Monocular 3D Scene Modeling and Inference: Understanding Multi-Object Traffic Scenes

Christian Wojek[1,2], Stefan Roth[1], Konrad Schindler[1,3], and Bernt Schiele[1,2]

[1]Computer Science Department, TU Darmstadt
[2]MPI Informatics, Saarbrücken
[3]Photogrammetry and Remote Sensing Group, ETH Zürich

Abstract. Scene understanding has (again) become a focus of computer vision research, leveraging advances in detection, context modeling, and tracking. In this paper, we present a novel probabilistic 3D scene model that encompasses multi-class object detection, object tracking, scene labeling, and 3D geometric relations. This integrated 3D model is able to represent complex interactions like inter-object occlusion, physical exclusion between objects, and geometric context. Inference allows to recover 3D scene context and perform 3D multi-object tracking from a mobile observer, for objects of multiple categories, using only monocular video as input. In particular, we show that a joint scene tracklet model for the evidence collected over multiple frames substantially improves performance. The approach is evaluated for two different types of challenging on-board sequences. We first show a substantial improvement to the state-of-the-art in 3D multi-people tracking. Moreover, a similar performance gain is achieved for multi-class 3D tracking of cars and trucks on a new, challenging dataset.

1 Introduction

Robustly tracking objects from a moving observer is an active research area due to its importance for driver assistance, traffic safety, and autonomous navigation [1, 2]. Dynamically changing backgrounds, varying lighting conditions, and the low viewpoint of vehicle-mounted cameras all contribute to the difficulty of the problem. Furthermore, to support navigation, object locations should be estimated in a global 3D coordinate frame rather than in image coordinates.

The main goal of this paper is to address this important and challenging problem by proposing a new *probabilistic 3D scene model*. Our model builds upon several important lessons from previous research: (1) robust tracking performance is currently best achieved with a *tracking-by-detection* framework [3]; (2) short term evidence aggregation, typically termed *tracklets* [4], allows for increased tracking robustness; (3) the objects should not be modeled in isolation, but in their *3D scene context*, which puts strong constraints on the position and motion of tracked objects [1, 5]; and (4) *multi-cue combination* of scene labels and object detectors allows to strengthen weak detections, but also to prune inconsistent false detections [5]. While all these different components have been shown to boost performance individually, in the present work we for the first time integrate them all in a single system. As our experiments show, the proposed probabilistic 3D scene model significantly outperforms the current state-of-the-art. Fig. 1

K. Daniilidis, P. Maragos, and N. Paragios (Eds.): ECCV 2010, Part ??, LNCS ????, pp. ???–???, 2010.

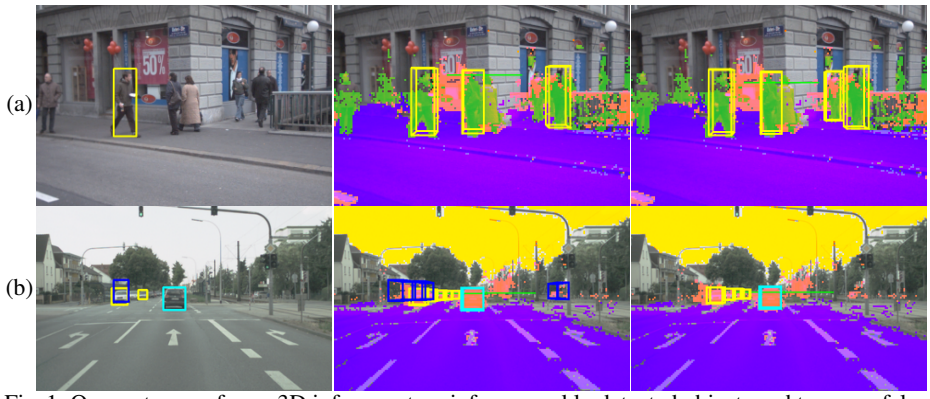


Fig. 1: Our system performs 3D inference to reinforce weakly detected objects and to prune false positive detections by exploiting evidence from scene labeling and an object detector. *(left)* Detector input; *(middle)* single-frame 3D inference with overlaid scene labeling and horizon estimate; *(right)* multi-frame tracking results (all results at 0.1 FPPI). See Sec. 6 for a detailed discussion.

shows example results for two different types of challenging onboard sequences. Our system is able to robustly track a varying number of targets in 3D world coordinates in highly dynamic scenes. This enables us to use a single camera only instead of relying on stereo cameras as in previous work (e.g., [1, 2]).

Despite using only monocular input, the proposed model allows to constrain object detections to geometrically feasible locations and enforces physically plausible 3D dynamics. This improves object detection results by pruning physically implausible false positives and strengthening weak detections along an object's trajectory. We demonstrate that accumulating scene evidence over a small number of frames with help of a 3D scene model significantly improves performance. As exact inference is intractable we employ reversible-jump Markov Chain Monte Carlo (RJMCMC) sampling to approximate per-frame distributions. Further improvement can be achieved by performing long-term data association with a Hidden Markov Model (HMM).

2 Related Work

Our work builds on recent advances in scene understanding by pixel-wise labeling, 3D scene analysis and tracking. The use of scene context has been investigated in the computer vision literature in several ways. Torralba [6] proposes to employ Gabor filter-bank responses in a bottom-up fashion in order to gain prior information on likely 2D object positions. More recently, Shotton et al. [7] use a strong joint-boosting classifier with context reasoning based on a CRF framework to provide a local, per pixel classification of image content. Ess et al. [8] and Brostow et al. [9] particularly address traffic scene understanding. [8] uses 2D Walsh-Hadamard filter-bank responses together with stereo depth information to infer traffic situations, while [9] leverages 3D point clouds to improve 2D scene segmentation. Tu et al. [10] use MCMC sampling techniques to combine top-down discriminative classifiers with bottom-up generative models for 2D image understanding. Common to these approaches is the goal of 2D image understand-

ing. Our work includes scene labeling as a cue, but its ultimate goal is to obtain a 3D model of the observed world.

