Ground Plane Estimation using a Hidden Markov Model

Ralf Dragon and Luc Van Gool
Computer Vision Lab, ETH Zurich
{dragon,vangool}@vision.ee.ethz.ch

Abstract

We focus on the problem of estimating the ground plane orientation and location in monocular video sequences from a moving observer. Our only assumptions are that the 3D ego motion $\vec{t}$ and the ground plane normal $\vec{n}$ are orthogonal, and that $\vec{n}$ and $\vec{t}$ are smooth over time. We formulate the problem as a state-continuous Hidden Markov Model (HMM) where the hidden state contains $\vec{t}$ and $\vec{n}$ and may be estimated by sampling and decomposing homographies. We show that using blocked Gibbs sampling, we can infer the hidden state with high robustness towards outliers, drifting trajectories, rolling shutter and an imprecise intrinsic calibration. Since our approach does not need any initial orientation prior, it works for arbitrary camera orientations in which the ground is visible.

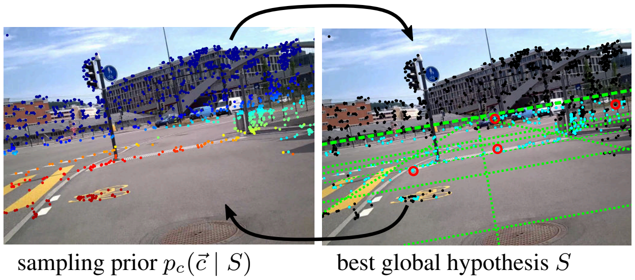
sampling prior $p_c(\vec{c} \mid S)$
best global hypothesis S

Figure 1. Overview over our approach. Given the prior $p_c(\vec{c} \mid S)$ (left, color-coded dots) that a correspondence $\vec{c}$ lies on the ground plane S, homographies are sampled (right, estimation from the four red circles), decomposed and used as ground plane hypotheses. Our Hidden Markov Model finds the best path over time through many ground plane hypotheses. By this, S is refined and thus the sampling prior p_c.

1. Introduction

With the wide spread of cheap and light consumer cameras, new applications are developing such as dashboard or ego cameras (attached to helmets or glasses), or cameras attached to bikes or remote controlled vehicles. The major difference to traditional recording techniques is that there is only weak human guidance: There is no object being watched or scene being captured, but a path being documented. In this paper, we focus on paths on the ground plane and tackle the problem of estimating the orientation and the path of the camera — or from camera's perspective: Where is the ground and how do we move on it?

Such knowledge about the ground plane orientation and offset is an important prior for many computer vision applications, e.g. tracking [7], semantic segmentation [1], free space estimation [14], and scene analysis [5, 6]. If we use a monocular setup, the knowledge about the ground plane is particularly useful since it a allows to measure distances by projecting foot points onto the ground plane. Furthermore, if we have a moving camera, we can project the camera position onto the ground in order to relate observations from different frames with each other.

Such a motion of a moving monocular camera is constrained since it occurs on the ground plane. An obvious idea is to determine the plane from homographies which are established between pairs of frames. However, these planes are not consistently connected over time and sampling points from within one plane and from the right plane is a problem. In this work, we jointly solve the problem of estimating the motion and finding the ground plane. We formulate both unknowns as continuous hidden states in a Hidden Markov Model (HMM) which allows finding a smooth solution over time. Having determined a smooth solution, we use blocked Gibbs sampling to refine our solution. Since our method has no orientation constraint but just enforces that the motion vector $\vec{t}$ should be orthogonal to the ground plane normal $\vec{n}$ over many frames, it can be used to determine the orientation in a large variety of video sequences.

This paper is structured as follows. In Sec. 2, we give an overview over related work. In Sec. 3, we describe the homography decomposition which is used to estimate the ground plane and ego motion. In Sec. 4, we derive the Hidden Markov Model and we show how to initialize and iteratively refine the solution using blocked Gibbs sampling. In Sec. 5, experimental results are presented, and in Sec. 6, we will have a conclusion.

2. Related Work

Existing approaches of ground plane (GP) estimation can be classified into which and how many sensors are used, and into the restrictions which are applied on the motion (e.g. planar or nonholonomic) and on the camera orientation.

An obvious solution of the estimation of the GP orientation is to extract the 3D scene structure. Assuming a dominant ground plane, RANSAC-like approaches can be used for robust parameter estimation. Since there might be other planes besides the GP in the 3D data, estimating GP parameters from 3D data is a multi-structure fitting task which can be solved by approaches like J-Linkage [19]. The 3D structure can be obtained from depth sensors as LIDAR [10] or TOF [12]. A less expensive alternative to generate 3D point clouds is a stereo camera setup in which the ground plane can be estimated from disparity [17]. Assuming that the scene is static, monocular approaches for simultaneous localizing and mapping (SLAM) can also be used to extract the 3D shape and then the GP [18, 11]. However, in contrast to our work, these approaches use wide-angle or omnidirectional cameras for enhanced robustness. In our work, we tackle the problem of estimating the GP without determining the 3D structure. The advantages are lower computational complexity, lower sensitivity to degenerate configurations or small field of view, and higher robustness towards geometrical imperfections of the camera projection.

Our approach was inspired by the approach of [21] who also assume a freely-moving monocular camera setup. They use plane estimates from a homography decomposition as initialization for a bundle adjustment of 3D structure and ego-motion. However, these estimates are not necessarily the ground plane, but façades or even planes on moving vehicles. Furthermore, since their optimization approach, dubbed TRASAC, maximizes a number of inlier trajectories, the GP with only few features is often not considered.

