o_k^{t_k} \in H_i} \big((1-\epsilon_2) + \epsilon_2 \mathcal{U}(o_i^{t_i} | H_j, \mathcal{I}^{t_i})\big) \\ q_{ij} &= -\,\frac{1}{2}\epsilon_3 O(H_i, H_j) - \frac{1}{2} \sum_{o_k^{t_k} \in H_i \cap H_j} \big((1-\epsilon_2) + \epsilon_2 \mathcal{U}(o_i^{t_i} | H_\ell, \mathcal{I}^{t_i})\big) \ , \end{aligned} \tag{12}$$

where $H_\ell \in \{H_i, H_j\}$ denotes the weaker of the two candidates; $O(H_i, H_j)$ measures the physical overlap between the candidates based on average object dimen-

[4] Disregarding higher-order interactions results in too high penalties in cases where more than two trajectories compete for the space and/or detections; if interaction penalties are high enough to enforce complete exclusion, this will not alter the result.

sions; ϵ_1 is the "model penalty" chosen such that it neutralises the utility of ≈ 2 strong detections (to suppress erratic false detections); ϵ_2 is a regulariser to guarantee a minimal utility for each explained detection – smaller ϵ_2 reduces the influence of the goodness-of-fit, and puts more weight on the fact that a detection could be associated with the candidate at all; ϵ_3 is the scaling coefficient of the overlap penalty, and should be chosen large enough to prevent simultaneous selection of trajectories with significant overlap. The maximisation problem Eq. (11) is NP-hard, but due to its special structure strong local maxima can be found efficiently. Details about the optimisation algorithm are given in appendix Appendix C.

Besides establishing 3D trajectories, tracking also acts as a temporal smoothing filter: false detections consistent with the scene geometry are weeded out, if they lack support in nearby frames, and conversely missed detections on good trajectories are filled in. Note that starting from an exhaustive set of candidates by definition solves the initialisation of new trajectories (usually after 2-3 detections), and allows one to recover from temporary track loss and occlusion.

Person Identities. Trajectory selection is repeated at every frame. The selected set offers the most likely explanation of the observed data in the current frame *and in the past*. It is hence possible to follow trajectories back in time and determine where a person came from, even if that person had previously been missed. On the downside, the new explanation is not guaranteed to be consistent with the one selected previously. Identities hence have to be propagated by checking the overlap between trajectories found at consecutive time steps.

5. Experimental Evaluation

We present experimental results on four different sequences. In all cases, the sensors were a pair of forward-looking AVT Marlin F033C cameras, which deliver synchronised video streams of resolution 640×480 pixels at 12–14 frames per second. Sequences BAHNHOFSTRASSE (999 frames) and LINTHESCHER (1208 frames) have been recorded with a child stroller (baseline ≈ 0.4 m, sensor height ≈ 1 m, aperture angle $\approx 65°$) in busy pedestrian zones, with people and street furniture frequently obstructing portions of the field of view. LOEWENPLATZ (800 frames) and BELLEVUE (1500 frames) have been recorded from a car (baseline ≈ 0.6 m, sensor height ≈ 1.3 m, aperture angle $\approx 50°$) driving on inner-city streets among other vehicles. Pedestrians appear mostly on sidewalks and crossings, and are observed only for short time spans. The sequences were recorded in autumn and winter and exhibit realistic lighting and contrast. Videos of tracking results are available as supplementary material.

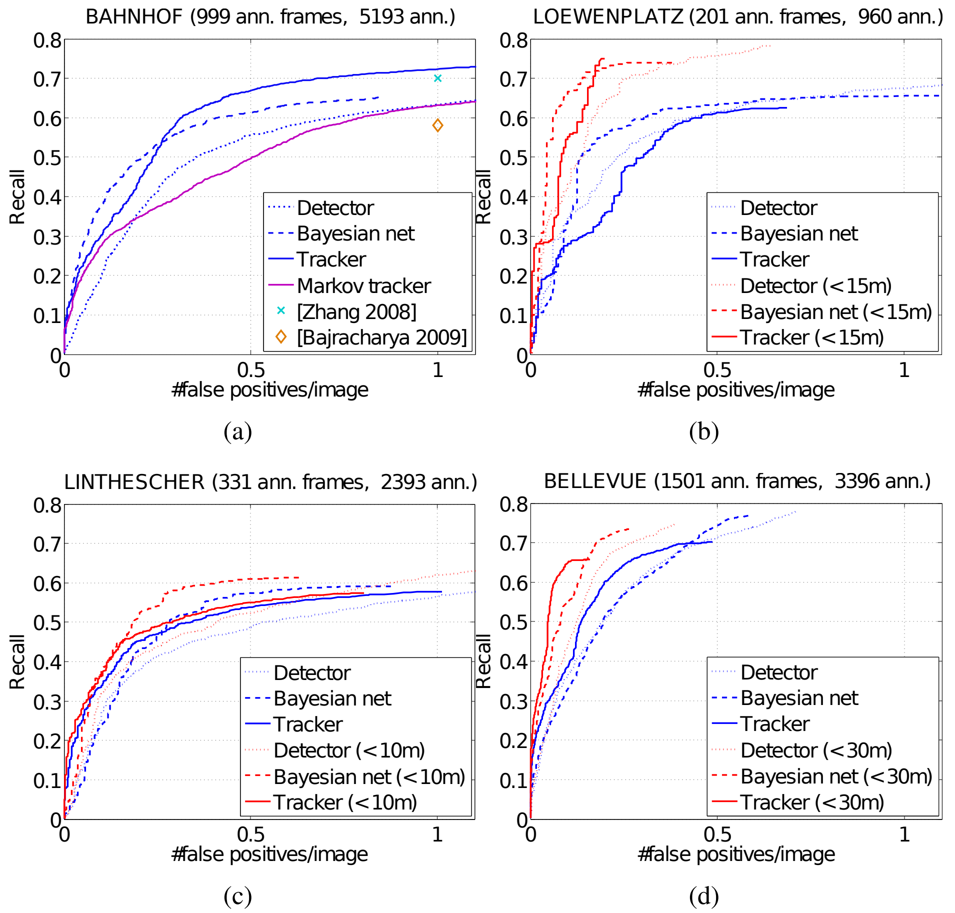


Figure 9: Single-frame performance evaluation. See text for details.

