Depth and Appearance for Mobile Scene Analysis

Andreas Ess[1] Bastian Leibe[1] Luc Van Gool[1,2]

[1]ETH Zurich
Zurich, Switzerland
{aess,leibe}@vision.ee.ethz.ch

[2]KU Leuven
Leuven, Belgium
vangool@esat.kuleuven.be

Abstract

In this paper, we address the challenging problem of simultaneous pedestrian detection and ground-plane estimation from video while walking through a busy pedestrian zone. Our proposed system integrates robust stereo depth cues, ground-plane estimation, and appearance-based object detection in a principled fashion using a graphical model. Object-object occlusions lead to complex interactions in this model that make an exact solution computationally intractable. We therefore propose a novel iterative approach that first infers scene geometry using Belief Propagation and then resolves interactions between objects using a global optimization procedure. This approach leads to a robust solution in few iterations, while allowing object detection to benefit from geometry estimation and vice versa. We quantitatively evaluate the performance of our proposed approach on several challenging test sequences showing strolls through busy shopping streets. Comparisons to various baseline systems show that it outperforms both a system using no scene geometry and one just relying on Structure-from-Motion without dense stereo.

1. Introduction

Detecting pedestrians reliably from a moving platform is a fundamental asset for obstacle avoidance and path planning with numerous applications in autonomous driving and mobile robotics. In this paper, we consider a moving platform, equipped with a stereo pair of forward-looking cameras, driving through a busy pedestrian zone (Figure 1). The detection task in this scenario is extremely challenging due to a variety of factors. Firstly, images from unconstrained video streams exhibit a much lower quality than their photographed counterparts due to motion blur, unbayering artifacts, and varying lighting conditions. Secondly, the large number of independently moving objects, covering sometimes up to 50% of the image, leads to frequent partial occlusions between pedestrians, which is problematic for standard object detection and tracking techniques. Thirdly, even state-of-the-art pedestrian detectors are challenged by the large range of scales, the multitude of viewpoints, and the ambiguity of side vs. semi-frontal views of pedestrians. Finally, the suboptimal camera placement dictated by constraints on the platform (approx. 90 cm above ground) has adverse effects on the accuracy of depth measurements (*e.g.* for an object 20 meters away, a localization error of 1 pixel in y direction equals about 1 meter in depth in the scene). Building a reliable system in such highly dynamic scenes thus calls for a tight interaction between multiple cues.

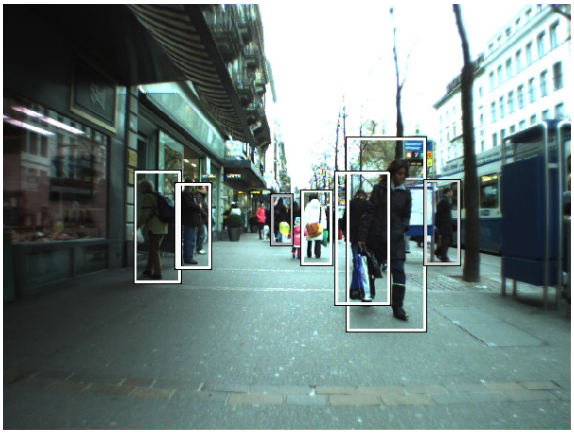

Figure 1. Pedestrian detection obtained using our system in a video stream taken from a moving platform. Our system has to cope with suboptimal imaging conditions and frequent partial occlusions.

In this paper, we will focus on the robust detection of pedestrians in a single video frame. Using input from pedestrian detection and dense stereo, we want to jointly estimate scene geometry and object locations. The different cues are integrated in a graphical model, allowing inference in all directions (Figure 2). However, to correctly model the problem, interactions between object detections have to be taken into account. As overlapping detections may be based on the same image pixels, they lead to implicit loops over the image $\mathcal{I}$ that are very hard to resolve using Belief Propagation. Thus inference may be biased by an accumulation of strong detections in some image region, even though only a subset of them would be consistent with each other. Resolving those interactions in the graphical model would require an infeasible modelling of the scene on the pixel level. More

importantly, hard exclusion cannot be adequately modelled in such a framework.

We suggest to solve this problem using an iterative two-step process: Belief Propagation for inference over the ground plane and object bounding boxes supported by it, followed by a global optimization stage that selects a consistent set of hypotheses under the constraint that each pixel can only contribute to a single object hypothesis.

Our main contributions are: 1) We simultaneously estimate scene geometry and detect objects in a challenging real-world scenario (from video input), in particular integrating cues from dense stereo, object detection, and ground-plane estimation. 2) We model this integration in a principled fashion using a graphical model that allows depth measurements to benefit from object detection and vice versa. 3) For inference in this model, we propose a novel iterative procedure that combines Belief Propagation and global optimization to account for object-object interactions. 4) We experimentally validate the proposed approach on challenging real-world data with a total of 2,293 video frames containing 10,958 pedestrian annotations. We make this data available to the community.[1]

The structure of the remaining paper is as follows: after summarizing related work, we present our scenario and motivate our choice for its solution in Section 2. Section 3 details the construction of the graphical model used for formalizing the problem. All important parameters are obtained through training, as shown in Section 4. A novel two-stage process for inferring the MAP estimate is introduced in Section 5. The paper is concluded by an extensive set of experiments on challenging real-world data in Section 6.