Since GP estimation has manly been used for obstacle detection for robots and cars, the parameterization of the motion is usually chosen problem specific. [17] assume a fixed camera orientation with respect to the GP. Further, they assume a downwards-looking camera which, to our experience, simplifies the problem since multi-structure fitting is not needed. With a similar setup, [9] estimate the GP orientation given the motion from dedicated odometry sensors.

Since a byproduct of our ground plane estimation is the ego motion, our work is related to visual odometry in which most approaches do not need a 3D reconstruction step. However, there exist only few monocular approaches. Since the reduction to a single camera leads to ambiguities and higher noise sensitivity, constraints on the type of motion are important. [13] propose an approach in which camera rotations are restricted to occur around the vertical axis. [16] show that if additionally the motion is nonholonomic, ego motion estimates can be found from only one correspondence which drastically reduces the sampling complexity. Both methods use a decomposition of the framewise Fundamental Matrix $\mathbf{F}$ which only allows to find the translation direction – the distance is found from dedicated odometry sensors. Our work is more general since we train the usual camera motion from data. Besides, instead of $\mathbf{F}$, we decompose the ground plane homography $\mathbf{H}$ which allows to recover the translation distance up to a constant global scale. The monocular approach in the Viso software [4] decomposes $\mathbf{F}$ first and then finds the translation distance by a decomposition of $\mathbf{H}$ of the dominant plane. However, this plane does not have to be the GP and, as visible during the experiments, switching between different planes heavily perturbs the visual odometry. In contrast, our HMM reliably selects the true GP by enforcing orthogonal ego motion and GP normal over long time spans. To the best of our knowledge, although very simple, this constraint has not been proposed before for visual odometry or GP estimation. Closest is the work of [20] who used the orthogonality as additional linear constraint to synchronize the scales of different moving objects in monocular multibody SfM.

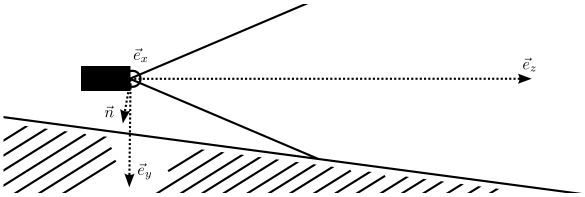

Figure 2. The camera coordinate system with the z-axis $\vec{e}_z$ pointing into the direction of view. Although n_3 is obviously negative, the ground plane is still visible.

3. Sampling of Ground Plane Orientation

In order to estimate the ground plane orientation $\vec{n}$ and ego motion $\vec{t}$ from an image pair, we decompose a homography $\mathbf{H}$ which has been estimated using a minimal sampling set of four correspondences. Let $\bar{x}$ be the homogeneous representation of a 2D point in image coordinates and $\vec{x}$ the corresponding 3D point in the camera coordinate system (cf. Fig. 2). The rigid coordinate transformation

$$\vec{x}' = \mathbf{R}\vec{x} + \vec{s} \tag{1}$$

of 3D points $\vec{x}$ with shift $\vec{s}$ and rotation $\mathbf{R}$ is constrained as follows if $\vec{x}$ are located on a plane defined by $\vec{n}^\intercal \vec{x} = d$ ($\vec{n}$ pointing towards the plane): The image projections $\bar{x}$ are transformed according to

$$\bar{x}' = \mathbf{H}\bar{x} = \mathbf{K}(\mathbf{R} + \vec{t} \cdot \vec{n}^\intercal)\mathbf{K}^{-1}\bar{x}, \tag{2}$$

where $\mathbf{K}$ is the camera matrix with $\bar{x} = \mathbf{K}\vec{x}$, and

$$\vec{t} = \frac{\vec{s}}{d} \tag{3}$$

is the translation of the camera, normalized by the absolute distance d between the GP and the camera center.

We use the standard technique from [8, Ch. 5] to decompose $\mathbf{H}$ into $D = (\vec{n}, \vec{t}, \vec{r})$, where $\vec{r}$ is an axis/angle representation of $\mathbf{R}$.

For the decomposition, the camera matrix $\mathbf{K}$ is assumed to be known. As we will show in the experiments, an approximate

$$\mathbf{K} = \begin{bmatrix} f & 0 & \frac{w}{2} \\ 0 & f & \frac{h}{2} \\ 0 & 0 & 1 \end{bmatrix} \text{ with } f = \frac{w}{2} \arctan\left(\frac{60}{180}\frac{\pi}{2}\right) \quad (4)$$

at an image with w and height h, assuming a horizontal field of view of 60 degrees, is sufficiently accurate.

For a general homography $\mathbf{H}$, there are four possible decomposition D_j, out of which $\vec{n}_1 = -\vec{n}_2$ and $\vec{n}_3 = -\vec{n}_4$. It is often stated (e.g. in [8, Ch. 5]) that a positive z-coordinate n_{j3} can be used to determine upon visibility and to rule out one of D_1 and D_2, and of D_3 and D_4, respectively. However when the horizon is visible in the camera, planes with $n_3 < 0$ are still visible (Fig. 2). Thus, we directly project that points $\bar{x}_i$ onto the plane, which were used to estimate $\mathbf{H}$:

$$\bar{x}_i = \mathbf{K}\vec{x}_i \quad \wedge \quad \vec{n}_j^\intercal \vec{x}_i = 1\,. \quad (5)$$

We choose those two decompositions D_j which yield $x_{i3} > 0$ in all the resulting projected 3D points $\vec{x}_i$. If there is no solution, then the points $\bar{x}_i$ were located on different sides of the horizon and we discard $\mathbf{H}$.

