Estimating Vehicle Ego-Motion and Piecewise Planar Scene Structure from Optical Flow in a Continuous Framework

Andreas Neufeld, Johannes Berger, Florian Becker,
Frank Lenzen, and Christoph Schnörr

IPA & HCI, University of Heidelberg, Germany
{neufeld,becker,schnoerr}@math.uni-heidelberg.de
{johannes.berger,frank.lenzen}@iwr.uni-heidelberg.de

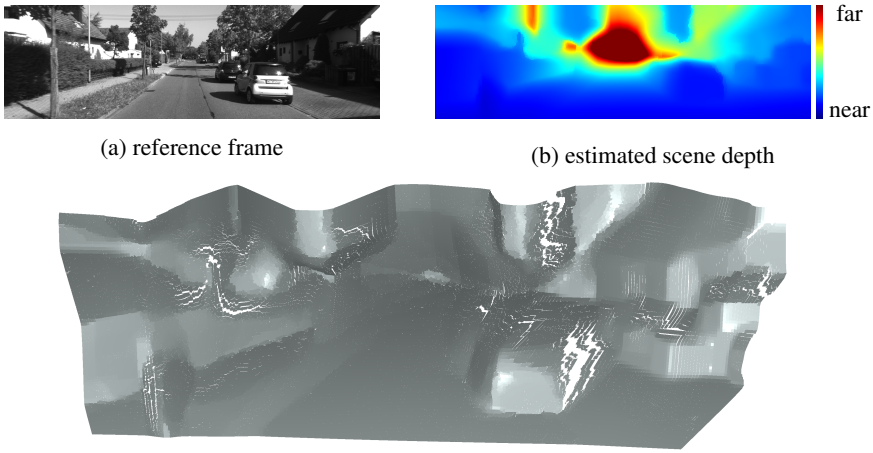
Abstract. We propose a variational approach for estimating egomotion and structure of a static scene from a pair of images recorded by a single moving camera. In our approach the scene structure is described by a set of 3D planar surfaces, which are linked to a SLIC superpixel decomposition of the image domain. The continuously parametrized planes are determined along with the extrinsic camera parameters by jointly minimizing a non-convex smooth objective function, that comprises a data term based on the pre-calculated optical flow between the input images and suitable priors on the scene variables. Our experiments demonstrate that our approach estimates egomotion and scene structure with a high quality, that reaches the accuracy of state-of-the-art stereo methods, but relies on a single sensor that is more cost-efficient for autonomous systems.

1 Introduction

1.1 Overview

For the scenario of a camera moving through a static scene, e.g. in an automotive environment, we present an approach for jointly estimating the scene structure and the camera egomotion. In a preprocessing step the optical flow between these two frames together with a confidence map is estimated, and serves as input data. Moreover, for one of the frames, a partition of the image domain into superpixels is determined. The main part (and main contribution) of our method consists of a variational approach with a non-convex smooth objective function, which includes suitable chosen priors on the scene depth and plane parameters to guarantee a consistent scene representation with only a sparse set of depth discontinuities. By minimizing this objective function we obtain an estimate of the egomotion in terms of rotation and translation together with a description of the scene by one 3D plane per superpixel. Fig. 1 depicts a typical scene reconstruction. From the plane parameters both scene depth and surface normals can be determined directly.

We stress that, due to the *monocular* nature of the considered problem with a less favorable motion parallax, the task is more difficult than stereo setups studied in this context. However, industry favors more cost- and energy-efficient sensor solutions.



(a) reference frame

(b) estimated scene depth

(c) surface visualization of reconstructed scene

Fig. 1: *Best viewed in color.* (a) first frame of an image pair from the KITTI stereo benchmark; (b) depth map derived from the piecewise planar scene structures computed by our monocular approach jointly with the camera motion; (c) shaded visualization of the piecewise planar structure.

1.2 Related work

Scene reconstruction in the automotive context poses an important foundation for higher-level reasoning e.g. in advanced driver assistant systems. For vision based outdoor scene reconstruction stereo based systems currently dominate, as this well-posed problem setting with a known calibrated stereo camera setup leads to highly accurate results. This is substantiated by the enormous popularity of the KITTI benchmark [7].

