Dynamic 3D Scene Analysis from a Moving Vehicle

Bastian Leibe[1] Nico Cornelis[2] Kurt Cornelis[2] Luc Van Gool[1,2]

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

[2]KU Leuven
Leuven, Belgium
{firstname.lastname}@esat.kuleuven.be

Abstract

In this paper, we present a system that integrates fully automatic scene geometry estimation, 2D object detection, 3D localization, trajectory estimation, and tracking for dynamic scene interpretation from a moving vehicle. Our sole input are two video streams from a calibrated stereo rig on top of a car. From these streams, we estimate Structure-from-Motion (SfM) and scene geometry in real-time. In parallel, we perform multi-view/multi-category object recognition to detect cars and pedestrians in both camera images. Using the SfM self-localization, 2D object detections are converted to 3D observations, which are accumulated in a world coordinate frame. A subsequent tracking module analyzes the resulting 3D observations to find physically plausible spacetime trajectories. Finally, a global optimization criterion takes object-object interactions into account to arrive at accurate 3D localization and trajectory estimates for both cars and pedestrians. We demonstrate the performance of our integrated system on challenging real-world data showing car passages through crowded city areas.

1. Introduction

The task we address in this paper is dynamic scene analysis from a moving, camera-equipped vehicle. At any point in time, we want to detect other traffic participants in the environment (cars, bicyclists, and pedestrians), localize them in 3D, estimate their past trajectories, and predict their future motion (as shown in Fig. 1). Such a capability has obvious applications in driver assistance systems, but it also serves as a testbed for many interesting research challenges.

Scene analysis of this sort requires multi-viewpoint, multi-category object detection. Since we cannot control the vehicle's path, nor the environment it passes through, the detectors need to be robust to a large range of lighting variations, noise, clutter, and partial occlusion. For 3D localization, an accurate estimate of the scene geometry is necessary. The ability to integrate such measurements over time additionally requires continuous self-localization and recalibration. In order to finally make predictions about future states, powerful tracking is needed that can cope with a changing background. On the other hand, each object will typically persist in the vehicle's field of view only for a few seconds. It is thus not as important to uniquely track a person's identity as in classic surveillance scenarios.

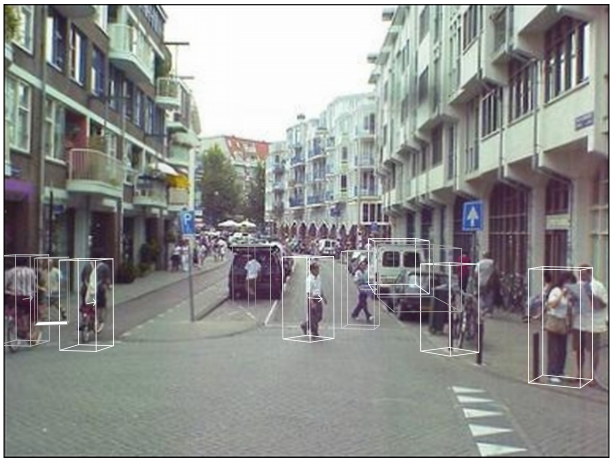


Figure 1. *Online 3D localization and trajectory estimation results of our system obtained from inside a moving vehicle. The different bounding box intensities correspond to our system's confidence level in its estimates.*

In this paper, we present a system which addresses those challenges by integrating recognition, reconstruction, and tracking in a collaborative ensemble. Namely, SfM yields scene geometry for each image pair, which greatly helps the other modules. Recognition picks out objects of interest and separates them from the dynamically changing background. Tracking adds a temporal context to individual object detections and provides them with a history supporting their presence in the current video frame. Detected object trajectories, finally, are extrapolated to future frames and are constantly reevaluated in the light of new evidence.

The paper contains the following contributions. 1) We present an integrated system for dynamic scene analysis on a mobile platform. We demonstrate how its individual components can benefit from each other's continuous input and how the transferred knowledge can be used to improve scene analysis. 2) In particular, we present a multi-view/multi-category object detection module that can reli-

1-4244-1180-7/07/$25.00 ©2007 IEEE

ably detect cars and pedestrians in crowded real-world traffic scenes. We show how knowledge about the scene geometry can be used in such a system both to improve recognition performance and to fuse the outputs of multiple detectors. 3) We demonstrate how the resulting 2D detections can be integrated over time to arrive at accurate 3D localization and orientation estimates of static objects. 4) In order to deal with moving objects, we propose a novel tracking approach which formulates the tracking problem as spacetime trajectory analysis followed by hypothesis selection. This approach is capable of tracking a large and variable number of objects through complex outdoor scenes with a moving camera. In addition, it can model physical object-object interactions to arrive at a globally optimal scene interpretation. 5) Finally, we demonstrate the performance of our integrated system on two challenging video sequences of car passages through crowded city centers showing accurate 3D localization and trajectory estimation results for cars, bicyclists, and pedestrians.

The paper is structured as follows. The following sections discuss related work and give a general overview of our system. After that, Sections 2, 3, and 4 describe our scene geometry estimation, object detection, and tracking approaches in detail. Section 5 presents experimental results. A final discussion concludes our work.

Related Work. Scene analysis with a moving camera is a notoriously difficult task because of the combined effects of egomotion, blur, and rapidly changing lighting conditions [3, 6]. In addition, the introduction of a moving camera invalidates many simplifying techniques we have grown fond of, such as background subtraction and a constant ground plane assumption. Such techniques have been routinely used in surveillance and tracking applications from static cameras (*e.g.* [2, 12]), but they are no longer applicable here. While object tracking under such conditions has been demonstrated in clean highway situations [3], reliable performance in urban areas is still an open challenge [7].