For testing, all system parameters were kept constant throughout all sequences except for the platform-dependent parameters: camera calibration, camera height, and ground plane prior (which depends on the wheelbase and suspension of the platform).

5.1. *Quantitative Results.*

Per-frame evaluation. To assess single-frame performance, the bounding boxes estimated with different thresholds are compared to manually annotated ground truth, plotting the recall (fraction of correctly found pedestrians) over false positives per image (FPPI). A bounding box is deemed correct if its intersection with the ground truth box is $> 50\%$ of their union.

In Fig. 9(a) we evaluate the single-frame performance of different variants of the system on Seq. BAHNHOFSTRASSE, and also compare to competing methods. The HOG detector without any scene information already performs reasonably well ("Detector"). Single-frame inference with added 3D geometry improves performance by 5–15% ("Bayesian net"). Multi-frame tracking ("Tracker") further improves the reachable recall, but loses recall in the high-precision regime, which is largely an effect of per-frame evaluation: since the tracker needs to accumulate 2–3 detections before starting a trajectory, it loses recall every time a new person enters the scene. At the often quoted operating point of 1 FPPI, the tracker achieves 73% recall.

To assess the gain of multi-hypothesis tracking, we reduce our system to a 1^{st}-order Markov tracker, which follows each trajectory independently, starting new trajectories from unassigned detections. It reaches 55% recall, similar to the raw detector, which underpins the need for multi-frame interaction reasoning.

On the same sequence Zhang et al. (2008) report 70% recall at 1 FPPI.[5] They do not use stereo, but track in image coordinates in batch mode, i.e. trajectories are only found once the detections over the entire video sequence are available to the tracker. Bajracharya et al. (2009) report 58% recall on this sequence at 1 FPPI with a tracker that does use stereo (and 42% recall on Seq. LINTHESCHER, see below).

In Fig. 9(b)-(d), we show results on further sequences. As above, single-frame detection performance is measured, hence the tracker suffers from an initial latency – the effect is more pronounced for Seq. LOEWENPLATZ, because it contains many briefly visible pedestrians. Note however, tracking is indispensable for motion prediction and dynamic path planning. Ground truth annotations cover all pedestrians, including those in the far distance. We show both the performance on all annotated pedestrians (blue curves) and the performance in the near and midrange (red curves).

In Table 1, we compare the influence of the employed stereo algorithm on the result, since existing algorithms differ considerably in terms of quality and runtime (c.f. Sec. 2.2). Specifically, we compare two GPU-based dense stereo matchers: fast plane sweep stereo ("PS", Cornelis and Van Gool, 2005), and a recent top-of-the-line method ("Zach", Zach et al., 2009). Modern algorithms indeed improve the performance of both scene analysis and tracking, but the gains

[5] All data is publicly available at `http://www.vision.ee.ethz.ch/~aess/dataset/`.

	FPPI	full depth range		restricted to 15 m	
		Detector	Tracker	Detector	Tracker
no depth	0.5	—	0.19	—	0.32
	1.0	—	0.29	—	0.47
PS	0.5	0.63	0.60	0.66	0.66
	1.0	0.68	0.70	0.67	0.74
Zach	0.5	0.65	0.64	0.67	0.73
	1.0	0.67	0.73	0.67	0.78

Table 1: Single-frame results on Seq. BAHNHOFSTRASSE with different stereo matchers. Better depth maps improve localisation and tracking in the near field. Since we use robust statistics on depth, elaborate stereo algorithms bring little improvement at the system level.

are modest and come at the cost of much higher runtime (20 ms for PS vs. >1 s for Zach). It appears that when depth maps are treated as intermediate result and processed with robust statistics, high-end stereo does not help much, in spite of visibly better depth maps. On the contrary, it is indispensable to measure depth, even though the last bit of accuracy is not crucial: bypassing stereo all together and estimating depth from bounding box size gives abysmal results.

Track-level Evaluation. To also evaluate the tracking in more detail, we quantitatively evaluate on the trajectory level in Table 2. There are still no satisfactory methods for automatic track-level evaluation, hence the correspondence between estimated and actual trajectories had to be verified interactively.

	BAHNHOFSTRASSE	LOEWENPLATZ
ground truth	89	107
tracker	125	126
mostly tracked	0.55	0.48
partially tracked	0.30	0.27
mostly missed	0.15	0.25
false alarms	0.62	1.09
ID switches	16	6
mean/median latency	9.9 / 1.5	0.3 / 2.0

Table 2: Trajectory-based evaluation on Seq. BAHNHOFSTRASSE and Seq. LOEWENPLATZ.

Following other trajectory-level evaluations (Wu and Nevatia, 2007; Li et al., 2009), we examine all ground truth subjects and classify them in one of three categories: *mostly tracked* (covered to >80% by the best estimated trajectory),

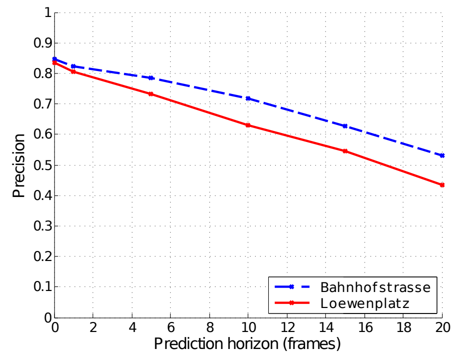


Figure 10: Precision of the tracker prediction for increasing prediction horizon. Data was recorded at 12–14 fps.

partially tracked (covered 20-80%), or *mostly missed* (covered <20%). Furthermore we report the number of ground truth trajectories and the number of trajectories output by the tracker, the average number of false alarms per frame, the total number of identity switches (cases where a new trajectory is started although the subject is still the same), and the mean and median latency (the number of frames until a trajectory is initialised for a new subject).

In both cases, few false alarms occur, and few trajectories are *mostly missed*. The fraction of *partially tracked* subjects, and for Seq. BAHNHOFSTRASSE the mean latency, are high: it happens frequently that a distant pedestrian is visible for a few frames, then disappears into occlusion and reappears at smaller distance, where he is picked up by the tracker. In the best case this will produce an identity switch (since the occlusion lasts too long to associate the two trajectories), but more often the subject will be picked up for the first time only after the occlusion is then reported as *mostly missed* or *partially tracked*. For this reason the mean latency is a lot higher than the median on Seq.BAHNHOFSTRASSE: the entire track before and during occlusion counts as latency. 9 out of 76 persons fall into this category and have latencies >30 frames. In fact, most other *partially tracked* subjects are quite well covered – 17% of them lie between 70 and 80%.