In the recent years *monocular* scene reconstruction approaches became increasingly popular although they have to additionally determine the unknown relative camera position between two frames. This has been proved to be feasible also in real-time both for indoor [13,9,12,15,16] and the even more challenging task of outdoor setups, where a world map is aggregated over an entire image sequence (Simultaneous Localization And Mapping, SLAM) [5,22]. Despite the higher computational effort compared to stereo setups, monocular camera systems feature reduced calibration effort which is interesting from the industrial point of view. Results presented e.g. in [3] demonstrate that depth accuracy comparable to stereo methods can be achieved even in an automotive context. Similar to the methods above we consider the case of a monocular camera setup, however, do not accumulate information over an image sequence but only resort to *two* consecutive image frames to estimate scene and egomotion. In [25,24] epipolar geometry is pre-computed and flow is restricted to fixed epipolar lines. We implement a *joint* estimation approach of egomotion and scene description, as is also done in [13,3].

A few algorithms rely on independent matches for scene reconstruction [18], but most algorithms incorporate a prior on the regularity of the depth map to cope with

ambiguities and distortions in the data. Piecewise constant depth maps seem to be a reasonable assumption in connection with modeling shallow objects and occlusions present in indoor scenes. For street scenes however, slanted planes such as the street or house fronts dominate, and providing an accurate reconstruction is important for subsequent reasoning steps. Stereo methods [20,21,24,23] implementing this prior rank at top positions in the according KITTI benchmark. While the above methods work with a (partially) discretized parameter space, we consider continuous variables, which results in a differentiable objective function, for which established and soundly studied numerical method are available. The objective function enables us to perform a joint optimization in all variables.

Since our approach utilizes a scene description by piecewise planar surfaces, it closely relates to estimating multiple homographies explaining the optical flow induced by the motion of a camera relative to planar surfaces. The seminal works [14,26] showed that the set of homographies of any number of views is embedded in a four dimensional subspace which also carries a manifold structure [6]. Recent approaches [4,17] are based on inter-homography constraints and do not require camera calibration. In contrast, our method assumes the *intrinsic camera parameters* to be known. This requirement comes with the advantage, that the planes can be estimated physically correctly (up to a global scale).

The approach presented in this work builds upon an accurate estimation of the optical flow for which we can resort to existing and *publicly available* methods that have proven to be accurate in the considered scenario. We choose to the top ranked monocular optical flow method [19] in the KITTI benchmark with source code available.

2 Approach Overview

Preliminaries, Notation. Throughout this paper, we consider scenarios where a 3D scene is recorded by a projective camera from two different perspectives. We denote 3D points by $X \in \mathbb{R}^3$. W.l.o.g. we assume the first camera position to be $(0,0,0)^\top$ with viewing direction $(0,0,1)^\top$ and refer to the image recorded from this position by I_1. We denote the projection of a point X onto the first image plane by $x = \pi(X) \in \Omega$ with *image domain* $\Omega \subset \mathbb{R}^2$. Assuming the *intrinsic* camera parameter to be known we can w.l.o.g. utilize normalized image coordinates, i.e. $\pi(X) := X_3^{-1} \left(\begin{smallmatrix} X_1 \\ X_2 \end{smallmatrix}\right)$.

For the second recording, the camera is rotated with rotation matrix $R \in \mathsf{SO}(3)$ and translated by vector $t \in \mathsf{S}^2$. We refer to (R,t) as the *extrinsic* camera parameters. The translation is constrained to unit norm, since scene scale cannot be determined from monocular images. The projection of a point X onto the second image plane then is given as $x' := \pi(R^\top(X-t))$ and the acquired image is denoted by I_2.

We aim at representing the reconstructed scene by a number of space planes which we parametrize by $v \in \mathbb{R}^3$, such that any space point $X \in \mathbb{R}^3$ lying on the plane fulfills $\langle v, X\rangle = 1$. Assuming that the scene can be (locally) represented by plane parameters v, the apparent motion induced by the camera movement is described by

$$x' = \pi\left(H(R,t,v)\left(\begin{smallmatrix} x \\ 1 \end{smallmatrix}\right)\right), \tag{1}$$

with the homography $H(R,t,v) := R^\top(I - tv^\top)$ (cf. e.g. [10, Chap. 13]).