In order to allow tracking with a moving camera, several approaches have started to explore the possibilities of combining tracking with detection [1, 8, 21, 23]. At the same time, object detection itself has made tremendous progress over the last few years [5, 15, 18, 22, 23], to an extent that state-of-the-art detectors are becoming applicable in complex outdoor scenes. [10] have shown that geometric scene context can greatly help recognition and have proposed a method to estimate it from a single image. More recently, [13] have combined recognition and SfM, however only for the purpose of localizing static objects.

In our approach, we integrate geometry estimation and tracking-by-detection in a combined system that searches for the best scene interpretation by global optimization. [2] also perform global trajectory optimization to track up to six mutually occluding individuals by modelling their positions on a discrete occupancy grid. However, their approach requires static cameras, and optimization is performed only for one individual at a time. In contrast, our approach models object positions continuously while moving through a 3D world and allows to find a combined optimal solution.

System Overview. Our input data are two video streams from a calibrated stereo rig mounted on top of a vehicle. From this data, an SfM module computes a camera pose and ground plane estimate for each image. This information is fed to an object detection module, which processes both camera images to detect cars and pedestrians in the vehicle's field of view. The necessary reliability of the detection module is achieved by integrating multiple local cues, fusing the output of several single-view detectors, and making use of the continuously updated ground plane estimate. Using the estimated camera pose, 2D detections are then converted to 3D observations and passed to the subsequent tracking module. This module analyzes the incoming 3D observations to find plausible spacetime trajectories and selects the best explanation for each frame pair by a global optimization criterion.

2. Real-Time Scene Geometry Estimation

Real-Time Structure-from-Motion (SfM). Our SfM module is based on the approach by [4], which is highly optimized and runs at 26-30 fps. It takes the green channel of each camera as input and extracts image feature points by finding local maxima of a simple feature measure based on average intensities of four subregions. The extracted features are matched between consecutive images and then fed into a classic SfM pipeline [9], which reconstructs feature tracks and refines 3D point locations by triangulation. Bundle adjustment is running in parallel with the main SfM algorithm to refine camera poses and 3D feature locations for previous frames and thus reduce drift.

Online Ground Plane Estimation. For each image pair, SfM delivers an updated camera calibration. In addition, we obtain an online ground plane estimate by computing local normals on a set of trapezoidal road strips between the reconstructed wheel contact points of adjacent frames and averaging those local measurements over a larger window. In order to do this reliably, it is necessary to find a good compromise for the window size this estimate is based on. We experimentally found a window size of 3m, roughly corresponding to the ground patch beneath the vehicle, to be optimal for a variety of different cases. Note that this procedure automatically adjusts for the driving speed. A lower driving speed leads to more accurate reconstruction, so that the smaller strip sizes are sufficient. Conversely, higher speed (or lower frame rate) reduces reconstruction quality, but this is compensated for by the larger strip size between frames.

Figure 2 highlights the importance of this continuous

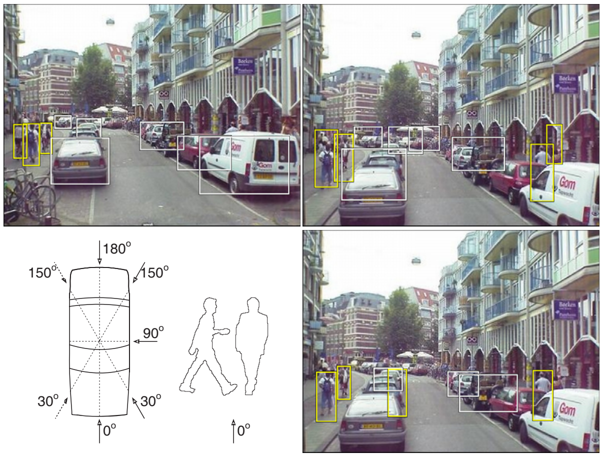


Figure 2. *(top and right) Illustration for the importance of a continuous reestimation of scene geometry. The images show the effect on object detection when the vehicle hits a speedbump (top) if using an unchanged ground plane estimate; (bottom) if using the online reestimate. (bottom left) Training viewpoints used for cars and pedestrians.*

reestimation step if later stages are to trust its results. In this example, the camera vehicle hits a speedbump, causing a massive jolt in camera perspective. The top row of Fig. 2 shows the resulting detections when the ground plane estimate from the previous frame is simply taken over. As can be seen, this results in several false positives at improbable locations and scales. The bottom image displays the detections when the reestimated ground plane is used instead. Here, the negative effect is considerably lessened.

3. Object Detection

The recognition system is based on a battery of single-view, single-category ISM detectors [15]. This approach lets local features, extracted around interest regions, vote for the object center in a 3-dimensional Hough space, followed by a top-down segmentation and verification step. For our application, we use the robust multi-cue extension from [14], which integrates local *Shape Context* descriptors [19] computed at *Harris-Laplace*, *Hessian-Laplace*, and *DoG* interest regions [17, 19]. The main contribution of this section is how to fuse those different detectors and how to integrate scene geometry into the recognition system.