This paper is most similar to work by Hoiem et al. [5] and Ess et al. [1]. [5] combines image segmentation and object detections in order to infer the objects' positions in 3D. Their work, however, is limited to single images and does not exploit temporal information available in video. [1] extends [5], but requires a stereo camera setup to achieve robust tracking of pedestrians from a mobile platform. Similarly, [2] tracks pedestrians for driver assistance applications and employs a stereo camera to find regions of interest and to suppress false detections. Note, however, that stereo will yield only little improvement in the far field, because a stereo rig with a realistic baseline will have negligible disparity. Thus, further constraints are needed, since appearance-based object detection is unreliable at very small scales. Therefore, we investigate the feasibility of a monocular camera setup for mobile scene understanding. Another system that uses monocular sequences is [11]. Contrary to this work, we tightly couple our scene model and the hypothesized positions of objects with the notion of scene tracklets, and exploit constraints given by a-priori information (e.g., approximate object heights and camera pitch). Our experiments show that these short-term associations substantially stabilize 3D inference and improve robustness beyond what has previously been reported. Our experimental results show that the proposed approach outperforms the stereo system by Ess et al. [1].

Tracking-by-detection, with an offline learned appearance model, is a popular approach for tracking objects in challenging environments. Breitenstein et al. [12], for instance, track humans based on a number of different detectors in image coordinates. Similarly, Okuma et al. [3] track hockey players in television broadcasts. Huang et al. [13] track people in a surveillance scenario from a static camera, grouping detections in neighboring frames into *tracklets*. Similar ideas have been exploited by Kaucic et al. [4] to track vehicles from a helicopter, and by Li et al. [14] to track pedestrians with a static surveillance camera. However, none of these tracklet approaches exploit the strong constraints given by the size and position of other objects, and instead build up individual tracks for each object. In this paper we contribute a probabilistic scene model that allows to jointly infer the camera parameters and the position of *all* objects in 3D world coordinates by using only monocular video and odometry information. Increased robustness is achieved by extending the tracklet idea to *entire scenes* toward the inference of a global scene model.

Realistic, but complex models for tracking including ours are often not amenable to closed-form inference. Thus, several approaches resort to MCMC sampling. Khan et al. [15] track ants and incorporate their social behavior by means of an MRF. Zhao et al. [16] use MCMC sampling to track people from a static camera. Isard&MacCormick [17] track people in front of relatively uncluttered backgrounds from a static indoor camera. All three approaches use rather weak appearance models, which prove sufficient for static cameras. Our model employs a strong object detector and pixel-wise scene labeling to cope with highly dynamic scenes recorded from a moving platform.

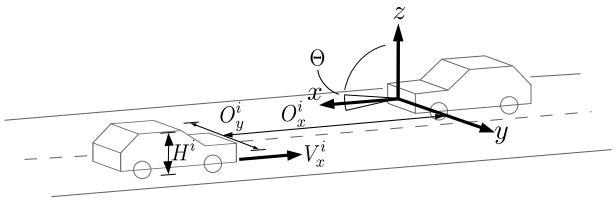


Fig. 2: Visualization of the 3D scene state $\mathbf{X}$ in the world coordinate system. The camera is mounted to the vehicle on the right.

3 Single-Frame 3D Scene Model

We begin by describing our 3D scene model for a *single image*, which aims at combining available prior knowledge with image evidence in order to reconstruct the 3D positions of all objects in the scene. For clarity, the time index t is omitted when referring to a single time step only. Variables in image coordinates are printed in lower case, variables in 3D world coordinates in upper case; vectors are printed in bold face.

The posterior distribution for the 3D scene state $\mathbf{X}$ given image evidence $\mathcal{E}$ is defined in the usual way, in terms of a prior and an observation model:

$$P(\mathbf{X}|\mathcal{E}) \propto P(\mathcal{E}|\mathbf{X})P(\mathbf{X}) \tag{1}$$

The 3D state $\mathbf{X}$ consists of the individual states of all objects $\mathbf{O}^i$, described by their relative 3D position $(O_x^i, O_y^i, O_z^i)^\top$ w.r.t. the observer and by their height H^i. Moreover, $\mathbf{X}$ includes the internal camera parameters $\mathbf{K}$ and the camera orientation $\mathbf{R}$.

The goal of this work is to infer the 3D state $\mathbf{X}$ from video data of a monocular, forward facing camera (see Fig. 2). While in general this is an under-constrained problem, in robotic and automotive applications we can make the following assumptions that are expressed in the prior $P(\mathbf{X})$: The camera undergoes no roll and yaw w.r.t. the platform, its intrinsics $\mathbf{K}$ are constant and have been calibrated off-line, and the speed and turn rate of the platform are estimated from odometer readings. Furthermore, the platform as well as all objects of interest are constrained to stand on a common ground plane (i.e., $O_z^i = 0$). Note that under these assumptions the ground plane in camera-centric coordinates is fully determined by the pitch angle Θ. As the camera is rigidly mounted to the vehicle, it can only pitch a few degrees. To avoid degenerate camera configurations, the pitch angle is therefore modeled as normally distributed around the pitch of the resting platform as observed during calibration: $\mathcal{N}(\Theta; \mu_\Theta, \sigma_\Theta)$. This prior allows deviations arising from acceleration and braking of the observer. This is particularly important for the estimation of distant objects as, due to the low camera viewpoint, even minor changes in the pitch may cause a large error for distance estimation in the far field.