4. Maximum Likelihood Time-Consistent Ground Plane Estimation

The course over the frames f of the ground plane normal $\vec{n}^f$ as well as the ego motion $(\vec{t}^f, \vec{r}^f)$ during a whole sequence is estimated using a state continuous Hidden Markov Model (HMM) formulation defined by continuous underlying states S, observations $\mathcal{O}$, and observation and transition probabilities $p_o(\mathcal{O} \mid S)$ and $p_t(S' \mid S)$, respectively. Denote by $S^f = (\vec{n}^f, \vec{t}^f, \vec{r}^f)$ the ground plane orientation and ego motion at frame f. The transition probability $p_t(S^{f+1} \mid S^f)$, short $p_t(S^{f+1})$, describes the likelihood of a state change due to motion or orientation changes of the camera. The observation probability $p_o(\mathcal{O}^f \mid S^f)$, short $p_o(\mathcal{O}^f)$, describes the probability of observing the minimal sampling set $\mathcal{O}^f = \{(\bar{x}_1, \bar{x}'_1), \ldots, (\bar{x}_4, \bar{x}'_4)\}$ from which $\mathbf{H}$ and its decomposition can be estimated, given that the underlying state is S^f.

The model has the layered structure displayed in Fig. 3. At each frame f, decomposition estimates from different minimal sampling sets $\mathcal{O}_i^f$ form competing states S_i^f. Transitions are possible from every state at frame f to every state at $f+1$. After we observed a sequence (estimated many decompositions from different $\mathcal{O}_i^f$, cf. Sec. 4.3), and having learned the distributions p_o and p_t (cf. Sec. 4.2), the HMM is used to find the most likely sequence of decompositions through F frames:

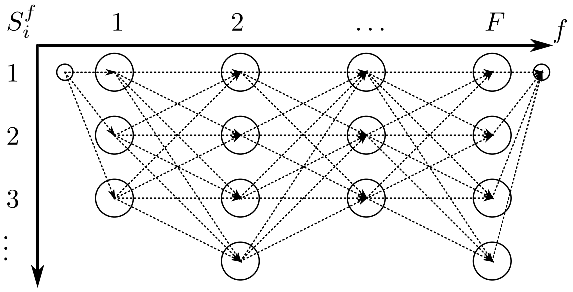

Figure 3. The structure of the HMM. Nodes represent states S_i^f. Each layer of states at a frame f is densely connected with the layer of the consecutive frame $f+1$. The left and rightmost nodes are the enter and exit nodes. If the nodes are assigned an observation likelihood and the edges a transition likelihood, the shortest path according to Eq. (6) maximizes the HMM likelihood.

$$\begin{aligned} \vec{p}^* &= \arg\min_{\vec{p}} l_{\text{HMM}}(\vec{p}) \\ &= \arg\min_{\vec{p}} l_{\text{HMM}}(S_{p_1}^1, \ldots, S_{p_F}^F, \mathcal{O}_{p_1}^1, \ldots, \mathcal{O}_{p_F}^F) \\ &= \arg\min_{\vec{p}} \sum_{f=1}^{F-1} \left(l_o(\mathcal{O}_{p_f}^f) + l_t(S_{p_{f+1}}^{f+1}) \right) + l_o(\mathcal{O}_{p_F}^F), \end{aligned} \quad (6)$$

where l denotes the negative log likelihood of a corresponding probability p. The path $\vec{p}^*$ is found by searching the shortest path through the corresponding graph (Fig. 3).

4.1. Blockwise Linearity Assumption

In order to reduce the computational cost for establishing the Markov model, to reduce measurement noise and to eliminate the false of the two remaining $\mathbf{H}$ decompositions, we process the sequence blockwise. Instead of each pair of consecutive frames being decomposed into D^f, a whole block of length N receives a decomposition B^f. Assigning multiple frames in a block the same decomposition is justified by the demand to receive a smooth decomposition regarding $\vec{t}$ and $\vec{n}$. Furthermore, more stable block decompositions can be found since estimation errors can be averaged out and since the homographies can be estimated over multiple time spans.

In order to estimate one decomposition for a block of length N, we use the first frame as reference frame: We decompose homographies $\mathbf{H}_{1i}$ between the first and the i^{th} frame in the block. For each $\mathbf{H}_{1i}$, there are two possible decompositions D_1^i and D_2^i. Since the first frame is the same for all $\mathbf{H}_{1i}$, the camera coordinate system is also the same for all D^i. Thus, we can compare two decompositions and easily determine an energy

$$E(D_1, D_2) = \angle(\vec{n}_1, \vec{n}_2)^2 + \frac{1}{2}\left(\left(\angle(\vec{t}_1, \vec{n}_2) - \frac{\pi}{2}\right)^2 + \left(\angle(\vec{t}_2, \vec{n}_1) - \frac{\pi}{2}\right)^2\right) \tag{7}$$

describing the dissimilarity between two decompositions in terms of parallel $\vec{n}$ and orthogonal $\vec{n}$ and $\vec{t}$. The best-matching decompositions are found by minimizing the error E_{blk}:

$$\vec{c}^* = \arg\min_{\vec{c}} \sum_{i,j \geq 2, i \neq j} E(D^i_{c_i}, D^j_{c_j}) = \arg\min_{\vec{c}} E_{\text{blk}}, \tag{8}$$

where $\vec{c} \in \{1,2\}^N$ is a binary indicator variable. We minimize E_{blk} with QPBO [15]. Although the energy is not necessary submodular, finding $\vec{c}^*$ is usually successful in a very short period if the underlying homographies were sampled from a planar structure.