Our system uses a set of 5 single-view detectors for the different car orientations and one additional pedestrian detector (Fig. 2). We do not differentiate between pedestrians and bicyclists here, as they are often indistinguishable from a distance and our detector responds well to both categories. We start by running all detectors in parallel on both camera images and collect their hypotheses (without the final verification step). For each such hypothesis h, we compute two per-pixel probability maps $p(\mathbf{p} = \mathit{figure}|h)$ and $p(\mathbf{p} = \mathit{ground}|h)$, as described in [15].

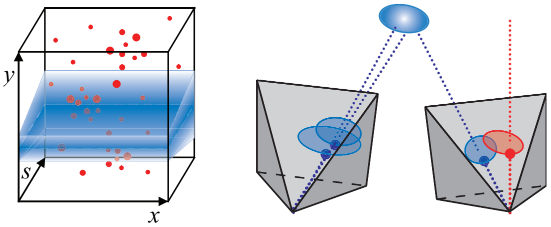


Figure 3. *Benefits of scene geometry for object detection. (left) A ground plane significantly reduces the search volume for Hough voting. A Gaussian size prior additionally "pulls" object hypotheses towards the right locations. (right) The responses of multiple detectors are combined if they refer to the same scene object.*

Integration of Scene Geometry Constraints. Given the camera calibration and ground plane estimate from SfM, we can associate each image-plane hypothesis h with a 3D location by projecting a ray from the camera center through the base point of its detection bounding box. If the ray intersects the ground plane, we can estimate the object's real-world size by projecting a second ray through the bounding box top point and intersecting it with a vertical plane through its 3D base. Using this information, we can express the likelihood for a real-world object H given image I entirely by the image-plane hypotheses h according to the following marginalization:

$$p(H|I) = \sum_{h} p(H|h)p(h|I) \sim \sum_{h} p(h|H)p(H)p(h|I) \quad (1)$$

The following paragraphs describe each of those three factors in detail and explain how they are used in our recognition system.

2D Recognition Score. The last term in eq. (1) is the likelihood of hypothesis h given the image. Using the top-down segmentation of h, we express this likelihood in terms of the pixels h occupies:

$$p(h|I) = \sum_{\mathbf{p}\in I} p(h|\mathbf{p}) \approx \sum_{\mathbf{p}\in Seg(h)} p(\mathbf{p} = \mathit{figure}|h)p(h), \quad (2)$$

where $Seg(h)$ denotes the segmentation area of h, i.e. the pixels for which $p(\mathbf{p} = \mathit{figure}|h) > p(\mathbf{p} = \mathit{ground}|h)$.

Ground Plane Constraints. The middle term $p(H)$ expresses a 3D prior for finding an object at location H, which we split into separate priors for the object size and distance given its category.

$$p(H) = p(H_{size}|H_{categ})p(H_{dist}|H_{categ})p(H_{categ}) \quad (3)$$

In our application, we assume a uniform distance prior and model the size prior by a Gaussian (similar to [10]). This effective coupling between object distance and size through a ground plane assumption has several beneficial effects. First, it significantly reduces the search volume during voting to a corridor in Hough space (Fig. 3(left)). In addition, the Gaussian size prior serves to "pull" object hypotheses towards the correct locations, thus improving also recognition quality.

Multi-Detector Integration. The third factor in eq. (1), finally, is a 2D/3D transfer function $p(h|H)$, which relates the image-plane hypothesis h to the 3D object hypothesis H. This term is of particular interest in combination with the sum over all h, since it allows to effectively fuse the results of the different single-view detectors by clustering the inferred world states. The intuition behind this step is that two image-plane detections are consistent if they correspond to the same 3D object (Fig. 3(right)). Thus, we can disambiguate between overlapping responses from different detectors on the basis of the world state they would infer, which is done in the following global optimization step.

Multi-Category Hypothesis Selection. In order to obtain the final interpretation for the current image pair, we search for the combination of hypotheses that together best explain the observed evidence. In [15], this is done by adopting an MDL formulation and expressing the *savings* [16] of a particular hypothesis h as

$$S_h = K_0 S_{data} - K_1 S_{model} - K_2 S_{error}, \qquad (4)$$

where S_{data} corresponds to the number N of data points or pixels that are explained this way; S_{model} denotes the model cost, usually a constant; and S_{error} describes a cost for the error that is made by this representation. More generally, it can be shown that if the error term is chosen as the sum over all data points x assigned to a hypothesis h of the probabilities that the point assignment is wrong

$$S_{error} = \sum_{x \in h} (1 - p(x|h)), \qquad (5)$$

then the savings reduce to the merit term

$$S_h = -\kappa_1 + \sum_{x \in h} \left((1-\kappa_2) + \kappa_2 p(x|h)\right), \qquad (6)$$

which is effectively just the sum over the data assignment likelihoods, together with a regularization term to compensate for unequal sampling. When hypotheses overlap, they compete for data points, resulting in interaction costs. As shown in [16], the optimal hypothesis selection can then be formulated as a Quadratic Boolean Optimization Problem

$$\max_m m^\mathsf{T} Q m = \max_m m^\mathsf{T} \begin{bmatrix} q_{11} & \cdots & q_{1M} \\ \vdots & \ddots & \vdots \\ q_{M1} & \cdots & q_{MM} \end{bmatrix} m \qquad (7)$$