Finally, we estimate planes on a pre-computed connected partition $\{\Omega^i\}_i$ (superpixels) of the first image using the SLIC (Simple Linear Iterative Clustering) method [2]. We further define the common boundary of superpixel i and j by $\partial^{ij} := \overline{\Omega^i} \cap \overline{\Omega^j}$. The set of all neighboring superpixel pairs is denoted by $\mathcal{N}_\Omega := \{(i,j) | i,j \in \{1,\dots,n\}, \partial^{ij} \neq \emptyset\}$. We *assume* that all space points $X \in \mathbb{R}^3$ projected to superpixel $i \in \{1,\dots,n\}$, i.e. $\pi(X) \in \Omega^i$, lie on a plane parametrized by $v^i \in \mathbb{R}^3$, see Fig. 2 for an illustration. Using (1) we gain a low-parametric *model* for the optical flow

$$u(x; R, t, v) := x' - x = \pi(H(R,t,v)\,(\tfrac{x}{1})) - x. \tag{2}$$

Then, for an *observed* optical flow $\hat{u} : \Omega \mapsto \mathbb{R}^2$ which approximately transports I_1 to I_2 we formulate the inverse problem of determining the piecewise planar scene description $v := (v^1,\dots,v^n) \in \mathbb{R}^{3n}$ and camera motion (R,t), which explains $\hat{u}$, as finding a solution to the problem

$$\min_{R\in\mathsf{SO}(3), t\in\mathsf{S}^2, v\in\mathbb{R}^{3n}} E(R,t,v). \tag{3}$$

The energy function $E(R,t,v)$ furthermore incorporates priors on the scene structure and is detailed in Sect. 3.

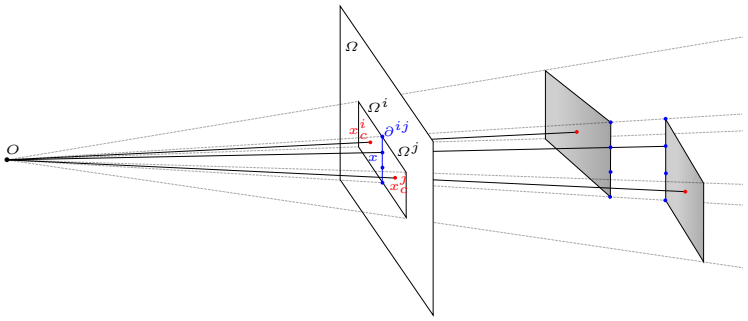


Fig. 2: *Best viewed in color.* Projective camera and discretization. Two rectangular superpixels Ω^i, Ω^j in the image domain Ω and two space planes parametrized by v^i, v^j and restricted to the cone defined by the camera origin O and the superpixel coverage. Regularity of depth is evaluated at all positions $x \in \partial^{ij}$ (blue dots) along common superpixel boundaries – see Sect. 3.2. The non-negativity prior on depth is evaluated on superpixel centers x_c^i, x_c^j (red dots) – see Sect. 3.4.

3 Variational Approach

Our energy function $E(R,t,v)$ decomposes into

$$E(R,t,v) = E_u(R,t,v) + \lambda_z E_z(v) + \lambda_v E_v(v) + \lambda_p E_p(v), \tag{4}$$

where E_u is the data fidelity term, E_z and E_v are priors on the depth and the plane parameters, respectively and E_p is a term penalizing negative depth values. We detail all four terms in Sects. 3.1–3.4. The terms are coupled via the positive weighting parameters λ_z, λ_v and λ_p. Our choice for these parameters is provided in the experimental section, cf. Sect. 4. Our numerical approach to minimize (4) is presented in Sect. 3.5.

3.1 Data Fidelity

The fidelity term $E_u(R,t,v)$ in our optimization problem is the deviation of an *observed* optical flow $\hat{u}(x)$ from our model (2) and is defined as

$$E_u(R,t,v) := \sum_{i=1}^{n} \sum_{x \in \Omega^i} w_{\hat{u}}(x) \|u(x;R,t,v^i) - \hat{u}(x)\|_2^2 \,. \tag{5}$$

Here, $w_{\hat{u}}(x) \geq 0$ denotes a spatially varying weighting of the data term which is provided by a confidence measure of the optical flow algorithm as detailed next.

