Variational Recursive Joint Estimation of Dense Scene Structure and Camera Motion from Monocular High Speed Traffic Sequences

Florian Becker, Frank Lenzen, Jörg H. Kappes and Christoph Schnörr
HCI and IPA, Heidelberg University, Speyerer Str. 6, 69115 Heidelberg, Germany
{becker,kappes,schnoerr}@math.uni-heidelberg.de, frank.lenzen@iwr.uni-heidelberg.de

Abstract

We present an approach to jointly estimating camera motion and dense scene structure in terms of depth maps from monocular image sequences in driver-assistance scenarios. For two consecutive frames of a sequence taken with a single fast moving camera, the approach combines numerical estimation of egomotion on the Euclidean manifold of motion parameters with variational regularization of dense depth map estimation. Embedding this online joint estimator into a recursive framework achieves a pronounced spatio-temporal filtering effect and robustness. We report the evaluation of thousands of images taken from a car moving at speed up to 100 km/h. The results compare favorably with two alternative settings that require more input data: stereo based scene reconstruction and camera motion estimation in batch mode using multiple frames. The employed benchmark dataset is publicly available.

1. Introduction

Overview and Motivation. Computer vision research has a strong impact on driver assistance technology. Besides designing dedicated detectors for specific object classes [4, 7], current major trends include low-level estimation of dense scene structure from stereo sequences [22], the transition to monocular imaging sensors [23, 15], and context-based 3D scene representation and labeling supported by high-level assumptions and constraints [24].

This paper focuses on the low-level task to jointly estimate dense scene structure and egomotion under minimal assumptions, adverse conditions and requirements, that are typical for driver assistance scenarios – see Fig. 1:

- Online joint estimation from only two consecutive frames in view of on-board implementations later on;
- No assumptions about scene structure in order to cope with arbitrary scenes;
- No additional input (e.g. odometer readings) besides internal camera parameters estimated offline (calib.);

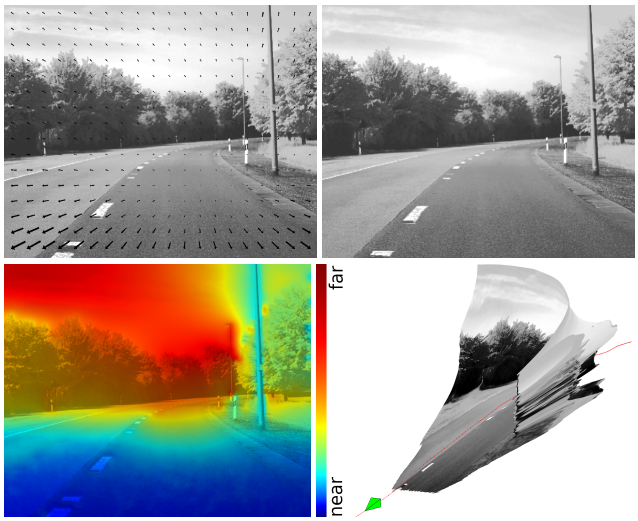


Figure 1. **Best viewed in color. Upper row:** Two consecutive frames (size 656×541 pixels) of the *Bend* sequence with large displacements up to 35 pixels induced by a fast moving camera. **Lower row:** Our approach jointly estimates, from sparse noisy displacement estimates, *dense depth maps* (**left**) *and camera motion* in an online recursive framework. **Right:** Reconstruction of dense scene structure based on the depth maps from the camera's viewpoint, and the corresponding camera track.

- Ability to cope with large displacements induced by a fast moving camera;
- Comprehensive evaluation using image sequences recorded in real scenarios.

In this connection, the major issue to be addressed concerns the design of an integrated approach that ensures sufficient regularization to achieve robust and accurate estimation, without compromising real-time capability through unrealistically complex computations.

Our approach therefore combines *highly accurate* numerics on the *low-dimensional* Euclidean manifold in order to disambiguate and track translational and rotational egomotion from ill-posed two-frame displacement estimates, with *less accurate* variation models for estimating *high-*

 Published by the IEEE Computer Society

dimensional scene structure, leading to efficient overall inference. Applying the resulting online joint estimator within a recursive prediction-estimation loop to an image sequence achieves favorable spatio-temporal filtering and increased robustness.

The estimates computed with our approach provide a basis for subsequent tasks like obstacle and collision warning, and further related problems of advanced scene analysis, to be considered in future work.

Related Work. Most approaches to scene reconstruction rely on stereo imaging or multiple view reconstructions in batch mode.

Stereo set-ups [22, 6] are only relevant for sensing close-up ranges at low speeds, due to the small baseline in driver assistance scenarios, and are less attractive than just a single camera from the technological system oriented viewpoint.

Factorization [19] and bundle adjustment [21] have become a mature technology for jointly determining camera and scene structure from tracked features. While this requires to accumulate several frames and more expensive numerics, recent local and more efficient approaches, e.g. for visual odometry [12, 14], entail only sparse representations of scene structure.

Further recent work on the reconstruction of accurate depth maps from arbitrary multiple views includes [15, 18]. These works however require the camera motion to be determined in a preceding step using feature tracking. Other related approaches only allow for camera translation but no rotation [23], or estimate the epipole but require images to be aligned with respect to a common reference plane [10].

An attractive alternative would employ direct feature-to-depth mappings, learned offline from ground truth databases [17]. Besides the tremendous effort necessary to compile a sufficiently large set of – in particular, far field – ground truth data, we don't currently know how such an approach generalizes to *arbitrary* scenes, and if it can compete with reconstructions that rely on measurements efficiently estimated online, as in our case.

Contribution and Organization. We present an approach that estimates from a monocular high-speed image sequence of arbitrary static scenes both camera motion and *dense* scene structure (depth maps), using noisy sparse displacements computed from two consecutive frames at each instant of time. The approach combines, by joint optimization, geometric integration over the Euclidean manifold SE_3 for incremental motion parameter estimation, with large-scale variational depth map estimation, subject to spatial and short-time temporal regularization. The novelty of our approach is due to the ability to recover *dense* scene structure and egomotion from *monocular* sparse displacement estimates within a truly recursive *online* estimation framework.