Related Work. In recent years, object detection has reached a level where it becomes interesting for practical applications, *e.g*. for detecting pedestrians in real-world scenes [1, 9, 11, 19, 18, 20]. Still, pedestrian detection remains a very difficult task due to the large degree of intra-category variability, changing scale, articulation, and frequent partial occlusion. The importance of context for reliable object detection has therefore been widely recgonized [17, 13, 15]. In a recent publication, Hoiem *et al*. [6] showed how geometric context can be inferred from a single image in conjunction with object detection. We build upon these ideas and extend them for our scenario, with a considerable scale range in detections, frequent partial occlusions, and integrating stereo depth. Most importantly, our novel two-step process resolves ambiguities between interacting pedestrians that cannot be handled in Hoiem's framework.

In another related system, Leibe *et al*. [8] detect objects from a moving vehicle, integrating detection and Structure-from-Motion. In their imagery, objects typically appear in a well-contained scale range. They fit a ground plane through past wheel contact points and then fix it for the detection stages. Such a fitting required temporal look-ahead in our experiments. In contrast, our framework integrates multiple cues to explain the scene causally, *i.e*. using only information from the current and previous frames.

The use of depth cues for improving detections suggests itself in systems equipped with camera pairs. The most notable recent systems taking advantage of depth include the ones by Giebel *et al*. [5] and Gavrila and Munder [4]. However, their pedestrian tracking systems work with the assumption of a fixed groundplane, interactions between pedestrians are not modelled, and no results for busy scenes are shown. In the test sequence of [4], only 1,000 out of 20,000 test images contain pedestrians—often, only one. Here, we consider among others a sequence with 5,500 annotated pedestrians in 1,000 frames.

2. Problem Formulation

Given pedestrian detections in a single video frame $\mathcal{I}$ and its corresponding depth map $\mathcal{D}$ as inputs, we are interested in simultaneously inferring the ground plane and the set of valid pedestrian hypotheses. The hypotheses o_i are obtained from a standard pedestrian detector. Geometric reasoning is conducted over the scene's ground plane $\boldsymbol{\pi}$ and the object bounding boxes. The latter are automatically adapted by our algorithm for better scene explanation. Depth cues d_i are introduced in a robust way, such that the system can cope with faulty depth maps. These components and their interactions are formalized in a graphical model, which is generated on a per-frame basis. To handle interactions between objects in a computationally feasible way, we split the reasoning into two stages. We first obtain an initial estimation of the scene geometry, disregarding overlapping hypotheses. Next, the obtained MAP estimates are passed to a global optimization stage that handles interactions on a pixel level. The obtained results can then be fused between the stereo cameras and used for tracking-by-detection; this is however not a subject of this paper.

3. Graphical Model

Figure 2 shows the graphical model we use for inference over object hypotheses o_i, depth cues d_i, and the ground plane $\boldsymbol{\pi}$ using evidence from the image $\mathcal{I}$, the depth map $\mathcal{D}$, and the depth map's ground plane evidence $\boldsymbol{\pi}_{\mathcal{D}}$. The dashed box indicates repetition of the contained parts for the number of objects n. Inference in this model is performed as follows:

$$P(\boldsymbol{\pi}, o_i, d_i, \boldsymbol{\pi}_{\mathcal{D}}, \mathcal{D}, \mathcal{I}) = P(\boldsymbol{\pi}) P(\boldsymbol{\pi}_{\mathcal{D}}|\boldsymbol{\pi}) \prod_i^n O_i \qquad (1)$$

$$O_i = P(o_i|\boldsymbol{\pi}) P(\mathcal{I}|o_i) P(d_i|o_i, \boldsymbol{\pi}) P(\mathcal{D}|d_i).$$

[1] http://www.vision.ee.ethz.ch/~aess/iccv2007/

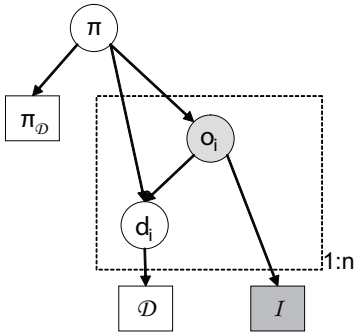

Figure 2. Graphical model for the fusion of object and ground plane detection with the aid of a depth map. The shaded areas indicate implicit loops that are solved in a novel two-stage approach.

Note that the model is loopy: first, there are the obvious cycles between $\boldsymbol{\pi}$, o_i and d_i. More important, however, are the interactions between overlapping object hypotheses o_i that are implicitly introduced by the reliance on a common image $\mathcal{I}$. Those are hard to solve using Belief Propagation, where there is no principle of exclusion. See Section 5 for the method we propose to resolve this problem.

