IJCV manuscript No.
(will be inserted by the editor)

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

Florian Becker · Frank Lenzen · Jörg H. Kappes · Christoph Schnörr

Received: date / Accepted: date

Abstract We present an approach to jointly estimating camera motion and dense structure of a static scene in terms of depth maps from monocular image sequences in driver-assistance scenarios. At each instant of time, only two consecutive frames are processed as input data of a joint estimator that fully exploits second-order information of the corresponding optimization problem and effectively copes with the non-convexity due to both the imaging geometry and the manifold of motion parameters. Additionally, carefully designed Gaussian approximations enable probabilistic inference based on locally varying confidence and globally varying sensitivity due to the epipolar geometry, with respect to the high-dimensional depth map estimation. Embedding the resulting joint estimator in an online recursive framework achieves a pronounced spatio-temporal filtering effect and robustness.

We evaluate hundreds of images taken from a car moving at speed up to 100 km/h and being part of a publicly available benchmark data set. The results compare favorably with two alternative settings: stereo based scene reconstruction and camera motion estimation in batch mode using multiple frames. They, however, require a calibrated camera pair or storage for more than two frames, which is less attractive from a technical viewpoint than the proposed monocular and recursive approach. In addition to real data, a synthetic sequence is considered which provides reliable ground truth.

Keywords structure from motion · variational approach · recursive formulation · dense depth map

F. Becker
Heidelberg Collaboratory for Image Processing, University of Heidelberg, Speyerer Str. 6, 69115 Heidelberg, Germany
E-mail: becker@math.uni-heidelberg.de
http://hci.iwr.uni-heidelberg.de

F. Lenzen
Heidelberg Collaboratory for Image Processing, University of Heidelberg, Speyerer Str. 6, 69115 Heidelberg, Germany
E-mail: frank.lenzen@iwr.uni-heidelberg.de
http://hci.iwr.uni-heidelberg.de

J. H. Kappes
Image and Pattern Analysis Group, University of Heidelberg, Speyerer Str. 6, 69115 Heidelberg, Germany
E-mail: kappes@math.uni-heidelberg.de
http://ipa.iwr.uni-heidelberg.de

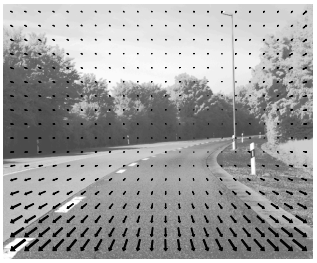
C. Schnörr
Image and Pattern Analysis Group, University of Heidelberg, Speyerer Str. 6, 69115 Heidelberg, Germany
E-mail: schnoerr@math.uni-heidelberg.de
http://ipa.iwr.uni-heidelberg.de

1 Introduction

1.1 Overview and Motivation

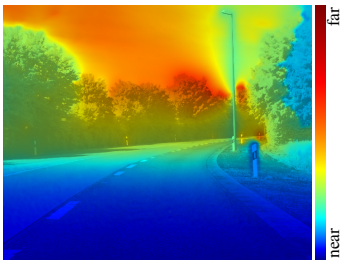
Computer vision research has a strong impact on driver assistance technology. Besides designing dedicated detectors for specific object classes (Enzweiler and Gavrila 2009; Gerónimo et al 2010), current major trends include low-level estimation of dense scene structure from stereo sequences (Wedel et al 2008), the transition to monocular imaging sensors (Weishaupt et al 2010; Newcombe and Davison 2010), and context-based 3D scene representation and labeling supported by high-level assumptions and constraints (Wojek et al 2010).

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:

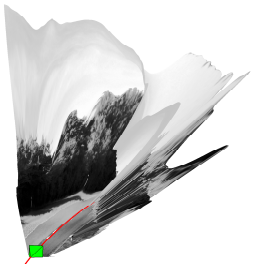


(a) frame 99 and optical flow

(b) frame 100



(c) frame and dense depth map superimposed



(d) reconstructed scene and camera track

Fig. 1 **(a)–(b)** Two consecutive frames of the *Bend* sequence (see Sect. 5.1) with large displacements up to 35 pixels induced by a fast moving camera. Our approach jointly estimates, from sparse noisy displacement measurements, **(c)** *dense depth maps* and **(d)** *camera motion* in an online recursive framework. Reconstruction of dense scene structure based on the depth maps from the camera's viewpoint (green), and the corresponding camera track (red).

- 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 (calibration);
- Ability to cope with large displacements induced by a fast moving camera;
- Thoroughly annotation of all results by confidence estimates;
- 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* variational models for estimating *high-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.

1.2 Scope

The task of deriving scene structure and camera movement from a monocular camera is – compared to a stereo setup – a much more challenging one (see discussion in Sect. 3.1.2). However, the considered application in an automotive scenario in particular requires very reliable information. Thus, a sound theoretical basis is as important as real-time capabilities. In this work we therefore focus on the *theoretical justification* of the proposed method. Probabilistic formulation helps to interpret handling of inaccurate data (Sect. 3) and the variational point of view provides insights and guarantees concerning convergence of the numerics (Sect. 4).

In favor of a comprehensive discussion of the approach, we restrict ourselves to *static* scenes. We consider this a reasonable intermediate step towards the even more challenging case of dynamic scenes being subject of future work.

Furthermore, for now we refrain from refining our *research implementation* towards real-time speed, although the approach is designed with respect to real-time-applications due to its recursive nature and restriction to two frames, see Sect. 5.2.1.

The *unknown global scale* of the scene and camera translation can not be measured from observed motion in this setup. Although important for real applications, we here do not investigate approaches to reconstruct this scalar e.g. from known dimensions of objects. However, the probabilistic framework easily allows to integrate even weak and sporadic sensor data (e.g. the scalar velocity) that are available in practice, see Sect. 3.1.3.

In general, the design resorts to few *established and well understood components* to reduce the number of factors which influence the quality of the results. The experimental section aims to demonstrate that the proposed concept is correct and the results provide a baseline for more enhanced implementations tuned for more specific applications.

1.3 Related Work

Most approaches to scene reconstruction rely on stereo imaging or multiple view reconstructions in batch mode. In the automotive context stereo set-ups dominate for estimating the scene structure (Yamaguchi et al 2012; Geiger et al 2010; Hirschmüller 2008). They directly extend to measure optical flow to derive scene flow information (Rabe et al 2010; Wedel et al 2008). However, they are only relevant for sensing close-up ranges at low speeds, due to the small baseline in driver assistance scenarios, require extensive calibration and are less attractive than just a single camera from the technological system oriented viewpoint.

Factorization (Sturm and Triggs 1996) and bundle adjustment (Triggs et al 2000) 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 (Mouragnona et al 2009; Konolige and Agrawal 2008), entail only sparse representations of the scene structure. A more recent approach (Lin et al 2011) jointly estimates camera pose and a point cloud based on two frames only. For improved robustness, it makes use of edge features and imposes smoothness constraints on the observed motion.

Dense methods for egomotion estimation from monocular (Newcombe et al 2011; Sheikh et al 2007) or stereo data (Comport et al 2007; Valgaerts et al 2010) make use of all image pixels and thus potentially provide higher accuracy than feature-based approaches, e.g. PTAM (Klein and Murray 2007) or (Nister et al 2004), see Valgaerts et al (2012) for an excellent comparison.

Work on the reconstruction of accurate *dense* depth maps from arbitrary multiple views includes Wendel et al (2012); Graber et al (2011); Newcombe and Davison (2010); Stühmer et al (2010). These works, however, require the camera motion to be determined in a preceding step using feature tracking, e.g. PTAM (Klein and Murray 2007). Other related approaches only allow for camera translation but no rotation (Weishaupt et al 2010), or estimate the epipole but require images to be aligned with respect to a common reference plane (Irani et al 2002). Instead of relying on tracked features, the recent approach by Newcombe et al (2011) determines the camera pose by matching the recorded image frame with the dense reconstructed scene model. Bagnato et al (2011) jointly estimate a depth map and the camera motion from two images of an omnidirectional image sequence.

An attractive alternative employs direct feature-to-depth mappings, learned offline from ground truth databases (Saxena et al 2008; Liu et al 2010). 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. Hadsell et al (2009) propose to train *online* based on stereo depth estimations with the aim to improve the depth estimation in future frames. More complex line segment features are sufficient to derive indoor scene geometry from a single frame as show by Lee et al (2009).

Mester (2011) postulated a number of requirements for monocular reconstruction approaches, including annotation of motion measurements by covariance matrices gained by analyzing the structure tensor, using dense image registration instead of feature matching and carefully exploiting temporal consistency, which includes incorporating reliability measures.

The previous work in Becker et al (2011) fulfills the aforementioned demands. Here we further refine the mathematical foundation and improve the inference step considerably by introducing a *fully joint* and *second-order* update scheme which now *guarantees* a decrease of the objective function. Depth map and camera motion are now both *completely annotated* with confidence information.

We complement the theoretical part by discussing the achievable depth map accuracy and pointing out the relation to variational optical flow approaches and to epipolar geometry. The experimental verification is based on a considerably larger variety of image data. Comparison to results from stereo, bundle adjustment and synthetic ground truth were extended and refined.

1.4 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.

Section 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. Section 4 covers the numerical details of the inference step.

We report in Sect. 5 results of an evaluation of our approach using more than 2700 real images provided by a novel publicly available database, that aims at providing a benchmark for computer vision algorithms in the context of automotive applications.

Moreover, we show that our approach compares favorably to results computed with less restricted approaches. Using public implementations of stereo depth estimation (Rhemann et al 2011; Szeliski et al 2008; Geiger et al 2010) and the Voodoo Camera Tracker[1] (VCT) ensure reproducibility of all results.

The experiments with real image sequences are complemented by results for a synthetic image sequences which features reliable ground truth.

1.5 Notation

We provide an overview of notations used in this work.

$\mathbb{R}$, $\mathbb{R}_+$	set of real numbers, non-negative real numbers
$\mathcal{M} = \mathcal{M}^d \times \mathcal{M}^C$, $\mathrm{G} = \mathrm{G}^d \times \mathrm{G}^C$	manifold $\mathcal{M}$, Lie group G jointly representing depth map and camera motion
SE_n, se_n	n-dimensional special Euclidean group, associated Lie algebra
SO_n, so_n	n-dimensional orthogonal group, associated Lie algebra
$\mathcal{L}$	basis vector of a Lie algebra
$T_X\mathrm{G}$	tangential space of Lie group G at $X \in \mathrm{G}$
$\langle \cdot, \cdot \rangle$	inner product
$\| \cdot \|$	Euclidean norm
Id	identity operator
∇, $\overline{\nabla}$, H	gradient, affine connection, Hessian
exp, log	scalar-valued exponential and logarithm
Exp, Log	matrix-valued exponential and logarithm
$\mathrm{G}_\rho(x)$	Gaussian filter mask with variance ρ
$\mathrm{p}(x)$, $\mathrm{p}(x\|y)$	probability, conditional probability
$\mathcal{N}(x; \mu, \Sigma)$, $\mathcal{N}_{\mathcal{M}}(x; \mu, \Sigma)$	normal distribution on the real vector space, on the manifold $\mathcal{M}$
$\succ$, $\succeq$	matrix relation for positive (semi-)definiteness
$[\cdot]_i$	i-th component of a compound vector expression
$[x]_\times$	3×3-matrix representing the cross product with $x \in \mathbb{R}^3$
X^k	state in iteration k of the recursive update loop
$X^{(i)}$	state in iteration i of the inference step optimization

[1] http://www.digilab.uni-hannover.de/docs/manual.html, v1.2.0b

2 Problem Statement, Approach (Overview)

2.1 Preliminaries

We adopt the common concepts of multiple view geometry (Hartley and Zisserman 2000). 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 $\mathcal{M}^C := \mathsf{SE}_3$ of Euclidean transformations $C = (R, h) \in \mathcal{M}^C$, parametrized by rotations R and translations h, is identified with the matrix Lie group

$$\mathrm{G}^C := \left\{ Q = \begin{pmatrix} R & h \\ 0^\top & 1 \end{pmatrix} : R \in \mathsf{SO}_3,\ h \in \mathbb{R}^3 \right\}, \tag{1}$$

where SO_3 denotes the group of proper rotations.

For any $x \in \mathbb{R}^3$ the 3×3-matrix

$$[x]_\times := \begin{pmatrix} 0 & -x_3 & x_2 \\ x_3 & 0 & -x_1 \\ -x_2 & x_1 & 0 \end{pmatrix}, \tag{2}$$

represents the cross product linear operator such that $[x]_\times y = x \times y$ for all $y \in \mathbb{R}^3$.