Finally, we assign the block the decomposition $B = (\vec{n}, \vec{t}, \vec{r})$ by combining the individual sub-decompositions $D^i_{c^*_i} = (\vec{n}^i_{c^*_i}, \vec{t}^i_{c^*_i}, \vec{r}^i_{c^*_i})$. While combining, we assume that the stability of a homography $\mathbf{H}_{1i}$ grows linearly with the span $s_i = i - 1$. Thus, a decomposition receives a weight w_i proportionally to s_i. We combine vectors $\vec{n}^i$ (and analog $\vec{t}^i$, and $\vec{r}^i$) to $\vec{n}$ by a weighted average of direction $\vec{n}^i/|\vec{n}^i|$ and normalized length $|\vec{n}^i|/s_i$ independently:

$$\vec{n} = \left(\sum_{i=2}^{N} w_i \frac{|\vec{n}^i|}{s_i}\right) \cdot \left(\sum_{i=2}^{N} w_i \frac{\vec{n}^i}{|\vec{n}^i|}\right). \tag{9}$$

Please note that in contrast to sampling correspondences to estimate a decomposition D between a pair of frames, in order to estimate a block decompositions, four *trajectories* are sampled. Since the trajectories may not last over the full block length, some decompositions may be unavailable. For convenience, this is not expressed in equations (8) and (9), but the implementation for estimating a block estimation with missing data is straightforward.

4.2. Observation and Transition Probabilities

4.2.1 Modeling

The observation and transition probabilities $p_o(\mathcal{O} \mid S)$ and $p_t(S' \mid S)$ from Eq. (6) are assumed to be stationary. We model both as Gaussian mixtures over angular deviations δ_o, and $\vec{\delta}$ respectively.

The angular deviation for the observation probability is simply the root mean square error of E in Eq. (8):

$$\delta_o = \sqrt{E_{\text{blk}}}. \tag{10}$$

The transition probability depends on four angular deviations from a linear motion:

$$\delta_1(S' \mid S) = \angle(\vec{n}, \vec{n}') \tag{11}$$

$$\delta_2(S' \mid S) = \angle(\vec{t}, \vec{t}') \tag{12}$$

$$\delta_3(S' \mid S) = |\vec{r}| + |\vec{r}'| \cdot \operatorname{sgn}\left(\angle(\vec{r}, \vec{r}') - \frac{\pi}{2}\right) \tag{13}$$

$$\delta_4(S' \mid S) = \arctan\frac{1}{|\vec{t}|} - \arctan\frac{1}{|\vec{t}'|}. \tag{14}$$

The definition of Eq. (11) and (12) is straightforward. Eq. (13) is more complicated in order to handle orientation changes of $\vec{r}$, e.g. between right and left curves. δ_4 in Eq. (14) needs some more explanation. The motion $\vec{t}$ has been defined as a fraction in Eq. (3), thus

$$\alpha = \arctan\frac{1}{|\vec{t}|} = \arctan\frac{d|\vec{n}|}{|\vec{s}|}. \tag{15}$$

Accordingly, in the triangle spanned by the orthogonal vectors $d\vec{n}$ and $\vec{s}$, α is that angle having $\vec{s}$ as adjacent and $d\vec{n}$ as opposite leg. δ_4 is the deviation of this angle in two consecutive states. The fact that we can trade off a distance change to the ground plane with a change in speed is inherent to the homography decomposition. However, by the ground plane orthogonality constraint in δ_o, we encourage that d remains constant.

4.2.2 Training

We train the mixture models for p_o and p_t using all blocks of all city sequences of the Kitti dataset [3]. Here, different kinds of motion patterns occur (sharp curves, acceleration and braking) observed from a camera on a car roof. We use GPS to compute ground truth poses and from this underlying ground truth states S^f_{gt} and GP homographies $\mathbf{H}^f_{\text{gt}}$ (2). Please note that this motion is quite restricted, since, e.g., a car does not lean into curves. However, it seems ground truth for less restricted motions is not available.

Regarding the observation probability p_o, we collect samples of δ_o using many decompositions B^f sampled from inlier trajectories $\mathcal{I}(\mathbf{H}_{\text{gt}})$ of different blocks and sequences. However, although sampled from inlier trajectories, B^f still might be estimated from a degenerate sampling set. We have many ground truth cues at hand to select such bad observations (e.g., the deviations $\delta_i(B^f | S^f_{\text{gt}}), i = 1 \ldots 4$), but their weighting is unknown and using a single cue leads to poor results. The sum of relative deviations

$$\delta^f_{\text{rel}} = \sum_{i=1}^{4} \frac{\delta_i(S^{f+1}_{\text{gt}} \mid B^f)}{\delta_i(S^{f+1}_{\text{gt}} \mid S^f_{\text{gt}})} \tag{16}$$

allows to relate the angular deviation over time between the ground truth and the decomposition, assuming that the following state is given by the ground truth. For each block, we sample 100 different inlier decompositions B^f and select that one which minimizes δ^f_{rel}.

In order to estimate the transition probability p_t, the ground truth decompositions are extracted from S^f_{gt} and the deviations $\vec{\delta}^f_{\text{gt}}$ are computed. However, there is no normal associated with the GPS data – it is assumed that the camera is always perfectly horizontally aligned with the horizon, thus $\vec{n} = (0, 1, 0)^\intercal$. But then $\delta_1(S' \mid S)$ would always be zero. Instead, we assume that δ_1 is uncorrelated to $\delta_2 \dots \delta_4$, and has the same distribution as δ_2.

Having many samples of δ_o and $\vec{\delta}$, the mixtures for p_t and p_o are estimated using 10 components and weights according to the relative occurrence of the components.

4.3. Sampling Decompositions

Since only a small fraction of the ground plane may be visible, random sampling results in a small fraction of minimal sampling sets $\mathcal{O}$ which are entirely sampled from the ground plane. Since the number of samples of each frame quadratically raises the complexity of the HMM, we use an iterative approach which alternates finding the underlying HMM state and sampling for 20 times. By this, inlier-only minimal sampling sets (MSS) are sampled with increasing probability over the iterations (cf. Fig. 4).