Prediction. Since the aim of people tracking in traffic is to react to their behaviour, either by appropriate path planning, or by an emergency manoeuvre, we also assess how well future locations can be predicted from the estimated trajectories. To this end, we count the precision of bounding boxes extrapolated to future frames, and plot them for varying time horizon in Fig. 10. As expected, precision drops with increasing look-ahead, but remains acceptable up to ≈ 1 second (12 frames).

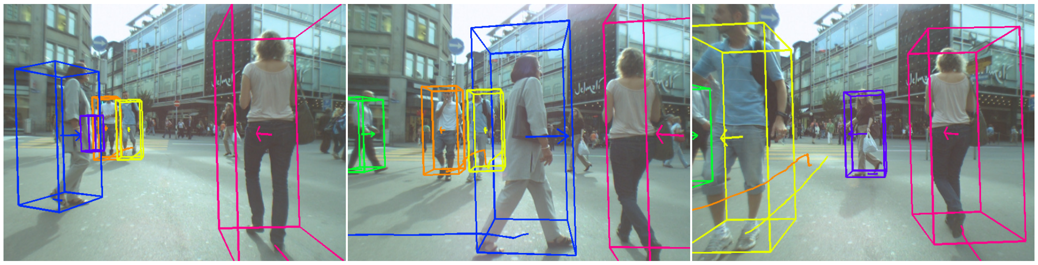


Figure 11: Example results for Seq. LINTHESCHER.

The plot illustrates the worst-case scenario: usually a precision of 0.9 does *not* imply that re-planning fails in 1 out of 10 frames, because not all pedestrians affect the planned path.

5.2. Qualitative Results

In this section we illustrate the behaviour of the described tracking method with example images. Fig. 11 shows an example from Seq. LINTHESCHER featuring multiple full occlusions (the woman crossing in the foreground temporarily occludes every other person).

Fig. 12 shows how both adults and children are tracked in Seq. BAHNHOFSTRASSE, although the latter deviate significantly from the typical height and aspect ratio. In the bottom row, a situation is shown where tracking is superior to mere object detection: without motion prediction, the man in the pink bounding box would possibly cause an unnecessary avoidance manoeuvre.

In Fig. 13 pedestrians are tracked from a driving car, with the camera rig mounted on the roof. People are visible for shorter periods, since they are either passed at high speed or cross the street in front of the vehicle.

6. Conclusion

6.1. Summary

We have described a system for detection and 3D localisation of people in street scenes recorded from a mobile stereo rig. The system is able to track multiple people in 3D world coordinates, based on image-based pedestrian detection and dense stereo depth. Robustness is achieved by

1. treating automatic image measurements as noisy observations, from which a per-frame estimate of the 3D environment is derived by MAP estimation in a probabilistic scene model, and

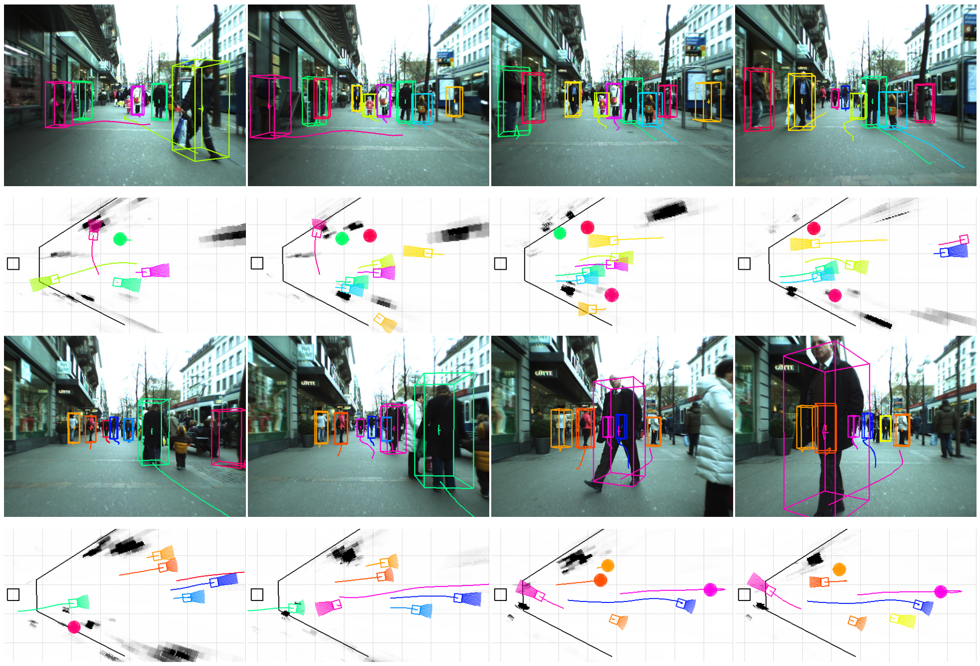


Figure 12: Example results for Seq. BAHNHOFSTRASSE.

2. multi-hypothesis tracking based on the per-frame estimates to find, in each time step, the people trajectories which best explain all past and present evidence.

The system has been tested on several realistic stereo sequences, including both quantitative comparisons to ground truth annotations and qualitative analysis of the system's ability to track and predict pedestrian motion.

6.2. Outlook

The presented work should be seen as an initial attempt to combine close-range photogrammetry of dynamic scenes and automatic image understanding. There are several promising directions for future research in this area.

Obviously, sensor fusion will play an important role. Due to the variety of tasks, a large array of sensors could be useful, from the obvious GPS/IMU for self-localisation to more exotic sensors such as thermal cameras for human detection, or terrestrial multi-spectral sensing for the more ambitious goal of dense scene understanding.