In the following, all 3D calculations are executed in camera coordinates, *i.e.* the projection matrix is $\mathtt{P} = [\mathtt{K}|\mathbf{0}]$. This simplifies parameterizations and keeps the set of possible ground planes in a range that can be trained meaningfully.

Ground Plane. As shown in previous publications [6, 8, 4], the ground plane helps substantially in constraining object detection to meaningful locations. It is defined in the current camera frame as $\boldsymbol{\pi} = (\mathbf{n}, \pi_4)$, with the normal vector $\mathbf{n}(\theta, \phi)$ parameterized by spherical coordinates. We consider a combination of a prior, object bounding boxes, and the ground plane evidence $\boldsymbol{\pi}_{\mathcal{D}}$ for inferring $\boldsymbol{\pi}$. Therefore, our system is not critically dependent on any of these cues individually.

Inference is adapated on a per-frame basis using information obtained from $\mathcal{D}$. Specifically, we consider the robust, depth-weighted median residual between $\boldsymbol{\pi}$ and $\mathcal{D}$,

$$r(\boldsymbol{\pi}, \mathcal{D}) = \mathrm{med}_{\mathbf{x}\in\mathcal{D}} \|d_\perp(\boldsymbol{\pi}, \mathbf{x})\|_{\mathtt{C}_d}, \tag{2}$$

with $\mathbf{x} \in \mathcal{D}$ the 3D points inferred from $\mathcal{D}$, $d_\perp(\cdot,\cdot)$ the signed plane-point distance, and $\|\cdot\|_{\mathtt{C}_d}$ the Mahalanobis distance for $d_\perp(\cdot,\cdot)$, inferred from a 3D point's uncertainty. We formulate the probability that the real ground plane generated the depth map evidence by means of a 1D Gaussian,

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

where $\sigma_{\mathcal{D}}$ models the trust in $\mathcal{D}$. Again, even though less than 50% of $\mathcal{D}$ might belong to $\boldsymbol{\pi}$, we are not dependent on this cue alone. The prior $P(\boldsymbol{\pi})$ is learned from a training set (see Section 4).

Object Hypotheses. Object hypotheses $o_i = \{v_i, \mathbf{c}_i\}$ $(i = 1 \ldots n)$ are created from the output of a detector on a per-frame basis (typically, $n \approx$ 70–90). They consist of a validity flag $v_i \in \{0, 1\}$ and a 2D center point with scale $\mathbf{c}_i = \{x, y, s\}$. 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's base point $\mathbf{b} = (x, y + s\frac{h}{2}, 1)$ in homogeneous image coordinates, its world counterpart is found as

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

The object's depth is thus $d(o_i) = \|\mathbf{B}_i\|$. Its height $\mathbf{B}^h_i$ is obtained in a similar fashion. Because of the large viewpoint variability, the detector's output for center and scale are only taken as estimates, denoted $\hat{x}_i$, $\hat{y}_i$, and $\hat{s}_i$. Taking these directly may yield misaligned bounding boxes, which in turn can result in wrong estimates for distance and size. We therefore try to compensate for detection inaccuracies by considering a set of possible bounding boxes $bbox_i^{\{k,l\}}$ for each o_i. These boxes are constructed from a set of possible real centers $\mathbf{c}_i = \{y_i, s_i\}$ (fixing $x_i = \hat{x}_i$ due to its negligible influence), which are obtained by sampling around the detection, $y_i = \hat{y}_i + k\sigma_y \hat{s}_i$, $s_i = \hat{s}_i + l\sigma_s \hat{s}_i$. The number of samples, *i.e.* the range of $\{k, l\}$, is the same for every object. 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), $P(\mathbf{c}_i|\boldsymbol{\pi}) \propto P(\mathbf{B}^h_i) P(d(o_i))$ is expressed as the product of a distance and a size prior for the corresponding 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 fail frequently, especially in untextured regions. We thus integrate depth cues into our framework in a robust manner. For each object hypothesis, we consider a depth flag $d_i \in \{0, 1\}$, indicating whether the depth map for a bounding box is reliable $(d_i = 1)$ or not. The depth cue term $P(d_i|o_i, \boldsymbol{\pi})$ is rearranged as follows:

$$P(d_i|o_i, \boldsymbol{\pi}) \propto P(v_i|\mathbf{c}_i, d_i, \boldsymbol{\pi}) P(\mathbf{c}_i|d_i, \boldsymbol{\pi}) P(d_i), \tag{6}$$

assuming a uniform object prior $P(o_i)$. We consider two measurements. Firstly, we evaluate the depth measured inside $bbox_i$ and its consistency with $d(o_i)$ as an indicator for $P(\mathbf{c}_i|d_i = 1, \boldsymbol{\pi})$. Secondly, we test whether this depth information can be considered largely uniform, which reflects our expectation in case a pedestrian is present. This can be used for defining $P(v_i = 1|\mathbf{c}_i, d_i = 1, \boldsymbol{\pi})$.