Sect. 2 provides an overview of the overall approach and specifies underlying assumptions and approximations, followed by detailing each component of our method in Sect. 3. We report in Sect. 4 results of an evaluation of our approach using thousands of real images provided by a novel database [13], that aims at providing a benchmark for computer vision algorithms in the context of automotive applications. All image data is available online[1].

Moreover, we show that our approach compares favorably to results computed with less restricted approaches (stereo, bundle adjustment) using public implementations (Voodoo Camera Tracker[2] v1.1.0b) and [20, 6], to ensure reproducibility of all results.

2. Problem Statement, Approach (Overview)

Preliminaries. We adopt the common concepts of multiple view geometry [8]. We assume the *internal* camera parameters to be known (offline calibration) and denote *incremental external* parameters corresponding to frame k by $C^k = (R^k, h^k)$, moving the camera from its position at time $k-1$, see Fig. 2.

The manifold $\mathcal{M} := SE_3$ of Euclidean transformations $C = (R, h) \in \mathcal{M}$, parametrized by rotations R and translations h, is identified with the matrix Lie group

$$G := \left\{ Q = \begin{pmatrix} R & h \\ 0^\top & 1 \end{pmatrix} : R \in SO_3,\ h \in \mathbb{R}^3 \right\}, \qquad (1)$$

and SO_3 denotes the group of proper rotations. $T_C\mathcal{M}$ and T_QG denote the tangent spaces of $\mathcal{M}$, G at $C \in \mathcal{M}$, $Q \in G$, respectively. The Lie algebra se_3 of SE_3 is given by

$$se_3 = \left\{ W = \begin{pmatrix} \hat{\omega} & v \\ 0^\top & 0 \end{pmatrix} : \omega, v \in \mathbb{R}^3 \right\}, \hat{\omega} := \begin{pmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{pmatrix}, \qquad (2)$$

where $so_3 \ni \hat{\omega}$ denotes the Lie algebra of SO_3 identified with the linear subspace of skew-symmetric matrices. We equip SE_3 with the Riemannian metric

$$\langle W^1, W^2 \rangle_G := \langle \hat{\omega}^1, \hat{\omega}^2 \rangle + c_G \langle v^1, v^2 \rangle, \quad c_G > 0, \qquad (3)$$

for all $W^i \in se_3$, $i = 1, 2$, where $\langle \cdot, \cdot \rangle$ on the right-hand side denotes the canonical matrix and vector inner product, respectively, and c_G is a constant parameter scaling the rotational vs. the translational part. Note that unlike for general Riemannian metrics, the metric (3) does not depend on $Q \in G$, hence is the same for all tangent spaces T_QG, justifying the notation $\langle \cdot, \cdot \rangle_G$.

The exponential mapping $\mathrm{Exp}\colon se_3 \to SE_3$ that diffeomorphically maps tangent vectors close to 0 onto the manifold within a neighborhood of the group identity I, can be

[1] http://hci.iwr.uni-heidelberg.de/VSFM

[2] http://www.digilab.uni-hannover.de/docs/manual.html

Published by the IEEE Computer Society

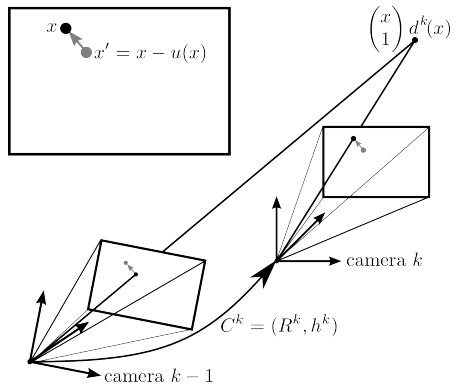


Figure 2. A scene point is defined by image coordinates x and depth $d^k(x)$ in the coordinate system of camera k. Its projection moves by $-u^k(x)$ to x' when the camera is rotated and translated *backward* by $C^k = (R^k, h^k)$.

computed in closed form,

$$\mathrm{Exp}\left[\begin{pmatrix} \hat{\omega} & v \\ 0^\top & 0 \end{pmatrix}\right] = \begin{pmatrix} R(\omega) & Q(\omega)v \\ 0^\top & 1 \end{pmatrix}, \tag{4}$$

$$R(\omega) = I + \frac{\sin(\|\omega\|)}{\|\omega\|}\hat{\omega} + \frac{1-\cos(\|\omega\|)}{\|\omega\|^2}\hat{\omega}^2, \tag{5}$$

$$Q(\omega) = I + \frac{1-\cos(\|\omega\|)}{\|\omega\|^2}\hat{\omega} + \frac{\|\omega\| - \sin(\|\omega\|)}{\|\omega\|^3}\hat{\omega}^2. \tag{6}$$

Problem Statement. Let $\Omega \subset \mathbb{R}^2$ be the image domain and $I^{0:k} := \{I^0, I^1, \ldots, I^k\}$ a given image sequence of frames $I^i \colon \Omega \to \mathbb{R}$, measured at times $i \in \{0, \ldots, k\}$ with cameras $C^{0:k}$. We wish to *jointly estimate in a recursive manner* both $C^{0:k}$ and a sequence $d^{0:k}$ of *depth maps* $d^i \colon \Omega \to \mathbb{R}_+$ that assign to each image point $x \in \Omega$ its depth $d^i(x)$ along the viewing ray, up to a common global unknown scale factor – see Fig. 2.

The difficulty of this problem is (i) due to a monocular driver assistance scenario (see Fig. 1) inducing less favorable motion parallax, (ii) a fast moving camera leading to displacements of consecutive frames up to 35 pixels (frame size 656×541 pix.), and (iii) a recursive online processing mode that updates the camera parameters and depth map based on two consecutive frames only.

Approach: Overview. The natural approach to this problem is to consider sequences of state variables $X^{0:k} = (C^{0:k}, d^{0:k})$ and observations $Y^{0:k} = u^{0:k}$, together with probabilistic models of state transitions $p(X^k|X^{k-1})$ and the observation process $p(Y^k|X^k)$ under Markovian assumptions, in order to recursively estimate X^k based on the posterior marginal distribution $p(X^k|Y^{0:k})$ (cf., e.g. [2]).