In terms of algorithms, multi-class detection is still an open research question. It may be possible to extend the current detection (and/or pixel labelling)

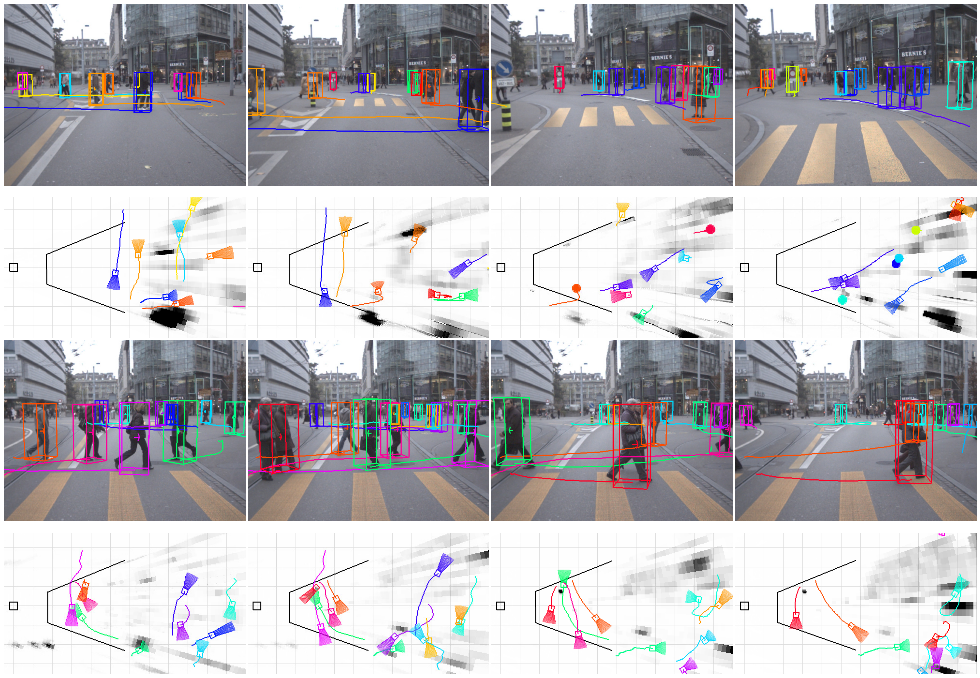
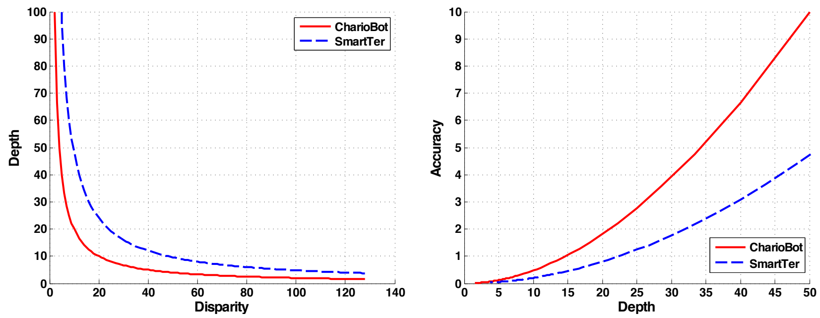


Figure 13: Example results for Seq. LOEWENPLATZ.

paradigms to a handful of classes, but scalable classification of the large variety of objects in our environment at a reasonable level of abstraction is still out of reach.

For the specific case of humans, more detailed modelling is of interest to better describe and predict behaviour: on one hand, also estimating a person's articulation (rather than only their position in space) may improve prediction of their future motion, and has also been shown to improve detection itself in certain situations (Andriluka et al., 2008); on the other hand, the motion planning of real people is not independent of their environment, so including models of social behaviour such as those developed for crowd simulation (e.g. Helbing and Molnár, 1995; Schadschneider, 2001) could potentially improve dynamic models (c.f. Pellegrini et al., 2009). Note, both tasks pose additional challenges, since significantly more parameters have to be determined from the same data.

In terms of photogrammetric methodology, the analysis of highly dynamic environments could well lead to a revival of dense stereo reconstruction, which has in recent years been somewhat over-shadowed by laser ranging, but offers the important advantage that strictly synchronous measurements can be acquired over a large solid angle. A comeback of dense stereo would be in line with another



(a) (b)

Figure A.1: The uncertainty of depth estimation depends on the focal length and the baseline. (a) Relation of disparity to depth for both setup types. (b) Localisation accuracy at given depths. *SmartTer*'s larger baseline allows accurate localisation at larger depths.

trend towards what could be called "low-precision photogrammetry" in the close-range domain: in many applications, including the one presented here, the critical factor is the *completeness* of the estimated model, whereas metric precision is – within reasonable bounds – less of an issue.

Acknowledgements

This project has been funded in parts by Toyota Motor Corporation and the EU projects DIRAC (IST-027787) and EUROPA (ICT-2008-231888). We thank Nico Cornelis and Christopher Zach for providing GPU implementations of their stereo matching algorithms. We also thank Kristijan Macek, Luciano Spinello, and Roland Siegwart for the opportunity to record from the *SmartTer* platform.

Appendix A. Accuracy of stereo depth

Given the focal length f_u in pixels, the camera baseline B in world units and the disparity d between the two images of a stereo pair, the depth z of a point w.r.t. the camera is

$$z = \frac{f_u B}{d} \quad . \tag{A.1}$$

Thus, the working range, respectively accuracy, of a stereo rig is primarily determined by the focal lengths of the cameras and the baseline. Fig. A.1 illustrates the relationship between disparity and depth for our *CharioBot* and *SmartTer* setups, as well as the corresponding depth uncertainty. If we define the "working range" to be that part of the common field of view for which the localisation error σ_z is below 1 m, we get a range of 1.5 to 15 m for the *CharioBot* platforms. The pedestrian detector theoretically can pick up people up to a distance of 19 m, corresponding to a localisation uncertainty of 1.5 m. For the longer *SmartTer* baseline the working range is 3.8 to 22 m, with the detector reaching 30 m / $\sigma_z = 1.8$ m.

Using error propagation, the localisation uncertainty of the stereo system can be inferred from the measurement uncertainty (σ_u, σ_v) of a pixel with position (u, v) and the uncertainty σ_d of the disparity estimate d. We can write the back-projection as

$$f(u, v, d) = \frac{B}{d} \left(u, v \cdot s_{vu}, f_u\right)^\top \quad . \tag{A.2}$$

Forward error propagation (taking into account that in practice $s_{vu} \approx 1$) yields the uncertainty covariance of a reconstructed 3D point as

$$\mathtt{C} = \left(\frac{\partial f}{\partial \mathbf{u}}\right)^\top \begin{pmatrix} \sigma_u & 0 & 0 \\ 0 & \sigma_v & 0 \\ 0 & 0 & \sigma_d \end{pmatrix} \left(\frac{\partial f}{\partial \mathbf{u}}\right) = \begin{pmatrix} \sigma_u + \sigma_d b^2 u & \sigma_d b^2 uv & \sigma_d f b^2 u \\ \sigma_d b^2 uv & \sigma_v + \sigma_d b^2 v & \sigma_d f b^2 v \\ \sigma_d f b^2 u & \sigma_d f b^2 v & \sigma_d f^2 b^2 \end{pmatrix} , \tag{A.3}$$

with $b = \frac{B}{d^2}$. The uncertainty grows quadratically with increasing depth. Increasing baseline or image resolution will linearly decrease the uncertainty.