The measurements are defined as follows: the median depth inside a bounding box, $d(d_i) = \mathrm{med}_{\mathbf{p}\in bbox_i} \mathcal{D}(\mathbf{p})$, yields a robust estimate of the contained object's depth. $\mathcal{D}(\mathbf{p})$ denotes the depth of pixel $\mathbf{p}$. Assuming additive white noise with covariance $\mathtt{C}_{2\mathrm{D}}$ on pixel measurements, we find the variance σ^2_{di} of $d(d_i)$ using error backpropagation,

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

where $\mathtt{F}_i^j$ are the Jacobians of a projection using camera matrix j. Thus, $\sigma_{di}^2 = C_i(3,3)$. This yields

$$P_{di}(x) \propto \mathcal{N}(x; d(d_i), \sigma_{di}^2). \tag{8}$$

For reasoning about depth uniformity, we consider the variation of depth in $bbox_i$, $V = \{\mathcal{D}(\mathbf{p}) - d(d_i) | \mathbf{p} \in bbox_i\}$, specifically taking V's interquartile range $[LQ, UQ]$ as a robust estimator. Our measure of uniformity is the normalized count of pixels that are within the depth confidence σ_{di},

$$q_i = \frac{\left|\{x \in [LQ, UQ]\right| - \sigma_{di} < x < \sigma_{di}\}|}{UQ - LQ}. \tag{9}$$

This robust "depth inlier percentage" serves as basis for learning $P(v_i|\mathbf{c}_i, \mathbf{d}_i = 1, \boldsymbol{\pi})$, as is shown in Section 4.

$P(o_i|d_i = 0)$ is assumed to be uniform, as an inaccurate depth map gives no information about the object's presence. We force $d_i = 0$ on image borders, where there is no image overlap and hence no depth information.

4. Training

Data is recorded at a resolution of 640×480 pixels (Bayered) at 15 FPS, with a camera baseline of 0.4 meters. The system's parameters have been trained on a video with 490 frames, containing 1,578 annotations. For the ground plane prior, we consider an 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 [7]. 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 reasons of 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. In our tests, the errors ensuing the discretization of $\boldsymbol{\pi}$ were below 0.05 meters in depth for a pedestrian 15 meters away.

The training sequences are also used for constructing the prior distribution $P(\boldsymbol{\pi})$. Figure 3 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 [9], trained on frontal and side views of pedestrians. The detector is run without the final global optimization stage, thus retaining flexibility in our system. The range of detected object scales is 60–400 pixels. Other detectors can be included in our system, as long as they provide confidence maps.

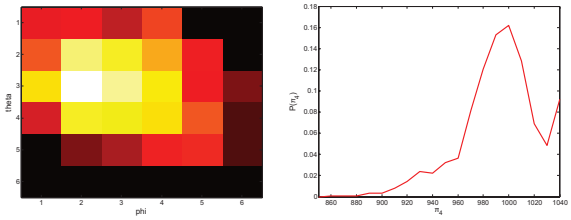

Figure 3. Learned priors for (θ, ϕ) (left) and π_4 (right), marginalized over π_4 and (θ, ϕ), respectively.

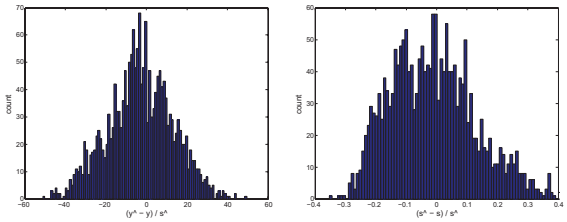

Figure 4. Center distributions normalized by detected scale $\hat{s}$ (left: center $(\hat{y} - y)/\hat{s}$, right: scale $(\hat{s} - s)/\hat{s}$), learned from 1,578 annotations. We approximate these using normal distributions.

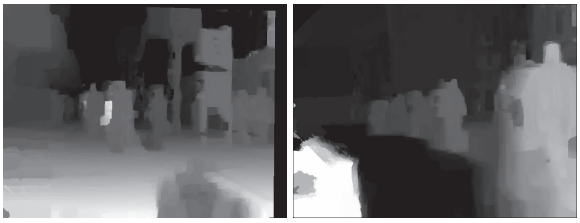

Figure 5. Example depth maps. Often, useful cues can be inferred (left), but robust measures have to account for faulty depth maps, *e.g.* missing ground plane (right).

As the original detection centers $\hat{x}, \hat{y}, \hat{s}$, 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. For this, we collect detections over the training sequence and compare them to ground-truth annotations. Figure 4 shows the resulting scale-normalized measurements $(\hat{y} - y)/\hat{s}$, $(\hat{s} - s)/\hat{s}$ used to learn (σ_y, σ_s). As can be seen from the figure, the Gaussian approximation is justified.