Approximations to this general approach are inevitable, however, due to the nonlinearity of the underlying processes, due to the high dimensionality of depth maps d^k and displacement fields u^k (cf. Fig. 2), and due to a strict requirement for computational efficiency imposed by the scenario shown by Fig. 1. We adopt therefore the variational modeling perspective as accepted alternative in situations where sampling based approaches are too time consuming (cf., e.g. [11]).

Accordingly, as detailed in Sect. 3, we devise Gaussian approximations $p(Y^k|X^k) = \mathcal{N}(u^k; \mu_u^k, \Sigma_u^k)$ and $\mathcal{N}(d^k; \hat{\mu}_d^k, \hat{\Sigma}_d^k)$ for the high-dimensional observations $Y^k = u^k$ and states d^k, respectively, that sufficiently take into account uncertainties due to the aperture problem and the viewing geometry (regions around the epipole). Evaluating the former Gaussian entails routine parallel coarse-to-fine signal processing, whereas the latter additionally takes into account *spatial and temporal* context (regularization) in term of predictions $\hat{\mu}_d^k$, $\hat{\Sigma}_d^k$.

$\mathcal{N}(d^k; \hat{\mu}_d^k, \hat{\Sigma}_d^k)$ is complemented by a *local* Gaussian model $\mathcal{N}_{\mathcal{M}}(C^k; C^{k-1}, \sigma_C^2 I)$ of motion parameters on the tangent space of the Euclidean manifold $\mathcal{M}$ at C^{k-1} (cf. [16]), to form together an approximation of the state transition $p(X^k|X^{k-1})$, $X^k = (C^k, d^k) = (R^k, h^k, d^k)$. Putting all components together, we define and compute our update as mode of the posterior marginal approximation $p(X^k|Y^{0:k}) \propto p(Y^k|X^k)p(X^k|X^{k-1})$. Concerning the motion parameters C^k, we prefer working directly on $\mathcal{M}$ using established concepts of numerics [1], rather than to represent the two-view geometry by the essential matrix and to recover C by additional factorization [9].

3. Approach: Details

Our approach *jointly* estimates egomotion and a *dense* depth map from a monocular image sequence. The *recursive* formulation requires constant amount of storage and aims at real-time applications. Large displacements inevitable in the considered scenario are handled in the common coarse-to-fine manner [5]. Uncertainty of observations and depth estimates are handled by probabilistic models.

3.1. Observation Process

We detail the observation process $p(Y^k|X^k)$, with state variables $X^k = (C^k, d^k)$ (camera, depth map) and the camera C^k given by C^{k-1} in the previous frame and the egomotion parameters (R^k, h^k). To simplify notation, we refer to frame $k-1$ with primes (e.g. C') and temporarily drop indices k and $k-1$.

Using the known internal camera parameters, we undo the corresponding affine transformation of the image plane (cf. [8]) and denote the normalized image coordinates by $x \in \Omega \subset \mathbb{R}^2$. Note that all related quantities like displacements, means and covariance matrices have to be transformed as well. To keep the notation simple, however, we only refer to *normalized quantities* in what follows.

Any scene point $d(x)\binom{x}{1}$ at depth d along the viewing ray $\binom{x}{1}$ projects to the image point with inhomogeneous coordinates x. We denote this projection of scene points $(X_1, X_2, X_3)^\top$ in camera C by P_C,

$$P_C(X_1, X_2, X_3) := \frac{1}{X_3}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}. \tag{7}$$

Consider any two subsequent points in time and the camera motion $C' \to C$ given by parameters (R, h). The motion induces an apparent motion $(R^\top, -R^\top h)$ of scene points

$$d'(x')\begin{pmatrix} x' \\ 1 \end{pmatrix} \;\to\; d(x)\begin{pmatrix} x \\ 1 \end{pmatrix} = R^\top\Big(d'(x')\begin{pmatrix} x' \\ 1 \end{pmatrix} - h\Big). \tag{8}$$

Now we *define* the displacement $u(x)$ in the image plane (see Fig. 2) by

$$x' = x - u(x). \tag{9}$$

Using eqns. (7) and (8) we obtain

$$u(x; R, h, d) = x - P_C\left(d(x)R\begin{pmatrix} x \\ 1 \end{pmatrix} + h\right). \tag{10}$$

Simple transformations show that x and x' are corresponding points w.r.t. the essential matrix (cf. [8]) $E = R^\top \hat{h}$, i.e.

$$\binom{x}{1}^\top E \binom{x'}{1} = 0\,, \tag{11}$$

which implicitly defines the epipolar line in C' corresponding to x for fixed R, h.

Observations Y correspond to *estimates* $\hat{u}(x)$ of the displacements (10) for all $x \in \Omega$ using the Lucas-Kanade method [3],

$$\hat{u}(x) := \Sigma_u(x)\Big(G_\rho(x) * \Big(\big(\partial_t I(x)\big)\big(\nabla I(x)\big)\Big)\Big), \tag{12a}$$

$$\Sigma_u(x) := \Big(G_\rho(x) * \Big(\big(\nabla I(x)\big)\big(\nabla I(x)\big)^\top\Big)\Big)^{-1}. \tag{12b}$$

Here, $G_\rho(x)*$ denotes element-wise Gaussian convolution of the subsequent matrix comprising partial derivatives ∂_t, $\nabla := \binom{\partial_{x_1}}{\partial_{x_2}}$ of the image sequence function $I(x,t)$. They are estimated by 3×3 binomial filters and first-order differences derived by linearizations at time k. Likewise, we choose a rather small smoothing kernel of size $\rho = 2$ pix., leading to a fast processing stage. We point out that stronger regularization (smoothing) is not necessary as the embedding multiscale framework and the state prediction (see Sect. 3.2) ensure *small* incremental displacements $u(x)$.

As for the unknown observation process $p(Y^k|X^k)$, our ansatz is

$$p(Y^k|X^k) = \mathcal{N}(\hat{u}^k; \mu_u^k, \Sigma_u^k), \tag{13}$$

where μ_u^k is composed position-wise of $\mu_u^k(x) := u(x; R^k, h^k, d^k(x))$ due to eqn. (10), and Σ_u^k is a block-diagonal covariance matrix with component matrices (12b).