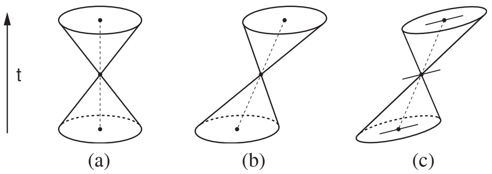
with an indicator vector $m = \{m_1, \ldots, m_M\}$, where $m_i = 1$ if h_i is selected and 0 otherwise; and an interaction matrix Q. Here, we pursue a similar approach. In contrast to [15], however, we perform the hypothesis selection not over image-plane hypotheses h_i, but over their corresponding world hypotheses H_i. Combining eqs. (1) and (6), we obtain the following merit terms

$$q_{ii} \;=\; S_{H_i} = -\kappa_1 + \sum_k p(h_k|H_i)p(H_i)f(h_k), \qquad (8)$$

$$f(h_k) \;=\; \frac{1}{A_{\sigma,v}(h_k)} \sum_{\mathbf{p} \in Seg(h)} ((1-\kappa_2) + \kappa_2 p(\mathbf{p}{=}fig.|h_k)) \qquad (9)$$

where $A_{\sigma,v}(h_k)$ acts as a normalization factor expressing the *expected area* of a 2D hypothesis at its detected scale and aspect. Two 3D hypotheses H_i and H_j interact if their supporting image-plane hypotheses h_{k_i} and h_{k_j} compete for the same pixels. In this case, we assume that the hypothesis $H^* \in \{H_i, H_j\}$ that is farther away from the camera is occluded and subtract its support in the overlapping image area. The interaction cost then becomes

$$q_{ij} = -\frac{1}{2} \sum_{k^*} p(h_{k^*}|H^*)p(H^*)f(h_{k^*}). \qquad (10)$$



(a) (b) (c)

Figure 4. *Visualization of example event cones for (a) a static object with unknown orientation; (b) a holonomically moving object; (c) a non-holonomically moving object.*

As a result of this procedure, we obtain a set of world hypotheses $\{H_i\}$, together with their supporting segmentations in the image. At the same time, the hypothesis selection procedure naturally integrates the contributions from the different single-view, single-category detectors. We perform the optimization separately for the two camera images and pass the resulting detections to the following temporal integration stage.

4. Temporal Integration and Tracking

In order to present our tracking approach, we introduce the concept of *event cones*. The event cone of an observation $H_{i,t} = \{\mathbf{x}_{i,t}, v_{i,t}$is the spacetime volume it can physically influence from its current position given its maximal velocity and turn rate. Figure 4 shows an illustration for several cases of this concept. If an object is static at time t and its orientation is unknown, all motion directions are equally probable, and the affected spacetime volume is a simple double cone reaching both forwards and backwards in time (Fig. 4(a)). If the object moves holonomically, *i.e.* without external constraints linking its speed and turn rate, the event cone becomes tilted in the motion direction (Fig. 4(b)). An example for this case would be a pedestrian at low speeds. In the case of nonholonomic motion, as in a car which can only move along its main axis and only turn while moving, the event cones get additionally deformed according to those (often nonlinear) constraints (Fig. 4(c)).

We thus search for plausible trajectories through the spacetime observation volume by linking up event cones. Starting from an observation $H_{i,t}$, we follow its event cone up and down the timeline and collect all observations that fall inside its volume in the adjoining time steps. Since we do not know the starting velocity $v_{i,t}$ yet, we begin with the case in Fig. 4(a). In all subsequent time steps, however, we

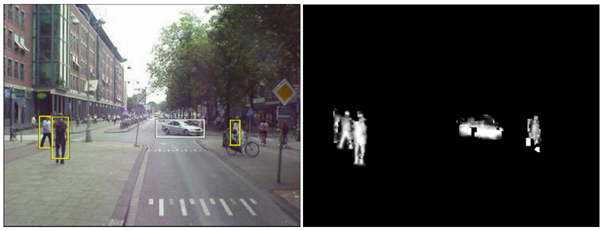


Figure 5. *Detections and corresponding top-down segmentations used to learn the object-specific color model.*

can reestimate the object state from the new evidence and adapt the growing trajectory accordingly.

It is important to point out that an individual event cone is not more powerful in its descriptive abilities than a bidirectional Extended Kalman Filter, since it is based on essentially the same equations. However, our approach goes beyond Kalman Filters in several important respects. First of all, we are no longer bound by a Markovian assumption. When reestimating the object state, we can take several previous time steps into account. In our approach, we aggregate the information from all previous time steps, weighted with a temporal discount λ. In addition, we are not restricted to tracking a single hypothesis. Instead, we start independent trajectory searches from all available observations (at all time steps) and collect the corresponding hypotheses. The final scene interpretation is then obtained by a global optimization criterion which selects the combination of trajectory hypotheses that best explains the observed data under the constraints that each observation may only belong to a single object and no two objects may occupy the same physical space at the same time. The following sections explain those steps in more detail.

Color Model. For each observation, we compute an object-specific color model a_i, using the top-down segmentations provided by the previous stage. Figure 5 shows an example of this input. For each detection $H_{i,t}$, we build an $8 \times 8 \times 8$ RGB color histogram over the segmentation area, weighted by the per-pixel confidence $\sum_k p(\mathbf{p} = \mathit{fig.}|h_k)p(h_k|H_{i,t})$ in this segmentation. Similar to [20], we compare color models by their Bhattacharyya coefficient

$$p(a_i|\mathcal{A}) \sim \sum_q \sqrt{a_i(q)\mathcal{A}(q)} \qquad (11)$$

Dynamic Model. Given a partially grown trajectory $\mathcal{H}_{t_0:t}$, we first select the subset of observations which fall inside its event cone. Using the simple motion models