Moreover, we assume the height of all scene objects to follow a normal distribution around a known mean value, which is specific for the respective object class c_i, $\mathcal{N}(H^i; \mu_H^{c_i}, \sigma_H^{c_i})$. This helps to prune false detections that are consistent with the ground plane, but are of the wrong height (e.g., background structures such as street lights). The overall prior is thus given as

$$P(\mathbf{X}) \propto \mathcal{N}(\Theta; \mu_\Theta, \sigma_\Theta) \cdot \prod_i \mathcal{N}(H^i; \mu_H^{c_i}, \sigma_H^{c_i}) \tag{2}$$

Next, we turn to the observation model $P(\mathcal{E}|\mathbf{X})$. The image evidence $\mathcal{E}$ is comprised of a set of potential object detections and a scene labeling, i.e., category labels densely estimated for every pixel. As we will see in the experiments, the combination of these two types of image evidence is beneficial as object detections give reliable but rather coarse bounding boxes, and low level cues enable more fine-grained data association by penalizing inconsistent associations and supporting consistent, but weak detections.

For each object our model fuses object appearance given by the object detector confidence, geometric constraints, and local evidence from bottom-up pixel-wise labeling:

$$P(\mathcal{E}|\mathbf{X}) \propto \prod_i \Psi_D\big(\mathbf{d}^{a(i)}\big) \cdot \Psi_G\big(\mathbf{O}^i, \Theta; \mathbf{d}^{a(i)}\big) \cdot \Psi_L^i\big(\mathbf{X}; \mathbf{l}\big) \tag{3}$$

Here, $a(i)$ denotes the association function, which assigns a candidate object detection $\mathbf{d}^{a(i)}$ to every 3D object hypothesis $\mathbf{O}^i$. Note that the associations between objects and detections are established as part of the MCMC sampling procedure (see Sec. 3.2). The appearance potential Ψ_D maps the appearance score of detection $\mathbf{d}^{a(i)}$ for object i into the positive range. Depending on the employed classifier, we use different mappings – see Sec. 5 for details.

The geometry potential Ψ_G models how well the estimated 3D state $\mathbf{O}^i$ satisfies the geometric constraints due to the ground plane specified by the camera pitch Θ. Denoting the projection of the 3D position $\mathbf{O}^i$ to the image plane as $\mathbf{o}^i$, the distance between $\mathbf{o}^i$ and the associated detection $\mathbf{d}^{a(i)}$ in x-y-scale-space serves as a measure of how much the geometric constraints are violated. We model Ψ_G using a Gaussian

$$\Psi_G(\mathbf{O}^i, \Theta; \mathbf{d}^{a(i)}) = \mathcal{N}(\mathbf{o}^i; \mathbf{d}^{a(i)}, \boldsymbol{\sigma}_G + \bar{\boldsymbol{\sigma}}_G) \;, \tag{4}$$

where we split the kernel bandwidth into a constant component $\boldsymbol{\sigma}_G$ and a scale-dependent component $\bar{\boldsymbol{\sigma}}_G$ to account for inaccuracies that arise from the scanning stride of the sliding-window detectors.

The scene labeling potential Ψ_L^i describes how well the projection $\mathbf{o}^i$ matches the bottom-up pixel labeling. For each pixel j and each class c the labeling yields a classification score $l^j(c)$. Similar to Ψ_D, the labeling scores are normalized pixel-wise by means of a softmax transformation in order to obtain positive values.

It is important to note that this cue demands 3D scene modeling: To determine the set of pixels that belong to each potential object, one needs to account for inter-object occlusions, and hence know the objects' depth ordering. Given that ordering, we proceed as follows: each object is back-projected to a bounding box $\mathbf{o}^i$, and that box is split into a visible region $\boldsymbol{\delta}^i$ and an occluded region $\boldsymbol{\omega}^i$. The object likelihood is then defined as the ratio between the cumulative score for the expected label e and the cumulative score of the pixel-wise best label $k \neq e$, evaluated over the visible part of $\mathbf{o}^i$:

$$\Psi_L^i(\mathbf{X}; \mathbf{l}) = \left(\frac{\sum_{j \in \boldsymbol{\delta}^i} l^j(e) + \tau}{\epsilon|\boldsymbol{\omega}^i| + \sum_{j \in \boldsymbol{\delta}^i} l^j(k) + \tau} \right)^{\alpha} \;, \tag{5}$$

where the constant τ corresponds to a weak Dirichlet prior; $\epsilon|\boldsymbol{\omega}^i|$ avoids highly occluded objects to have a large influence with little available evidence; and α balances the relative importance of detector score and pixel label likelihood.

Importantly, $P(\mathbf{X}|\mathcal{E})$ is not comparable across scene configurations with different numbers of objects. We address this with a reversible jump MCMC framework [18].

3.1 Inference framework

To perform inference in the above model, we simulate the posterior distribution $P(\mathbf{X}|\mathcal{E})$ in a Metropolis-Hastings MCMC framework [19]. At each iteration s new scene samples $\mathbf{X}'$ are proposed by different *moves* from the proposal density $Q(\mathbf{X}'; \mathbf{X}^{(s)})$. Since our goal is to sample from the equilibrium distribution, we discard the samples from an initial burn-in phase. Note that the normalization of the posterior does not have to be known, since it is independent of $\mathbf{X}$ and therefore cancels out in the posterior ratio.

3.2 Proposal moves

Proposal moves change the current state of the Markov chain. We employ three different move types: *diffusion moves* to update the last state's variables, *add moves* and *delete moves* to change the state's dimensionality by adding or removing objects from the scene. Add and delete moves are mutually reversible and trans-dimensional. At each iteration, the move type is selected randomly with fixed probabilities q_{Add}, q_{Del} and q_{Dif}.

Diffusion moves change the current state by sampling new values for the state variables. At each diffusion move, object variables are updated with a probability of $q_{\mathbf{O}}$, while Θ is updated with a probability of q_{Θ}.

To update objects we draw the index i of the object to update from a uniform distribution and then update $\mathbf{O}^i$. Proposals are drawn from a multi-variate normal distribution centered at the position of the previous state and with diagonal covariance.