Optical Flow Estimation. The optical flow $\hat{u}$ between images I_1 and I_2 as required by the data term (5) is computed in a pre-processing step using the algorithm *Data-Flow* being the highest ranked publicly available monocular implementation (cf. [19]) in the KITTI optical flow challenge.

We complement the output obtained from *Data-Flow* with a confidence map $w_{\hat{u}}(x)$, which avoids the influence of flow vectors which are considered incorrect. To this end we also estimate the *backward* flow between I_2 and I_1, providing an estimate $\hat{u}^{-1}(x)$ of the inverse mapping of $\hat{u}(x)$. Only points that are consistently mapped forth and back are considered correct and we define the confidence map as

$$w_{\hat{u}}(x) := \exp\left(-\tfrac{1}{2}\|x - (\hat{u}^{-1} \circ \hat{u})(x)\|_2^2 / \sigma_{\hat{u}}^2\right) \tag{6}$$

with value $\sigma_{\hat{u}} > 0$. Experimentally, we found the value $\sigma_{\hat{u}} = \frac{1}{2\sqrt{2}}$ to be suitable.

3.2 Smoothness Prior on Depth

In order to enforce that planes of neighboring superpixels form a seamlessly connected surface in most parts of the image, we introduce the prior $E_z(v)$ as follows. We consider points on the common boundary $x_\partial \in \partial^{ij}$ of superpixel i and j and penalize deviations of their inverse depth $z^{-1}(x_\partial, v) = x_\partial^\top v$ according to the two plane models v^i and v^j, see Fig. 2 for an illustration.

In order to encourage sharp depth edges we make use of the generalized Charbonnier functional

$$\rho_C(x) := (x^2 + \epsilon)^\alpha - \epsilon^\alpha. \tag{7}$$

We choose $\epsilon = 10^{-10}$ and $\alpha = 1/4$ throughout the work, so that $\rho_C^2(x)$ smoothly approximates $|x|$. Then the energy function for one boundary ∂^{ij} reads as

$$E_z^{ij}(v) := \sum_{x \in \partial^{ij}} \rho_C^2(x^\top v^i - x^\top v^j). \tag{8}$$

Note that we opted to compare *inverse* depth $z^{-1}(x, v)$ due to a superior numerical performance and reconstruction. Then the global smoothness term consists of a weighted sum of E_z^{ij} over all neighboring superpixels $(i, j) \in \mathcal{N}_\Omega$:

$$E_z(v) := \sum_{(i,j)\in\mathcal{N}_\Omega} w_\Omega^{ij} E_z^{ij}(v)\,. \tag{9}$$

The weights $w_\Omega^{ij} \geq 0$ are computed based on appearance differences , i.e.

$$w_\Omega^{ij} := \exp\left(-\tfrac{1}{2}(m_i - m_j)^2/\sigma_\Omega^2\right)\,, \tag{10}$$

where m_i and m_j are the mean gray values of frame I_1 in superpixel Ω^i and Ω^j, respectively. For parameter σ_Ω, we use a fixed value of 0.2.

3.3 Smoothness Prior on Plane Parameters

In addition to seamless surfaces on superpixel boundaries, we aim at plane parameters which up to a small set of discontinuities are constant over the image domain. This property encourages large connected planar structures.

For the plane smoothness prior we employ again the Charbonnier function ρ_C (see (7), here applied component-wise), and the boundary weights w_Ω^{ij} from (10):

$$E_v(v) = \sum_{(i,j)\in\mathcal{N}_\Omega} w_\Omega^{ij} \|\rho_C(v^i - v^j)\|_2^2\,. \tag{11}$$

3.4 Positive Depth Prior

As a further constraint, we require all observed space points to be in front of the camera. Thus, we introduce an additional prior E_p. We apply a soft hinge function

$$\rho_+(x) := \begin{cases} 1 - 2x & x \leq 0 \\ (1-x)^2 & 0 < x \leq 1 \\ 0 & 1 < x \end{cases}\,, \tag{12}$$

to the inverse depth given by $z^{-1}(x_c^i, v^i) = \left\langle \begin{pmatrix} x_c^i \\ 1 \end{pmatrix}, v^i \right\rangle$, evaluated at superpixel centers $x_c^i \in \Omega^i$, see Fig. 2. Summing over all superpixels, this leads to

$$E_p(v) = \sum_{i=1}^{n} \rho_+^2(z^{-1}(x_c^i, v^i))\,. \tag{13}$$

3.5 Optimization

The considered optimization task (3) comprises a non-convex smooth energy function (4) and manifold constraints $R \in \mathsf{SO}(3)$ and $t \in \mathsf{S}^2$. In order to find a local

minimum of $E(R,t,v)$, we choose the Levenberg-Marquardt method [11], which has been adapted to Riemannian manifolds in [1].