4.3.1 Initial Sampling

For each block, the trajectories are initially motion-segmented into 10 segments using multi-scale motion clustering [2]. Since trajectories from the same segments are more likely to be on the same plane, we sample such that an MSS is within one segment with 50% probability. Our observation is that by this, we need far less samples than we would need with random sampling. We sample 300 decompositions for each block and use the best 100 according to E_{blk} to find an initial estimate of the HMM's hidden states according to Eq. (6).

4.3.2 Blocked Gibbs Sampling

After we have an initial estimate of the underlying states $\mathcal{S} = (S^1, \dots, S^F)$ from all blocks f, we use blocked Gibbs sampling for refinement. Thus, we sample further MSSs containing trajectories to estimate further block decompositions. Since we cannot sample MSSs from the HMM's observation probability p_o, we introduce the prior $p_c(c^f \mid \mathcal{S})$ specifying the likelihood of a correspondence $c^f = (\bar{x}^f, \bar{x}'^f)$ in the f^{th} block given the underlying states $\mathcal{S}$. We define it as follows:

$$p_c(c^f \mid \mathcal{S}) \propto \sum_{\phi=1}^{F} p_o(\mathcal{O}^\phi) \cdot w(f-\phi) \cdot p_{\text{plane}}(c^f \mid S^\phi) \cdot p_{\text{stable}}(\bar{x}^f \mid S^\phi), \tag{17}$$

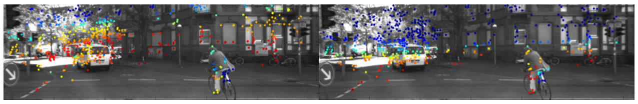

Figure 4. Distribution of p_c (cf. Eq. (17)) in the first and 20th iteration of sequence 2. Red denotes high and blue low probability. While p_{plane} enforces that points which do not fulfill the plane motion (e.g. the houses on the right), p_{stable} cancels out solutions over the horizon (cf. Fig. 5) and which would be located very far if on the ground plane.

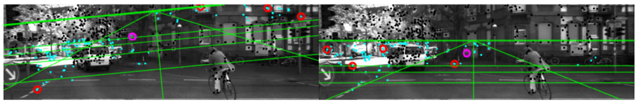

Figure 5. Minimal sampling sets (red circles) sampled according to p_c as displayed in Fig. 4. The green markings on the estimated ground plane are located at $z = \{2|\vec{t}|, \dots, 5|\vec{t}|, \infty\}$ as well as at $x = \{-2.5d, 0, 2.5d\}$. The magenta circle is the projection of the estimated ego motion $\vec{t}$ into the image.

where p_o is the observation probability of the MSS $\mathcal{O}^\phi$ at frame ϕ from the most likely HMM solution found so far, and w is a Blackman-Harris window with a radius corresponding to 10 s. These two weighting factors allow propagating likely solutions to nearby frames, while p_{plane} and p_{stable} guide the sampling within a frame as follows: p_{plane} specifies that the motion of c should fulfill the motion of S^ϕ (2), and p_{stable} prioritizes numerically stable solutions.

We define $p_{\text{plane}}(c \mid S)$ as distribution of the relative symmetric reprojection error

$$r = \frac{1}{2}\frac{\epsilon(\bar{x}', \mathbf{H}_S\bar{x}) + \epsilon(\bar{x}, \mathbf{H}_S^{-1}\bar{x}')}{\epsilon(\bar{x}, \bar{x}')}, \tag{18}$$

where ϵ specifies the Euclidean distance of the image coordinates of two homogeneous vectors. r is assumed to be Gaussian-distributed with vanishing mean and standard deviation 5. In order to compute $p_{\text{stable}}(\bar{x} \mid S)$, we project $\bar{x}$ on the plane (5) yielding a 3D position $\vec{x}$. In this paper, we assume that points with a large z-coordinate x_3 are unstable since a small perturbation in the image has a big effect on the ground plane orientation. Points near the horizon even might be projected from $x_3 = \infty$ to $x_3 = -\infty$. Thus, we prioritize using a Gaussian on x_3 having a vanishing mean and standard deviation $10d$.

4.3.3 Pruning

Although we guide the sampling, the quadratic complexity of the transitions in the HMM is intractable if we would consider all observations $\mathcal{O}$ sampled in all previous iterations. To prune possibly bad observations $\mathcal{O}$, we compare the negative log likelihoods $l_o(\mathcal{O})$ with $l_{\text{HMM}}(\vec{p}^*)$ (6). Since all l_o and l_t are positive, we can discard all $\mathcal{O}$ with

$$l_o(\mathcal{O}) > l_{\text{HMM}}(\vec{p}^*) \tag{19}$$

as they cannot improve a given HMM. Furthermore, all samples $\mathcal{O}_i^f$ are pruned which would raise the negative log likelihood of the HMM too much: Alternative paths $\vec{p}_i^{\,f}$ which differ from $\vec{p}^*$ at frame f by using state S_i^f are discarded iff

$$l_{\text{HMM}}(\vec{p}_i^{\,f}) > 2 \cdot l_{\text{HMM}}(\vec{p}^*). \tag{20}$$

Of course, this assumption leads to discarding alternative diverse paths and may prune good observations. However, similar to simulated annealing, during the first iterations, l_{HMM} is quite high and alternative paths can be explored.