$$\begin{aligned} \dot{x} &= v\cos\theta \\ \dot{y} &= v\sin\theta \\ \dot{\theta} &= K_c \end{aligned} \quad \text{and} \quad \begin{aligned} \dot{x} &= v\cos\theta \\ \dot{y} &= v\sin\theta \\ \dot{\theta} &= K_c v \end{aligned} \qquad (12)$$

for holonomic and nonholonomic motion on the ground plane, respectively, we compute predicted positions

$$\begin{aligned} x^p_{t+1} &= x_t + v\Delta t\cos\theta \\ y^p_{t+1} &= y_t + v\Delta t\sin\theta \\ \theta^p_{t+1} &= \theta_t + K_c\Delta t \end{aligned} \quad \text{and} \quad \begin{aligned} x^p_{t+1} &= x_t + v\Delta t\cos\theta \\ y^p_{t+1} &= y_t + v\Delta t\sin\theta \\ \theta^p_{t+1} &= \theta_t + K_c v\Delta t \end{aligned} \qquad (13)$$

and approximate the positional uncertainty by an oriented Gaussian to arrive at the dynamic model $\mathcal{D}$

$$\mathcal{D}: \begin{aligned} p\left(\begin{bmatrix} x_{t+1} \\ y_{t+1} \end{bmatrix}\right) &\sim \mathcal{N}\left(\begin{bmatrix} x^p_{t+1} \\ y^p_{t+1} \end{bmatrix}, R^{\mathsf{T}}\begin{bmatrix} \sigma^2_{mov} & 0 \\ 0 & \sigma^2_{turn} \end{bmatrix} R\right) \\ p(\theta_{t+1}) &\sim \mathcal{N}(\theta^p_{t+1}, \sigma^2_{steer}) \end{aligned} \qquad (14)$$

where R is the rotation matrix, K_c the path curvature, and the nonholonomic constraint is approximated by adapting the rotational uncertainty σ_{turn} as a function of v.

Spacetime Trajectory Search. Each candidate observation $H_{i,t+1}$ is then evaluated under the covariance of $\mathcal{D}$ and compared to the trajectory's appearance model $\mathcal{A}$ (its mean color histogram), yielding

$$p(H_{i,t+1}|\mathcal{H}_{t_0:t}) = p(H_{i,t+1}|\mathcal{A}_t)p(H_{i,t+1}|\mathcal{D}_t). \qquad (15)$$

After this, the trajectory is updated by the weighted mean of its predicted position and the supporting observations:

$$\mathbf{x}_{t+1} = \frac{1}{Z}\left(p(\mathcal{H}_{t+1}|\mathcal{H}_{t_0:t})\mathbf{x}^p_{t+1} + \sum_i p(H_{i,t+1}|\mathcal{H}_{t_0:t})\mathbf{x}_i\right). \qquad (16)$$

with $p(\mathcal{H}_{t+1}|\mathcal{H}_{t_0:t}) = e^{-\lambda}$ and normalization factor Z. Velocity, rotation, and appearance model are updated in the same fashion.

Static cars are treated as a special case, since their orientation cannot be inferred from their motion direction and our appearance-based detectors provide a too coarse orientation estimate for our goal to estimate a precise 3D bounding box. Instead, we accumulate detections over a longer time frame. Using the observation that the main localization uncertainty from our detectors occurs both along the car's main axis and along our vehicle's viewing direction, we then estimate the car orientation as the weighted mean between our detectors' orientation estimate and the cluster shape of all inlier observations projected onto the ground plane.

Global Trajectory Selection. We express the support $\mathcal{S}$ of a trajectory $\mathcal{H}_{t_0:t}$ reaching from time t_0 to t by the evidence collected from the images $I_{t_0:t}$ during that time span:

$$\begin{aligned} \mathcal{S}(\mathcal{H}_{t_0:t}|I_{t_0:t}) &= \sum_i p(\mathcal{H}_{t_0:t}|H_{i,t_i})p(H_{i,t_i}|I_{t_i}) \qquad (17) \\ &= p(\mathcal{H}_{t_0:t})\sum_i \frac{p(H_{i,t_i}|\mathcal{H}_{t_0:t})}{p(H_{i,t_i})}p(H_{i,t_i}|I_{t_i}) \\ &\sim p(\mathcal{H}_{t_0:t})\sum_i p(H_{i,t_i}|\mathcal{H}_{t_0:t})p(H_{i,t_i}|I_{t_i}) \end{aligned}$$

where $p(H_{i,t_i}) = \sum_j p(H_{i,t_i}|\mathcal{H}_j)$ is a normalization factor that can be omitted, since we are only interested in relative scores. Further, we define

$$\begin{aligned} p(H_{i,t_i}|\mathcal{H}_{t_0:t}) &= p(\mathcal{H}_{t_i}|\mathcal{H}_{t_0:t})p(H_{i,t_i}|\mathcal{H}_{t_i}) \qquad (18) \\ &= e^{-\lambda(t-t_i)}p(H_{i,t_i}|\mathcal{A}_{t_i})p(H_{i,t_i}|\mathcal{D}_{t_i}) \end{aligned}$$

that is, we express the likelihood of an observation H_{i,t_i} belonging to trajectory $\mathcal{H}_{t_0:t} = (\mathcal{A}, \mathcal{D})_{t_0:t}$ by evaluating it

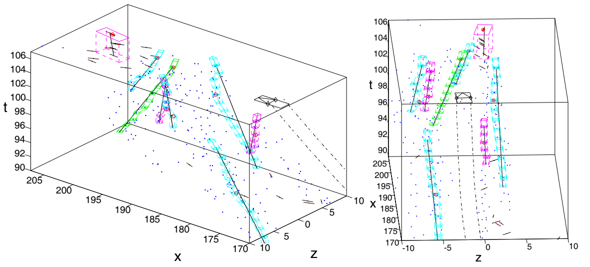


Figure 6. *Visualization of the estimated spacetime trajectories for cars and pedestrians from the scene in Fig. 1. Blue dots show pedestrian observations; red dots correspond to car observations.*

under the trajectory's appearance and dynamic model at that time, weighted with a temporal discount.