To update the camera pitch Θ proposals are generated from a mixture model. The first mixture component is a broad normal distribution centered at the calibrated pitch for the motionless platform. For the remaining mixture components, we assume distant objects associated with detections at small scales to have the class' mean height and use $\mathbf{d}^{a(i)}$ to compute their distance by means of the theorem of intersecting lines. Then the deviation between the detected bounding box and the object's projection in the image allows one to estimate the camera pitch. We place one mixture component around each pitch computed this way and assign mixture weights proportional to the detection scores to put more weight on more likely objects.

Add moves add a new object $\mathbf{O}^{N+1}$ to the chain's last state, where N is the number of objects contained in $\mathbf{X}^{(s)}$. As this move is trans-dimensional (i.e., the number of dimensions of $\mathbf{X}^{(s)}$ and $\mathbf{X}'$ do not match) special consideration needs to be taken when the posterior ratio $\frac{P(\mathbf{X}'|\mathcal{E})}{P(\mathbf{X}^{(s)}|\mathcal{E})}$ is evaluated. In particular, $P(\mathbf{X}^{(s)}|\mathcal{E})$ needs to be made comparable in the state space of $P(\mathbf{X}'|\mathcal{E})$. To this end, we assume a constant probability $\bar{P}(\mathbf{O}^{N+1})$ for each object to be part of the background. Hence, posteriors of states with different numbers of objects can be compared in the higher dimensional state space by transforming $P(\mathbf{X}^{(s)}|\mathcal{E})$ to

$$\hat{P}(\mathbf{X}^{(s)}|\mathcal{E}) = P(\mathbf{X}^{(s)}|\mathcal{E})\bar{P}(\mathbf{O}^{N+1}) \tag{6}$$

To efficiently explore high density regions of the posterior we use the detection scores in the proposal distribution. A new object index n is drawn from the discrete set of all K detections $\{\bar{d}\}$, which are not yet associated with an object in the scene, according to $Q(\mathbf{X}'; \mathbf{X}^{(s)}) = \frac{\psi_D(\bar{d}^n)}{\sum_k \psi_D(\bar{d}^k)}$. The data association function is updated by letting $a(N+1)$ associate the new object with the selected detection. For distant objects (i.e., detections

at small scales) we instantiate the new object at a distance given through the theorem of intersecting lines and the height prior, whereas for objects in the near-field a more accurate 3D position can be estimated from the ground plane and camera calibration.

Delete moves remove an object $\mathbf{O}^n$ from the last state and move the associated detection $\mathbf{d}^{a(n)}$ back to $\{\bar{d}\}$. Similar to the add move, the proposed lower dimensional state $\mathbf{X}'$ needs to be transformed. The object index n to be removed from the scene is drawn uniformly among all objects currently in the scene, thus $Q(\mathbf{X}'; \mathbf{X}^{(s)}) = \frac{1}{N}$.

3.3 Projective 3D to 2D marginalization

In order to obtain a score for a 2D position $\mathbf{u}$ (including scale) from our 3D scene model, the probabilistic framework suggests marginalizing over all possible 3D scenes $\mathbf{X}$ that contain an object that projects to that 2D position:

$$P(\mathbf{u}|\mathcal{E}) = \int \max_i \left(\left[\mathbf{u} = \mathbf{o}^i \right] \right) P(\mathbf{X}|\mathcal{E}) \, \mathrm{d}\mathbf{X} \, , \tag{7}$$

with $[expr]$ being the Iverson bracket: $[expr] = 1$ if the enclosed expression is true, and 0 otherwise. Hence, the binary function $\max_i([\cdot])$ detects whether there exists *any* 3D object in the scene that projects to image position $\mathbf{u}$. The marginal is approximated with samples $\mathbf{X}^{(s)}$ drawn using MCMC:

$$P(\mathbf{u}|\mathcal{E}) \approx \frac{1}{S} \sum_{s=1}^{S} \max_i \left(\left[\mathbf{u} = \mathbf{o}^{i,(s)} \right] \right) \, , \tag{8}$$

where $\mathbf{o}^{i,(s)}$ denotes the projection of object $\mathbf{O}^i$ of sample s to the image, and S is the number of samples. In practice $\max_i([\cdot])$ checks whether any of the 3D objects of sample s projects into a small neighborhood of the image position $\mathbf{u}$.

4 Multi-frame Scene Model and Inference

So far we have described our scene model for a single image in static scenes only. For the extension to video streams we pursue a two-stage tracking approach. First, we extend the model to neighboring frames by using greedy data association. Second, the resulting *scene tracklets* are used to extend our model towards long-term data association by performing *scene tracking* with an HMM.

4.1 Multi-frame 3D scene tracklet model

To apply our model to multiple frames, we first use the observer's estimated speed V_{ego} and turn (yaw) rate to roughly compensate the camera's ego-motion. Next, we use a coarse dynamic model for all moving objects to locally perform association, which is refined during tracking. For initial data associations objects that move substantially slower than the camera (e.g., people) are modeled as standing still, $V_x^i = 0$. For objects with a similar speed (e.g., cars and trucks), we distinguish those moving in the same direction as the observers from the oncoming traffic with the help of the detector's class label. The former are expected to move with a similar speed as the observer, $V_x^i = V_{ego}$,

whereas the latter are expected to move with a similar speed, but in opposite direction, $V_x^i = -V_{ego}$. The camera pitch Θ_t can be assumed constant for small time intervals.

For a given frame t we associate objects and detections as described in Sec. 3.2. In adjacent frames we perform association by finding the detection with maximum overlap to each predicted object. Missing evidence is compensated by assuming a minimum detection likelihood anywhere in the image. We define the scene tracklet posterior as

$$P(\mathbf{X}_t|\mathcal{E}_{-\delta t+t:t+\delta t}) \propto \prod_{r=t-\delta t}^{t+\delta t} P(\hat{\mathbf{X}}_r|\mathcal{E}_r), \tag{9}$$

where $\hat{\mathbf{X}}_r$ denotes the predicted scene configuration using the initial dynamic model just explained.