2.2 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 $I^l \colon \Omega \to \mathbb{R}$, measured at times $l \in \{0, \ldots, k\}$ with cameras $C^{0:k}$. From the induced projected motion u^k (optical flow) we wish to *jointly estimate in a recursive manner* both $C^{1:k}$ and a sequence $d^{1:k}$ of *depth maps* $d^l \colon \Omega \to \mathbb{R}_+$ that assign to each image point $x \in \Omega$ its depth $d^l(x)$ 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 (px) (with frame size 656 px × 541 px), and

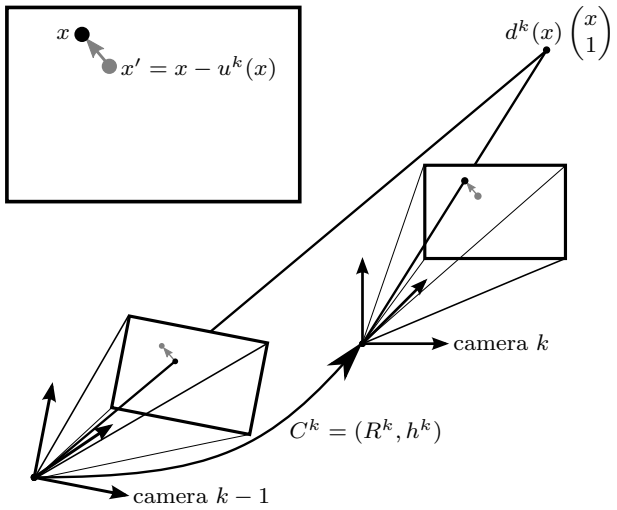


Fig. 2 A scene point is defined by image coordinates x and depth $d^k(x)$ in the coordinate system of camera k. Its projection moves by $-u^k(x)$ to x' when the camera is rotated and translated *backward* by $C^k = (R^k, h^k)$.

(iii) a recursive online processing mode that updates the camera parameters and depth map based on two consecutive frames only.

2.3 Design Decisions: Overview

Major design decisions are discussed only in this section while further details are addressed in the following sections.

The considered scenario consists in solving the challenging inverse problem to determine the parameters (depth, camera motion) which explain the observed optical flow. Subsequent image frames contain much inferior depth information than a stereo image pair as we will discuss in Sect. 3.1.2. Thus, it is essential to handle weak information correctly instead of discarding it, which further motivates the *probabilistic formulation* utilized in Sect. 3. All results are annotated with an *accuracy estimate* to appropriately consider them in future frames and higher-level reasoning steps.

The decision for a *recursive formulation* is a consequence of the requirement for both robustness and real-time capability of the approach. In contrast to batch methods, only the minimum number (i.e. two) of image frames for structure-from motion need to be stored and considered in computation. Nevertheless, the *temporal smoothness* component accumulates information from all previous frames in order to handle the reduced information due to the unfavorable camera setup.

Due to the high dimensionality of the variables we employ *Gaussian approximations* instead of less compact descriptions such as particle filters. On the other hand, accurate representation of the projective model and the manifold structure of the variables is required and rules out linear approximations and thus a classical Kalman filter design.

Instead of reconstructing w.r.t. a particular fixed camera view for a number of frames as e.g. in Graber et al (2011); Newcombe and Davison (2010); Stühmer et al (2010); Newcombe et al (2011), we always define the *latest camera position as reference frame*. This is the most relevant of all observed frames for comparing to future frames and the most natural choice for the automotive context.

Dense formulation allows to incorporate motion information for all pixels in the images to improve robustness and accuracy of both scene structure and camera pose, also confirmed by recent work (Valgaerts et al 2012; Newcombe et al 2011; Comport et al 2007). The regular grid structure is well suited for parallel processing e.g. on (embedded) GPU hardware.

For robustness, observations need to be explained accurately and deviations from the model need to be balanced between the error sources. As it is not clear how decoupled methods can accomplish this, *joint estimation* of depth and egomotion is the consequent choice followed here.

Dependency of the results on the accuracy of *external sensors* such as acceleration sensors is avoided by explicitly *not assuming their presence*. However, the probabilistic formulation readily allows to incorporate also additional noisy measurements, e.g. to determine the unknown global scale, see Sect. 3.1.3.

The numerical scheme for the inference step is based on *established mathematical tools* which provide guarantees on correctness and convergence. Details are discussed in Sect. 4.

2.4 Approach: Overview

The natural approach to the considered problem is to consider sequences of state variables $X^{0:k} = (d^{0:k}, C^{0:k})$ and observations $Y^{1:k} = u^{1:k}$, together with probabilistic models of state transitions $\mathrm{p}(X^k|X^{k-1})$ and the observation process $\mathrm{p}(Y^k|X^k)$ under Markovian assumptions, in order to recursively estimate X^k based on the posterior marginal distribution $\mathrm{p}(X^k|Y^{1:k})$ (cf., e.g. Bain and Crisan (2009)).

Approximations to this general approach are inevitable, however, due to the non-linearity 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. Jordan et al (1999)).

Accordingly, as detailed in Sect. 3, we devise Gaussian approximations $\mathrm{p}(Y^k|X^k) = \mathcal{N}(u^k; \mu_u^k, \Sigma_u^k)$ and $\mathcal{N}(d^k; \hat{\mu}_d^k, \hat{\Sigma}_d^k)$ for the high-dimensional observed motion $Y^k = u^k$ 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 terms of predictions $\hat{\mu}_d^k$, $\hat{\Sigma}_d^k$.

The prior $\mathcal{N}(d^k; \hat{\mu}_d^k, \hat{\Sigma}_d^k)$ is complemented by a *local* Gaussian model $\mathcal{N}_{\mathcal{M}^C}(C^k; C^{k-1}, \hat{\Sigma}_C^k)$ of the motion parameters on the tangent space of the Euclidean manifold $\mathcal{M}^C = \mathsf{SE}_3$ at C^{k-1} (cf. Pennec (2006)), to form an approximation of the state transition process $\mathrm{p}(X^k|X^{k-1})$ from frame $k-1$ to frame k, where $X^k = (d^k, C^k) = (d^k, R^k, h^k)$.

Putting all components together, we define and compute our update as mode of the posterior marginal approximation

$$\mathrm{p}(X^k|Y^{1:k}) \propto \mathrm{p}(Y^k|X^k)\mathrm{p}(X^k|X^{k-1}).$$

Concerning the motion parameters C^k, we prefer working directly on $\mathcal{M}^C$ using established concepts of numerics (Absil et al 2008), rather than to represent the two-view geometry by the essential matrix and to recover C by additional factorization (Helmke et al 2007).

3 Approach: Details

Our approach *jointly* estimates egomotion C and a *dense* depth map d 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 (Fleet and Weiss 2006). Uncertainty of observations and depth estimates are handled by probabilistic models.

3.1 Observation Process

We detail the observation process $\mathrm{p}(Y^k|X^k)$, with observed feature motion Y^k, state variables $X^k = (d^k, C^k)$ (camera motion, 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 parameter matrix $K \in \mathbb{R}^{3\times3}$, we undo the corresponding affine transformation of the image plane (cf. Hartley and Zisserman (2000)) 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 coordinate system of camera C by P_C,

$$P_C(X_1, X_2, X_3) := \frac{1}{X_3}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}. \tag{3}$$

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

$$d'(x')\begin{pmatrix} x' \\ 1 \end{pmatrix} \;\to\; d(x)\begin{pmatrix} x \\ 1 \end{pmatrix} = R^\top\Big(d'(x')\begin{pmatrix} x' \\ 1 \end{pmatrix} - h\Big). \tag{4}$$

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

$$x' = x - u(x). \tag{5}$$

Using eqns. (3) and (4) we obtain

$$u(x; R, h, d) = x - P_C\left(d(x)R\begin{pmatrix} x \\ 1 \end{pmatrix} + h\right) \tag{6}$$

which represents the model for optical flow between two consecutive frames.

Observations Y correspond to *estimates* $\hat{u}(x)$ of the displacements (6) for all $x \in \Omega$, accompanied by an (possibly anisotropic) accuracy estimate represented by a positive definite 2×2-matrix $\Sigma_u(x)$. Although any method which provides $(\hat{u}(x), \Sigma_u(x))$ is conceivable, we here resort to the well studied method by Lucas and Kanade (see Baker and Matthews (2004)),

$$\hat{u}(x) := -\Sigma_u(x)\Big(\mathrm{G}_\rho(x) * \Big(\big(\partial_t I(x)\big)\big(\nabla I(x)\big)\Big)\Big), \tag{7a}$$

$$\Sigma_u(x) := \Big(\mathrm{G}_\rho(x) * \Big(\big(\nabla I(x)\big)\big(\nabla I(x)\big)^\top\Big)\Big)^{-1}. \tag{7b}$$

Here, $\mathrm{G}_\rho(x)*$ denotes element-wise Gaussian convolution of the subsequent matrix comprising partial derivatives $\partial_t I$, $\nabla I := \binom{\partial_{x_1} I}{\partial_{x_2} I}$ of the image sequence function $I(x,t)$. They are estimated by 3×3 binomial filters

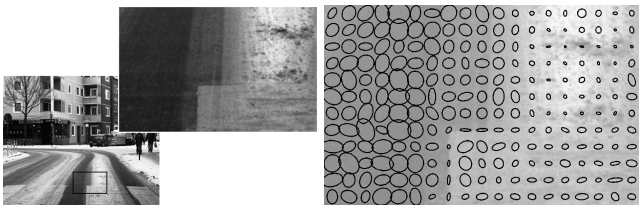


Fig. 3 Detailed view of an image frame and an ellipse representation of the estimated flow uncertainty $\Sigma_u^k(x)$. Highly textured regions (upper right) can be correctly distinguished from locations with low confidence due to low signal-to-noise ratio (left) and image edges (aperture problem; middle).

and first-order differences derived by linearizations at time k. Likewise, we choose a rather small smoothing kernel of size $\rho = 2\,\text{px}$, leading to a fast processing stage. We point out that stronger local regularization (smoothing) is not necessary as the embedding multi-scale framework and the state prediction (see Sect. 3.2) ensure *small* incremental displacements $u(x)$.

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

$$\mathrm{p}(Y^k|X^k) = \mathcal{N}(\hat{u}^k; \mu_u^k, \Sigma_u^k), \tag{8}$$

where μ_u^k is composed position-wise of $u(x; R, h, d)$, i.e.

$$\mu_u^k(x) := u(x; R^k, h^k, d^k(x)) \tag{9}$$

due to eqn. (6), and Σ_u^k is a block-diagonal covariance matrix with component matrices (7b). Note that the definition of μ_u^k makes explicit the conditioning on the state parameters $X^k = (d^k, C^k) = (d^k, R^k, h^k)$.

Model (8) only approximates the true unknown observation process (7a). Uncertainty of observations u^k is modeled by Σ_u^k and internally represented by the precision matrices $(\Sigma_u^k)^{-1}$ (cf. eqn. (7b)). Hence, motion estimation is acquired at *every* image position, but weighted according to its information content (see Fig. 3): homogeneous image regions are represented as rank-0 matrices. The reduced velocity information provided at image edges (aperture problem) is marked by rank-1-matrices and thus can be correctly accounted for within the overall recursive estimation framework – see Sect. 3.3.

3.1.1 Connection to Epipolar Geometry

The motion model (6) can be reformulated to emphasis the influence of the depth:

$$\begin{aligned} u(x; R, h, d) =& (1 - \gamma(d)) u_0(x, R, h) && \text{(10a)} \\ &+ \gamma(d) u_\infty(x, R, h) && \text{(10b)} \end{aligned}$$

with asymptotic optical flow

$$u_0(x, R, h) := \lim_{d \to 0} u(x; R, h, d) = x - P_C(h)\,,$$
$$u_\infty(x, R, h) := \lim_{d \to \infty} u(x; R, h, d) = x - P_C(R \begin{pmatrix} x \\ 1 \end{pmatrix}),$$

and weight

$$\gamma(d) := \frac{[dR \begin{pmatrix} x \\ 1 \end{pmatrix}]_3}{[dR \begin{pmatrix} x \\ 1 \end{pmatrix} + h]_3}.$$

Under the mild assumption that the camera moves forward ($h_3 > 0$) and rotation is moderate ($[R \begin{pmatrix} x \\ 1 \end{pmatrix}]_3 \geq 0$) between two successive frames, we have $\gamma(d) \in [0, 1]$ for $d \in \mathbb{R}_+$ and u is a convex combination of the extreme values u_0 and u_∞. In particular, the dependency of u on d vanishes in the epipole point $x = e := P_C(R^\top h)$, i.e.

$$u(e; R, h, d) = u_0(e, R, h) = u_\infty(e, R, h) = e - P_C(h)\,.$$

As a consequence, at this point no depth information can be derived from the observed motion, while the flow itself is not necessarily zero.