In order to find the combination of trajectory hypotheses that together best explain the observed evidence, we again solve a Quadratic Boolean Optimization Problem $\max_{\widetilde{m}} \widetilde{m}^{\mathsf{T}}\widetilde{Q}\widetilde{m}$ with the additional constraint that no two objects may occupy the same space at the same time. With a similar derivation as in Section 3, we arrive at

$$\begin{aligned} \widetilde{q}_{ii} &= -\widetilde{\kappa}_1 + \sum_{H_{k,t_k}\in\mathcal{H}_i} ((1-\widetilde{\kappa}_2) + \widetilde{\kappa}_2\; g_{k,i}) && (19)\\ \widetilde{q}_{ij} &= -\frac{1}{2}\sum_{H_{k,t_k}\in\mathcal{H}_i\cap\mathcal{H}_j} ((1-\widetilde{\kappa}_2) + \widetilde{\kappa}_2\; g_{k,*} + \widetilde{\kappa}_3\; O_{ij}) && (20)\\ g_{k,i} &= p(H_{k,t_k}|\mathcal{H}_i)p(H_{k,t_k}|I_{t_k}). \end{aligned}$$

where again $\mathcal{H}^* \in \{\mathcal{H}_i, \mathcal{H}_j\}$ denotes the weaker of the two hypotheses and the additional penalty term O_{ij} measures the physical overlap between the spacetime trajectory volumes of $\mathcal{H}_i$ and $\mathcal{H}_j$ given average object dimensions.

The hypothesis selection procedure always searches for the best explanation of the current world state *given all evidence available up to now*. It is not guaranteed that this explanation is consistent with the one we got for the previous frame. However, as soon as it is selected, it explains the whole past, as if it had always existed. We can thus follow a trajectory back in time to determine where a pedestrian came from when he first stepped into view, even though no hypothesis was selected for him back then. Fig. 6 visualizes the estimated spacetime trajectories for such a case.

Efficiency Considerations. The main computational cost in this stage comes from three factors: the cost to find trajectories, to build the quadratic interaction matrix $\widetilde{Q}$, and to solve the final optimization problem. However, the first two steps can reuse information from previous time steps. Thus, instead of building up trajectories from scratch at each time step t, we merely check for each of the existing hypotheses $\mathcal{H}_{t_0:t-1}$ if it can be extended by the new observations. In addition, we start new trajectory searches down the time line from each new observation $H_{i,t}$. Similarly, most entries of the previous interaction matrix $\widetilde{Q}_{t-1}$ can be reused and just need to be weighted with the temporal discount $e^{-\lambda}$.

The cost of the optimization problem depends on the connectedness of the matrix $\widetilde{Q}$, *i.e.* on the number of non-zero interactions between hypotheses. For static cars and for the 2D case in Section 3, this number is typically very low, since only few hypotheses overlap. For pedestrian trajectories, the number of interactions may however grow quite large. In this paper, we therefore just compute a greedy approximation for both optimization problems. However, a range of efficient relaxation techniques have become available in recent years which can be used to compute more exact solutions (*e.g.* [11]).

5. Experimental Results

Data Sets. In the following, we evaluate our integrated approach on two challenging video sequences. The first test sequence consists of 1175 image pairs recorded at 25fps and a resolution of 360×288 pixels over a distance of about 500m. It contains a total of 77 (sufficiently visible) static cars parked on both sides of the street, 4 moving cars, but almost no pedestrians at sufficiently high resolutions. The main difficulties for object detection here lie in the relatively low resolution, strong partial occlusion between parked cars, frequently encountered motion blur, and extreme contrast changes between brightly lit areas and dark shadows. Only the car detectors are used for this sequence.

The second sequence consists of 290 image pairs captured over the course of about 400m at the very sparse frame rate of 3fps and a resolution of 384×288 pixels. This very challenging sequence shows a vehicle passage through a crowded city center, with parked cars and bicycles on both street sides, numerous pedestrians and bicyclists travelling on the side walks and crossing the street, and several speed bumps. Apart from the difficulties mentioned above, this sequence poses the additional challenge of detecting and separating many mutually occluding pedestrians at very low resolutions while simultaneously limiting the number of false positives on background clutter. In addition, temporal integration is further complicated by the low frame rate.

In the following sections, we present experimental results for object detection and tracking performance on both sequences. However, it would clearly be unrealistic to expect perfect detection and tracking results under such difficult conditions, which may make the quantitative results hard to interpret. We therefore provide the result videos at `http://www.vision.ethz.ch/bleibe/cvpr07`.