Appendix B. Probabilities in the Bayes net

The basic building blocks of the probabilistic scene model are the probability distributions of the single variables in Eq. (4). This section describes in detail, how these distributions are modelled.

Appendix B.1. Depth evidence

The depth map $\mathcal{D}$ is regarded as noisy observation to account for inaccuracies and gross errors of stereo matching. Using the confidence map, we make use of the observed depth in a robust manner: each object hypothesis is assigned a depth flag $d_i \in \{0, 1\}$, which indicates whether the depth map for its bounding box is reliable ($d_i = 1$) or not. This flag's evidence is inferred from the confidence map $\mathcal{C}$ and is encoded in $P(d_i|\mathcal{C})$.

The consistency between the stereo depth $z(\mathcal{D}, \mathbf{b}_i)$ measured inside the bounding box $\mathbf{b}_i$ and the depth $z(o_i)$ obtained by projecting the bounding box to the ground plane serves as an indicator for $P(\mathbf{c}_i|\boldsymbol{\pi}, \mathcal{D}, d_i = 1)$. Second, we test the depth variation inside the box and define $P(v_i = 1|d_i = 1)$ to reflect the expectation

Var.	**Meaning**	**Domain**
Observed		
$\mathcal{I}$	Images of camera pair	
$\mathcal{D}$	Depth maps of camera pair	
$\mathcal{C}$	Confidence maps	
Output / Hidden		
$\mathbf{c}_i$	Object centre point and scale	$\{\{k,l\} \mid k=1\dots K, l=1\dots L\}$
v_i	Object validity	$\{0,1\}$
d_i	Validity of depth per object	$\{0,1\}$
$\boldsymbol{\pi}$	Ground plane	$\left\{\{\phi,\theta,\pi^{(4)}\} \middle\| \begin{matrix}\phi,\theta=1\dots6,\\ \pi^{(4)}=1\dots20\end{matrix}\right\}$

Table B.1: Variables of the model, along with their domains.

that the depth is largely uniform when a pedestrian is present. In detail, the two terms are defined as follows: the median depth inside a bounding box $\mathbf{b}_i$,

$$z(\mathcal{D},\mathbf{b}_i) = \underset{\text{pixel } \mathbf{p}\in\mathbf{b}_i}{\text{med}} \mathcal{D}(\mathbf{p})^{(3)} \quad , \tag{B.1}$$

yields a robust estimate of the corresponding object's depth. With the measurement uncertainty $\sigma^2_{(z),i} = \mathtt{C}_i^{(3,3)}$ from Eq. (A.3), this yields

$$P_{(z),i}(z(o_i)) \sim \mathcal{N}(z(o_i); z(\mathcal{D},\mathbf{b}_i), \sigma^2_{(z),i}) \quad . \tag{B.2}$$

$P_{(z),i}(z(o_i))$ thus models the probability that a given object distance $z(o_i)$ corresponds to the measured depth of the bounding box. As described later under heading *detection evidence*, it is used to measure the consistency between a detected bounding box and the depth map.

To measure depth uniformity, we compute the depth variation of the pixels $\mathbf{p}$ within $\mathbf{b}_i$, $V = \{\mathcal{D}(\mathbf{p})^{(3)} - z(\mathcal{D},\mathbf{b}_i) \mid \mathbf{p}\in\mathbf{b}_i\}$. Depth uniformity is measured by the normalised count of pixels in the confidence interval $\pm\sigma_{(z),i}$, disregarding values outside the inter-quartile range $[LQ(V), UQ(V)]$ to be robust against outliers and points outside a person's silhouette:

$$\eta_i = \frac{\left|\{a \in [LQ, UQ] \mid a^2 < \sigma^2_{(z),i}\}\right|}{UQ - LQ} \ . \tag{B.3}$$

This robust "depth inlier fraction" serves as basis for learning $P(v_i|d_i = 1)$, as described below in Sec. Appendix B.3. The probability $P(v_i|d_i = 0)$ is assumed uniform, since incorrect regions in the depth map hold no information about object presence. $P(d_i|\mathcal{C})$ is learnt from a training set with annotated ground truth pedestrians.

Appendix B.2. Ground plane

To keep computations tractable, the range for the ground plane parameters $(\theta, \phi, \pi^{(4)})$ is restricted to the intervals observed in the training sequences, and discretised to a $(6{\times}6{\times}20)$ grid. The discretisation is chosen to keep quantisation errors < 0.05 for θ and < 0.01 for ϕ, resulting in errors $< 5 \cdot 10^{-7}$ in the entries of $\mathbf{n}$. The depth errors ensuing from the discretisation are < 0.2 m in depth for a pedestrian at a distance of 15 m. The prior distribution $P(\boldsymbol{\pi})$ is also estimated from the same training sequences: in input images with few objects, Least-Median-of-Squares (LMedS, Rousseeuw and Leroy, 1987) fitting to the depth map $\mathcal{D}$ yields correct estimates of the ground plane. with the robust median residual of Eq. (2)

$$\boldsymbol{\pi} = \min_{\boldsymbol{\pi}_i} r(\boldsymbol{\pi}_i, \mathcal{D}) \quad . \tag{B.4}$$

Fitting ground plane parameters to all images of the training set with Eq. (B.4) we learn $P(\boldsymbol{\pi})$.