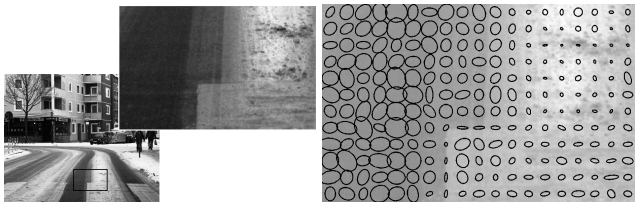


Figure 3. Detailed view of an image frame and an ellipse representation of the estimated flow uncertainty $\Sigma_u^k(x)$. Highly textured regions (upper right) can be correctly distinguished from locations with low confidence due to low signal-to-noise ratio (left) and image edges (aperture problem; middle).

Note that the definition of μ_u^k makes explicit the conditioning on the state parameters $X^k = (C^k, d^k) = (R^k, h^k, d^k)$.

Model (13) only approximates the true unknown observation process (12a). The uncertainty of observations u^k is modelled by Σ_u^k and internally represented by the precision matrices $(\Sigma_u^k)^{-1}$ (cf. eqn. (12b)). Hence, homogeneous image regions and the aperture problem are represented as rank-0 and rank-1 matrices, respectively, (see Fig. 3) and thus can be correctly accounted for within the overall recursive estimation framework – see Sect. 3.3.

3.2. State Transition and Prediction

We detail the state transition model $p(X^k|X^{k-1})$ for the state variables $X = (C, d)$.

Camera. We take $C^{k-1} =: \hat{C}^k$ both as prediction $\hat{C}^k$ of C^k and as mean of the probabilistic model

$$C^k \sim p(C^k|C^{k-1}) = \mathcal{N}_\mathcal{M}(C^k; C^{k-1}, \sigma_C^2 I) \tag{14}$$

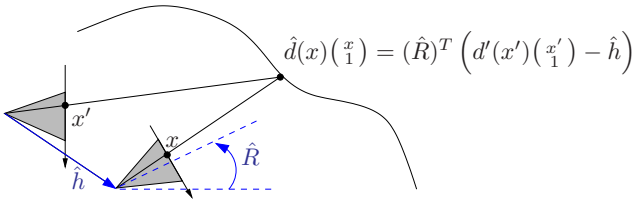
$$\propto \exp\Big(-\frac{1}{2\sigma_C^2}\mathrm{dist}_\mathcal{M}(C^{k-1}, C^k)^2\Big), \tag{15}$$

where $\mathrm{dist}_\mathcal{M}(\cdot,\cdot)$ denotes the geodesic distance on the Euclidean manifold $\mathcal{M} = SE_3$. The prediction $\hat{C}^k = C^{k-1}$ is justified by the fast frame rate. Distribution (14) represents an isotropic Gaussian distribution of random points $C \in \mathcal{M}$ around $C^{k-1} \in \mathcal{M}$ (cf. [16]). σ_C is a user parameter that we set and keep constant throughout all experiments. Model (14) will be further detailed in connection with inference in Sect. 3.3.

Depth Map. The predicted depth map $\hat{d}^k$ is computed by transporting d^{k-1} by the motion parameters $(\hat{R}^k, \hat{h}^k) = (R^{k-1}, h^{k-1})$. To obtain predicted depth values $\hat{d}^k(x)$ at grid positions x in frame k, we approximately infer corresponding positions x' in frame $k-1$ using eqns. (9) and (10),

$$x' \approx P_C\left(d^{k-1}(x)R^{k-1}\begin{pmatrix} x \\ 1 \end{pmatrix} + h^{k-1}\right). \tag{16}$$

We bilinearly interpolate d^{k-1} at x' to obtain $d'(x')$ and the according space point $d'(x')\binom{x'}{1}$ in camera C'. Its



1. Step: map x to $x' \approx P_C\big(d'(x)R'\begin{pmatrix}x\\1\end{pmatrix} + h'\big)$
2. Step: estimate $\hat{d}(x)\begin{pmatrix}x\\1\end{pmatrix} = (\hat{R})^T\big(d'(x')\begin{pmatrix}x'\\1\end{pmatrix} - \hat{h}\big)$

Figure 4. Prediction $\hat{d}$ of the state variable d. Using $(\hat{R}, \hat{h}) := (R', h')$ and depth estimation d', we can map the coordinate system x of the current camera set up to the previous x'. With this mapping, we approximate $\hat{d}$ using (8) with $(R, h) = (\hat{R}, \hat{h})$.

transition to camera C is given by (8), and we define the depth $d(x)$ as prediction $\hat{d}^k(x)$. Figure 4 illustrates this process. Note that eqn. (16) only is an approximation because we do not know the correct argument $d^k(x)$ as required by eqn. (10), and that

$$\hat{d}^k = \hat{d}^k(x; X^{k-1}) = \hat{d}^k(x; R^{k-1}, h^{k-1}, d^{k-1}) \tag{17}$$

is a function of $X^{k-1} = (C^{k-1}, d^{k-1})$.

We assume that a local variance map σ_d^{k-1} of d^{k-1} in the previous frame is known. In Sect. 3.3 we will detail on how this information is obtained. Prediction errors of the depth map are accounted for by assuming a constant increase σ_d of the local variance, which is transported identical to d^{k-1}, i.e. $(\hat{\sigma}_d^k(x))^2 = (\sigma_d^{k-1}(x'))^2 + \sigma_d^2$. Experiments confirm this assumption, see Fig. 7.

Based on this relationship, we make a Gaussian ansatz as approximate probabilistic model of d^k,

$$p(d^k|X^{k-1}) \propto \exp\big(-f_d(d^k; \hat{d}^k, \hat{\sigma}_d^k)\big). \tag{18}$$

The energy functional f_d includes a prior penalizing the deviation from the prediction $\hat{d}^k$ and a spatial smoothness prior,

$$f_d(d^k; \hat{d}^k, \hat{\sigma}_d^k) = \frac{1}{2}\int_\Omega \left(\frac{d^k(x) - \hat{d}^k(x)}{\hat{\sigma}_d^k(x)}\right)^2 + \|\nabla d^k(x)\|^2 dx. \tag{19}$$

Here, we used continuous notation to facilitate interpretation of the terms. After discretization, $d^k, \hat{d}^k, \hat{\sigma}_d^k \in \mathbb{R}^{|\Omega|}$ are vectors indexed by grid positions $x \in \Omega$, and we re-use the symbol ∇ to denote the matrix $\nabla\colon \mathbb{R}^{|\Omega|} \to \mathbb{R}^{2|\Omega|}$ approximating the gradient mapping. Furthermore, we define the predicted covariance matrix of $\hat{d}^k$ as $\hat{S}_d^k := \mathrm{Diag}(\hat{\sigma}_d^k(x))^2$.