5. Experiments

For the evaluation, we use sequences from the Kitti dataset [3] (Fig. 6) as well as own data containing a large variety of different camera motions and setups (Fig. 7). The first class of sequences is taken with high quality cameras and there are GPS tracks available. Furthermore, the cameras are fixed to a car, so the ground plane orientation does not change significantly over time. In contrast to these sequences, the own videos are not aligned with the horizon and the optical axis does not coincide with the motion direction. In order to test the impact of an imprecise camera matrix, we use the generic $\mathbf{K}$ from Eq. (4) for our own data. An additional difficulty is that the sequences 9-12 show different, untrained motion patterns recorded by a low-quality keychain camera with major rolling shutter distortions[1].

5.1. Ground Plane Orientation Accuracy

In this experiment, the accuracy of the ground plane estimation is measured using the Kitti sequences. We use two quality criteria for the evaluation: The comparison of inlier correspondences and the direct comparison of the angle between estimated and ground truth ground plane orientation.

The GPS data allows us to extract ground truth motion $(\vec{s}, \vec{r})$. d and the camera orientation are static and given by the extrinsic camera calibration to the vehicle. Assuming that the vehicle is always located upright on the ground plane, $\vec{n}$ is also given. Thus, we can compute ground truth homographies $\mathbf{H}_{\text{gt}}$ (2). Similarly, we can compute estimated ground plane homographies $\mathbf{H}_{\mathcal{O}}$ using the most likely observations from the HMM. Using an outlier threshold of $r = 10$ (18), we compare the set of inliers $\mathcal{I}(\mathbf{H}_{\text{gt}})$ with $\mathcal{I}(\mathbf{H}_{\mathcal{O}})$ using the intersection over union metric

$$q_{\text{iou}} = \frac{|\mathcal{I}(\mathbf{H}_{\text{gt}}) \cap \mathcal{I}(\mathbf{H}_{\mathcal{O}})|}{|\mathcal{I}(\mathbf{H}_{\text{gt}}) \cup \mathcal{I}(\mathbf{H}_{\mathcal{O}})|}. \tag{21}$$

In Fig. 8, it can be observed that q_{iou} can be siginificantly raised by the Gibbs sampling.

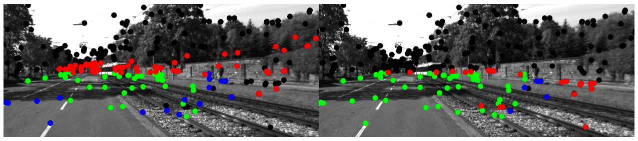

Figure 8. Comparison between the ground truth inlier set $\mathcal{I}(\mathbf{H}_{\text{gt}})$ and the estimated one $\mathcal{I}(\mathbf{H}_{\mathcal{O}})$. The intersection is drawn in green, missing ground plane points in blue and superfluous in red. Left: the result after the first iteration ($q_{\text{iou}} = 0.32$). Right: the improved results after the last iteration ($q_{\text{iou}} = 0.53$).

	our error / deg	our iou	inl error / deg	inl iou
1	1.68	0.58	20.71	0.27
2	3.27	0.53	35.68	0.18
3	3.86	0.47	74.79	0.21
4	7.48	0.39	48.01	0.44
5	3.83	0.59	62.07	0.24
6	5.98	0.52	73.55	0.20
7	2.37	0.73	78.86	0.20
8	4.33	0.46	103.29	0.18

Table 1. The average angular error of $\vec{n}$ and q_{iou} over the different sequences, using our original HMM formulation and an inlier-preferring one.

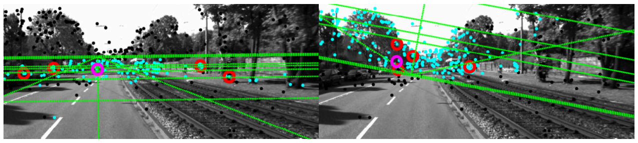

Figure 9. Results in sequence 1 of our original HMM formulation (left), and a formulation containing an inlier-related energy.

As baseline, we use our HMM with modified transition probabilities. As in RANSAC-like approaches, we add a term which makes solutions with many inliers more probable: We change Eq. (6) and set

$$l'_o = l_o - \frac{1}{2}\frac{|\mathcal{I}(\mathbf{H}_{\mathcal{O}})|^2}{\sigma^2_{\text{inlier}}} \quad \text{and} \tag{22}$$

$$l'_t = l_t - \frac{1}{2}\frac{|\mathcal{I}(\mathbf{H}_{\mathcal{O}'}) \cap \mathcal{I}(\mathbf{H}_{\mathcal{O}})|^2}{\sigma^2_{\text{inlier}}}. \tag{23}$$

Even if we make the variance σ^2_{inlier} large (1000^2), our results degrade drastically: As it can be extracted from Table 1, the errors rises a lot. Fig. 9 reveals that the added term leads to finding false ground planes that maximize inlier counts.

In order to qualitatively evaluate sequences 9-16 without GPS ground truth, we plot the estimated horizon, inliers to the ground plane homographies as well as orientation lines in the ground plane coordinate system. As it can be observed from Fig. 10, our results are quite accurate although the generic camera matrix $\mathbf{K}$ is used[1].

5.2. Monocular Visual Odometry

Although the principle purpose is ground plane estimation, we can also use the approach for monocular visual odometry. We compare our approach to the implementation in Viso [4]. To measure the performance, we compute

[1] Original sequences and videos with qualitative results are available at www.vision.ee.ethz.ch/~dragonr/1401.