Appendix B.3. Detection evidence

Pedestrian hypotheses $o_i = \{v_i \mathbf{c}_i\}, (i = 1 \ldots n)$ for each stereo pair are generated with the HOG detector, set to a low threshold to ensure maximum recall. This typically yields 10–100 putative detections per time step. These consist of a centre point in 2D image coordinates, along with a scale: $\mathbf{c}_i = \{x, y, s\}$; and of a binary flag $v_i \in \{0, 1\}$ indicating the presence or absence of a person at that position. Given a specific $\mathbf{c}$ and a standard object size (w, h) at scale $s = 1$, a bounding box can be constructed. The box base point in homogeneous image coordinates $\mathbf{g} = (x, y + \frac{1}{2}sh, 1)$ is projected to 3D by casting a ray and intersecting it with the ground plane, yielding the point

$$\mathbf{G} = -\frac{\pi^{(4)} \mathtt{K}^{-1} \mathbf{g}}{\mathbf{n}^{\top} \mathtt{K}^{-1} \mathbf{g}} \quad . \tag{B.5}$$

$\mathtt{K}$ denotes the camera's internal calibration. 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, by intersecting another ray through the bounding box's top point with a fronto-parallel plane, orthogonal to the ground.

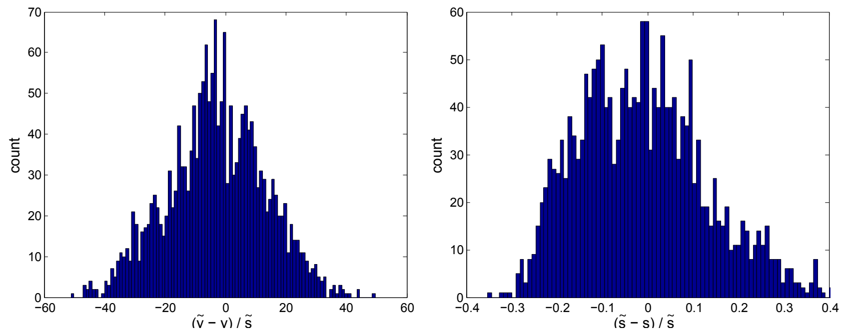


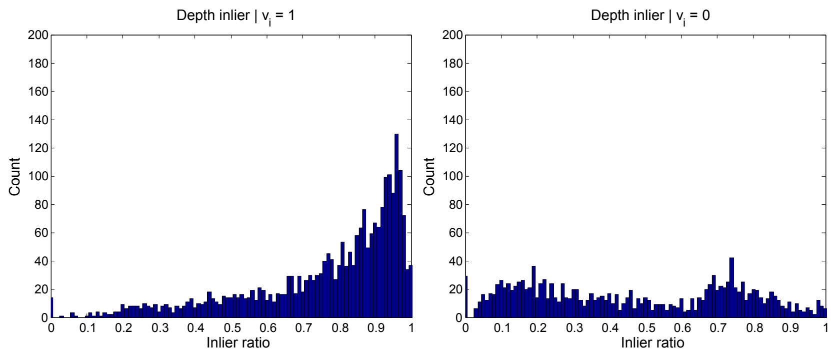
Figure B.1: Centre distributions normalised by detected scale $\tilde{s}$ (left: centre $(\tilde{y}-y)/\tilde{s}$, right: scale $(\tilde{s}-s)/\tilde{s}$), learnt from 1,578 annotations. We approximate these using normal distributions.

Because of the large localisation uncertainty of appearance-based detection, the detector estimates for centre and scale are again only considered as observations $(\tilde{x}_i, \tilde{y}_i, \tilde{s}_i)$. Using the detector output directly would often yield misaligned bounding boxes, which in turn lead to wrong estimates for distance and size. Instead, we estimate the centre and scale, by considering a set of possible bounding boxes $\mathbf{b}_i^{\{k,\ell\}}$ for each o_i. Around the detector output, a discrete set of bounding boxes are sampled: $y_i = \tilde{y}_i + k\sigma_y\tilde{s}_i$, $s_i = \tilde{s}_i + \ell\sigma_s\tilde{s}_i$ ($x_i = \tilde{x}_i$ is fixed due to its negligible influence). The step sizes σ_y and σ_s are inferred from detections and ground truth annotations on a training set. Fig. B.1 shows the resulting scale-normalised measurements $(\tilde{y}-y)/\tilde{s}$, $(\tilde{s}-s)/\tilde{s}$. The distributions are represented by zero-mean Gaussians, which is a reasonable approximation, as can be seen in the figure.

The number of samples, i.e. the range of $\{k, \ell\}$, is fixed for all objects. An object can thus be assigned one out of a discrete set of position/scale pairs (in our implementation 3×3 steps worked best), each corresponding to a different 3D height and distance. In the following, we omit the superscripts for readability.

By means of Eq. (B.5), $P(v_i = 1|\mathbf{c}_i, \boldsymbol{\pi}) \sim P(\mathbf{G}_i^h)P\big(z(o_i)\big)$ is expressed as the product of a prior on object distance $P\big(z(o_i)\big)$, and a prior on object size $P(\mathbf{G}_i^h)$. The object size distribution is assumed to be Gaussian, $P(\mathbf{G}^h) \sim \mathcal{N}(1.7, 0.085^2)$ [m]. The distance distribution $P\big(z(o_i)\big)$ is assumed uniform in the system's operating range (2–30 m for *CharioBot* and *CharioBot II*; 3–50 m for *SmartTer*).

It is difficult to model the dependence of $P(\mathbf{c}_i|\boldsymbol{\pi}, \mathcal{D}, d_i = 1)$ on the forward-projected object depth $z(o_i)$ and the depth map measurement $z(\mathcal{D}, \mathbf{b}_i)$ exactly. In practice, we model only the dominant factor $P_{(z),i}\big(z(o_i)\big)$ from Eq. (B.2). We found that modelling both factors with a learnt non-parametric distribution yields inferior results. $P(\mathbf{c}_i|\boldsymbol{\pi}, \mathcal{D}, d_i = 0)$ is assumed uniform.



(a) (b)

Figure B.2: Distribution of depth inliers for correct (a) and incorrect (b) detections, learnt from 1,578 annotations and 1,478 negative examples. Based on these distributions, we learn a classifier using logistic regression.

The detector reliability $P(\mathcal{I}|v_i)$ is learnt by logistic regression on a training set of correct and incorrect detections with their associated detection scores. A similar procedure is used to learn $P(v_i|d_i = 1)$: we measure the fraction η_i of pixels that are uniform in depth for bounding boxes in the training data, using Eq. (B.3). We do this for both correct and incorrect bounding boxes and fit a sigmoid with logistic regression for $P(v_i|d_i = 1)$. Fig. B.2 illustrates that η_i is a reasonable indicator of object presence.