4.2 Long term data association with scene tracking

While the above model extension to scene tracklets is feasible for small time intervals, it does not scale well to longer sequences, because greedy data association in combination with a simplistic motion model will eventually fail. Moreover, the greedy formalism cannot handle objects leaving or entering the scene.

We therefore introduce an explicit data association variable $\mathcal{A}_t$, which assigns objects to detections in frame t. With this explicit mapping, long-term tracking is performed by modeling associations over time in a hidden Markov model (HMM). Inference is performed in a sliding window of length w to avoid latency as required by an online setting:

$$\begin{aligned} P(\mathbf{X}_{1:w}, \mathcal{A}_{1:w}|\mathcal{E}_{-\delta t+1:w+\delta t}) = P(\mathbf{X}_1|\mathcal{A}_1, \mathcal{E}_{-\delta t+1:1+\delta t}) \\ \prod_{k=2}^{w} P(\mathcal{A}_k|\mathcal{A}_{k-1}) P(\mathbf{X}_k|\mathcal{A}_k, \mathcal{E}_{-\delta t+k:k+\delta t}) \end{aligned} \tag{10}$$

The emission model is the scene tracklet model from Sec. 4.1, but with explicit data association $\mathcal{A}_k$. The transition probabilities are defined as $P(\mathcal{A}_k|\mathcal{A}_{k-1}) \propto P_e^{\eta} P_l^{\lambda}$. Thus, P_e is the probability for an object to enter the scene, while P_l denotes the probability for an object to leave the scene. To determine the number η of objects entering the scene, respectively the number λ of objects leaving the scene, we again perform frame-by-frame greedy maximum overlap matching. In Eq. (10) the marginals $P(\mathbf{X}_k, \mathcal{A}_k|\mathcal{E}_{-\delta t+1:w+\delta t})$ can be computed with the sum-product algorithm. Finally, the probability of an object being part of the scene is computed by marginalization over all other variables (cf. Sec. 3.3):

$$P(\mathbf{u}_k|\mathcal{E}_{-\delta t+1:w+\delta t}) = \sum_{\mathcal{A}_k} \int \max_i \left(\left[\mathbf{u}_k = \mathbf{o}_k^i \right] \right) P(\mathbf{X}_k, \mathcal{A}_k|\mathcal{E}_{-\delta t+1:w+\delta t}) \, \mathrm{d}\mathbf{X}_k \tag{11}$$

In practice we approximate the integral with MCMC samples as above, however this time only using those that correspond to the data association $\mathcal{A}_k$. Note that the summation over $\mathcal{A}_k$ only requires to consider associations that occur in the sample set.

5 Datasets and Implementation Details

For our experiments we use two datasets: (1) *ETH-Loewenplatz*, which was introduced by [1] to benchmark pedestrian tracking from a moving observer; and (2) a new multi-class dataset we recorded with an onboard camera to specifically evaluate the challenges targeted by our work including realistic traffic scenarios with a large number of small objects, objects of interest from different categories, and higher driving speed.

ETH-Loewenplatz. This publicly available pedestrian benchmark contains 802 frames overall at a resolution of 640×480 pixels of which every 4$^{\text{th}}$ frame is annotated. The sequence, which has been recorded from a driving car in urban traffic at ≈15 fps, comes with a total of 2631 annotated bounding boxes. Fig. 4 shows some examples.

MPI-VehicleScenes. As the above dataset is restricted to pedestrians observed at low driving speeds, we recorded a new multi-class test set consisting of 674 images. The data is subdivided into 5 sequences and has been recorded at a resolution of 752×480 pixels from a driving car at ≈15 fps. Additionally ego-speed and turn rate are obtained from the car's ESP module. See Fig. 5 for sample images. 1331 front view of cars, 156 rear view of cars, and 422 front views of trucks are annotated with bounding boxes. Vehicles appear over a large range of scales from as small as 20 pixels to as large as 270 pixels. 46% of the objects have a height of ≤ 30 pixels, and are thus hard to detect.

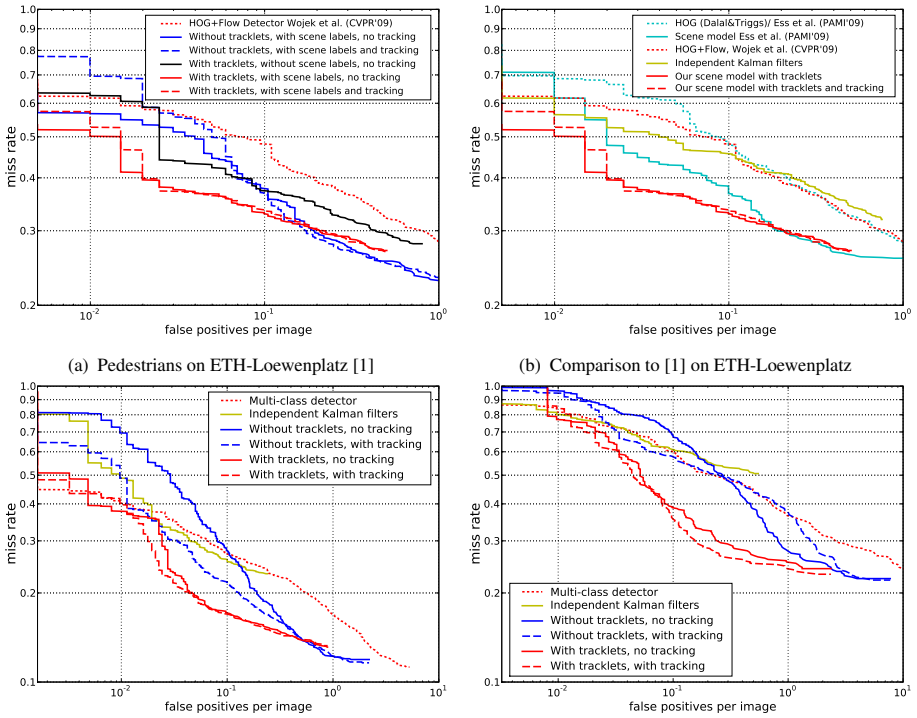
Object detectors. To detect potential object instances, we use state-of-the-art object detectors. For *ETH-Loewenplatz* we use our motion feature enhanced variant of the HOG framework [20]. SVM margins are mapped to positive values with a soft-clipping function [21].