Furthermore, any point pair $(x, x') = (x, x - u(x))$ as defined by (5) is connected via the essential matrix $E := [h]_\times R$ through the constraint (Hartley and Zisserman 2000)

$$0 = \begin{pmatrix} x' \\ 1 \end{pmatrix}^\top E \begin{pmatrix} x \\ 1 \end{pmatrix}, \tag{11}$$

see Appendix A.1 for a detailed verification.

Thus, eqn. (11) implicitly defines the epipolar line in camera C' corresponding to x for fixed camera parameters R and h. Consequently, an one-dimensional correspondence search between the previous and current image frame along d implicitly respects the epipolar constraint. Furthermore, from the representation (10) it becomes clear, that only the segment of the epipolar lines is considered which is geometrically reasonable.

The connection to the fundamental matrix F is provided by the camera calibration matrix K and $F = K^{-\top} E K^{-1}$.

3.1.2 Depth Reconstruction Accuracy

In order to motivate the upcoming formulation and to ease the interpretation of experimental results we provide a theoretical analysis of the accuracy of the depth measurement. For completeness we also compare to stereo setups which will provide reference depth maps in the experimental section.

Let us fix the camera movement (R, h) and some depth $d(x)$. Furthermore, we assume that the optical flow can be measured at pixel x with the correct mean

$\hat{u}(d) = u(x; R, h, d)$ and a unit isotropic accuracy $\Sigma_u = \sigma_u^2 I$, both expressed in normalized coordinates. Then we employ a Gaussian approximation of the distribution $\mathrm{p}(\hat{d}|\hat{u}(d))$ of the *reconstructed* depth $\hat{d}$ and for $\hat{d} \approx d$ we have

$$\mathrm{p}(\hat{d}|\hat{u}(d)) \sim \mathrm{p}(\hat{u}(d)|\hat{d}) \approx \mathcal{N}(\hat{d}; d, \sigma_{\hat{d}}^2)\,. \tag{12}$$

A suitable choice (Tierney and Kadane 1986) for $\sigma_{\hat{d}}$ is

$$\sigma_{\hat{d}}^{-2} = \frac{\partial^2}{\partial \hat{d}^2}(-\log \mathrm{p}(\hat{d}|\hat{u}(d)))\Big|_{\hat{d}=d}\,. \tag{13}$$

Then we define the (approximate) standard deviation of the depth measure *relative to* σ_u as

$$\sigma_g(d, x) := \frac{\sigma_{\hat{d}}}{\sigma_u} = \frac{[R\left(\begin{smallmatrix} x \\ 1 \end{smallmatrix}\right) d + h]_3^2}{\left\| HR\left(\begin{smallmatrix} x-e \\ 0 \end{smallmatrix}\right)\right\|} \tag{14}$$

with $H := \left(\begin{smallmatrix} -h_3 & 0 & h_1 \\ 0 & -h_3 & h_2 \end{smallmatrix}\right)$ and epipole e. This quantity models the dependence of the expected depth measurement error on *geometric* factors.

We exemplarily compare two simple scenarios: a simple forward movement ($R_\mathrm{m} = I$, $h_\mathrm{m} = b(0, 0, 1)^\top$, $b \in \mathbb{R}$) typical for the considered *monocular setup* and a sidewise translation ($R_\mathrm{s} = I$, $h_\mathrm{s} = b(1, 0, 0)^\top$), which is essentially a *rectified stereo setup* with baseline $b \in \mathbb{R}$. The measures simplify to:

$$\text{monocular: } \sigma_{g,\mathrm{m}} = \frac{(d+b)^2}{b}\frac{1}{\|x\|} \tag{15}$$

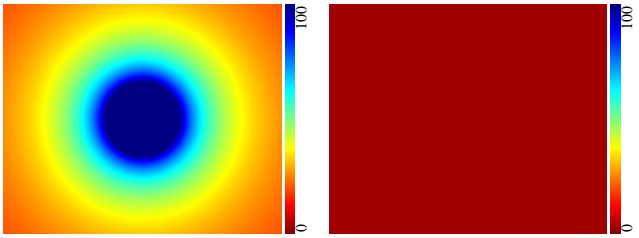
$$\text{stereo: } \sigma_{g,\mathrm{s}} = \frac{d^2}{b}\,. \tag{16}$$

For both cases the expected error increases quadratically in depth d (assuming $b \ll d$), but more determining is the spatial dependency of the monocular setup. As a consequence, the potential accuracy is limited in regions near the image center while the accuracy of a stereo method is independent of the image position. This generalizes to general rotation and translation parameters (see (14)) where the distance to the epipole has considerable impact on the accuracy.

Figure 4(a) shows the characteristic shape of $\sigma_g(x)$ for the monocular scenario. It becomes clear that extracting depth information near the epipole requires an accurate model which motivates the elaborate formulation and numerics in Sect. 3 and Sect. 4, respectively. In comparison, a stereo image pair (Fig. 4(b)) provides much more information at any image position.

3.1.3 Global Scale Estimation

Due to the probabilistic formulation, further information can easily be incorporated to fix the unknown global scale of the scene and camera translation.



(a) monocular uncertainty (b) stereo uncertainty

Fig. 4 Expected depth measurement error $\sigma_g(x, d)$ for **(a)** forward motion typical for a monocular setup (see (15)) and **(b)** a stereo setup (see (16)). For both cases we set $b = 0.3$, $d = 1$ and use the same color encoding (clipping at 100). The stereo setup has a spatially constant uncertainty ($\sigma_g = \frac{10}{3}$). In contrast, for the considered monocular setup σ_g is generally larger and decreases considerably near the epipole (image center). This emphasizes the requirement for an accurate model and numerics as propagated here.

For example odometry sensor data can be included as an *additional observation term* in the form of a Gaussian probability distribution of C^k. This formulation allows to describe uncertain and (partly) missing information, e.g. the vertical motion component. For more details we refer to Sect. 3.2.1, where the same technique is used to incorporate the prediction prior on C^k.

In the same manner, range sensors or detected objects of known size introduce observation terms incorporating the depth map variables.

Note, that in this work we do not assume the presence of additional sensor data and thus we will not further investigate this issue.

3.2 State Transition and Prediction

Next we detail the state transition model $\mathrm{p}(X^k|X^{k-1})$ for the state variables $X = (d, C)$. The previous state $X^{k-1} = (d^{k-1}, C^{k-1})$ is equipped with a variance estimation σ_d^{k-1} and Σ_C^{k-1} which will be addressed in Sect. 3.4.

3.2.1 Camera

We take $C^{k-1} =: \hat{C}^k$ both as prediction $\hat{C}^k$ of C^k and as mean of C^k, which is justified by the fast frame rate. The variance $\hat{\Sigma}_C^k$ of the camera parameters prediction is propagated by defining $\hat{\Sigma}_C^k := \Sigma_C^{k-1} + \Sigma_C$. The parameters σ_R^2 and σ_h^2 on the main diagonal of diagonal matrix Σ_C account for the uncertainty of the rotational and translational prediction, respectively. Then the probabilistic model for C^k reads as

$$C^k \sim \mathrm{p}(C^k|C^{k-1}) = \mathcal{N}_{\mathcal{M}^C}(C^k; C^{k-1}, \hat{\Sigma}_C^k). \tag{17}$$

Here, $\mathcal{N}_{\mathcal{M}^C}(x; y, \Sigma)$ is the normal distribution on the manifold $\mathcal{M}^C = \mathsf{SE}_3$ and is defined in Pennec (2006, Theorem 3) as

$$\mathcal{N}_{\mathcal{M}^C}(x; y, \Sigma) \propto \exp\left(-\frac{1}{2}\mathrm{dist}^2_{\mathcal{M}^C}(x, y; \Sigma)\right), \tag{18}$$

$$\mathrm{dist}^2_{\mathcal{M}^C}(x, y; \Sigma) := (\mathrm{Log}_y x)^\top \Gamma(\Sigma)(\mathrm{Log}_y x), \tag{19}$$

with concentration matrix $\Gamma(\Sigma)$, $\mathrm{Log}_y x := \mathrm{Log}(y^{-1}x)$ and $\mathrm{Log}(x)$ as defined in Appendix B.1.1.

The relation of $\Gamma(\Sigma)$ to Σ is given by a non-closed form and thus we use the approximation proposed in Pennec (2006, Theorem 5):

$$\Gamma(\Sigma) \approx \left(\Sigma^{-1} - \frac{1}{3}\mathrm{Ric}\right)_{\geq \epsilon} \tag{20}$$

where Ric is the Ricci curvature matrix. Furthermore, $(X)_{\geq\epsilon}$ ensures positive definiteness of X by replacing the eigenvalues $\lambda_1(X), \ldots, \lambda_6(X)$ with $\max\{\epsilon, \lambda_1(X)\}$, $\ldots$, $\max\{\epsilon, \lambda_6(X)\}$ and some small $\epsilon > 0$.

3.2.2 Depth Map

The predicted depth map $\hat{d}^k$ is computed by transporting d^{k-1} by the motion parameters $\hat{C}^k = (\hat{R}^k, \hat{h}^k) = (R^{k-1}, h^{k-1})$. 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. (5) and (6),

$$x' \approx x - u(x; R^{k-1}, h^{k-1}, d^{k-1}(x)) \tag{21}$$

$$= P_C\left(d^{k-1}(x) R^{k-1}\begin{pmatrix} x \\ 1 \end{pmatrix} + h^{k-1}\right). \tag{22}$$

We bi-linearly interpolate d^{k-1} at x' to obtain $d'(x')$ and the according space point $d'(x')\begin{pmatrix} x' \\ 1 \end{pmatrix}$ in camera C'. Its transition to camera C is given by (4), and we define the depth $d(x)$ as prediction $\hat{d}^k(x)$, i.e.

$$\hat{d}^k(x) = \hat{d}^k(x; X^{k-1}) = \hat{d}^k(x; d^{k-1}, R^{k-1}, h^{k-1}). \tag{23}$$

Figure 5 illustrates this process. Note that eqn. (21) only is an approximation because we do not know the correct argument $d^k(x)$ as required by eqn. (6), and that (23) is a function of $X^{k-1} = (d^{k-1}, C^{k-1})$.

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 identically 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. 11. Based on this relationship, we make a Gaussian ansatz as approximate probabilistic model of d^k,

$$\mathrm{p}(d^k | X^{k-1}) \propto \exp\big(-f_d(d^k; \hat{d}^k, \hat{\sigma}_d^k)\big). \tag{24}$$

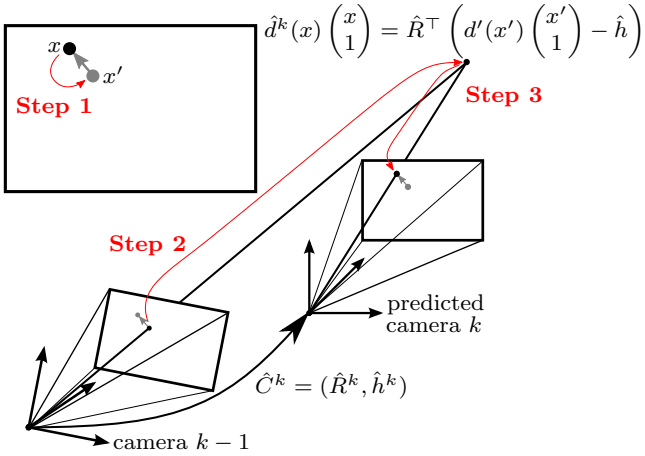


Fig. 5 Prediction $\hat{d}^k$ of the state variable d: We assume the predicted camera motion $(\hat{R}, \hat{h}) := (R^{k-1}, h^{k-1})$. Step 1: Map $x \in \Omega$ in (predicted) camera k to x' in camera $k-1$ using (21). Step 2: Transfer corresponding space point $(x', 1)^\top d'(x')$ from coordinate system of camera $k-1$ to (predicted) camera k. Step 3: Use depth of this space point as predicted depth $\hat{d}^k(x)$.

The energy functional f_d includes a prior penalizing the deviation from the prediction $\hat{d}^k$ and a spatial smoothness prior with variance parameter $\sigma_s \in \mathbb{R}$

$$f_d(d^k; \hat{d}^k, \hat{\sigma}_d^k) := \tag{25a}$$

$$\frac{1}{2}\int_\Omega \left(\frac{d^k(x) - \hat{d}^k(x)}{\hat{\sigma}_d^k(x)}\right)^2 + \frac{1}{\sigma_s^2}\|\nabla d^k(x)\|^2 \mathrm{d}x. \tag{25b}$$

Here, we used continuous notation to facilitate interpretation of the terms. The decision in favor for a quadratic spatial smoothness term and alternatives are discussed by the end of this section.