The depth validity flag d_i is derived from the confidence map $\mathcal{C}$. Let $\mathcal{C}_>$ denote the case that $> 50\%$ of the pixels inside the bounding box are marked "confident", i.e. the depth information is reliable. With the same training set as above we obtain $P(d_i\!=\!1|\mathcal{C}_>)\!\approx\!0.96$. The probability of a valid depth in a non-confident region is set to $P(d_i\!=\!1|\neg\,\mathcal{C}_>)\!=\!0$.

Appendix C. Optimal Trajectory Selection

Due to the pairwise constraints between different candidate trajectories, the multi-person tracking problem translates to the quadratic pseudo-Boolean maximisation problem Eq. (11).

The complexity of maximising Eq. (11) w.r.t. $\mathbf{m}$ is combinatorial. Luckily, heuristics exist to find strong local maxima (Schindler et al., 2006; Rother et al., 2007). The non-zero entries of $\mathbf{m}$ corresponding to the maximum $\widehat{\mathcal{D}}(\mathbf{m})$ indicate which candidates form the best set of trajectories for the current frame.

CPT	Description
Ground plane	
$P(\boldsymbol{\pi})$	prior learnt from sequence
$P(\mathcal{D}\|\boldsymbol{\pi})$	diagonal entries only, Eq. (3)
Objects	
$P(\mathbf{c}_i\|\boldsymbol{\pi},\mathcal{D},d_i=1)$	distance correspondence, Eq. (B.2)
$P(\mathbf{c}_i\|\boldsymbol{\pi},\mathcal{D},d_i=0)$	uniform distribution, no comparison possible
$P(v_i\|d_i=1)$	assumption of object flatness, Eq. (B.3)
$P(v_i\|d_i=0)$	uniform distribution
$P(v_i\|\mathbf{c}_i,\boldsymbol{\pi})$	height and distance assumptions, $P(\mathbf{G}_i^h)P(z(o_i))$
$P(\mathcal{I}\|v_i)$	object detection probability
Depth	
$P(d_i\|\mathcal{C})$	depth validity depending on confidence map

Table B.2: Summary of conditional probability tables (CPTs) employed in the model, with their respective factors.

To find the optimum, we use an extended version of the multi-branch search of Schindler et al. (2006). The method exploits the fact that the number of actual trajectories is comparatively small, and performs a "controlled combinatorial explosion" by reducing the number of branches to be followed in each step in geometric progression. Since the function $\mathcal{D}$ is submodular ($q_{ii} > 0$, and $q_{ij} \leq 0\ \forall i \neq j$), the path to the global maximum can never contains descending steps: starting from some vector $\mathbf{m}'$, the next inspected solution $\mathbf{m}''$ must fulfil $\mathcal{D}(\mathbf{m}'') > \mathcal{D}(\mathbf{m}')$.

We point out a tighter bound: given $\mathbf{m}'$, let $\mathcal{L}'$ be the set of candidates currently not selected, $\{\forall i \in \mathcal{L}' : m'_i = 0\}$, and denote by $\mathbf{1}_i$ a vector that contains all 0s except at entry i. Starting from $\mathbf{m}'$, the maximally reachable score is bounded above by

$$s = \mathcal{D}(\mathbf{m}') + \max\big[0, \sum_{i \in \mathcal{L}'} \left(\mathcal{D}(\mathbf{m}' + \mathbf{1}_i) - \mathcal{D}(\mathbf{m}')\right)\big] \,. \tag{C.1}$$

The intuitive meaning of this is that the benefit of adding *all unselected* candidates to the current solution would be highest if they all were independent, and can only go down if there are any interactions among them.[6] It follows that one can quit a search branch as soon as the best objective value found so far exceeds the upper bound Eq. (C.1). This early bail-out reduces the number of search steps by 37%.

[6]This is the defining property of submodularity; c.f. Boros and Hammer (2002).

Andriluka, M., Roth, S., Schiele, B., 2008. People-tracking-by-detection and people-detection-by-tracking. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition (on CDROM).

Avidan, S., 2005. Ensemble tracking. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition 2, 494–501.

Bajracharya, M., Moghaddam, B., Howard, A., Brennan, S., Matthies, L. H., 2009. A fast stereo-based system for detecting and tracking pedestrians from a moving vehicle. International Journal of Robotics Research 28, 1466–1485.

Bay, H., Ess, A., Tuytelaars, T., Van Gool, L., 2008. Speeded-up robust features (SURF). Computer Vision and Image Understanding 110(3), 346–359.

Bibby, C., Reid, I., 2008. Robust real-time visual tracking using pixel-wise posteriors. In: Proceedings 10th European Conference on Computer Vision 2, 831–844.

Bishop, C. M., 2006. Pattern Recognition and Machine Learning. Springer.

Boros, E., Hammer, P. L., 2002. Pseudo-boolean optimization. Discrete Applied Mathematics 123 (1-3), 155–225.

Cornelis, N., Van Gool, L., 2005. Real-time connectivity constrained depth map computation using programmable graphics hardware. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition 1, 1009–1104.

Cornelis, N., Van Gool, L., 2008. Fast scale invariant feature detection and matching on programmable graphics hardware. In: Proceedings Workshop on Computer Vision on GPUs (on CDROM).

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition 1, 886–893.

Davison, A. J., 2003. Real-time simultaneous localization and mapping with a single camera. In: Proceedings 9th International Conference on Computer Vision, 1403–1410.

Dollar, P., Wojek, C., Schiele, B., Perona, P., 2009. Pedestrian detection: A benchmark. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition (on CDROM).

Enzweiler, M., Gavrila, D. M., 2009. Monocular pedestrian detection: Survey and experiments. IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (12), 2179–2195.

Ess, A., Leibe, B., Schindler, K., Van Gool, L., 2008. A mobile vision system for robust multi-person tracking. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition (on CDROM).

Ess, A., Leibe, B., Schindler, K., Van Gool, L., 2009a. Moving obstacle detection in highly dynamic scenes. In: Proceedings International Conference on Robotics and Automation, 56–63.

Ess, A., Leibe, B., Schindler, K., Van Gool, L., 2009b. Robust multi-person tracking from a mobile platform. IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (10), 1831–1846.

Ess, A., Leibe, B., Van Gool, L., 2007. Depth and appearance for mobile scene analysis. In: Proceedings 11th International Conference on Computer Vision, 1–8.

Felzenszwalb, P., McAllester, D., Ramanan, D., 2008. A discriminatively trained, multiscale, deformable part model. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition (on CDROM).