For our new *MPI-VehicleScenes* test set we employ a multi-class detector based on traditional HOG-features and joint boosting [22] as classifier. It can detect the four object classes *car front, car back, truck front* or *truck back*. The scores are mapped to positive values by means of class-wise sigmoid functions. Note that for our application it is important to explicitly separate front from back views, because the motion model is dependent on the heading direction. This detector was trained on a separate dataset recorded from a driving car, with a similar viewpoint as in the test data.

Scene labeling. Every pixel is assigned to the classes *pedestrian, vehicle, street, lane marking, sky* or *void* to obtain a scene labeling. As features we use the first 16 coefficients of the Walsh-Hadamard transform extracted at five scales (4-64 pixels), along with the pixels' (x, y)-coordinates to account for their location in the image. This algorithm is trained on external data and also employs joint boosting as classifier [23].

Experimental setup. For both datasets and all object classes we use the same set of parameters for our MCMC sampler: $q_{Add} = 0.1$, $q_{Del} = 0.1$, $q_{Dif} = 0.8$, $q_{\mathbf{O}} = 0.8$, $q_{\Theta} = 0.2$. For the HMM's sliding window of Eqn. 10 we choose a length of $W = 7$ frames. Our sampler uses 3,000 samples for burn-in and 20,000 samples to approximate the posterior and runs without parallelization at about 1 fps on recent hardware. By running multiple Markov chains in parallel we expect a possible speed-up of one or two orders of magnitude. As we do not have 3D ground truth to assess 3D performance, we project the results back to the images and match them to ground truth annotations with the PASCAL criterion ($intersection/union > 50\%$).

`http://www.vision.ee.ethz.ch/~aess/dataset/`
The data is publicly available at `http://www.mpi-inf.mpg.de/departments/d2`



(a) Pedestrians on ETH-Loewenplatz [1]

(b) Comparison to [1] on ETH-Loewenplatz

(c) Cars on MPI-VehicleScenes

(d) Trucks on MPI-VehicleScenes

Fig. 3: Results obtained with our system. The first row shows results for *pedestrians* on ETH-Loewenplatz and compares to the state-of-the-art. The second row shows results for *truck* and *car* on our new *MPI-VehicleScenes* dataset. Figure best viewed in color.

Baselines. As baselines we report both the performance of the object detectors as well as the result of an extended Kalman filter (EKF) atop the detections. The EKFs track the objects independently, but work in 3D state space with the same dynamic models as our MCMC sampler. To reduce false alarms in the absence of an explicit model for new objects entering, tracks are declared valid only after three successive associations. Analogous to our system, the camera ego-motion is compensated using odometry. Best results were obtained, when the last detection's score was used as confidence measure.

6 Experimental Results

We start by reporting our system's performance for pedestrians on *ETH-Loewenplatz*. Following [1] we consider only people with a height of at least 60 pixels. The authors kindly provided us with their original results to allow for a fair comparison.

In the following we analyze the performance at a constant error rate of 0.1 false positive per image (FPPI). At this error rate the detector (dotted red curve) achieves a miss

The original results published in [1] were biased *against* Ess et al., because they did not allow detections slightly < 60 pixels to match true pedestrians ≥ 60 pixels, discarding many correct detections. We therefore regenerated all FPPI-curves.

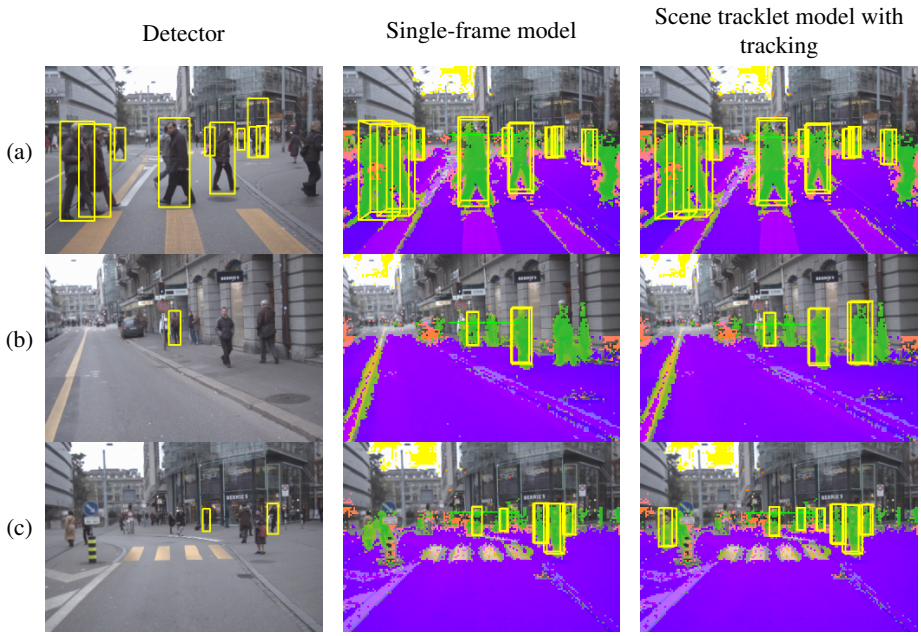


Fig. 4: Sample images showing typical results of our model along with MAP scene labels at a constant error rate of 0.1 false positives per image. *Street* pixels appear in purple, *lane markings* in light purple, *sky* in yellow, *pedestrians* in green and *vehicles* in orange. *Void* (background) pixels are not overlayed. The light green line denotes the estimated horizon.