Inserting the discretized functional f_d (19) into (18) and ignoring normalizing constants, we obtain after multiplying out and rearranging terms using some basic matrix algebra,

$$p(d^k|X^{k-1}) \propto \mathcal{N}(d^k; \hat{\mu}_d^k, \hat{\Sigma}_d^k), \qquad \text{with} \tag{20a}$$

$$\hat{\mu}_d^k = \hat{\Sigma}_d^k(\hat{S}_d^k)^{-1}\hat{d}^k,\ \hat{\Sigma}_d^k = \left((\hat{S}_d^k)^{-1} + \nabla^\top\nabla\right)^{-1}. \tag{20b}$$

Notice that the prior $\hat{d}^k$ fixes a single, but arbitrary global scale of d and h that cannot be inferred from monocular sequences.

3.3. State Update

Having observed $Y^k = \hat{u}^k$ in terms of the displacement vector field (13) that depends on the unknown state variables $X^k = (C^k, d^k) = (R^k, h^k, d^k)$, we update the state by estimating X^k as mode of the distribution

$$\begin{aligned} p(X^k|Y^{0:k}) &\propto p(Y^k|X^k)p(X^k|X^{k-1}) \\ &= \mathcal{N}(\hat{u}^k; \mu_u^k, \Sigma_u^k)\, \mathcal{N}_{\mathcal{M}}(C^k; C^{k-1}, \sigma_C^2 I)\, \mathcal{N}(d^k; \hat{\mu}_d^k, \hat{\Sigma}_d^k) \end{aligned}$$

based on eqns. (13), (14) and (20). Accordingly, the objective function $f(d, C) := -\log p(X^k|Y^{0:k})$ decomposes into $f(d, C) = f_u(d, C) + f_C(C) + f_d(d)$ with

$$f_u(d, C) = \frac{1}{2}(\hat{u}^k - \mu_u^k)^\top(\Sigma_u^k)^{-1}(\hat{u}^k - \mu_u^k), \tag{22a}$$

$$f_C(C) = \frac{1}{2\sigma_C^2}\mathrm{dist}_{\mathcal{M}}(C^{k-1}, C^k)^2, \tag{22b}$$

$$f_d(d) = \frac{1}{2}(d^k - \hat{\mu}_d^k)^\top(\hat{\Sigma}_d^k)^{-1}(d^k - \hat{\mu}_d^k). \tag{22c}$$

Note that $\mu_u^k(x) = u(x; R^k, h^k, d^k(x))$ depends nonlinearly on R^k, h^k and d^k.

Our approach to solving

$$\min_{C,d}\ f(d, C), \quad C \in SE_3,\ d \in \mathbb{R}^{|\Omega|}_{\geq 0} \tag{23}$$

consists in alternating update steps for C and d, detailed below, embedded into a multiscale framework.

Camera Motion Parameter. We consider the partial minimization of any functional f specified by (22) with respect to the motion parameters $C = (R, h)$. Adopting the identification (1), the problem reads

$$\min_Q f(Q), \quad Q \in G. \tag{24}$$

Let $Q^{(i)}$, $i = 0, 1, 2, \ldots$ denote the minimizing sequence to be determined. Given $Q^{(i)}$ for some i, we determine $Q^{(i+1)} = \varphi(\bar{t}; Q^{(i)})$ in terms of a step size $\bar{t}$ and an one-parameter flow $\varphi(t; Q^i) \subset G$, with $\varphi(0; Q^i) = Q^{(i)}$ and $\dot{\varphi}(0; Q^{(i)}) \in T_{Q^{(i)}}G$, given by

$$\varphi(t; Q^{(i)}) := Q^{(i)}\mathrm{Exp}(tW^{(i)}), \quad W^{(i)} \in se_3. \tag{25}$$

Step size $\bar{t}$ is computed by line search along $\varphi(t; Q^i)$ (cf. [1]) in order to locally minimize f,

$$\bar{t} \approx \arg\min_{t>0} f\big(\varphi(t; Q^{(i)})\big). \tag{26}$$

Matrix $W^{(i)}$ determining the flow (25) is computed by considering

$$\begin{aligned} \frac{d}{dt} f\big(\varphi(t; Q^{(i)})\big)\Big|_{t=0} &= \big\langle \nabla f\big(\varphi(0; Q^{(i)})\big), \dot{\varphi}(0; Q^{(i)})\big\rangle \\ &= \langle \nabla f(Q^{(i)}), Q^{(i)}W^{(i)}\rangle, \end{aligned} \tag{27}$$

where ∇f denotes the (ordinary) gradient of f with respect to the ambient matrix space $\mathbb{R}^{4\times4}$. The gradient $\nabla_G f$ on the manifold G is obtained from the equation (cf. [1])

$$\langle \nabla_G f(Q^{(i)}), Q^{(i)}W\rangle_G = \langle \nabla f(Q^{(i)}), Q^{(i)}W\rangle, \forall W \in se_3, \tag{28}$$

which is equivalent to $\langle {Q^{(i)}}^\top \nabla_G f(Q^{(i)}), W\rangle_G = \langle {Q^{(i)}}^\top \nabla f(Q^{(i)}), W\rangle$ for *arbitrary* $W \in se_3$. Hence,

$$\nabla_G f(Q^{(i)}) = (Q^{(i)})^{-\top} \begin{pmatrix} \hat{a} & \frac{1}{c_G} b \\ 0^\top & 0 \end{pmatrix}, \tag{29}$$

$$\begin{pmatrix} \hat{a} & b \\ 0^\top & 0 \end{pmatrix} = \Pi_{se_3}\big({Q^{(i)}}^\top \nabla f(Q^{(i)})\big), \tag{30}$$

where Π_{se_3} denotes the orthogonal projection onto the linear subspace (2). Choosing $W^{(i)} = -\begin{pmatrix} \hat{a} & \frac{1}{c_G} b \\ 0^\top & 0 \end{pmatrix}$ in (27) shows due to the equalities in (28) and (29),

$$\frac{d}{dt} f\big(\varphi(t; Q^{(i)})\big)\Big|_{t=0} = -\left\| \begin{pmatrix} \hat{a} & \frac{1}{c_G} b \\ 0^\top & 0 \end{pmatrix} \right\|_G^2 \leq 0, \tag{31}$$

i.e. that the flow $\varphi(t; Q^{(i)})$ given by (25) decreases f.