Förstner, W., Gülch, E., 1987. A fast operator for detection and precise location of distinct points, corners and centres of circular features. In: Proceedings ISPRS Intercommission Workshop on Fast Processing of Photogrammetric Data, 281–305.

Gavrila, D. M., Munder, S., 2007. Multi-cue pedestrian detection and tracking from a moving vehicle. International Journal of Computer Vision 73 (1), 41–59.

Gelb, A., 1996. Applied Optimal Estimation. MIT Press.

Havlena, M., Ess, A., Moreau, W., Torii, A., Jancosek, M., Pajdla, T., Van Gool, L., 2009. AWEAR 2.0 system: Omni-directional audio-visual data acquisition and processing. In: Proceedings 1st Workshop on Egocentric Vision (on CDROM).

Helbing, D., Molnár, P., 1995. Social force model for pedestrian dynamics. Physics Review E 51(5), 4282–4286.

Hoiem, D., Efros, A. A., Hebert, M., 2006. Putting objects in perspective. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition 2, 2137–2144.

Isard, M., Blake, A., 1998. CONDENSATION–conditional density propagation for visual tracking. In: International Journal of Computer Vision. Vol. 29(1), 5–28.

Leibe, B., Schindler, K., Cornelis, N., Van Gool, L., 2008. Coupled detection and tracking from static cameras and moving vehicles. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(10), 1683–1698.

Leibe, B., Seemann, E., Schiele, B., 2005. Pedestrian detection in crowded scenes. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition 1, 878–885.

Li, Y., Huang, C., Nevatia, R., 2009. Learning to associate: HybridBoosted multi-target tracker for crowded scene. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition (on CDROM).

Lin, Z., Davis, L. S., 2008. A pose-invariant descriptor for human detection and segmentation. In: Proceedings 10th European Conference on Computer Vision 4, 423–436.

Mei, C., Sibley, G., Cummins, M., Newman, P., Reid, I., 2009. A constant-time efficient stereo SLAM system. In: Proceedings British Machine Vision Conference (on CDROM).

Murphy, K. P., Weiss, Y., Jordan, M. I., 1999. Loopy belief propagation for approximate inference: An empirical study. In: Proceedings Uncertainty in Artifical Intelligence, 467–475.

Nistér, D., Naroditsky, O., Bergen, J. R., 2004. Visual odometry. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition 1, 652–659.

Nummiaro, K., Koller-Meier, E., Van Gool, L., 2003. An adaptive color-based particle filter. Image and Vision Computing 21 (1), 99–110.

Okuma, K., Taleghani, A., de Freitas, N., Little, J., Lowe, D., 2004. A boosted particle filter: Multitarget detection and tracking. In: Proceedings 8th European Conference on Computer Vision 1, 28–39.

Papageorgiou, C., Poggio, T., 2000. A trainable system for object detection. International Journal of Computer Vision 38 (1), 15–33.

Pearl, J., 1988. Probabilistic Reasoning in Intelligen Systems. Morgan Kaufmann Publishers Inc.

Pellegrini, S., Ess, A., Schindler, K., Van Gool, L., 2009. You'll never walk alone: modeling social behavior for multi-target tracking. In: Proceedings 12th International Conference on Computer Vision, 261–268.

Rother, C., Kolmogorov, V., Lempitsky, V. S., Szummer, M., 2007. Optimizing binary MRFs via extended roof duality. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition (on CDROM).

Rousseeuw, P. J., Leroy, A. M., 1987. Robust Regression and Outlier Detection. John Wiley and Sons.

Sabzmeydani, P., Mori, G., 2007. Detecting pedestrians by learning shapelet features. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition (on CDROM).

Schadschneider, A., 2001. Cellular automaton approach to pedestrian dynamics – theory. In: Proceedings Pedestrian and Evacuation Dynamics, 75–86.

Scharstein, D., Szeliski, R., 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision 47 (1-3), 7–42.

Schindler, K., U, J., Wang, H., 2006. Perspective n-view multibody structure-and-motion through model selection. In: Proceedings 9th European Conference on Computer Vision 1, 606–619.

Shashua, A., Gdalyahu, Y., Hayun, G., 2004. Pedestrian detection for driving assistance systems: Single-frame classification and system level performance. In: Proceedings Intelligent Vehicle Symposium, 1–6.

Stauffer, C., Grimson, W. E. L., 1999. Adaptive background mixture models for real-time tracking. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition, 2246–2252.

Toyama, K., Krumm, J., Brumitt, B., Meyers, B., 1999. Wallflower: principles and practice of background maintenance. In: Proceedings 7th International Conference on Computer Vision, 255–261.

Viola, P., Jones, M., Snow, D., 2003. Detecting pedestrians using patterns of motion and appearance. In: Proceedings 9th International Conference on Computer Vision, 734–741.

Walk, S., Majer, N., Schindler, K., Schiele, B., 2010. New features and insights for pedestrian detetion. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition (on CDROM).

Wang, X., Han, T. X., Yan, S., 2009. A HOG-LBP human detector with partial occlusion handling. In: Proceedings 12th International Conference on Computer Vision, 32–39.

Wedel, A., Rabe, C., Vaudrey, T., Brox, T., Franke, U., Cremers, D., 2008. Efficient dense scene flow from sparse or dense stereo data. In: Proceedings 10th European Conference on Computer Vision 1, 739–751.

Wojek, C., Dorkó, G., Schulz, A., Schiele, B., 2008. Sliding-windows for rapid object class localization: A parallel technique. In: Pattern Recognition – Proceedings 30th DAGM Symposium, 71–81.

Wojek, C., Walk, S., Schiele, B., 2009. Multi-cue onboard pedestrian detection. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition (on CDROM).

Wu, B., Nevatia, R., 2007. Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet part detectors. International Journal of Computer Vision 75 (2), 247–266.

Zach, C., Frahm, J.-M., Niethammer, M., 2009. Continuous maximal flows and Wulff shapes: Application to MRFs. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition (on CDROM).

Zhang, L., Li, Y., Nevatia, R., 2008. Global data association for multi-object tracking using network flows. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition (on CDROM).

Zhu, Q., Yeh, M.-C., Cheng, K.-T., Avidan, S., 2006. Fast human detection using a cascade of histograms of oriented gradients. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition 2, 1491–1498.