rate of 48.0%, cf. Fig. 3(a). False detections typically appear on background structures (such as trees or street signs, cf. Fig. 4(a)) or on pedestrians' body parts. When we perform single frame inference (solid blue curve) with our model we improve by 10.4%; additionally adding tracking (dashed blue curve) performs similarly (improvement of 11.6%; see Fig. 4, Fig. 1(a)), but some false positives in the high precision regime are reinforced. When we omit scene labeling but use scene tracklets (black curve) of two adjacent frames our model achieves an improvement of 10.8% compared to the detector. When pixel-labeling information is added to obtain the full model (solid red curve), we observe best results with an improvement of 15.2%. Additionally performing long-term data association (dashed red curve) does not further improve the performance for this dataset: recall has already saturated due to the good performance of the detector, whereas the precision cannot be boosted because the remaining false positives happen to be consistent with the scene model (e.g., human-sized street signs).

Fig. 3(b) compares the system's performance to EKFs and state-of-the-art results by Ess et al. [1]. When we track detections with EKFs (yellow curve) we gain 2.5% compared to the detector, but add additional false detections in the high precision regime, as high-scoring false positives on background structures are further strengthened. Compared to their detector (HOG, [21], dotted cyan curve), the system in [1] achieves an improvement of 11.1% using stereo vision (solid cyan curve), while our monocular approach gains 15.2% over the detector used in our system [20]. We obtain a miss rate of 32.8% using monocular video, which clearly demonstrates the power of the proposed

approach using multi-frame scene tracklets in conjunction with local pixel-labeling. Some example results of our system are depicted in Fig. 1 and 4. Our scene tracklet model allows to stabilize horizon estimation compared to a single-frame model, see Fig. 1(a). Moreover, consistent detections and scene labels boost performance, especially when geometry estimation is more difficult, such as for example in the absence of a sufficient number of objects with confident detections, cf. Fig. 4(b),(c).

Next, we turn to the evaluation on our new *MPI-VehicleScenes* dataset. We note, that *cars rear* are detected almost perfectly, due to the fact that there are only few instances at rather similar scales. Moreover, the test dataset does not contain rear views of trucks. Hence, we will focus on the classes *car front* and *truck front*. In the following, when we refer to cars or trucks this always concerns front views.

For cars the detector achieves a miss rate of 27.0% (see Fig. 3(c)). Independent EKFs improve results by 1.1% to a miss rate of 25.9%. However, in the high precision regime some recall is lost. False positives mainly occur on parts of actual cars, such as on head lights of cars in the near-field, and on rear views of cars – see Fig. 5(a). Thus, in the single-frame case of our approach the false detections are often strengthened rather than weakened by the scene geometry, cf. Fig. 1(b), and in some cases even wrongly bias geometry estimation, thus lowering the scores for correct objects. A drop in high precision performance is the result (27.8% miss rate at 0.1 FPPI). This drop can partially be recovered to a miss rate of 21.8%, when an HMM is added for longer-term tracking.

When scene tracklets are employed, many false hypotheses are discarded because of the gross mismatch between their expected and observed dynamics. Consequently, scene tracklets boost performance significantly, resulting in an improvement of 9.9% in miss rate. Adding long-term tracking with the HMM again only slightly improves result over scene tracklets (by 0.1%). Therefore we conclude that the critical source of improvement is *not* to track objects over extended periods of time, but to enforce a *consistent scene interpretation with short tracklets*, by tightly coupling tracklet estimation with geometry fitting and scene labeling.

Finally, we also report results for trucks, cf. Fig. 3(d). For this class our detector has a higher miss rate of 59.4%. This is caused by a significantly higher intra-class variation among trucks and by the fact that the frontal truck detector often fires on cars due to the high visual similarity of the lower part – an example is shown in Fig. 1(b). As a consequence, independent EKFs do not yield an improvement (miss rate 60.9%), as already observed for cars. Similarly, our model using single-frame evidence is not able to disambiguate the classes when both detectors fire, resulting in a miss rate of 67.9%. Though HMM tracking improves this to 57.6%.

As in the previous examples, our scene tracklet model is able to suppress many false detections through evidence aggregation across a small number of frames (miss rate 38.6%). Also, weak detections on small scale objects are strengthened, thus recall is improved – cf. Fig. 5(a),(b). Compared to the detector, we improve the miss rate by 20.8%, respectively by 23.9% when also adding HMM tracking.

Discussion. Overall, our experiments for two datasets and four different object classes indicate that our scene tracklet model is able to exploit scene context to robustly infer both the 3D scene geometry and the presence of objects in that scene from a monocular camera. This performance is mainly due to the use of a strong *tracking-by-detection* framework which employs *tracklets* on a scene level thereby leveraging evidence from

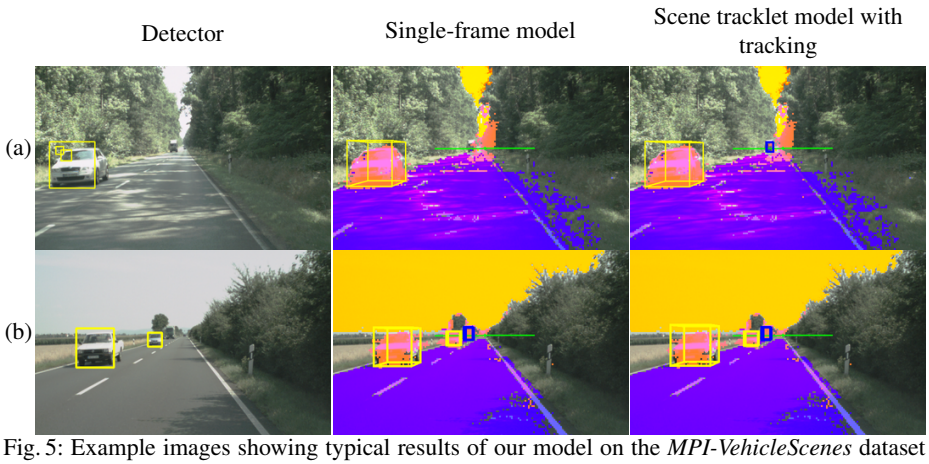


Fig. 5: Example images showing typical results of our model on the *MPI-VehicleScenes* dataset at a constant error rate of 0.1 false positives per image. For color description see Fig. 4.