After discretization, $d^k, \hat{d}^k, \hat{\sigma}_d^k \in \mathbb{R}^n$ are vectors indexed by n grid positions $x \in \Omega$, and we re-use the symbol ∇ to denote the matrix $\nabla\colon \mathbb{R}^n \to \mathbb{R}^{2n}$ approximating the gradient mapping. Furthermore, we define the predicted positive semi-definite 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 (25) into (24) and ignoring normalizing constants, we obtain after multiplying out and rearranging the terms using some basic matrix algebra (see, e.g. Rasmussen and Williams (2006, App. A.2)),

$$\mathrm{p}(d^k | X^{k-1}) \propto \mathcal{N}(d^k; \hat{\mu}_d^k, \hat{\Sigma}_d^k), \qquad \text{with} \tag{26a}$$

$$\hat{\mu}_d^k = \hat{\Sigma}_d^k \left(\hat{S}_d^k\right)^{-1} \hat{d}^k, \tag{26b}$$

$$\hat{\Sigma}_d^k = \left(\left(\hat{S}_d^k\right)^{-1} + \sigma_s^{-2}\nabla^\top\nabla\right)^{-1}. \tag{26c}$$

Due to the specific form, $\hat{\Sigma}_d^k$ is positive semi-definite.

Notice that the prior (24) and the translational part of (17) fix a single, but arbitrary global scale of d and h that cannot be inferred from monocular sequences.

Alternative Spatial Regularization. Depth map regularization implicitly enforces a specific prior on the scene structure. First order total-variation (TV) regularization (Rudin et al 1992) is known to result in visually more crisp results than quadratic regularization. However, TV enforces piece-wise constant depth which does not fit well the given application: Even the correct reconstruction of a slanted plane (flat road, house wall) requires a regularization term which is aware of the projective nature of the variables. Second order TV (Lenzen et al 2013; Bredies et al 2010) might be an interesting approximation to be investigated in further work. However, here we resort to quadratic regularization (see (25)) as it provides an immediate probabilistic interpretation while reasonably allowing for slanted structures.

3.3 State Update

Having observed $Y^k = \hat{u}^k$ in terms of the displacement vector field (8) that depends on the unknown state variables $X^k = (d^k, C^k) = (d^k, R^k, h^k)$, we update the state by estimating $X^k = (d^k, C^k)$ as mode of the distribution

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

based on eqns. (8), (17) and (26). We define the *objective function* $f(d, C)$ as

$$\begin{aligned}f(d,C) :=& -\log \mathrm{p}(X^k|Y^{1:k}) && \text{(28a)}\\ =& f_u(d,C) + f_C(C) + f_d(d), && \text{(28b)}\end{aligned}$$

which is composed of

$$\begin{aligned}f_u(d^k, C^k) =& \frac{1}{2}(\hat{u}^k - \mu_u^k)^\top \left(\Sigma_u^k\right)^{-1}(\hat{u}^k - \mu_u^k), && \text{(29a)}\\ f_C(C^k) =& \frac{1}{2}\mathrm{dist}^2_{\mathcal{M}^C}(C^k, C^{k-1}; \hat{\Sigma}_C^k), && \text{(29b)}\\ f_d(d^k) =& \frac{1}{2}(d^k - \hat{\mu}_d^k)^\top \left(\hat{\Sigma}_d^k\right)^{-1}(d^k - \hat{\mu}_d^k). && \text{(29c)}\end{aligned}$$

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

$$(d^k, C^k) := \arg\min_{d,C}\ f(d,C), \quad C \in \mathsf{SE}_3,\ d \in \mathbb{R}^n_+ \qquad (30)$$

consists in *joint* second order update steps for C and d on a manifold, embedded into a multiscale framework. In Sect. 4 we give a detailed description of the optimization and numerics.

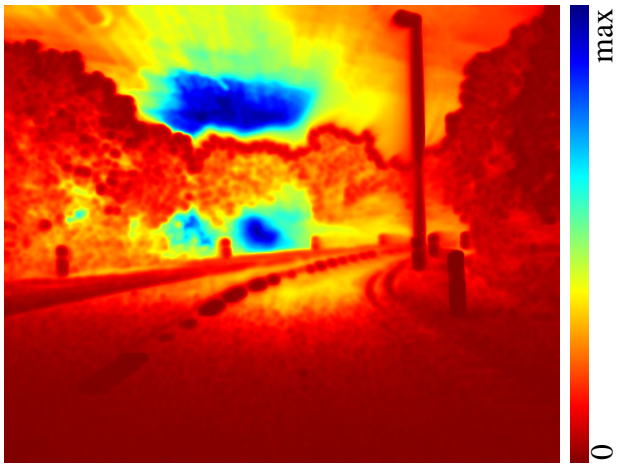


Fig. 6 Estimated depth variance σ_d^k for the frame depicted in Fig. 1. Two sources of high uncertainty ($\sigma_d^k(x)$ large) can be identified: Near the epipole (image center), no information can be derived from the observed motion, see Sect. 3.1.1. In the texture-less region of the sky (above epipole), it is not possible to measure the optical flow, see Fig. 3.

3.4 Estimation of Variable Variance

It is essential to propagate variance information along with the results (d^k, C^k) to ensure that only the accurate information components are incorporated as prior (cf. Sect. 3.2) in the estimation of (d^{k+1}, C^{k+1}).

3.4.1 Local Depth Variance

Similar to σ_g in Sect. 3.1.2 we approximate the local variance $(\sigma_d^k)^2$ of the depth map by the second derivatives of f in (d^k, C^k) and restricted to non-negative values,

$$\begin{aligned}(\sigma_d^k(x))^{-2} :=& \max\left\{0, \frac{\partial^2}{\partial d(x)^2} f_u(d^k, C^k)\right\} && \text{(31a)}\\ &+ \frac{\partial^2}{\partial d(x)^2} f_d(d^k, C^k)\ . && \text{(31b)}\end{aligned}$$

This quantity allows to identify regions with high uncertainty caused by the parallax and/or the lack of image features, see Fig. 6 for an example.

3.4.2 Camera Motion Variance

The joint co-variance matrix of the camera motion parameter is approximated by the second derivatives of f in Q,

$$(\Sigma_C^k)^{-1} = (\nabla_{QQ} f(C^k, d^k))_+\ , \qquad (32)$$

where $(\cdot)_+$ takes the non-negative part of the eigenvalues, which can be accomplished using a Cholesky-decomposition with low computational effort due to the small dimension of Q. Figure 7 visualizes an exemplary camera distribution.

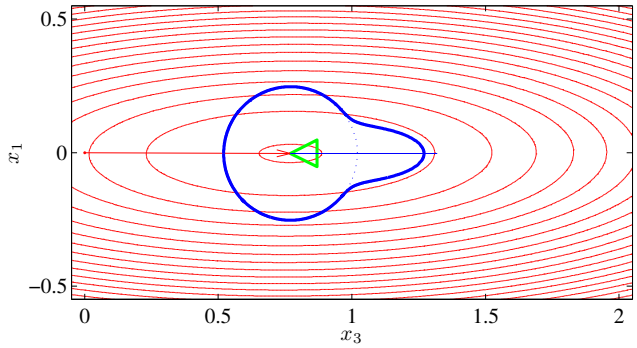


Fig. 7 2d-projection (top view) of estimated variance of camera pose (green): contour lines and mode (arrow) of probability density function of the estimated camera translation h^k (red). Probability density function of rotation R^k as polar plot (blue) with radius offset for better visibility (dotted). Translation accuracy is lower along ($\approx x_3$) than perpendicular to ($\approx x_1$) the camera principal axis.

3.5 Interpretation as Variational Optical Flow Estimation

The proposed ansatz can be interpreted as a variational optical flow estimation method. To this end we consider the minimization of the objective function (28) w.r.t. d and C. The prediction prior terms introduce ordinary biases on the variable to impose temporal smoothness. For clarity, in the following considerations we omit the camera and depth prediction term by setting $f_C(d, C) = 0$ and $(\hat{S}_d^k)^{-1} = 0$ in (25). Then (28) simplifies to

$$
\begin{aligned}
& f(d, C) = f_u(d, C) + f_d(d) \\
=& \frac{1}{2}(\hat{u} - \mu_u)^\top \Sigma_u^{-1}(\hat{u} - \mu_u) + \frac{1}{2\sigma_s^2} d^\top (\nabla^\top \nabla) d \\
=& \frac{1}{2} \sum_{i=1}^{n} \underbrace{(\hat{u}(x_i) - \mu_u(x_i))^\top \Sigma_u^{-1}(x_i)(\hat{u}(x_i) - \mu_u(x_i))}_{\text{addend } i} \\
& + \frac{1}{2\sigma_s^2} \sum_{i=1}^{n} \|\nabla d(x_i)\|^2.
\end{aligned}
$$

Inserting the definition of $\hat{u}$ and Σ_u in (7), addend i of the first term of $f(d, c)$ can be rewritten as (omitting the local dependency on x_i)

$$
\begin{aligned}
& (\hat{u} - \mu_u)^\top \Sigma_u^{-1}(\hat{u} - \mu_u) && (33) \\
=& \begin{pmatrix} \mu_u \\ 1 \end{pmatrix}^\top \left(\mathrm{G}_\rho * \begin{pmatrix} \nabla I \nabla I^\top & \partial_t I \nabla I \\ \partial_t I (\nabla I)^\top & (\partial_t I)^2 \end{pmatrix} \right) \begin{pmatrix} \mu_u \\ 1 \end{pmatrix} && (34)
\end{aligned}
$$

which corresponds to the data term for flow μ_u introduced in the combined local-global variational optical flow approach in Bruhn et al (2005).

However, in this work the flow $\mu_u = \mu_u(x_i)$ is not represented by a functional $\Omega \mapsto \mathbb{R}^2$ but is parametrized by the depth map $d : \Omega \mapsto \mathbb{R}_+$ and camera motion C using the relation $\mu_u(x_i) = u(x_i; R, h, d(x_i))$ (see (9)). Furthermore, the term f_d imposes smoothness on the *scene representation* d instead of the flow vector field μ_u.

Thus, the proposed approach implicitly and globally estimates an optical flow field which is consistent with the geometrical model (6) and includes a geometrically motivated regularization term. Figure 1(a) visualizes such a reconstructed flow field.

4 Optimization, Numerics

4.1 Overview

The inference step in Sect. 3.3 requires a joint optimization (30) for depth map d and camera parameter C which possesses several challenges that we address with approved and theoretically founded methods.

The interaction of camera and scene variables is non-linear and involved due to the projective model in the observation term and leads to an objective function which is non-convex in general. The highly detailed depth map and their spatial connection due to the smoothness prior requires numerics capable of large-scale problems while camera pose estimation has to respect the manifold nature. However, the camera setup requires joint estimation of d and C to accurately explain the observations and to evenly balance deviations from the model between the error sources.

In this section we propose a Newton-like *second-order iterative method* for efficiently coping with non-linearity. The task of choosing a joint descent direction for (d, C) is *reduced to a system of linear equations* which is a well understood problem also for a large number of variables. The *Lie group formulation* respects the manifold nature. A decrease of the objective function is guaranteed until convergence by adding carefully chosen proximity terms.

Section 4.2 introduces and defines the mathematical concepts and their details required for a compact description of the algorithm. Algorithm 1 summarizes the essential steps for determining the minimizing sequence and also links to Sections 4.4–4.10 providing motivation and details.

The embedding multiscale framework (Sect. 4.10) enables the local optical flow estimator to handle the large dynamics of the displacement magnitude caused by the monocular camera setup and the high camera speed.

Algorithm 1 Overview over the second-order iterative update of the variables $X^{(i)}$ which respects their manifold structure and guarantees decrease of objective function $f(X)$. Comments point to sections which provide details and motivation.

initialize $X^{(0)}$, $i = 0$	▷ Sect. 4.10
repeat	
$i \leftarrow i + 1$	
determine gradient $\nabla_T f$ of f at $X^{(i-1)}$	▷ Sect. 4.4
determine second order information A of f	▷ Sect. 4.5
choose M, s.t. $B^{(i)} := (A + M)^{-1} \succ 0$	▷ Sect. 4.6
compute search dir. $W^{(i)} = -B^{(i)} \nabla_T f$	▷ Sect. 4.7
determine step size $t^{(i)}$ (line search)	▷ Sect. 4.8
update: $X^{(i)} \leftarrow \phi(t^{(i)}, X^{(i-1)}, W^{(i)})$	▷ Sect. 4.3
until convergence	▷ Sect. 4.9

4.2 Preliminaries

For a compact and clear representation, and in order to focus on the actual method instead of implementation details in the following sections, we here define and detail the required concepts connected to Lie groups. In particular we consider the variable domain $\mathbb{R}^n \times \mathsf{SE}_3$ as a single manifold. To this end we summarize the Lie group interpretation of $\mathbb{R}^n$ in Sect. 4.2.2.