Object Detection Performance. Figure 7(left) displays example detection results of our system on difficult images from the two test sequences. All images have been processed at their original resolution by SfM and bilinearly interpolated to twice their initial size for object detection. For a quantitative evaluation we annotated one video stream for each sequence and marked all objects that were within 50m distance and visible by at least 30-50%. It is important

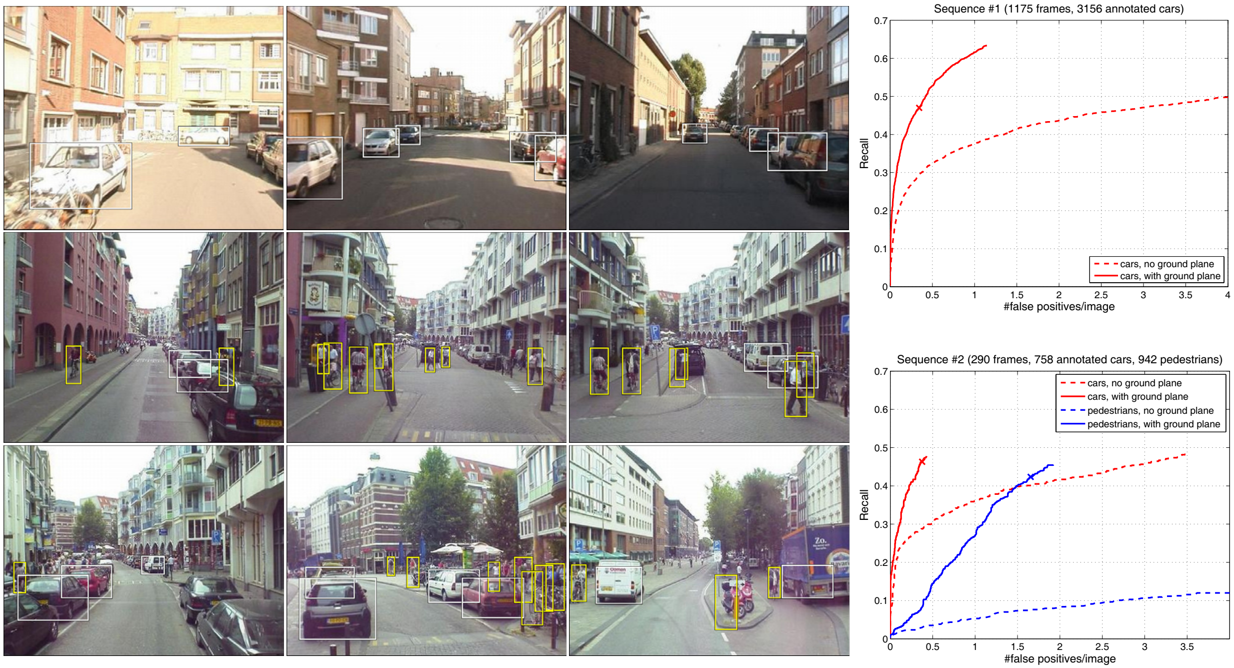


Figure 7. *(left) Example car and pedestrian detections of our system on difficult images from the two test sequences. (right) Quantitative comparison of the detection performance with and without scene geometry constraints (the crosses mark the operating point for tracking).*

to note that this includes many cases with partial visibility. Fig 7(right) shows the resulting detection performance with and without ground plane constraints. As can be seen from the plots, both recall and precision are greatly improved by the inclusion of scene geometry, up to an operating point of 0.34 fp/frame for cars and 1.65 fp/frame for pedestrians.

Tracking Performance. Figure 8 shows online tracking results of our system (using only detections from previous frames) for both sequences. As can be seen, our system manages to localize and track other traffic participants despite significant egomotion and dynamic scene changes. The 3D localization and orientation estimates typically converge at a distance of 15-30m and lead to accurate 3D bounding boxes for cars and pedestrians. A major challenge for sequence #2 is to filter out false positives from incorrect detections. At 3fps, this is not always possible. However, false positives typically get only low confidence ratings and quickly fade out again as they fail to get continuous support.

6. Conclusion

In this paper, we have presented an integrated system for dynamic 3D scene analysis from a moving platform. We have proposed a novel method to fuse the output of multiple single-view object detectors and to integrate continuously reestimated scene geometry constraints. In order to aggregate detections over time, we have further proposed a novel tracking approach that can localize and track a variable number of objects with a moving camera and that arrives at a consistent scene interpretation by global optimization. The resulting system obtains an accurate analysis of dynamic scenes, even at very low frame rates.

One of the key points we want to make here is convergence. The different fields of Computer Vision have advanced tremendously in recent years. While all modalities considered in this paper, SfM, object detection, and tracking, are far from being solved yet individually, all three have become sufficiently mature to be useful in combination with the others. As we have demonstrated here, the individual tasks can benefit considerably by the integration and the close collaboration with the other modalities. and novel capabilities can emerge as a consequence. Many more such cross-links can be exploited. For example, stereo depth estimates can directly be used to extract foci of attention for object detection [6]. Similarly, results from tracking could be used to guide feature extraction and speed up recognition considerably. It is reasonable to expect that those additions will increase system performance, and we will investigate them in future work.

Acknowledgments: This work is supported, in parts, by EU projects DIRAC (IST-027787) and HERMES (IST-027110). We also wish to acknowledge the support of Toyota Motor Corporation/Toyota Motor Europe, KU Leuven Research Fund's GOA project MARVEL, and TeleAtlas for providing additional survey videos to test on.

References

[1] S. Avidan. Ensemble tracking. In *CVPR'05*, 2005.

[2] J. Berclaz, F. Fleuret, and P. Fua. Robust people tracking with global trajectory optimization. In *CVPR'06*, 2006.