The height distribution is chosen as $P(\mathbf{B}^h|h) \sim \mathcal{N}(1.7, 0.12^2)$ (meters), though we consider different standard deviations σ_h in a first systematic experiment in Section 6. This is mainly to account for remaining discretization errors due to the sampling of $\mathbf{c}_i$, as well as the occasional occurence of children in the sequences. The depth distribution $P(d(o_i))$ is assumed uniform in our system's operating range, *i.e.* 0.5–30 meters distance.

Depth Cues. The depth map $\mathcal{D}$ for each frame is obtained using a publicly available version of Belief Propagation-based stereo [3]. See Figure 5 for two example depth maps.

The true distribution of $P(\mathbf{c}_i|d_i = 1, \boldsymbol{\pi})$, given the object's depth $d(o_i)$ and the robust depth map estimate $d(d_i)$,

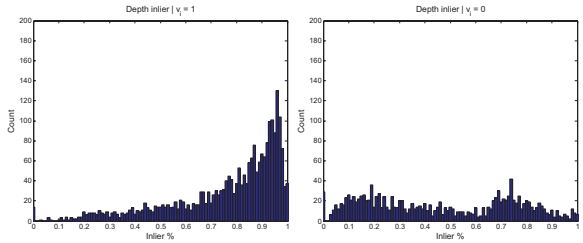

Figure 6. Distribution of robust depth inliers for correct (left) and incorrect (right) detections, learned from 1,578 annotations and 1,478 negative examples.

is very intricate to find. It involves many factors: firstly, the uncertainty of the object's center propagated to its distance. Due to the sampling of $\mathbf{c}_i$, we can neglect this factor. Secondly, it depends on P_{di} 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_{di}.

In Section 6, we compare two ways for modelling $P(\mathbf{c}_i|d_i = 1, \boldsymbol{\pi})$. The first option uses a non-parametric distribution $P(v_i|d(o_i) - d(d_i))$, learned from the training sequence. The second option models it using the dominating factor $P_{di}(d(o_i))$ only.

For learning $P(v_i|\mathbf{c}_i, d_i = 1, \boldsymbol{\pi})$, 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 Figure 6, q_i is a good indicator of an object's presence. Using logistic regression, we fit a sigmoid to arrive at $P(v_i|\mathbf{c}_i, d_i = 1, \boldsymbol{\pi})$. In Section 6, we also test the use of $P(v_i = 1|\mathbf{c}_i, d_i = 1, \boldsymbol{\pi}) = \max P(\mathbf{c}_i|d_i = 1, \boldsymbol{\pi})$. Using the training set, we found $P(d_i = 1) \approx 0.96$.

5. Inference

The exact modelling of interactions between different hypotheses is of paramount importance in highly dynamic scenarios, where pedestrians overlap frequently and compete for the same pixels. To overcome the missing notion of exclusion in a Belief Propagation framework, we suggest a novel two-stage procedure that first infers geometric context using possibly overlapping bounding boxes and then applies a global optimization step that models interactions between different objects on a pixel level, using the detector's confidence maps.

Belief Propagation. The graph of Figure 2 is constructed in Matlab using the BayesNet toolbox [12], with all variables modelled as discrete entities and their conditional probability tables defined as described above. Inference is conducted using Pearl's Belief Propagation [16]. Due to the loopy nature of our model, this yields only an approximate solution. We found this to be more than sufficient in our application, which is also confirmed by other researchers' experience [14].

Global Optimization. As stated above, the reliance of object hypotheses on a common image introduces implicit loops in our graphical model, since overlapping detections cannot be considered independent. Intuitively, each image pixel can only be explained by a single object, therefore some detections are mutually exclusive. The idea of our approach is to make this dependence explicit and use a Quadratic Boolean Optimization formulation to select a subset of object detections that are mutually consistent.

Starting from the validity flags $v_i \in \{0, 1\}$, we want to optimize the function

$$\max_{\mathbf{v}} \mathbf{v}^\top \mathtt{Q}\mathbf{v} = \max_{\mathbf{v}} \mathbf{v}^\top \begin{bmatrix} q_{11} & \cdots & q_{1n} \\ \vdots & \ddots & \vdots \\ q_{n1} & \cdots & q_{nn} \end{bmatrix} \mathbf{v}, \qquad (11)$$

where the interaction matrix $\mathtt{Q}$ contains individual merit terms in the diagonal elements q_{ii} and (negative) interaction terms in the off-diagonal elements $\{q_{ij}, q_{ji}\}$.