Depth Map. A descent direction of $f(d) = f(C, d)$ in (23) is given by

$$\delta_d^{(i)} := -\big(B^{(i)}\big)^{-1} \nabla_d f(d^{(i)}) \tag{32}$$

with some positive definite, symmetric perturbation matrix $B^{(i)}$. Then for any $\|\nabla_d f\|_2 > 0$, there is a $t > 0$, such that $f(d^{(i)} + t\delta_d^{(i)}) < f(d^{(i)})$, see e.g. [1].

To ensure the element-wise non-negativity of $d^{(i)}$, we define a quality function $q(d) := f(d) + \chi_{\geq 0}(d)$, where $\chi_{\geq 0}(d)$ is the characteristic function of the set $\mathbb{R}^{|\Omega|}_{\geq 0}$, and utilize a line search method to determine a near-optimal $\bar{t}$ such that $\bar{t} \approx \arg\min_{t>0} q(d^{(i)} + t\delta_d^{(i)})$.

The choice of $B^{(i)}$ is crucial. Gradient descent, i.e. $B^{(i)} = I$, turned out to be unsatisfactorily slow due to the high dimension of d and the spatial variable interactions due to (20). Positive definiteness of the Hessian $\mathrm{H}f$ of the objective is not guaranteed due the non-convexity of the projections (10) in the observation term (13). Thus, a standard Newton method ($B^{(i)} = \mathrm{H}f$) might diverge. Therefore, we propose the choice (with $\epsilon > 0$)

$$B^{(i)} := \gamma(\mathrm{H}f_u(d^{(i)}, C^{(i)})) + \mathrm{H}f_d(d^{(i)}) + \epsilon I, \tag{33}$$

where $\gamma(S)$ adjusts the eigenvalues of the matrix S by replacing them by their absolute values to guarantee positive semi-definiteness. Due to the form of (20) and the third term in (33), the positive definiteness of $B^{(i)}$ is guaranteed. Figure 5 demonstrates the fast convergence of our method within only few iterations under real conditions. Along with the results in Sect. 4, this justifies the choice (33).

Note that due to (11), searching along d parameterizes a search along the epipolar lines defined by fixed R, h.

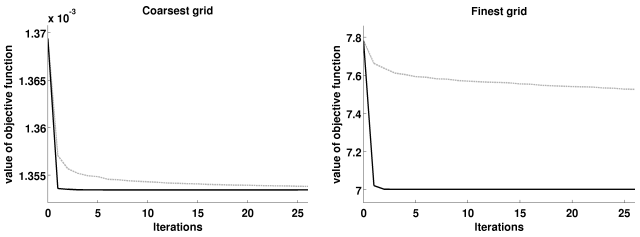


Figure 5. Representative performance of the minimization approach (32), (33) in a real scenario. The objective function is effectively minimized after few iterations at each resolution level (solid curve), as opposed to gradient descent that converges slowly (dashed curve).

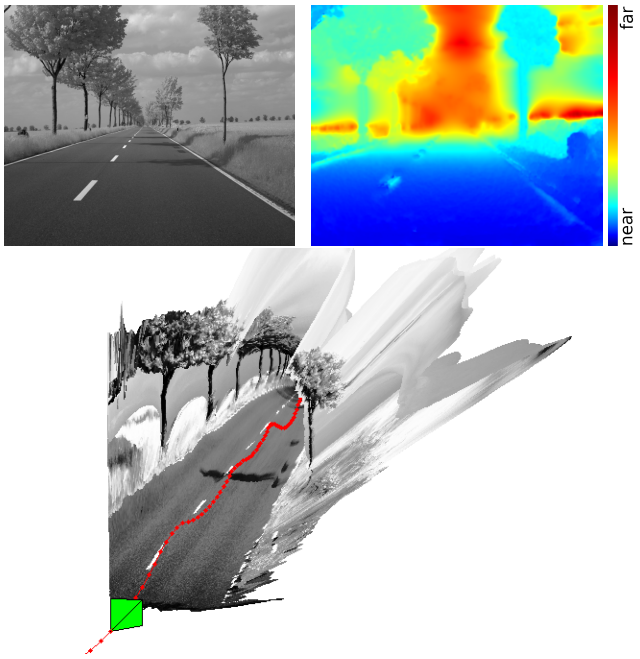


Figure 6. **Best viewed in color. Left:** input frame, **right:** calculated depth map, **bottom:** triangulation of the scene, camera position and direction (green camera symbol), camera track (red, dots indicate subsequent positions). Due to the moving trees, *homogeneous* sky regions *between* trees are naturally assigned to the trees (in terms of depth) in the temporal context, rather than to the horizon and in contrast to the center region.

Local Depth Variance. We approximate the local variance σ_d^k of the depth map by the second derivatives of f in (C^k, d^k), bounded to ≥ 0,

$$(\sigma_d^k(x))^2 = \max\{0, \frac{\partial^2}{\partial d(x)^2} f_u(C^k, d^k)\} + \frac{\partial^2}{\partial d(x)^2} f_d(d^k)\ .$$

Along with the computed d^k, this information is incorporated as prior (cf. Sect. 3.2) in the estimation of d^{k+1} to identify regions with high uncertainty caused by the parallax and/or the lack of image features.