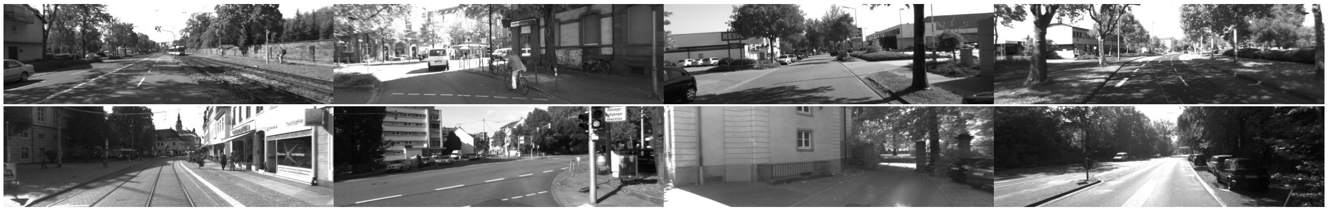

Figure 6. Overview of the sequences 1-8 (row-wise starting top left), taken from the Kitti dataset.

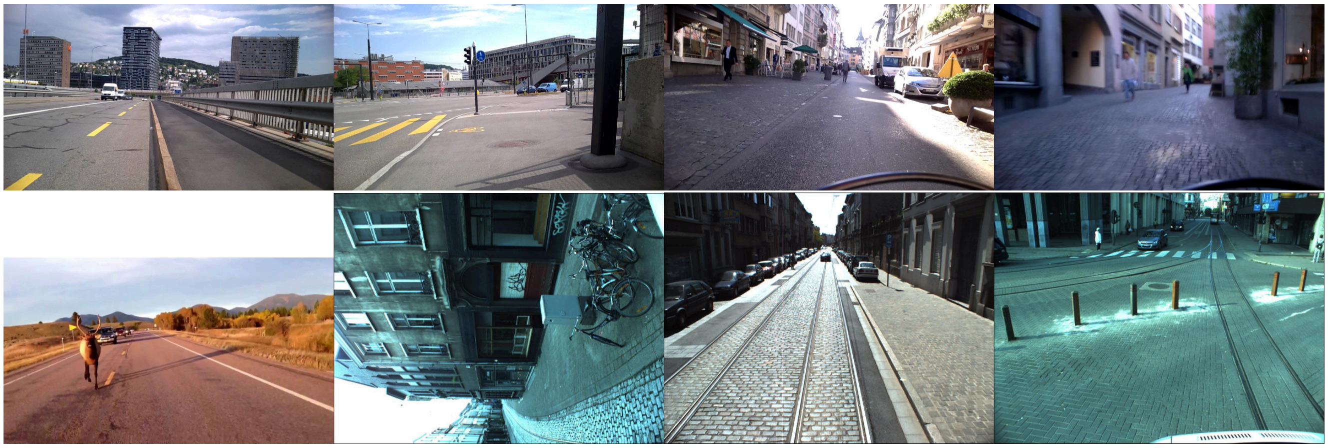

Figure 7. Overview over ego-motion sequences 9-16 (row-wise starting top left). Sequences 9-12 are taken from a camera attached to a bike, sequence 13 from a motorbike with a freely-moving camera, and 14-16 from a car-mounted camera.

	our ang / deg	our dist / %	Viso ang / deg	Viso dist / %
1	0.02	0.59	0.80	0.69
2	0.07	0.75	1.22	0.40
3	0.04	0.72	0.27	0.23
4	0.23	1.99	0.92	0.33
5	0.01	0.34	0.41	0.28
6	0.07	0.74	0.39	0.17
7	0.20	1.65	3.06	4.95
8	0.11	2.13	1.68	1.11

Table 2. The average relative angular and distance errors of our GP-based approach (left two columns) comparing with Viso [4].

camera poses and compare them with the given GPS ground truth using the translational and angular errors, normalized by the path length. The results are given in Table 2.

As it can be observed, our approach receives comparable distance errors although we did not include any visual odometry specific optimizations as degeneracy handling. However, our angular errors are even one magnitude better than the one of Viso. Further qualitative results[1] in sequences 9-16 indicate that our approach is far more robust towards rolling shutter and camera blur.

5.3. Discussion

The ability to determine the horizon (Sec. 5.1) and to measure orientation changes (Sec. 5.2) allows us to use our approach as a *visual Gyroscope*. Although this seems not useful as support for drivers or pilots, the horizon is a very important cue in image analysis, especially in a monocular setup where distances have to be measured by foot points on the ground plane.

Regarding the runtime, our non-optimized current implementation is far from realtime. However, for online GP estimation, we could reformulate our approach in order to iteratively estimate the hidden HMM states while using the Gibbs sampling only from new frames. With this, we could easily reach realtime performance. Furthermore, we did not take care about handling degenerate cases, e.g. if there is no motion. As this is an essential step in visual odometry, we can expect to boost our results by this.

Since the HMM does not only help us to find the best ground plane but also returns a probability for this interpretation, we could learn to distinguish different motion patterns with multiple HMMs and use it for the classification of the ego motion. Finally, the HMM formulation has shown to be a powerful tool to select the best from multiple hypotheses which are connected by inertia constraints. It can be expected that it outperforms the common handling of outliers via RANSAC or robust loss functions in structure-from-motion or SLAM approaches.