The proposed energy function $E(R,t,v)$ can be decomposed into a sum of m squared functions $f_j(R,t,v)$, where $m = 2|\Omega| + \sum_{(i,j)\in\mathcal{N}_\Omega} |\partial_{ij}| + 3|\mathcal{N}_\Omega| + n$, i.e.

$$E(R,t,v) = \sum_{j=1}^{m} (f_j(R,t,v))^2 = \|f(R,t,v)\|_2^2\,, \tag{14}$$

with $f(R,t,v) := (f_1(R,t,v), \ldots, f_m(R,t,v))^\top \in \mathbb{R}^m$.

We combine the variables into a joint vector $Y := (R,t,v)$ and locally re-parametrize Y near (R^k, t^k, v^k) by parameters $\eta := (\omega, \delta t, \delta v)^\top \in \mathbb{R}^{3+3+3n}$ as

$$Y(\eta) := (R^k \operatorname{Exp}([\omega]_\times), \Pi_{\mathsf{S}^2}(t^k + \delta_t), v^k + \delta_v). \tag{15}$$

Here, $\operatorname{Exp}(\cdot)$ is the matrix exponential function applied to the skew-symmetric matrix $[\omega]_\times := \begin{pmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{pmatrix}$, which can be efficiently evaluated using the Rodrigues' rotation formula, c.f. [10]. Furthermore, $\Pi_{\mathsf{S}^2}(t) := t/\|t\|_2$ denotes the orthogonal projection of t to S^2. Using first order Taylor expansion we obtain an approximation of $f(Y(\eta))$ in $Y = (R^k, t^k, v^k)$,

$$\tilde{f}^k(\eta) := f^k(0) + \left(\mathrm{J}\, f^k(\eta)\big|_{\eta=0}\right)\eta\,, \tag{16}$$

with Jacobian $\mathrm{J}\, f^k$ of f^k. The Jacobian is obtained for the rotation and translation by differentiating the function compositions $\frac{\partial}{\partial\omega}(f \circ \operatorname{Exp})(\omega)$ and $\frac{\partial}{\partial t}(f \circ \Pi_{\mathsf{S}^2})(t)$, respectively. Substituting this approximation in (14) yields a model of the actual energy function $\tilde{E}^k(\eta)$. However, we augment this objective function by a step regularization term in order to cope with strongly non-linear terms:

$$\min_\eta \tilde{E}^k(\eta) + \mu^k \|\eta\|_2^2. \tag{17}$$

The resulting objective is quadratic in η and thus can be solved efficiently. The update rule for the damping parameter μ^k is described in [1]. A limit of 80 iterations was used as stopping criterion which was sufficient for most of the considered data. We again stress the fact that the minimization of $E(R,t,v)$ is performed jointly w.r.t. R,t,v.

4 Experiments

Evaluation Methodology. In the following we evaluate the quality of scene description and egomotion estimate separately, see paragraphs *Plane Parameter Evaluation* and *Camera Motion Evaluation* below. The KITTI benchmark database [7] provides a suitable image data source as it is annotated with accurate depth and egomotion estimates. As reference surface normal information is not available in these data sets and no *monocular* approach with publicly available code can be compared to, we resort to a state-of-the-art *stereo* method [24], which is highly ranked as *SPS-St* in the KITTI

stereo benchmark. It provides scene depth as well as a surface normals and can be assumed to be very accurate due to the well-posed stereo setup.

The KITTI odometry benchmark contains reference camera poses for a small number of sequences. Based on this reference data, we compare the *odometry results* of our approach to those of the freely available monocular approach *VISO2-M* [8].

Parameter Choice. In order to reduce errors caused by optical flow vectors pointing outside the image area, we apply our method to an image pair in inverse temporal order. The camera motion is thus initialized by a trivial *backward* motion $R = I$, $t = (0, 0, -1)^\top$ and flat scene $v = (0, 0, 0.001)^\top$ everywhere. Furthermore, we chose $\lambda_z = 0.05$, $\lambda_v = 0.001$ and $\lambda_p = 0.1$ – see (4) – throughout the experiments.