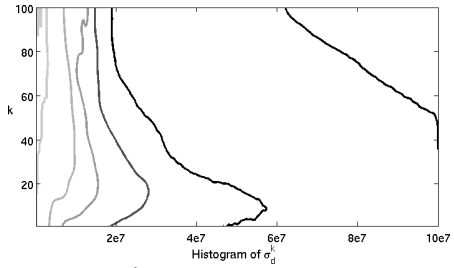


Figure 7. Histogram $\widehat{\mathrm{pdf}}_\Omega(\sigma_d^k)$ of depth variance σ_d^k for increasing k, displayed as contour plot. The lines tend to the left and thus indicate that estimation uncertainty is effectively reduced, despite online processing with only two frames at each instant of time.

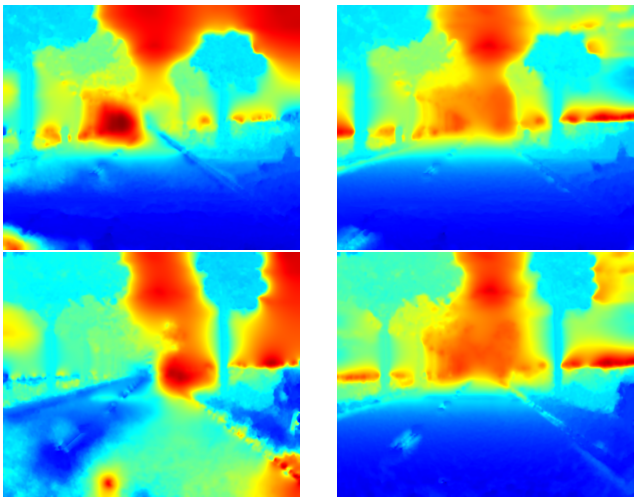


Figure 8. **Best viewed in color.** Importance of regularization over time (prediction prior) demonstrated for two different scenes, **top and bottom row**; **left** without prior, **right** with prior. Exploiting temporal context through the prior is essential, in particular around the epipole. Omitting the prior leads to corrupted estimates.

4. Experiments

4.1. Data Sets and Performance Measures

Data Sets. We evaluated the algorithm proposed in Sect. 3 by means of a novel database [13] of image sequences recorded from of a car driving at high speed in an everyday environment. This database was compiled for the evaluation of stereo reconstruction algorithms and is available online. Thus, two camera recordings from a stereo rig are available, one of which only was used as input to our algorithm. Each image sequence consists of up to 400 gray-value frames of size 656×541 pix., taken at 25 Hz.

Below, we refer to the *'Bend'* (Fig. 1) and to the *'Avenue'* sequence (Fig. 6). They are available online[3] along with further sequences and results compiled as videos.

Performance Measures. As a baseline for *depth estimation*, we applied two different stereo reconstruction methods to our dataset, see Sect. 4.3. Concerning *egomotion*, we applied the 'Voodoo Camera Tracker' (VCT) that, unlike our approach, works in batch mode and therefore can be expected to return more precise results, see Sect. 4.4.

4.2. Results

Depth Maps and Egomotion. Figures 1 and 6 show representative reconstructions. Comments are given in the captions. Further results are available online[3] as videos.

Uncertainty Reduction. Figure 7 depicts several fixed levels of the histogram $\widehat{\mathrm{pdf}}_\Omega(\sigma_d^k)$ of the variance σ_d^k of depth d^k, taken over the whole image plane, as a function of the frame index k (ordinate). The resulting lines tend to the left and thus demonstrate that our approach significantly reduces uncertainty of the depth estimation within a period of about 50 frames.

Temporal Filtering. The prediction step exploited by $(d^k - \hat{d}^k)^2$ in (19) is essential for robust depth estimation. Figure 8 shows for two different scenes depth maps estimated without and with prior in the left and right column, respectively. The differences are striking and show that just relying on the observations yields corrupted estimates.

4.3. Comparison to Stereo

As a baseline for depth estimation, we used an implementation of [20] included in the *Middlebury MRF Library*[4] (parameters: Birchfield-Tomasi method with α-β-swap, 80 disparities, L_1-regularization parameter 0.5), and alternatively the LIBELAS library, see [6] (default param.).

Figure 9 (and further results[3]) reveals that our *monocular* approach provides a competitive reconstruction.

4.4. Comparison to Voodoo Camera Tracker (VCT)

To evaluate egomotion estimation, we applied the VCT (parameters: free move, bundle adjustment with previous images, fixed internal parameters), that estimates camera parameters using bundle adjustment and tracked features. Figure 10 demonstrates for the *Bend* and *Avenue* sequences a remarkable agreement of our *monocular online* estimates.

5. Conclusion and Further Work

We presented an approach to the estimation of dense scene structure and camera motion from monocular image sequences, taken from a camera positioned inside a fast moving car. The approach optimizes the tradeoff between model expressiveness and computational efficiency. In particular, it works in an online two-frame mode and competes well with less desirable settings (stereo, bundle adjustment), as demonstrated by a comprehensive evaluation using real data in different scenarios.

[3] http://hci.iwr.uni-heidelberg.de/VSFM

[4] http://vision.middlebury.edu/MRF/code/

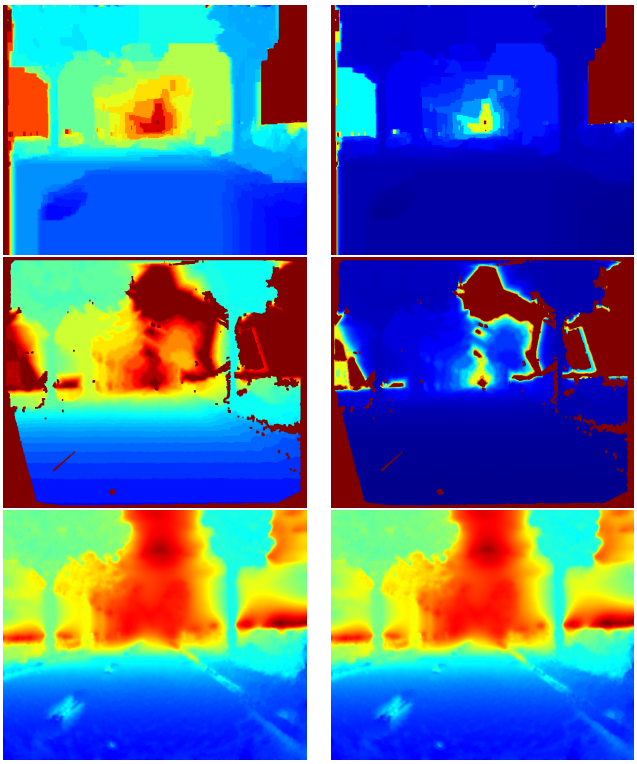


Figure 9. **Best viewed in color.** Comparison of two *stereo* implementations ([20]: **top row**, [6]: **center row**) to our *monocular* approach (**bottom row**). **Left column:** *individually rescaled* color maps for best depth visualization. **Right column:** *same* color maps for all three approaches. Stereo approaches suffer from low resolution (discrete displacements), resulting in erroneous partitions (e.g. trees and sky) and stair-casing effects (road) that may cause problems at subsequent processing stages (scene analysis).