Using a similar derivation as in [9, 10], we express a detection's score in terms of the pixels $\mathbf{p}$ it occupies (normalized by the detection scale),

$$\begin{aligned} P(o_i|\boldsymbol{\pi}, d_i, \mathcal{I}) &= P(o_i|\boldsymbol{\pi}, d_i) P(o_i|\mathcal{I}) \qquad (12) \\ &\sim P(o_i|\boldsymbol{\pi}, d_i) \prod_{\mathbf{p} \in o_i} P(\mathbf{p}|o_i), \end{aligned}$$

with $P(o_i|\boldsymbol{\pi}, d_i)$ the MAP estimate from Belief Propagation. Let $L_i = \log P(o_i|\boldsymbol{\pi}, d_i)$ and $F_i(\mathbf{p}) = P(\mathbf{p}|o_i)$. We define the cost of a detection by its log-likelihood in a first-order approximation,

$$\begin{aligned} \mathcal{S} &= \log \left[P(o_i|\boldsymbol{\pi}, d_i) \prod_{\mathbf{p} \in o_i} F_i(\mathbf{p}) \right] = L_i + \sum_{\mathbf{p} \in o_i} \log F_i(\mathbf{p}) \qquad (13) \\ &= L_i - \sum_{\mathbf{p} \in o_i} \sum_{n=1}^{\infty} \frac{1}{n} (1 - F_i(\mathbf{p}))^n \approx L_i - N + \sum_{\mathbf{p} \in o_i} F_i(\mathbf{p}) \, . \end{aligned}$$

Following [9], we thus arrive at the following merit terms

$$q_{ii} = -\kappa_1 + \sum_{\mathbf{p} \in o_i} ((1 - \kappa_2) + \kappa_2 F_i(\mathbf{p})) - \kappa_2 L_i, \qquad (14)$$

where κ_2 is a regularization term to compensate for unequal sampling and κ_1 is a counterweight. Two object detections o_i and o_j interact if they compete for the same pixels. In this case, we subtract the support of the detection $o_k \in \{o_i, o_j\}$ that is farther away from the camera in the overlapping image area, assuming it is partially occluded:

$$q_{ij} = -\frac{1}{2} \Big(\sum_{\mathbf{p} \in o_i \cap o_j} ((1 - \kappa_2) + \kappa_2 F_k(\mathbf{p})) + \kappa_2 L_k \Big). \qquad (15)$$

Using this formulation, we implement the following iterative procedure. We initialize the Quadratic Problem with the MAP estimate $P(o_i|\boldsymbol{\pi}, d_i)$ and then solve it using standard optimization techniques [9, 10]. This results in a subset of mutually consistent detections $\{o_i^*\}$, which are then again used to obtain a more stable ground plane estimate. In our experiments, this procedure converged to a stable solution in only few iterations.

6. Experimental Results

We experimentally validate our system on 3 test sequences of busy shopping streets, taken on different days and under different weather conditions. In the following, we describe the data, perform systematic experiments to underline some design choices and then apply the system with fixed parameters. For a detection to be counted as correct, it has to overlap with an annotation by more than 50% using the intersection-over-union measure [2]. Only one detection per annotation is counted as correct. For the experiments, only the left camera is evaluated. Performance could be improved further by integrating the right camera and adding temporal smoothing [5, 8], which is not yet done here.

Data. Data has been recorded at a resolution of 640×480 pixels (Bayered) and 15 FPS using a stereo pair of cameras mounted on a children's stroller. The streams suffer from unbayering artifacts, slight motion blur, and sometimes missing contrast. All frames are completely annotated up to a distance of $\approx$25m, resulting in a total of 2,293 frames and 10,958 annotations. The training sequence contains 490 frames with 1,578 annotations, and shows a walk over a fairly busy square on a cloudy day. The first test sequence (999 frames and 5,193 annotations) was taken under similar weather conditions, strolling on a sidewalk. Its main challenges are a large number of trees and dust bins that result in false positives (FP) from object detection, persons getting off public transport, as well as reflections from shopping windows. The second test sequence (450 frames, 2,359 annotations) shows a stroll over a busy square, with people moving in all directions. The square lies in the shade, resulting in low contrast and thus increased difficulty for the detector. Furthermore, the depth map is in many instances completely unusable for estimating $P(\boldsymbol{\pi}_{\mathcal{D}}|\boldsymbol{\pi})$, see Figure 5 (right). The third test sequence (354 frames, 1,828 annotations) was taken on a sunny day on a sidewalk, and contains a large number of shadows and reflections.

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.

Firstly, 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. Figure 7 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 misadjusted bounding boxes. A 3×3 sampling with $\sigma_h = 0.12$ thus seems a good compromise.

Secondly, we experimentally establish how to integrate

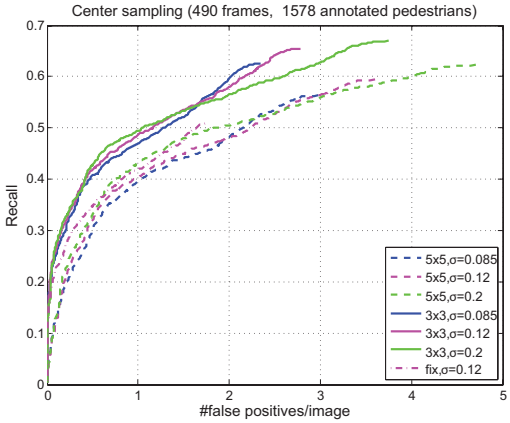

Figure 7. Influence of center/scale sampling and σ_h on performance. In all future experiments, we use 3×3 sampling and $\sigma_h = 0.12$.