Plane Parameter Evaluation. In contrast to stereo methods, the accuracy of *depth* estimates of monocular methods varies depending on the projected position in the image plane and camera motion. We adopt the error measure proposed in [3] between estimated depth $z(x)$ and reference depth $z_{\text{ref}}(x)$ which respects this varying sensitivity,

$$e(x) := F \frac{|z(x) - z_{\text{ref}}(x)|}{\sigma_g(z_{\text{ref}}(x), x)} \tag{18}$$

with F denoting the camera's focal length in pixels.

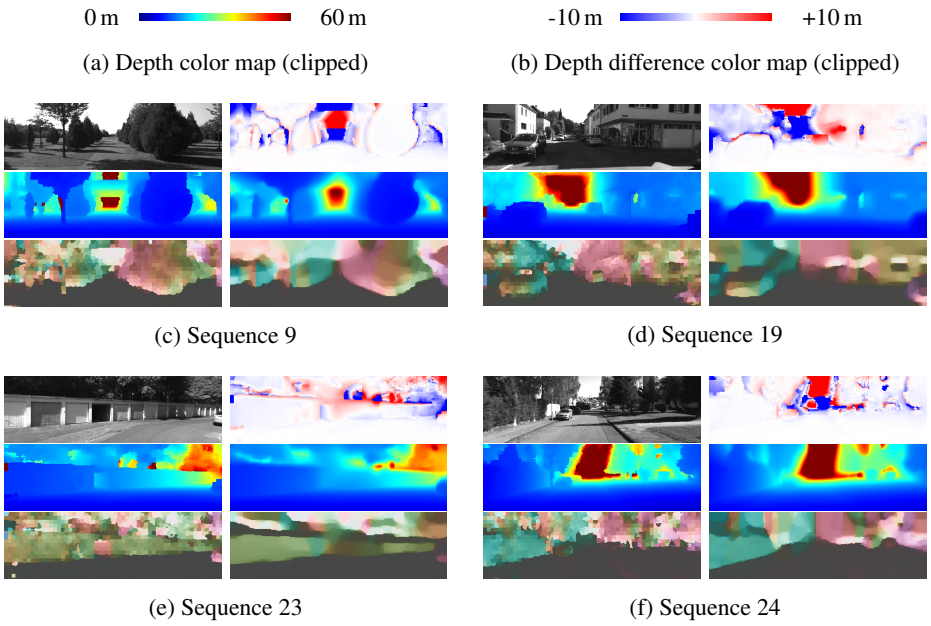
(a) (b) (c) (d) (e)

Fig. 3: *Best viewed in color.* (a) Approximate visually equidistant color scheme for plane normal visualization used throughout the work. (b) Exemplarily frame of a simple synthetic sequence and (c) ground truth normals, and normals as reconstructed by (d) *SPS-St* and (e) our method.

Estimating the global scale inherently unknown in a monocular setting allows a quantitative comparison to metric reference data. To this end we approximate the scale as the median of the depth ratios $z(x)/z_{\text{ref}}(x)$ on the most reliable 10% according to sensibility prediction similar as done in [3].

Table 1 lists summarizing statistics of errors $e(x)$ computed over all pixels with reference depth z_{ref} (calculated from disparities) for 194 frames. A qualitative comparison against the stereo method *SPS-St* is given in Fig. 4.

Plane normal parameters are qualitatively compared to those obtained from [24] in Fig. 4 and Fig. 5. For a quantitative comparison, we use 240 frame pairs (each with 1280×720 pixels) from four simple ray-traced scenes but with known ground truth normals, see Fig. 3 for an example. Results for our method and *SPS-St* are presented in Table 2. We use the same parameters as on the KITTI dataset with both methods. We observe that despite the less favorable monocular setup the error of the plane normals estimated by the proposed method is smaller than the errors from SPS-St.