a number of consecutive frames. The tight coupling with the observation model allows to exploit *3D scene context* as well as to *combine multiple cues* of a detector and from scene labeling. Long-term tracking with an HMM only results in minor additional improvement. In all cases, independent extended 3D Kalman filters cannot significantly improve the output of state-of-the-art object detectors on these datasets, and are greatly outperformed by the integrated state model. On the new multi-class *MPI-VehicleScenes* dataset we outperform state-of-the-art detection by 10.0% for cars, respectively 23.9% for trucks at 0.1 FPPI.

Comparing to other work that integrates detection and scene modeling, we also outperform [1] by 3.8% at 0.1 FPPI for the case of pedestrians, even though we do not use stereo information. At a recall of 60% our model reduces the number of false positives by almost a factor of 4.

7 Conclusion

We have presented a probabilistic 3D scene model, that enables multi-frame tracklet inference on a scene level in a tracking-by-detection framework. Our system performs monocular 3D scene geometry estimation in realistic traffic scenes, and leads to more reliable detection of objects such as pedestrians, cars, and trucks. We exploit information from object (category) detection and low-level scene labeling to obtain a *consistent 3D description of an observed scene*, even though we only use a single camera. Our experimental results show a clear improvement over top-performing state-of-the-art object detectors. Moreover, we significantly outperform basic Kalman filters and a state-of-the-art stereo camera system [1].

Our experiments underline the observation that objects are valuable constraints for the underlying 3D geometry, and vice versa (cf. [1, 5]), so that a joint estimation can improve detection performance.

In future work we plan to extend our model with a more elaborate tracking framework with long-term occlusion handling. Moreover, we aim to model further compo-

nents and objects of road scenes such as street markings and motorbikes. It would also be interesting to explore the fusion with complementary sensors such as RADAR or LIDAR, which should allow for further improvements.

Acknowledgement: We thank Andreas Ess for providing his data and results.

References

1. Ess, A., Leibe, B., Schindler, K., Van Gool, L.: Robust multi-person tracking from a mobile platform. PAMI **31** (2009)
2. Gavrila, D.M., Munder, S.: Multi-cue pedestrian detection and tracking from a moving vehicle. IJCV **73** (2007)
3. Okuma, K., Taleghani, A., de Freitas, N., Little, J., Lowe, D.: A boosted particle filter: Multitarget detection and tracking. In: ECCV. (2004)
4. Kaucic, R., Perera, A.G., Brooksby, G., Kaufhold, J., Hoogs, A.: A unified framework for tracking through occlusions and across sensor gaps. In: CVPR. (2005)
5. Hoiem, D., Efros, A.A., Hebert, M.: Putting objects in perspective. IJCV **80** (2008)
6. Torralba, A.: Contextual priming for object detection. IJCV **53** (2003)
7. Shotton, J., Winn, J., Rother, C., Criminisi, A.: *TextonBoost*: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In: ECCV. (2006)
8. Ess, A., Müller, T., Grabner, H., Van Gool, L.: Segmentation-based urban traffic scene understanding. In: BMVC. (2009)
9. Brostow, G., Shotton, J., Fauqueur, J., Cipolla, R.: Segmentation and recognition using SfM point clouds. In: ECCV. (2008)
10. Tu, Z., Chen, X., Yuille, A., Zhu, S.: Image parsing: Unifying segmentation, detection, and recognition. IJCV **63** (2005)
11. Shashua, A., Gdalyahu, Y., Hayun, G.: Pedestrian detection for driving assistance systems: Single-frame classification and system level performance. In: IVS. (2004)
12. Breitenstein, M.D., Reichlin, F., Leibe, B., Koller-Meier, E., Van Gool, L.: Robust tracking-by-detection using a detector confidence particle filter. In: ICCV. (2009)
13. Huang, C., Wu, B., Nevatia, R.: Robust object tracking by hierarchical association of detection responses. In: ECCV. (2008)
14. Li, Y., Huang, C., Nevatia, R.: Learning to associate: HybridBoosted multi-target tracker for crowded scene. In: CVPR. (2009)
15. Khan, Z., Balch, T., Dellaert, F.: MCMC-based particle filtering for tracking a variable number of interacting targets. PAMI **27** (2005)
16. Zhao, T., Nevatia, R., Wu, B.: Segmentation and tracking of multiple humans in crowded environments. PAMI **30** (2008)
17. Isard, M., MacCormick, J.: BraMBLe: a Bayesian multiple-blob tracker. In: ICCV. (2001)
18. Green, P.J.: Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **82** (1995)
19. Gilks, W., Richardson, S., Spiegelhalter, D., eds.: Markov Chain Monte Carlo in Practice. Chapman & Hall (1995)
20. Wojek, C., Walk, S., Schiele, B.: Multi-cue onboard pedestrian detection. In: CVPR. (2009)
21. Dalal, N.: Finding People in Images and Videos. PhD thesis, Institut National Polytechnique de Grenoble (2006)
22. Torralba, A., Murphy, K., Freeman, W.: Sharing visual features for multiclass and multiview object detection. PAMI **29** (2007)
23. Wojek, C., Schiele, B.: A dynamic CRF model for joint labeling of object and scene classes. In: ECCV. (2008)