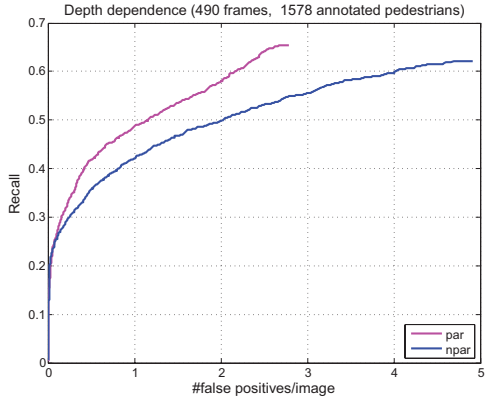

Figure 8. Influence of depth term choice on performance, a parametric distribution performs better.

the depth cues into our system. For $P(\mathbf{c}_i|d_i = 1, \boldsymbol{\pi})$, we consider either the learned, non-parametric distribution $P(v_i|d(o_i) - d(d_i))$ ("npar") or a normal distribution inferred from Eq. (8) ("par"). As can be seen from the result plot (Figure 8), 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.

Experimental Validation. With all parameter choices motivated in the previous sections, we now apply the proposed system to a set of challenging test sequences of strolls through busy pedestrian passages. In these experiments, we also compare our system to a set of baseline configurations emulating other approaches from the literature (Figure 10): "detector" refers to the output of the pedestrian detector, without its global optimization stage. This is the input to our system. "det.+opt" includes the optimization and is therefore a fair baseline comparison. Neither of these two approaches use scene geometry. The setup "det.+real GP" is motivated by [8]. It does not consider depth cues and is

obtained by preselecting a ground plane for each frame (determined using robust, LMedS-based plane fitting through reconstructed wheel contact points), with a temporal look-ahead corresponding to a travelling distance of ≈ 5m. Note that without this look-ahead, we could not get usable estimates for the ground plane using our hardware setup. "GM" stands for the MAP estimate obtained using Belief Propagation, "full sys." is the final output of our proposed approach, including the global optimization. For the first sequence, we also consider the full system without the depth uniformity cue, "full sys. (no dv)".

Figure 10 shows performance plots. On its own, the detector's precision is low, as its score is not distinctive enough. Slightly better results are obtained by including the global optimization. Substantially better results however ensue from including scene and depth information using the graphical model (resulting in an 8% gain in recall at 1.5 FP/image). This still disregards object-object interactions. The baseline "det.+real GP" considers these, but relies on a preselected groundplane, yielding an advantage compared to the detector, with a slight improvement over the MAP estimate. Compared to this baseline, the full system increases recall by a significant 19% at 1.5 FP/image. By replacing the depth uniformity cue with $P(v_i = 1|\mathbf{c}_i, d_i = 1, \boldsymbol{\pi}) = \max P(\mathbf{c}_i|d_i = 1, \boldsymbol{\pi})$, performance drops by 7%, showing that this additional depth information is indeed beneficial. The plots for sequences #2 and #3 corroborate the advantage of our approach. Some example detections on those test sequences are displayed in Figure 11.

Depth Dependency. Note that even at the chosen low threshold for the pedestrian detector, with many false positives, only a recall of about 70% is reached. The reason for this is our challenging test set with significant partial occlusions and many pedestrians appearing at small scales. To further investigate the influence of a pedestrian's distance on recognition performance, Figure 9 shows the average distance distribution of annotated pedestrians in the first test sequence together with the agreeing detections (top left/right). For distant pedestrians, the detector becomes less reliable. For the bottom plot, we fixed the operating point at 1 resp.1.5 FP/image on the global curve, and plot recall and FP/image over depth (full system). Recall is considerably higher for distances up to 15m and rapidly decreases after that, which coincides with the number of available detections.

7. Conclusion

We have presented a system that integrates depth and appearance information for robust pedestrian detection and simultaneous ground-plane estimation from video streams. Based on input cues from pedestrian detection and depth maps, it constructs a graphical model and resolves it in a novel two-stage process. We conducted a comprehensive set of experiments on challenging videos from busy shopping streets that underline the advantage of our integrative approach. The key message of our experiments is that, given a reasonable pedestrian detector, our algorithm gives it a considerable boost in performance. This is due to the integration of spatial constraints in the form of robust depth cues and the ground plane, and our system's ability to compensate for inaccuracies of the detector, and to resolve object-object interactions. So far, temporal context and the second camera are ignored, but the results could obviously be improved by taking advantage of those. In addition, we plan to evaluate depth information for recognizing other kinds of obstacles.