4.2.1 Special Euclidean Group SE_3

The group neutral element I^C of G^C is the 4×4 identity matrix. The Lie algebra $\mathrm{g}^C := \mathsf{se}_3$ of G^C is given by

$$\mathrm{g}^C = \left\{ W = \begin{pmatrix} [\omega]_\times & v \\ 0^\top & 0 \end{pmatrix} : \omega, v \in \mathbb{R}^3 \right\}, \tag{35}$$

where $\mathsf{so}_3 \ni [\omega]_\times$ denotes the Lie algebra of SO_3 identified with the linear subspace of skew-symmetric matrices and $[\omega]_\times$ as defined in (2).

The orthogonal basis $\{\mathcal{L}_1^C, \dots, \mathcal{L}_6^C\}$ spans g^C. We denote the tangent space of G^C at $Q \in \mathrm{G}^C$ by $T_Q\mathrm{G}^C$. Any $W \in T_{I^C}\mathrm{G}^C$ is transported to $T_Q\mathrm{G}^C$ by matrix-multiplication QW, here denoted as application of the linear operator $L_Q^C : T_{I^C}\mathrm{G}^C \mapsto T_Q\mathrm{G}^C$, i.e. $L_Q^C W = QW$. Its adjoint $(L_Q^C)^*$ is the matrix transpose, i.e. $(L_Q^C)^* W = Q^\top W$.

We equip SE_3 with the Riemannian metric

$$\langle W^1, W^2 \rangle_{\mathrm{G}^C} := \langle [\omega]_\times^1, [\omega]_\times^2 \rangle + \langle v^1, v^2 \rangle, \tag{36}$$

for all $W^1, W^2 \in \mathrm{g}^C$, where $\langle \cdot, \cdot \rangle$ on the right-hand side denotes the canonical matrix and vector inner product, respectively. Note that unlike for general Riemannian metrics, the metric (36) does not depend on $Q \in \mathrm{G}^C$, hence is the same for all tangent spaces $T_Q\mathrm{G}^C$, justifying the notation $\langle \cdot, \cdot \rangle_{\mathrm{G}^C}$.

The exponential mapping $\mathrm{Exp}\colon \mathrm{g}^C \mapsto \mathrm{G}^C$ that diffeomorphically maps tangent vectors close to 0 onto the manifold within a neighborhood of I^C as well as its inverse, $\mathrm{Log}\colon \mathrm{G}^C \mapsto \mathrm{g}^C$ can be computed in closed form, see Appendix B.1.1.

The orthogonal projection onto the tangent space $T_{I^C}\mathrm{G}^C$ can be computed in closed form: Using the representation $X = \left(\begin{smallmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{smallmatrix}\right) \in \mathbb{R}^{4\times4}$, $X_{11} \in \mathbb{R}^{3\times3}$ it is given by

$$\begin{aligned} \Pi_{T_{I^C}\mathrm{G}}(X) &= \arg\min_{W \in \mathrm{g}^C} \langle W - X, W - X \rangle_{\mathrm{G}^C} && (37) \\ &= \begin{pmatrix} \frac{1}{2}(X_{11} - X_{11}^\top) & X_{12} \\ 0^\top & 0 \end{pmatrix}. && (38) \end{aligned}$$

The Levi-Civita connection $\overline{\nabla}$ of an m-dimensional Riemannian manifold $\mathcal{M}$ is the unique torsion-free and metric-preserving affine connection

$$\overline{\nabla} : C^\infty(\mathcal{M}, T\mathcal{M}) \times C^\infty(\mathcal{M}, T\mathcal{M}) \mapsto C^\infty(\mathcal{M}, T\mathcal{M}),$$

which is defined (Absil et al 2008) by

$$\overline{\nabla}_W V := \lim_{t \to 0} \frac{L_{X(0)} L_{X(t)}^{-1} V(X(t)) - V(X(0))}{t}. \tag{39}$$

with trajectory $X(t) : \mathbb{R} \mapsto \mathcal{M}$ and $\frac{\mathrm{d}}{\mathrm{d}t} X(t) = W(X(t))$.

With an orthogonal base $\{\mathcal{L}_1, \dots, \mathcal{L}_m\}$ of the according tangential space $T\mathcal{M}$, the parametrization $W = \sum_{k=1}^m w_k \mathcal{L}_k$ and $V = \sum_{k=1}^m v_k \mathcal{L}_k$ and the Christoffel symbols $\Gamma_{ij}^k \in \mathbb{R}$, a general representation is given by

$$\overline{\nabla}_W V = \sum_{k=1}^m \left(\langle W, \partial v_k \rangle + \sum_{i,j=1}^m w_i v_j \Gamma_{ij}^k \right) \mathcal{L}_k \,. \tag{40}$$

Note that $\overline{\nabla}$ is linear in W.

The Christoffel symbols are defined uniquely by the equations (for all $1 \le i, j, k \le m$, with $[V, W] = VW - WV$)

$$\begin{aligned} \overline{\nabla}_{\mathcal{L}_i} \mathcal{L}_j - \overline{\nabla}_{\mathcal{L}_j} \mathcal{L}_i &= [\mathcal{L}_i, \mathcal{L}_j] && (41) \\ \text{and } \mathcal{L}_k \langle \mathcal{L}_i, \mathcal{L}_j \rangle &= \langle \overline{\nabla}_{\mathcal{L}_k} \mathcal{L}_i, \mathcal{L}_j \rangle + \langle \mathcal{L}_i, \overline{\nabla}_{\mathcal{L}_k} \mathcal{L}_j \rangle && (42) \end{aligned}$$

demanding symmetry and metric preservation, respectively.

For g^C, the Levi-Civita-connection $\overline{\nabla}^C$ is non-trivial and described by the Christoffel symbols and we list them in Appendix B.1.2.

4.2.2 Real Vector Space $\mathbb{R}^n$

The vector space $\mathbb{R}^n$ can be considered as a flat manifold $\mathcal{M}^d := \mathbb{R}^n$, e.g. by regarding $\mathbb{R}^n$ as the subgroup of SE_n representing the translational part. We identify $\mathcal{M}^d$ with the Lie group $\mathrm{G}^d = \mathbb{R}^n$ and adopt the natural vector representation of elements in $X \in \mathrm{G}^d$, $X = (x_1, \dots, x_n)$ with $x_i \in \mathbb{R}$. Then the group multiplication XY, inversion X^{-1} and neutral element I^d are the

element-wise addition ($XY = (x_1 + y_1, \ldots, x_n + y_n)$), element-wise additive inverse ($X^{-1} = (-x_1, \ldots, -x_n)$) and zero vector ($I^d = (0, \ldots, 0)$), respectively.

For any point $X \in \mathrm{G}^d$, we denote the attached tangential space by $T_X\mathrm{G}^d$. The exponential map $\mathrm{Exp} : T_{I^d}\mathrm{G}^d \mapsto \mathrm{G}^d$ connecting the associated Lie algebra g^d to the Lie group G^d, its inverse, $\mathrm{Log} : \mathrm{G}^d \mapsto T_{I^d}\mathrm{G}^d$, and the orthogonal projection $\mathit{\Pi}_{T_{I^d}\mathrm{G}^d} : \mathbb{R}^n \mapsto T_{I^d}\mathrm{G}^d$ are the identity operator. Thus, $T_X\mathrm{G}^d = \mathbb{R}^n$ with the canonical orthogonal basis for $\mathbb{R}^n$, $\{\mathcal{L}_1^d, \ldots, \mathcal{L}_n^d\}$. Also the linear operator L_X^d transporting $W \in T_{I^d}\mathrm{G}^d$ to $T_X\mathrm{G}^d$ for any $X \in \mathrm{G}^d$ is the identity and thus is self-adjoint, i.e. $L_X^d = (L_X^d)^* = \mathrm{Id}$. For any $V, W \in T_X\mathrm{G}^d$, the inner product is given by $\langle V, W\rangle_{\mathrm{G}^d} = V^\top W$.

On G^d, the Christoffel-symbols vanish, such that Levi-Civita-connection (40) simplifies to a directional derivative:

$$\overline{\nabla}_W^d V = \sum_{k=1}^{n} \langle W, \partial v_k \rangle \mathcal{L}_k^d \tag{43}$$

with $W = \sum_{k=1}^n w_k \mathcal{L}_k^d$ and $V = \sum_{k=1}^n v_k \mathcal{L}_k^d$. In particular, we have

$$\overline{\nabla}_W^d(\nabla f(X)) = (\mathrm{H}_w f(X))w, \tag{44}$$

where H_x is the Hessian matrix operator $(\frac{\mathrm{d}^2}{\mathrm{d}x_i \mathrm{d}x_j})_{i,j}$.

4.3 Joint Optimization

We consider an optimization problem equivalent to (30) over the manifold $\mathcal{M} := \mathcal{M}^d \times \mathcal{M}^C$, identified by the group

$$\mathrm{G} := \mathrm{G}^d \times \mathrm{G}^C = \{(X_d, X_C) \,|\, X_d \in \mathbb{R}^n, X_C \in \mathsf{SE}_3\} \,.$$

Both G^d and G^C are Lie groups and thus G is a Lie group with the associated Lie algebra denoted by g and we can adopt the framework for optimization on smooth manifolds for *jointly* optimizing over $\mathcal{M}$. All required concepts are defined by independently applying the corresponding definitions in Sect. 4.2.1 and Sect. 4.2.2 for the camera pose X_C and depth map X_d, respectively. In particular, the inner product is given by

$$\langle V, W\rangle_{\mathrm{G}} = \langle V_d, W_d\rangle_{\mathrm{G}^d} + \langle V_C, W_C\rangle_{\mathrm{G}^C} \tag{45}$$

for $V, W \in \mathrm{g}$.

The non-negativity constraint of d is moved into the objective function, i.e.

$$q(X) := f(X_d, X_C) + \delta_{\mathbb{R}^n_+}(X_d) \tag{46}$$

using the characteristic function

$$\delta_{\mathbb{R}^n_+}(X_d) := \begin{cases} 0 & \text{if } X_d \in \mathbb{R}^n_+ \\ \infty & \text{else} \end{cases} \,. \tag{47}$$

It remains to determine a minimizing sequence $X^{(i)} \in \mathrm{G}$, $i = 1, 2, \ldots$ for the problem

$$\min_{X \in \mathrm{G}} q(X) \,. \tag{48}$$

Given $X^{(i)}$, we determine $X^{(i+1)}$ by

$$X^{(i+1)} = \varphi(t^{(i)}, X^{(i)}, W^{(i)}) \tag{49}$$

with search direction $W^{(i)} \in T_I\mathrm{G}$, step size $t^{(i)} \in \mathbb{R}_+$ and

$$\varphi(t, X, W) := L_X \mathrm{Exp}\,(tW) \,. \tag{50}$$

4.4 Descent Step Direction

4.4.1 Manifold Gradient

For determining a search direction $W^{(i)}$, we consider the unconstrained objective function $f(X)$. Here, we denote the (ordinary) gradient of $f(X)$ in the ambient space at $X \in \mathrm{G}$ by $\nabla f(X)$. The gradient ∇_{G} on the manifold G is defined by (cf. Absil et al (2008))

$$\langle \nabla_{\mathrm{G}} f(X), V\rangle_{\mathrm{G}} = \langle \nabla f(X), V\rangle \quad \forall V \in T_X\mathrm{G} \tag{51}$$

which is equivalent to

$$\langle \nabla_{\mathrm{G}} f(X), L_X V\rangle_{\mathrm{G}} = \langle \nabla f(X), L_X V\rangle \quad \forall V \in T_I\mathrm{G} \tag{52}$$

with L_X as defined in Sect. 4.2, or equivalently

$$\langle L_X^* \nabla_{\mathrm{G}} f(X), V\rangle_{\mathrm{G}} = \langle L_X^* \nabla f(X), V\rangle \;\forall V \in T_I\mathrm{G} \,. \tag{53}$$

Then $\nabla_{\mathrm{G}} f(X)$ can be retrieved by the orthogonal projection $\mathit{\Pi}_{T_I\mathrm{G}}$ onto the tangent space,

$$\nabla_{\mathrm{G}} f(X) = (L_X^*)^{-1} \mathit{\Pi}_{T_I\mathrm{G}} (L_X^* \nabla f(X)) \tag{54}$$

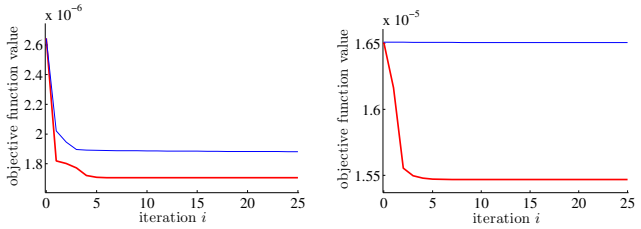
which can be computed explicitly for G (see Sects. 4.2.1, 4.2.2).