(a) Depth color map (clipped)

(b) Depth difference color map (clipped)

(c) Sequence 9

(d) Sequence 19

(e) Sequence 23

(f) Sequence 24

Fig. 4: *Best viewed in color.* Depth and plane normal comparison between our monocular and a reference stereo method [24]. From top to bottom, and left to right, each subfigure shows (top row) the reference frame and depth difference, (middle row) reference and estimated depth and (bottom row) reference and estimated normals. The depth values and depth differences are encoded as depicted in (a) and (b), respectively. The encoding of plane normals is illustrated in Fig. 3. Both depth and normal reconstructions mostly agree, but there is a loss in reconstruction detail near the epipole (near image center), see e.g. (d), which is an inherent problem of all monocular setups. Note that especially the ground surface is reconstructed well in most cases.

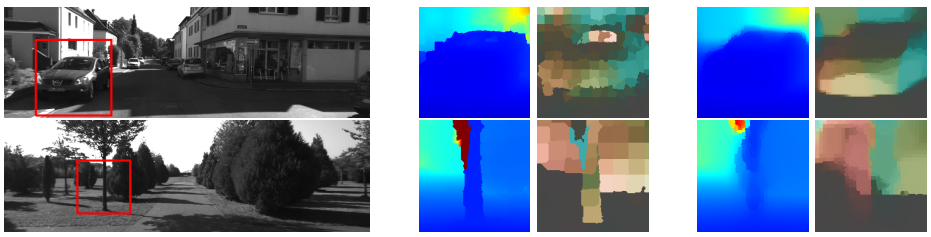


Fig. 5: *Best viewed in color.* Detailed views of scene reconstruction by (center) *SPS-St* and (right) our monocular method, showing depth and plane normals for both. Top row: our method uses less connected planes to explain the object. Lower row: the stereo method reconstructs the tree trunk overly wide but with sharp borders while our solution is more detailed but has smoother edges.

	noc			occ		
	mean e [px]	$p_{2\text{px}}$ [%]	$p_{3\text{px}}$ [%]	mean e [px]	$p_{2\text{px}}$ [%]	$p_{3\text{px}}$ [%]
our method	4.09	12.9	8.63	4.88	13.6	9.17
SPS-St	3.15	12.6	7.46	9.12	13.8	8.57

Table 1: Depth accuracy of our *monocular* method and *stereo* reference method *SPS-St*, evaluated on the KITTI stereo benchmark training data, distinguishing between areas without (*noc*) and with occluded areas (*occ*) as specified in the benchmark. Mean of depth error measurement $e(x)$ (see (18)) and percentage of pixels with error $e > 2\,\text{px}$ and $e > 3\,\text{px}$, respectively. Our approach shows similar performance as *SPS-St* despite the less beneficial parallax and unknown camera position.

	mean [deg.]	$p_{1\text{deg.}}$ [%]	$p_{2\text{deg.}}$ [%]	$p_{5\text{deg.}}$ [%]	$p_{10\text{deg.}}$ [%]
our method	11.5	58.4	45.5	31.1	22.7
SPS-St	14.8	79.4	66.6	46.4	33.4

Table 2: Plane normal errors for four synthetic sequences with known normals, see Fig. 3. The normal angle error w.r.t. ground truth is evaluated over 240 scene reconstructions. Note that we do not use a normalization scheme as in eq. (18). Our method outperforms the stereo method despite the less favourable monocular setup.

Egomotion Evaluation. We evaluate the egomotion accuracy of the proposed method as well as a reference method [8] on the first 100 frames of the first 11 KITTI odometry sequences which all provide ground truth camera poses. We determine the *angle* error of the camera rotation and – due to the ambiguity in global scale – also between the translation vectors. Our method has an average rotational error of $0.057°$ and translation error of $3.86°$, and performs better than the reference method *VISO2-M* [8] with errors $0.18°$ and $6.0°$, respectively.

5 Conclusion and Further Work

We presented a variational method for estimating relative camera positions and planar scene structure from two views of a static scene. An objective function over egomotion and scene planes defined on superpixels was formulated and minimized continuously. We demonstrated that our *monocular* approach provides a scene reconstruction with reasonable accuracy in depth and plane normals compared to an approach in the less challenging *stereo* setup. Egomotion estimates also show a slightly better performance than a state-of-the-art odometry method. Future directions are extension to multiple frames, explicitly handling depth discontinuities and simultaneous estimation of flow and scene parameters.