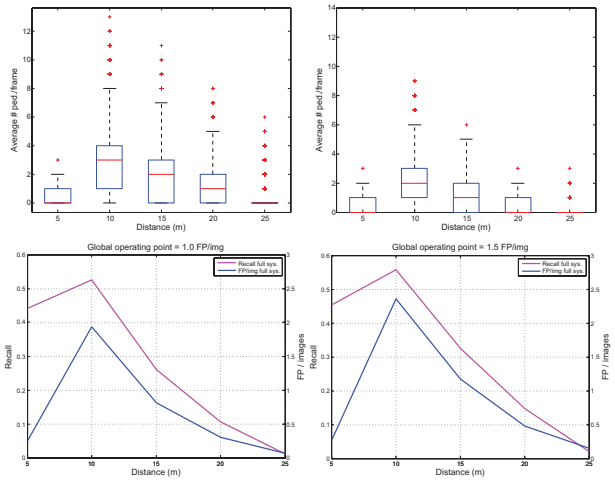

Figure 9. Top: Distribution of pedestrians over distance. Left: annotations, Right: correct detections. Bottom: Recall and FP/image at globally fixed operating points over distance in seq. #1.

Acknowledgments. This project has been funded in parts by Toyota Motor Corporation/Toyota Motor Europe and the EU project DIRAC (IST-027787). The authors thank Martin Vogt for doing the majority of the annotations.

References

[1] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[2] M. Everingham and others (34 authors). The 2005 pascal visual object class challenge. In *Selected Proceedings of the 1st PASCAL Challenges Workshop*, LNAI. Springer, 2006.

[3] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient belief propagation for early vision. *IJCV*, 70, 2006.

[4] D. Gavrila and S. Munder. Multi-cue pedestrian detection and tracking from a moving vehicle. *IJCV*, 73, 2007.

[5] J. Giebel, D. Gavrila, and C. Schnörr. A bayesian framework for multi-cue 3d object tracking. In *ECCV*, 2004.

[6] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. In *CVPR*, 2006.

[7] R. Labayrade, D. Aubert, and J.-P. Tarel. Real time obstacle detection on non flat road geometry through 'v-disparity' representation. In *IVS*, 2002.

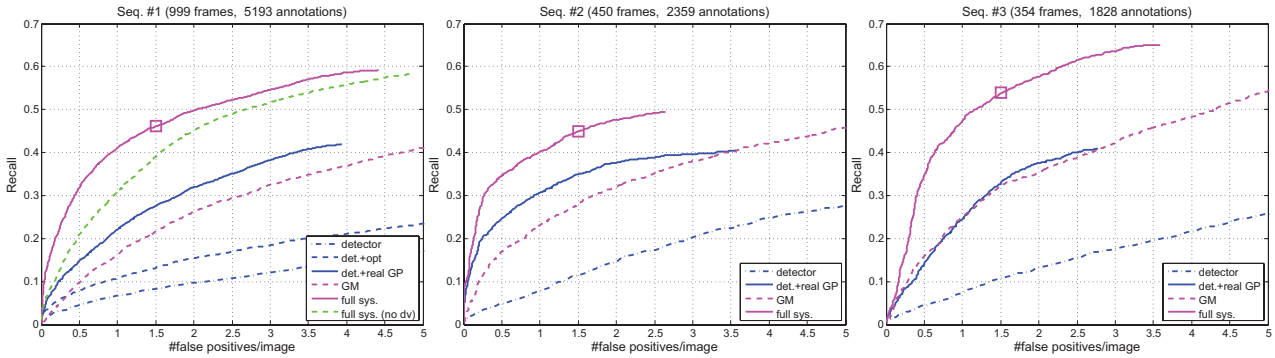

Figure 10. Performance of different system parts and baselines for the test sequences. The interaction of cues yields a substantial increase in performance. The square indicates the operating point used for sample images and video.

Figure 11. Experimental results obtained on 3 test sequences. Note the level of interaction between pedestrians (frequently overlapping bounding boxes). The images also show some typical false positives in red (trees, children's stroller, signs, mannequin).

[8] B. Leibe, N. Cornelis, K. Cornelis, and L. V. Gool. Dynamic 3d scene analysis from a moving vehicle. In *CVPR*, 2007.

[9] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *CVPR*, 2005.

[10] A. Leonardis, A. Gupta, and R. Bajcsy. Segmentation of range images as the search for geometric parametric models. *IJCV*, 14:253–277, 1995.

[11] K. Mikolajczyk, B. Leibe, and B. Schiele. Multiple object class detection with a generative model. In *CVPR*, 2006.

[12] K. Murphy. The bayes net toolbox for matlab. In *Computing Science and Statistics*, 2001.

[13] K. Murphy, A. Torralba, and W. Freeman. Using the forest to see the trees: A graphical model relating features, objects, and scenes. In *NIPS*, 2003.

[14] K. Murphy, Y. Weiss, and M. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *UAI*, pages 467–475, 1999.

[15] B. Ommer and J. M. Buhmann. Object categorization by compositional graphical models. In *EMMCVPR*, 2005.

[16] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[17] E. B. Sudderth, A. B. Torralba, W. T. Freeman, and A. S. Willsky. Learning hierarchical models of scenes, objects, and parts. In *ICCV*, 2005.

[18] P. Viola and M. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004.

[19] B. Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *ICCV*, 2005.

[20] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan. Fast human detection using a cascade of histograms of oriented gradients. In *CVPR*, 2006.