We introduce a function $\nabla_T f : \mathrm{G} \mapsto T\mathrm{G}$ which describes $\nabla_{\mathrm{G}} f(X)$ when moving from $X^{(i)}$ by $Y \in \mathrm{G}$ to $X^{(i)}Y$:

$$\begin{aligned} \nabla_T f(Y) :=& L_Y L_{X^{(i)}Y}^* \nabla_{\mathrm{G}} f(X^{(i)}Y) && (55)\\ =& L_Y \mathit{\Pi}_{T_I\mathrm{G}} \left(L_{X^{(i)}Y}^* \nabla f(X^{(i)}Y) \right) && (56) \end{aligned}$$

For the compactness of notation, we use

$$\nabla_T f := \nabla_T f(I). \tag{57}$$



(a) coarsest multiscale level (b) finest multiscale level

Fig. 8 Representative performance of the minimization approach in Sect. 4 in a real scenario. Comparison of the proposed second order method (red thick line) to a first order gradient descent ($B^{(i)} = \mathrm{Id}$, blue thin line) on the (a) coarsest and (b) finest multiscale level. The second order method effectively minimized the objective function after few iteration steps, as opposed to gradient descent which converges slowly.

4.4.2 General Descent Direction

Furthermore, we define a positive definite linear operator on the tangential space,

$$B : T_I\mathrm{G} \mapsto T_I\mathrm{G}$$
$$\langle V, BV\rangle_{\mathrm{G}} > 0, \forall V \in T_I\mathrm{G}, V \neq 0\,. \tag{58}$$

Then for any $\nabla_T f \neq 0$ and any positive definite linear operator $B^{(i)}$,

$$W^{(i)} := -B^{(i)}\nabla_T f \in T_I\mathrm{G} \tag{59}$$

is a descent direction, i.e.

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\varphi(t, X^{(i)}, W^{(i)}))\Big|_{t=0} < 0\,, \tag{60}$$

see Proposition 3 in Appendix B.2. Since $f(X)$ is smooth on G, there exists a $t^{(i)} > 0$ such that $f(X^{(i+1)}) < f(X^{(i)})$.

The choice of $B^{(i)}$ is crucial for performance. Due to the high dimension of X_d, the spatial variable interactions (see (26)) and the non-linear interactions between X_d and X_C in f_u (see (29a)), a gradient descent method (i.e. $B^{(i)} = \mathrm{Id}$) on the manifold turned out to be unsatisfactorily slow. In turn, a suitable choice of $B^{(i)}$ can improve the performance of the minimization procedure significantly. Figure 8 demonstrates this improvement for a real scenario and our choice of $B^{(i)}$ detailed next.

4.5 Second Order Information

For optimization on $\mathbb{R}^n$, the second derivatives (Hessian H) provide valuable information on the local shape of the objective function which can easily utilized in second-order approaches such as the Newton method. On a smooth manifold, the situation is usually more complex. Here, the role of the Hessian is taken on by the Levi-Civita connection $\overline{\nabla}$, see Sect. 4.2.

For any direction $W \in T_I\mathrm{G}$, $\overline{\nabla}_W \nabla_T f$ provides an approximation of the *change* of $\nabla_T f$ when moving from $X^{(i)}$ to $\varphi(1, X^{(i)}, W)$. We employ this description for finding an approximate critical point of $f(X)$ near $X^{(i)}$ which amounts to solving

$$\nabla_T f + \overline{\nabla}_W \nabla_T f = 0 \tag{61}$$

for $W \in T_I\mathrm{G}$.

The Levi-Civita connection $\overline{\nabla}_W$ is linear in W and thus we can define a linear operator $A : T_IG \mapsto T_IG$, such that

$$AW = \overline{\nabla}_W \nabla_T f. \tag{62}$$

Defining $b := \nabla_T f$, (61) can be written as

$$AW = -b\,. \tag{63}$$

Furthermore, if the inverse A^{-1} of A exists and is positive definite, then a step direction can be computed as $W = -A^{-1}b = -A^{-1}(\nabla_T f)$ and is guaranteed to be a descent direction due to the results in Sect. 4.4 and Proposition 3 in Appendix B.2.

However, due to the non-convexity of the objective function f, A is not positive definite in general. Thus, in the following, we propose an additive modification of A,

$$\overline{A} := A + M, \tag{64}$$

such that

$$B^{(i)} := (\overline{A})^{-1} \tag{65}$$

is positive definite and thus guarantees (59) to be a descent direction.

4.6 Choice of M

In this section we examine the structure of the objective function f in $X^{(i)}$ with the aim to determine a suitable modifying matrix M. Our choice is summarized in Propositions 1 and 2.

4.6.1 Function Approximation Interpretation

The linear equality system (63) can be interpreted as the optimality condition of a quadratic form

$$h(W) := f(X^{(i)}) + \langle b, W\rangle_{\mathrm{G}} + \frac{1}{2}\langle W, AW\rangle_{\mathrm{G}} \tag{66}$$

and $h(tW)$ is a local quadratic Taylor approximation of the objective function at $X^{(i)}$,

$$h(tW) \approx f(\varphi(t, X^{(i)}, W)), \tag{67}$$

see Appendix B.3 for a verification. Modification (64) by M relates to adding a quadratic term to $h(W)$ and we can define

$$\begin{aligned}\overline{h}(W) :=& h(W) + \frac{1}{2}\langle W, MW\rangle_{\mathrm{G}} && (68)\\ =& f(X^{(i)}) + \langle b, W\rangle_{\mathrm{G}} + \frac{1}{2}\langle W, \overline{A}W\rangle_{\mathrm{G}}\,. && (69)\end{aligned}$$

Furthermore, positive definiteness of $\overline{A}$ is equivalent to strict convexity of $\overline{h}(W)$ in W.

4.6.2 Linear Representation

For our further analysis we bring (63) into a matrix-vector representation using the orthogonal base $\{\mathcal{L}_1^d, \ldots, \mathcal{L}_n^d, \mathcal{L}_1^C, \ldots, \mathcal{L}_6^C\}$ for g (defined in Sect. 4.2). Then we can represent W uniquely by $w = (w_d, w_C) \in \mathbb{R}^{n+6}$ through

$$\begin{aligned}W(w) =& (W_d(w_d), W_C(w_C)) && (70)\\ =& \left(\sum_{k=1}^{n} w_{d,k}\mathcal{L}_k^d, \sum_{k=1}^{6} w_{C,k}\mathcal{L}_k^C\right)\,. && (71)\end{aligned}$$

We re-use the symbols A, $\overline{A}$, M and b for their corresponding matrix and vector representation $A, \overline{A}, M \in \mathbb{R}^{(n+6)\times(n+6)}$ and $b \in \mathbb{R}^{n+6}$, respectively. Then we can rewrite (63) as a linear equality system in matrix representation,

$$Aw = -b\,. \tag{72}$$

The involved matrices decompose according to the composition of the variables $w = (w_d, w_C)$ into

$$A = \begin{pmatrix} A_{dd} & A_{Cd}^\top \\ A_{Cd} & A_{CC}\end{pmatrix} \text{ and } b = \begin{pmatrix} b_d \\ b_C\end{pmatrix}. \tag{73}$$

We choose a block-diagonal modification matrix M,

$$M := \begin{pmatrix} M_{dd} & 0 \\ 0 & M_{CC}\end{pmatrix} \tag{74}$$

and obtain

$$\overline{A} = A + \begin{pmatrix} M_{dd} & 0 \\ 0 & M_{CC}\end{pmatrix} = \begin{pmatrix} \overline{A}_{dd} & A_{Cd}^\top \\ A_{Cd} & \overline{A}_{CC}\end{pmatrix}. \tag{75}$$

In Sect. 4.6.3 we state general requirements for a suitable $\overline{A}$ and motivate our choice of M_{dd} and M_{CC} in Sects. 4.6.4 and 4.6.5, respectively.

4.6.3 Conditions for $B^{(i)} \succ 0$

In the following we denote the symmetric part of a matrix M by $M^{\mathrm{sym}} := \frac{1}{2}(M + M^\top)$. For showing the positive definiteness of $B^{(i)} = \overline{A}^{-1}$ it is sufficient to proof $\overline{A} \succ 0$ and due to $x^\top \overline{A} x = x^\top \overline{A}^\top x = x^\top \overline{A}^{\mathrm{sym}} x$ we only need to consider the symmetric part of $\overline{A}$,

$$\overline{A}^{\mathrm{sym}} = \begin{pmatrix} \overline{A}_{dd} & A_{Cd}^\top \\ A_{Cd} & \overline{A}_{CC}^{\mathrm{sym}}\end{pmatrix}. \tag{76}$$

Identity $\overline{A}_{dd}^{\mathrm{sym}} = \overline{A}_{dd}$ follows from (44). The special structure of $\mathcal{M}^C$ renders A_{CC} non-symmetric in general.

Assuming $\overline{A}_{dd}$ is non-singular, the Schur complement (see Appendix B.4) of matrix $\overline{A}^{\mathrm{sym}}$ w.r.t. $\overline{A}_{dd}^{\mathrm{sym}}$,

$$\overline{S}^{\mathrm{sym}} := \overline{A}_{CC}^{\mathrm{sym}} - A_{Cd}\overline{A}_{dd}^{-1}A_{Cd}^\top \tag{77}$$

provides a condition on the positive definiteness of $\overline{A}$ which makes use of the block decomposition (76):

$$\overline{A} \succ 0 \quad \Leftrightarrow \quad \overline{A}_{dd} \succ 0 \quad \text{and} \quad \overline{S}^{\mathrm{sym}} \succ 0. \tag{78}$$

4.6.4 Choice of M_{dd}

To motivate our choice of M_{dd}, we consider the restriction of $\overline{h}(W)$ to $W = (W_d, 0) = (w_d, 0)$:

$$\overline{h}(w_d) := \overline{h}((w_d, 0)) = h((w_d, 0)) + \frac{1}{2}w_d^\top M_{dd} w_d \tag{79}$$

which decomposes according to (28) (and due to linearity of $\overline{\nabla}$ and ∇) into

$$\begin{aligned}\overline{h}(w_d) =& h_u((w_d, 0)) + h_d((w_d, 0)) + h_C((w_d, 0)) && (80)\\ &+ \frac{1}{2}w_d^\top M_{dd} w_d && (81)\end{aligned}$$

where $h_u(W)$, $h_d(W)$ and $h_C(W)$ are quadratic approximations in $X^{(i)}$ of $f_u(X)$, $f_d(X)$ and $f_C(X)$, respectively. In particular we have

$$\mathrm{H}\,\overline{h}(w_d) = A_{dd} + M_{dd} = \overline{A}_{dd} \tag{82}$$

and thus $\overline{A}_{dd} \succ 0$ is identical to the requirement that $\overline{h}(w_d)$ is a *strictly convex* model. We further analyze $\mathrm{H}\,\overline{h}(w_d)$, which decomposes due to (80) into

$$\begin{aligned}\mathrm{H}\,\overline{h}(w_d) =& \mathrm{H}\, f_u(X^{(i)}) + \mathrm{H}\, f_d(X^{(i)}) + \mathrm{H}\, f_C(X^{(i)}) + M_{dd}\\ =& \operatorname{diag}(s) + (\hat{\Sigma}_d^k)^{-1} + 0 + M_{dd}, && (83)\end{aligned}$$

where $s \in \mathbb{R}^n$, $s_j = \frac{\partial^2}{\partial d_j^2} f_u(X^{(i)})$. The diagonal form of the first term is due to the mutual independence of the observation terms. The second term is positive

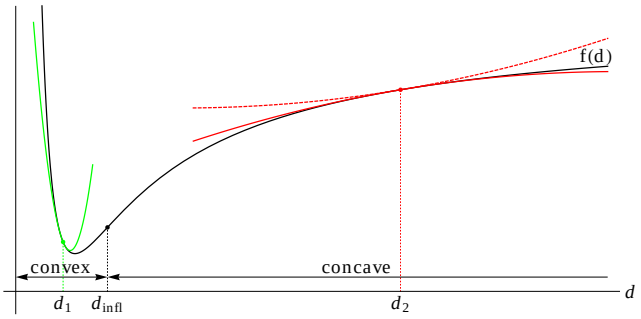


Fig. 9 Plot of $f_u(d, C)$, restricted to a single d_j for typical fixed C and two quadratic approximations in d_1 and d_2 (solid). The function and the local quadratic approximation is convex in $d_1 \in (0, d_{\text{infl}}]$ where d_{infl} is the inflection point. The function is concave in $d_2 \in [d_{\text{infl}}, \infty)$ and we choose an appropriate *convex* approximation there (dashed).

semi-definite, see Sect. 3.2.2. However, it is not diagonal because of the variable interconnections of the spatial regularization. Due to the high dimensionality of the matrix eigenvalue problem $\overline{A}_{dd} \succ 0$, we propose a coordinate-wise correction of A_{dd} by

$$M_{dd} := \operatorname{diag}(m) + \epsilon I \tag{84}$$

with $m \in \mathbb{R}^n$ and some small $\epsilon > 0$.