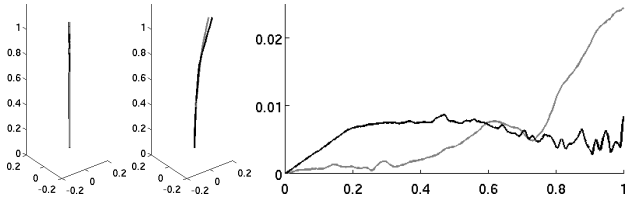


Figure 10. **Left&middle:** camera trajectories for *Avenue* and *Bend* sequences, resp., returned by VCT (gray) and our approach (black). The tracks are uniformly scaled to the overall camera movement (first to last frame). No rotational fitting was applied. **Right:** Euclidean distances between the trajectories, relative to the trajectory length. Differences are upper-bounded by 3%.

Our further work will focus on occlusion handling and reliable segmentation of independently moving objects, and related mid-level tasks of traffic scene analysis.

Acknowledgements

HCI is supported by the DFG, Heidelberg University and industrial partners. Authors thank Dr. W. Niehsen, Robert Bosch GmbH.

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. PUP, 2008. 3, 5, 6

[2] A. Bain and D. Crisan. *Fundamentals of Stochastic Filtering*. Springer, 2009. 3

[3] S. Baker and I. Matthews. Lucas-Kanade 20 Years On: A Unifying Framework. *IJCV*, 56(3), 2004. 4

[4] M. Enzweiler and D. Gavrila. Monocular Pedestrian Detection: Survey and Experiments. *PAMI*, 31(12), 2009. 1

[5] D. Fleet and Y. Weiss. *Optical Flow Estimation*, pages 239–257. Springer, 2006. 3

[6] A. Geiger, M. Roser, and R. Urtasun. Efficient Large-Scale Stereo Matching. In *Asian Conference on Computer Vision*, Queenstown, New Zealand, November 2010. 2, 7, 8

[7] D. Gerónimo, A. López, A. Sappa, and T. Graf. Survey of Pedestrian Detection for Advanced Driver Assistance Systems. *PAMI*, 32(7), 2010. 1

[8] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge Univ. Press, 2000. 2, 3, 4

[9] U. Helmke, K. Huper, P. Lee, and J. Moore. Essential Matrix Estimation Using Gauss-Newton Iterations on a Manifold. *IJCV*, 74(2), 2007. 3

[10] M. Irani, P. Anandan, and M. Cohen. Direct Recovery of Planar-parallax from Mult. Frames. *TPAMI*, 24(11), 2002. 2

[11] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. An Introduction to Variational Methods for Graphical Models. *Mach. Learning*, 37, 1999. 3

[12] K. Konolige and M. Agrawal. FrameSLAM: From Bundle Adjustment to Real-Time Visual Mapping. *IEEE Trans. Robotics*, 24(5):1066–1077, 2008. 2

[13] S. Meister, B. Jähne, and D. Kondermann. An Outdoor Stereo Camera System for the Generation of Real-World Benchmark Datasets with Ground Truth. Technical report, HCI, IWR, Heidelberg University, 2011. 2, 7

[14] E. Mouragnona, M. Lhuilliera, M. Dhomea, F. Dekeyserb, and P. Sayd. Generic and Real-Time Structure from Motion Using Local Bundle Adjustment. *Image Vis. Comp.*, 27(8):1178–1193, 2009. 2

[15] R. A. Newcombe and A. J. Davison. Live Dense Reconstruction with a Single Moving Camera. In *CVPR*, 2010. 1, 2

[16] X. Pennec. Intrinsic Statistics on Riemannian Manifolds: Basic Tools for Geom. Measurements. *JMIV*, 25, 2006. 3, 4

[17] A. Saxena, S. H. Chung, and A. Y. Ng. 3-D Depth Reconstruction from a Single Still Image. *IJCV*, 76, 2008. 2

[18] J. Stühmer, S. Gumhold, and D. Cremers. Parallel Generalized Thresholding Scheme for Live Dense Geometry from a Handheld Camera. In *CVGPU*, 2010. 2

[19] P. Sturm and B. Triggs. A Factorization Based Algorithm for Multi-Image Projective Structure and Motion. In *ECCV*, Cambridge, England, 1996. Springer. 2

[20] R. Szeliski, R. Zabih, and D. Scharstein et al. A Comparative Study of Energy Minimization Methods for Markov Random Fields with Smoothness-Based Priors. *TPAMI*, 2008. 2, 7, 8

[21] B. Triggs, P. F. Mclauchlan, R. I. Hartley, and A. W. Fitzgibbon. Bundle Adjustment – A Modern Synthesis. *LNCS*, 1883, 2000. 2

[22] A. Wedel, C. Rabe, and T. Vaudrey et al. Efficient Dense Scene Flow from Sparse or Dense Stereo Data. In *ECCV*, volume 3021 of *LNCS*, 2008. 1, 2

[23] A. Weishaupt, L. Bagnato, and P. Vandergheynst. Fast Structure from Motion for Planar Image Sequences. In *EUSIPCO*, 2010. 1, 2

[24] C. Wojek, S. Roth, K. Schindler, and B. Schiele. Monocular 3D Scene Modeling and Inference: Understanding Multi-Object Traffic Scenes. In *ECCV*, 2010. 1