References

1. Absil, P.A., Mahony, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press (2008) 7
2. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: SLIC Superpixels Compared to State-of-the-art Superpixel Methods. IEEE Trans Pattern Anal Mach Intell 34 (11), 2274 – 2282 (2012) 4
3. Becker, F., Lenzen, F., Kappes, J.H., Schnörr, C.: Variational Recursive Joint Estimation of Dense Scene Structure and Camera Motion from Monocular High Speed Traffic Sequences. Int J Comput Vision 105 (3), 269–297 (2013) 2, 8
4. Chojnacki, W., Szpak, Z.L., Brooks, M.J., van den Hengel, A.: Multiple Homography Estimation with Full Consistency Constraints. In: DICTA (2010) 3
5. Engel, J., Schöps, T., Cremers, D.: LSD-SLAM: Large-Scale Direct Monocular SLAM. In: ECCV (2014) 2
6. Eriksson, A., van den Hengel, A.: Optimization on the Manifold of Multiple Homographies. In: ICCV (2009) 3
7. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets Robotics: The KITTI Dataset. Int J Robot Res (2013) 2, 7
8. Geiger, A., Ziegler, J., Stiller, C.: Stereoscan: Dense 3D reconstruction in real-time. In: IEEE Intelligent Vehicles Symposium. pp. 963–968 (2011) 8, 10
9. Graber, G., Pock, T., Bischof, H.: Online 3D reconstruction using Convex Optimization. In: ICCV (2011) 2
10. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, second edn. (2004) 3, 7
11. Moré, J.J.: The Levenberg-Marquardt algorithm: Implementation and Theory. In: Numerical analysis, pp. 105–116. Springer (1978) 7
12. Newcombe, R.A., Davison, A.J.: Live dense reconstruction with a single moving camera. In: CVPR (2010) 2
13. Newcombe, R.A., Lovegrove, S.J., Davison, A.J.: DTAM: Dense Tracking and Mapping in Real-Time. In: ICCV. pp. 2320–2327 (2011) 2
14. Shashua, A., Avidan, S.: The Rank 4 Constraint in Multiple ($\geq$ 3) View Geometry. In: ECCV (1996) 3
15. Stühmer, J., Gumhold, S., Cremers, D.: Parallel Generalized Thresholding Scheme for Live Dense Geometry from a Handheld Camera. In: CVGPU (2010) 2
16. Stühmer, J., Gumhold, S., Cremers, D.: Real-time dense geometry from a handheld camera. In: Pattern Recognition (Proc. DAGM) (2010) 2
17. Szpak, Z.L., Chojnacki, W., Eriksson, A., van den Hengel, A.: Sampson Distance Based Joint Estimation of Multiple Homographies with Uncalibrated Cameras. Comput Vis Image Und 125, 200–213 (2014) 3
18. Tola, E., Strecha, C., Fua, P.: Efficient large-scale multi-view stereo for ultra high-resolution image sets. Machine Vision and Applications 23(5), 903–920 (Sep 2012) 2
19. Vogel, C., Roth, S., Schindler, K.: An Evaluation of Data Costs for Optical Flow. In: German Conference on Pattern Recognition (GCPR) (2013) 3, 5
20. Vogel, C., Roth, S., Schindler, K.: View-Consistent 3D Scene Flow Estimation over Multiple Frames. In: ECCV (2014) 3
21. Vogel, C., Schindler, K., Roth, S.: Piecewise Rigid Scene Flow. In: ICCV (2013) 3
22. Wendel, A., Maurer, M., Graber, G., Pock, T., Bischof, H.: Dense Reconstruction On-the-fly. In: CVPR (2012) 2
23. Yamaguchi, K., Hazan, T., McAllester, D., Urtasun, R.: Continuous Markov Random Fields for Robust Stereo Estimation. In: ECCV (2012) 3

24. Yamaguchi, K., McAllester, D.A., Urtasun, R.: Efficient Joint Segmentation, Occlusion Labeling, Stereo and Flow Estimation. In: ECCV (2014) 2, 3, 7, 8, 9
25. Yamaguchi, K., McAllester, D.A., Urtasun, R.: Robust Monocular Epipolar Flow Estimation. In: CVPR (2013) 2
26. Zelnik-Manor, L., Irani, M.: Multi-View Subspace Constraints on Homographies. In: ICCV (1999) 3