To motivate our choice of m, we consider the restriction of $h_u(W)$ to a single component j of W_d, i.e. $W = \delta(\mathcal{L}_j^d, 0)$ and

$$\begin{aligned} h_{u,j}(\delta) :=& h_u(\delta(\mathcal{L}_j^d, 0)) && (85)\\ =& f_u(X^{(i)}) + \delta \frac{\partial}{\partial d_j} f_{u,j}(X^{(i)}) + \frac{1}{2} s_j \delta^2 && (86)\\ \overset{(67)}{\approx}& f_u(X(\delta)) =: f_{u,j}(\delta_0 + \delta) && (87) \end{aligned}$$

with $X(\delta) := \varphi(\delta, X^{(i)}, (\mathcal{L}_j^d, 0))$ and $\delta_0 := X_{d_j}^{(i)}$. A typical shape of $f_{u,j}(\delta_0 + \delta)$ is plotted in Fig. 9.

Modification by M_{dd} amounts to adding a quadratic term, i.e.

$$\overline{h}_{u,j}(\delta) := \overline{h}_u(\delta(\mathcal{L}_j^d, 0)) = h_u(\delta(\mathcal{L}_j^d, 0)) + \frac{1}{2} m_j \delta^2. \tag{88}$$

For the choice of m_j we distinguish whether $f_{u,j}(\delta_0)$ is (locally) convex or concave:

Convex case. For those j where $f_{u,j}(\delta_0 + \delta)$ is convex near δ_0, i.e. $s_j \geq 0$, we consider $h_{u,j}(\delta)$ a suitable, convex local approximation and set $\overline{h}_{u,j}(\delta) = h_{u,j}(\delta)$, i.e. $m_j = 0$.

Concave case. Whenever we have $s_j < 0$, $h_{u,j}(\delta)$ is concave and we consider a *linear* model of $f_{u,j}$ instead and respect the reduced degree of detail by adding a quadratic proximity term, i.e.

$$\overline{h}_{u,j}(\delta) = f_{u,j}(X^{(i)}) + \delta \frac{\partial}{\partial d_j} f_{u,j}(X^{(i)}) + \frac{1}{2} \overline{s}_j \delta^2. \tag{89}$$

In our work, we set $\overline{s}_j := |\frac{\partial^2}{\partial d_j^2} f_u(X^{(i)})| = |s_j|$ as a measure for the deviation of the linear model from $f_u(X^{(i)})$ when moving along δ. Comparing (89) to (88) shows that $m_j = \overline{s}_j - s_j = -2s_j \geq 0$ for the concave case.

The choice of m_j for both the convex and concave case is summarized in the following.

Proposition 1 *Let M_{dd} be defined as in* (84) *and m given by*

$$m_j := \begin{cases} 0 & s_j \geq 0 \\ -2s_j & s_j < 0 \end{cases} = |s_j| - s_j. \tag{90}$$

Then $\overline{A}_{dd} = A_{dd} + M_{dd}$ is positive definite as required in (78).

Proof In particular we have $s_j + m_j = |s_j| \geq 0$ for all j and thus

$$\overline{A}_{dd} = \mathrm{H}\,\overline{h}(w_d) = \operatorname{diag}(s + m) + (\hat{\Sigma}_d^k)^{-1} + \epsilon I \succ 0. \tag{91}$$

□

4.6.5 Choice of M_{CC}

Let $\lambda_{\max}(M)$ denote the largest eigenvalue of a symmetric matrix M.

Proposition 2 *Choosing*

$$M_{CC} := \lambda_C I + \epsilon I \tag{92}$$

with some small $\epsilon > 0$ and

$$\lambda_C := \max\{0, \lambda_{\max}(-S^{\text{sym}})\} \tag{93}$$

$$\text{where } S := A_{CC} - A_{Cd} \overline{A}_{dd}^{-1} A_{Cd}^\top \tag{94}$$

is sufficient to fulfill $\overline{S}^{\text{sym}} \succ 0$ as required in (78).

Proof In particular, we have $\lambda_C I \succeq -S^{\text{sym}}$. Then the positive definiteness of the Schur complement (77) can be assessed by

$$\begin{aligned} \overline{S}^{\text{sym}} =& S^{\text{sym}} + M_{CC} = S^{\text{sym}} + \lambda_C I + \epsilon I && (95)\\ \succeq& S^{\text{sym}} - S^{\text{sym}} + \epsilon I \succ 0 && (96) \end{aligned}$$

□

Note that the eigenvalue problem (93) has dimension 6 only and thus λ_C can be computed efficiently.

4.7 Computing the Search Direction

Computing the search direction $W^{(i)}$ in (59) requires solving the linear equality system

$$\overline{A}w = -b \tag{97}$$

where $\overline{A}$ (as defined in (75)) is composed of a large, sparse matrix $\overline{A}_{dd}$ and a very small, dense 6×6-matrix A_{CC} as well as dense A_{Cd}. Furthermore, ensuring positive definiteness of $\overline{A}$ as described in Sect. 4.6 involves computation of matrix inverses. However, it is possible to *corporately* solve the linear equality system (97) and perform the required steps to ensure $\overline{A} \succ 0$ by using the Schur complement (see Appendix B.4). Algorithm 2 summarizes the proposed procedure.

Algorithm 2 Combined choice of M_{dd}, M_{CC} and computation of $w = (w_d, w_C)$ in (97).

choose M_{dd} s.t. $\overline{A}_{dd} = A_{dd} + M_{dd} \succ 0$ ▷ Proposition 1
solve $\overline{A}_{dd} U = A_{Cd}^\top$ for $U \in \mathbb{R}^{n \times 6}$
compute Schur complement $S = A_{CC} - A_{Cd} U$
choose M_{CC}, s.t. $\overline{S} = S + M_{CC} \succ 0$ ▷ Proposition 2
solve $\overline{S} w_C = -b_C + U^\top b_d$ for $w_C \in \mathbb{R}^6$
solve $\overline{A}_{dd} w_d = -b_d - A_{Cd}^\top w_C$ for $w_d \in \mathbb{R}^n$.
set $w = (w_d, w_C)$

Note that although we compute w_d and w_C separately, we solve the equality system (97), and thus determine a *joint* update direction for $f(X)$, see Appendix B.4.

Solutions to the involved linear equality systems are found by the bi-conjugate gradients stabilized method of MATLAB and using a preconditioning matrix which consists of the diagonal entries of $\overline{A}_{dd}$ only.

4.8 Line Search

The descent direction $W^{(i)}$ computed in the previous section decreases $f(\varphi(t, X^{(i)}, W^{(i)}))$ for some $t > 0$ if $\nabla_T f \neq 0$. To determine an approximate optimal step scaling, we perform a line search based on Wolfe's Rule (Bonnans et al 2003). We respect the componentwise non-negativity constraint on d by using the quality function $q(X)$ as defined in (46),

$$t^{(i)} \approx \arg\min_{t>0} q(\varphi(t, X^{(i)}, W^{(i)})) \tag{98}$$

and then choose $X^{(i+1)}$ as defined in (49).

4.9 Stopping Criterion

Within a multiscale level, the variables update is iterated as long as the relative improvement of the objective function value is sufficient, i.e.

$$\frac{f(X^{(i-1)}) - f(X^{(i)})}{f(X^{(i-1)})} > \epsilon. \tag{99}$$

Throughout this work we choose $\epsilon = 10^{-5}$.

4.10 Multiscale Framework

For handling large displacements we employ the state-of-the-art coarse-to-fine approach known from optical flow estimation (see e.g. Brox et al (2004)). On each resolution level, the single scale update described in Algorithm 1 is initialized by the result of the next coarser scale and is used for warping the previous to the current frame. The prediction step also provides the initialization for the coarsest scale which implicitly removes most of the large displacements.

Warping and image down-scaling use cubic spline interpolation for high accuracy. The non-negativity of the depth map values are preserved by bi-linear interpolation.

5 Experiments

The experimental section is organized as follows. We introduce the considered image sequences and their properties in Sect. 5.1. Algorithm implementation and parameter values are discussed in Sect. 5.2. Section 5.3 demonstrates the importance of the temporal smoothing prior.

While synthetic data comes with accurate ground truth, we employ two established geometric reconstruction approaches as a baseline for the performance of our method on real image sequences. Both can resort to more information than the proposed recursive monocular approach and thus are expected to provide more accurate results. A reference depth estimation is provided by stereo reconstruction methods in Sect. 5.4. Egomotion accuracy is compared to the camera track computed batch-like by a bundle adjustment approach in Sect. 5.5.

5.1 Data Sets

Freely available databases with automotive image data including the KITTI[2] and the enpeda[3] benchmark con-

[2] http://www.cvlibs.net/datasets/kitti/
[3] http://www.mi.auckland.ac.nz/

tain only few image sequences which (approximately) fulfill the constraint that the scene is static.

The novel HCI database[4] (Meister et al 2012) contains image sequences recorded from a car driving in an everyday environment. It was compiled for the evaluation of stereo reconstruction algorithms. Thus, two camera recordings from a stereo rig are available, of which we only used the left one as input to our algorithm. Here we resort to a publicly available sub-set[5] of five image sequences labeled *Avenue*, *Bend*, *City*, *Parking* and *Village* showing almost static real-life scenes. Each has up to 400 gray-value frames of size 656 px× 541 px, sampled with 25 frames/s and an intensity resolution of 12 bit.

Furthermore, we processed two sequences which considerably violate the static-scene assumption due to moving objects but come up with other interesting features. The *Junction* sequence from the database introduced above is 863 frames long and contains a very challenging 90° turn.

Sequence 2 of set 2 of the *enpeda* project (denoted as *enpeda-2-2*) provides reliable ground truth due to its synthetic nature. It consists of 396 rectified stereo image pairs, each with a resolution of 640 px×480 px and 12 bit, see Vaudrey et al (2008) for details.

5.2 Algorithm Details

5.2.1 Implementation and Runtime

The results presented in this section were obtained with our research implementation of the method described in Sects. 2–4. It is mostly MATLAB-based and is not parallelized. Consequently, the computing time in full resolution (656 px × 541 px) is about 2 minutes per frame.

However, the regular grid structure of the problem and the reduction to well understood numerical methods (linear equation solver) renders the method a suitable candidate for an efficient computation, e.g. on a (embedded) GPU.

5.2.2 Parameters

For all real and synthetic sequences we choose the same parameter set: temporal smoothness parameters $\sigma_R = 2$, $\sigma_h = 4$ and $\sigma_d = 10^6$ of camera rotation, translation and depth map, respectively, and spatial smoothness prior $\sigma_s = 10^6$. The multiscale framework uses 13 levels, where the resolution decreases by a factor of $\sqrt{2}$ between two adjacent levels. The sequence *enpeda-2-2* requires a slightly larger smoothing parameter for the optical flow (see (7)) due to its synthetic nature and we set $\rho = 3$ px.

Only minimal information is provided by the state initialization: we choose a constant depth $d^0(x) = 40$ for all $x \in \Omega$ and a forward movement of the camera, i.e. $R^0 = I$ and $h^0 = v_0(0, 0, 1)^\top$, where v_0 is a guess on the (scalar) vehicle velocity. Variances σ_d^0 and Σ_C^0 are set to large values.

5.3 Temporal Filtering

5.3.1 Relevance

The prediction prior on depth map and camera motion exploited in (17) and (25), respectively, are essential for robust depth estimation. Figure 10 presents depth maps estimated with smoothing parameters σ_R, σ_h, σ_s chosen considerably larger than the standard (see Sect. 5.2.2) which renders the temporal smoothness prior ineffective. The differences to an estimation with effective prior are striking and show that just relying on the observations yields corrupted estimates.

5.3.2 Uncertainty Reduction

Figure 11 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 level lines tend to the left and thus demonstrate that our approach significantly reduces the approximated uncertainty of the depth estimation within a period of about 25 frames.