6. Conclusion

In this paper, we proposed a new way of ground plane estimation. For different pairs of frames, we sample multiple hypotheses of the ground plane and ego motion via homo-

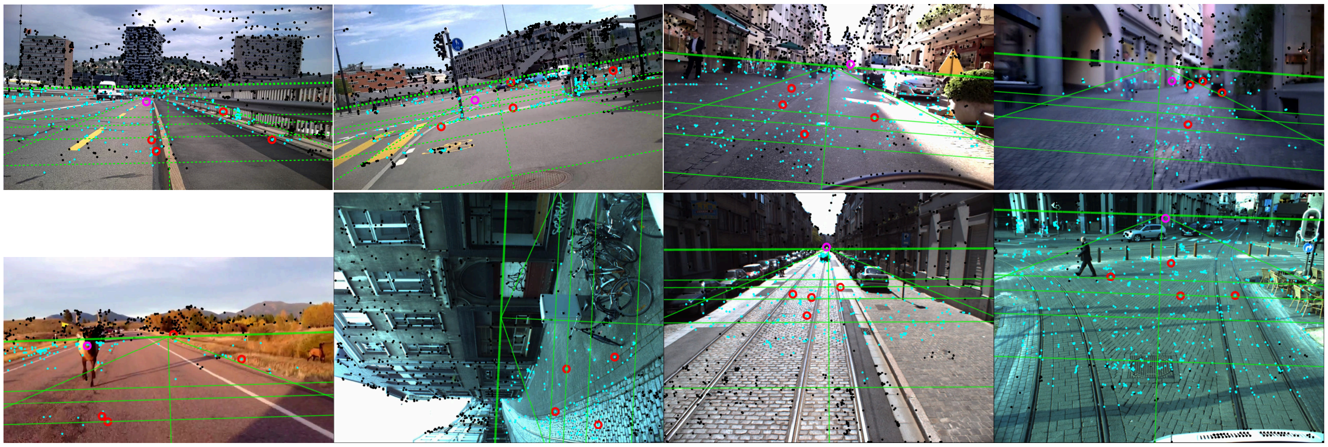

Figure 10. Qualitative results from our own sequences with diverse motions. The markings are explained in Fig. 5.

graphy decomposition. Our HMM formulation then allows to find the most likely set of hypotheses. This in turn is used to sample refined estimates with blocked Gibbs sampling.

We showed that our ground plane estimation approach works robustly in a large variety of sequences, including tilted cameras or heavily blurred and wobbling images. Using our approach for monocular visual odometry, we received state-of-the-art distance errors, but the angular error was one magnitude lower.

Since the approach is quite simple, it is applicable for diverse tasks: Motion patterns for the HMM transition probabilities can be learned depending on the application. Furthermore, by changing the HMM observation probabilities, even other kinds of structures like lines or vanishing points could be tracked over time.

Acknowledgement: This work was funded by the EU project VarCity.

References

[1] J. M. Alvarez, T. Gevers, Y. LeCun, and A. M. Lopez. Road scene segmentation from a single image. In *ECCV*, 2012. 1

[2] R. Dragon, B. Rosenhahn, and J. Ostermann. Multi-scale clustering of frame-to-frame correspondences for motion segmentation. In *ECCV*, Oct. 2012. 5

[3] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research*, 2013. 4, 6

[4] A. Geiger, J. Ziegler, and C. Stiller. Stereoscan: Dense 3D reconstruction in real-time. In *IEEE Intelligent Vehicles Symposium*, pages 963–968, June 2011. 2, 6, 7

[5] V. Hedau, D. Hoiem, and D. Forsyth. Thinking inside the box: Using appearance models and context based on room geometry. In *ECCV*, 2010. 1

[6] D. Hoiem, A. A. Efros, and M. Heber. Putting objects in perspective. *IJCV*, 80:3–15, 2008. 1

[7] S. M. Khan and M. Shah. A multiview approach to tracking people in crowded scenes using a planar homography constraint. In *ECCV*, 2006. 1

[8] Y. Ma, S. Soatto, J. Kosecka, and S. S. Sastry. *An Invitation to 3-D Vision*. Springer, 2004. 3

[9] D. Maier and M. Bennewitz. Appearance-based traversability classification in monocular images using iterative ground plane estimation. In *Intelligent Robots and Systems*, 2012. 2

[10] M. W. McDaniel, T. Nishihata, C. A. Brooks, and K. Iagnemma. Ground plane identification using LIDAR in forested environments. In *ICRA*, 2010. 2

[11] B. Micusik and J. Kosecka. Piecewise planar city 3D modeling from street view panoramic sequences. In *CVPR*, 2009. 2

[12] F. Mufti, R. Mahony, and J. Heinzmann. Robust estimation of planar surfaces using spatio-temporal RANSAC for applications in autonomous vehicle navigation. *Robotics and Autonomous Systems*, 60(1):16 – 28, 2012. 2

[13] D. Ortn and J. M. M. Montiel. Indoor robot motion based on monocular images. *Robotica*, 19:331–342, 2001. 2

[14] D. Pfeiffer and U. Franke. Efficient representation of traffic scenes by means of dynamic stixels. In *Proceedings of the IEEE Intelligent Vehicles Symposium*, pages 217–224, June 2010. 1

[15] C. Rother, V. Kolmogorov, V. Lempitsky, and M. Szummer. Optimizing binary MRFs via extended roof duality. In *CVPR*, 2007. 4

[16] D. Scaramuzza. 1-point-RANSAC structure from motion for vehicle-mounted cameras by exploiting non-holonomic constraints. *IJCV*, 95:74–85, 2011. 2

[17] S. Se and M. Brady. Ground plane estimation, error analysis and applications. *Robotics and Autonomous Systems*, 39:59–71, 2002. 2

[18] J.-P. Tardif, Y. Pavlidis, and K. Daniilidis. Monocular visual odometry in urban environments using an omnidirectional camera. In *IROS*, pages 2531–2538, 2008. 2

[19] R. Toldo and A. Fusiello. Robust multiple structures estimation with J-Linkage. In *ECCV*, pages 537–547, 2008. 2

[20] C. Yuan and G. Medioni. 3D reconstruction of background and objects moving on ground plane viewed from a moving camera. In *CVPR*, 2006. 2

[21] Z. Zhou, H. Jin, and Y. Ma. Robust plane-based structure from motion. In *CVPR*, pages 1482–1489, 2012. 2