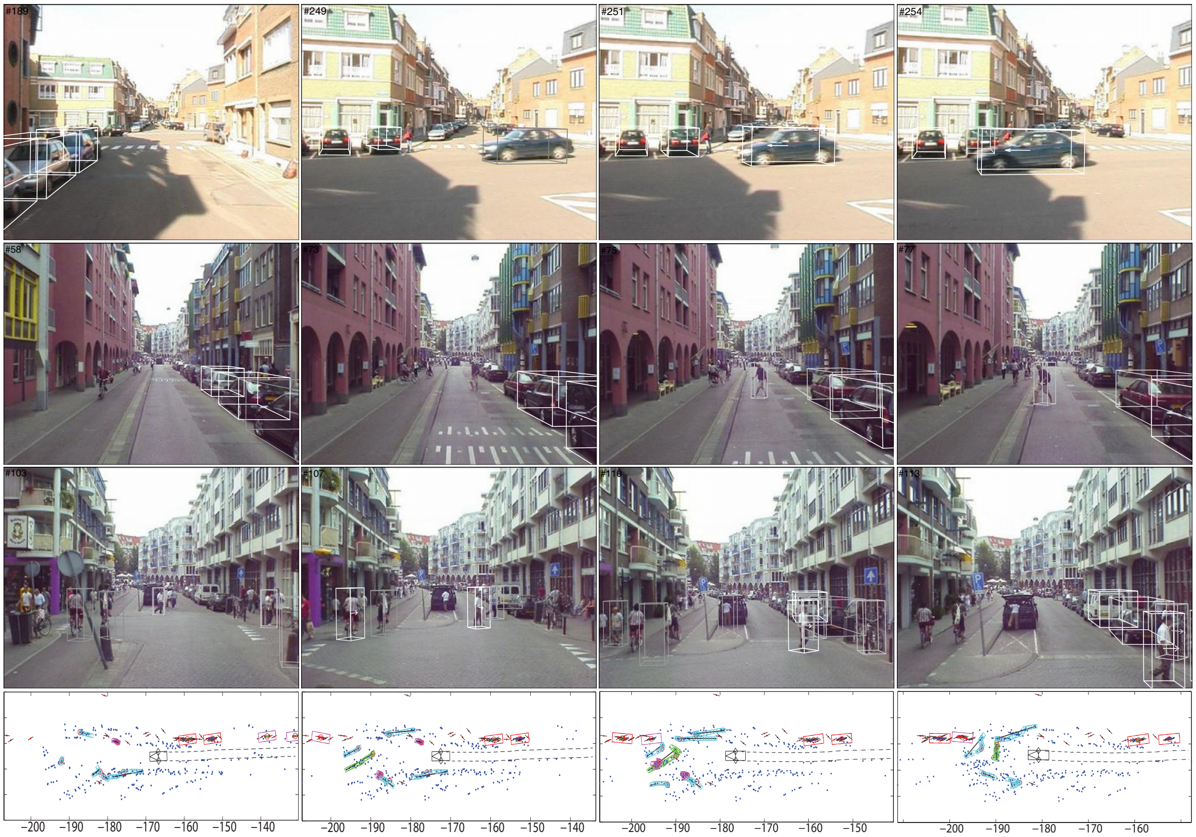


Figure 8. *3D localization and tracking results of our system. The bottom row shows a bird's eye view reconstruction of the third scene.*

[3] M. Betke, E. Haritaoglu, and L. Davis. Real-time multiple vehicle tracking from a moving vehicle. *MVA*, 12(2):69–83, 2000.

[4] N. Cornelis, K. Cornelis, and L. Van Gool. Fast compact city modeling for navigation pre-visualization. In *CVPR'06*, 2006.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR'05*, 2005.

[6] D. Gavrila and V. Philomin. Real-time object detection for smart vehicles. In *ICCV'99*, pages 87–93, 1999.

[7] J. Giebel, D. Gavrila, and C. Schnörr. A bayesian framework for multi-cue 3d object tracking. In *ECCV'04*, 2004.

[8] H. Grabner and H. Bischof. On-line boosting and vision. In *CVPR'06*, pages 260–267, 2006.

[9] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2000.

[10] D. Hoiem, A. Efros, and M. Hebert. Putting objects into perspective. In *CVPR'06*, 2006.

[11] J. Keuchel. Multiclass image labeling with semidefinite programming. In *ECCV'06*, pages 454–467, 2006.

[12] D. Koller, K. Daniilidis, and H.-H. Nagel. Model-based object tracking in monocular image sequences of road traffic scenes. *IJCV*, 10(3):257–281, 1993.

[13] B. Leibe, N. Cornelis, K. Cornelis, and L. V. Gool. Integrating recognition and reconstruction for cognitive traffic scene analysis from a moving vehicle. In *DAGM'06*, 2006.

[14] B. Leibe, K. Mikolajczyk, and B. Schiele. Segmentation based multi-cue integration for object detection. In *BMVC'06*, 2006.

[15] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *CVPR'05*, 2005.

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

[17] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

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

[19] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *Trans. PAMI*, 27(10), 2005.

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

[21] K. Okuma, A. Taleghani, N. de Freitas, J. Little, and D. Lowe. A boosted particle filter: Multitarget detection and tracking. In *ECCV'04*, 2004.

[22] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *ICCV'03*, 2003.

[23] B. Wu and R. Nevatia. Tracking of multiple, partially occluded humans based on static body part detections. In *CVPR'06*, 2006.