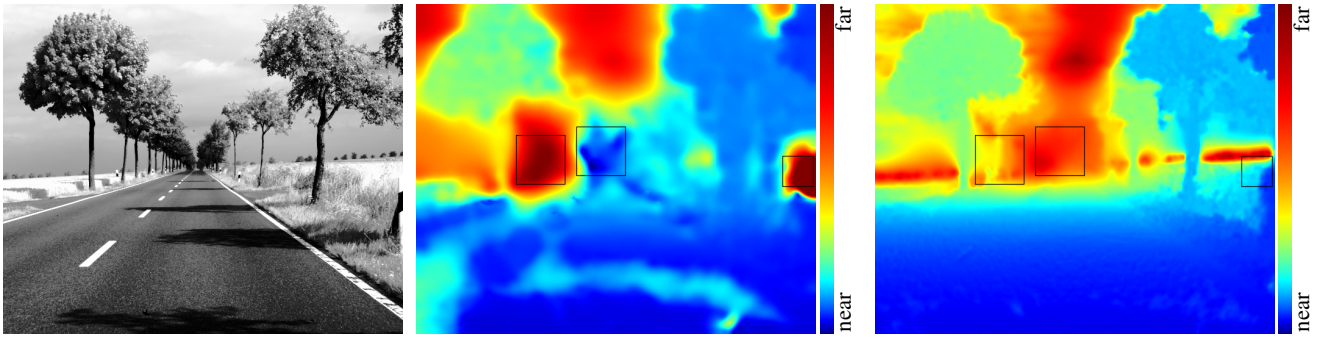
5.4 Depth Map Evaluation

We start with a detailed description of the acquisition of the reference depth maps in Sect. 5.4.1. Section 5.4.2 discusses the accuracy of the proposed monocular approach by means of exemplary results. The employed statistical error measure is motivated and quantitative results are discussed in Sect. 5.4.3.

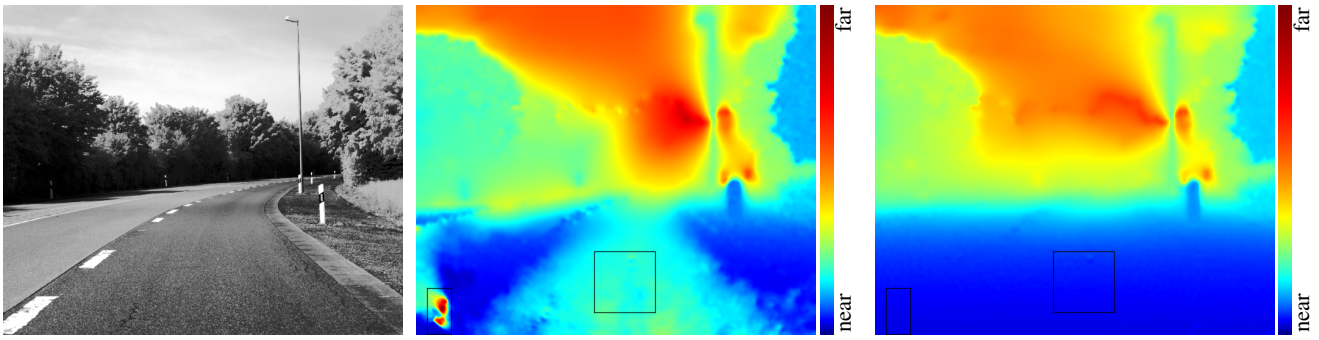
5.4.1 Reference Depth Map $\overline{d}$

The synthetic image sequence *enpeda-2-2* comes with accurate and almost dense ground truth. In contrast, the real image sequences lack ground truth but include rectified image pairs and we consider the result of a stereo method as reference depth map $\overline{d}$. Stereoscopic methods can be assumed to be much more accurate (see Sect. 3.1.2) as they can resort to much more information in the image pair.

[4] http://hci.iwr.uni-heidelberg.de/Benchmarks/
[5] http://hci.iwr.uni-heidelberg.de/VSFM/



(a) *Avenue* sequence, frame 53 (b) ineffective temporal smoothing prior (c) effective temporal smoothing prior



(d) *Bend* sequence, frame 98 (e) ineffective temporal smoothing prior (f) effective temporal smoothing prior

Fig. 10 Importance of temporal regularization (prediction prior) of depth map and camera motion demonstrated for the (a)–(c) *Avenue* and (d)–(f) *Bend* sequence. **(a), (d)** Image frame, **(b), (e)**, large σ_R, σ_h, σ_d render the temporal smoothness prior ineffective, **(c), (f)** with effective temporal smoothness prior. Exploiting temporal context through the prior is essential, in particular where only few information is available. With ineffective temporal smoothness, regions near the epipole (near image center) show severe under- (dark blue) and over-estimation (dark red) of depth. Also low-textured regions such as sky and road in (d) show this effect. Comparing to the image frames, results with effective temporal smoothing appear much more coherent.

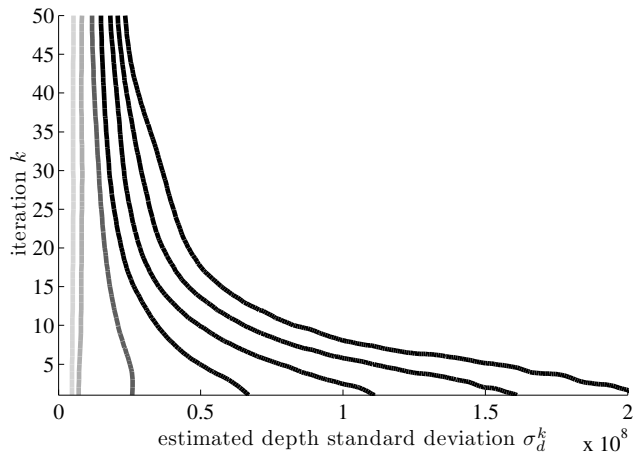


Fig. 11 Histogram $\widehat{\mathrm{pdf}}_\Omega(\sigma_d^k)$ of depth standard deviation σ_d^k over iteration k, visualized as contour plot. The lines uniformly tend to the left and thus indicate that estimation uncertainty is effectively reduced, despite online processing with only two frames at each instant of time.

Choice of reference stereo implementation. As reference we consider three *freely available implementations* of stereo algorithms:

Szeliski et al.: The authors' implementation[6] of Szeliski et al (2008) included in the Middlebury MRF Library.

Rhemann et al.: The authors' implementation[7] of Rhemann et al (2011), ranked 17 in the Middlebury Stereo Evaluation (as on March 22, 2012).

Geiger et al.: The LIBELAS library[8] by the authors of Geiger et al (2010) and developed for large-scale disparities.

Given the disparity measures $\overline{D}(x)$ of an arbitrary stereo approach as well as the focal length ($f = 25\,\mathrm{mm}$), stereo baseline ($b = 30\,\mathrm{cm}$) and pixel size ($p = 16\,\mu\mathrm{m}$), we can compute a calibrated depth map $\overline{d}(x)$ by

$$\overline{d}(x) := \frac{fb}{p}\overline{D}^{-1}(x). \tag{100}$$

[6] http://vision.middlebury.edu/MRF/code/
[7] http://www.ims.tuwien.ac.at/research/costFilter/
[8] http://www.rainsoft.de/software/libelas.html

Exemplarily, we provide in Fig. 12 for frame 100 of the *Avenue* sequence for each method the depth calibrated in meters. From the visual comparison of the depth maps with the corresponding image frame, it becomes apparent, that only the method by Geiger et al. is able to accurately reconstruct depth both at low and high distances with many details for real outdoor data. Thus for the evaluation of the depth map estimated by our method, we only consider the results of Geiger et al. as reference.

Unknown global scale s. For monocular approaches it is only possible to derive the scene geometry up to an unknown scale $s \in \mathbb{R}$. Thus, for each frame we estimate the scale s between d and reference $\overline{d}$ using a robust estimator which we detail in Appendix C.1. The resulting depth maps $s \cdot d$ are approximately calibrated to metric units.

5.4.2 Qualitative Evaluation

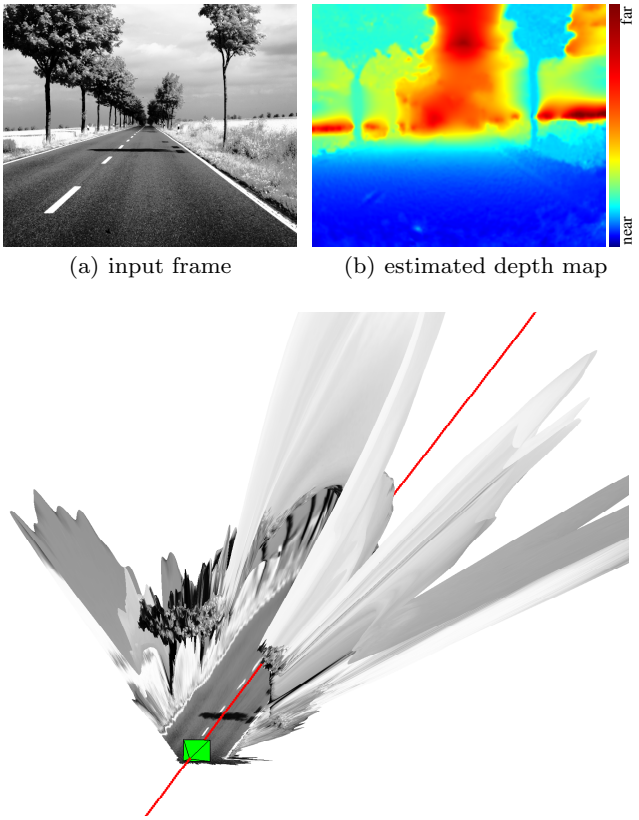
Static Real Sequences. Figures 1, 13 and 14 show representative reconstructions for sequences *Bend*, *Avenue* and *City*, respectively. Comments are given in the captions. All computed depth maps are encoded using a non-linear color scale. Histograms of image frames are equalized for visualization with improved contrast.

In Fig. 15, we provide the scaled depth map $s \cdot d(x)$ as well as the difference maps $s \cdot d(x) - \overline{d}(x)$, both calibrated to meters. The exemplary results demonstrate that the coarse structure of the scene can be reconstructed correctly. Objects can be resolved with a similar accuracy as by the stereo method if sufficient depth information is provided by the camera geometry, see Fig. 16(a). It deteriorates in the vicinity of the epipole which confirms the theoretical discussion in Sect. 3.1.2.

In region lacking texture information, both monocular and stereo methods have to apply some kind of prior to determine depth. Our monocular approach smoothly interpolates there, while stereo tends to imply large or infinite depth here which appears more crisp in the visualization, see Fig. 16(b).

Regions which do not permit accurate estimation due to geometry or lack of texture are annotated accordingly in the covariance map provided with the depth map, see Fig. 16 (right column), Fig. 6 and Sect. 3.4.1 for the definition. Hence, the depth information can be considered according to its accuracy by higher-level reasoning steps.

Dynamic Real Sequence. The *Junction* scene possesses several challenging properties, see Fig. 17. A car crosses the camera view around frame 595 which violates the assumption that the scene is static. Depth estimations are considerably wrong in this image region only and are corrected within few frames after the moving object has left the view. This demonstrates the robustness of the approach w.r.t. model violations.



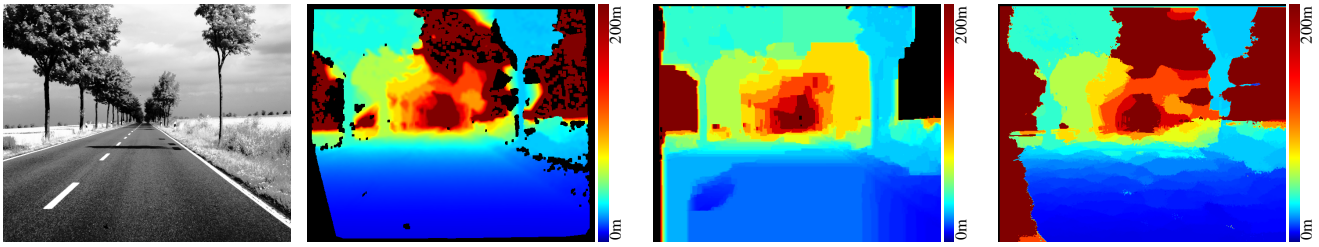
(a) input frame (b) estimated depth map

(c) reconstructed scene and camera track

Fig. 13 *Avenue* sequence: **(a)** input frame, **(b)** estimated depth map, color encoded, **(c)** triangulation of the scene, camera pose (green) and camera track (red). The *homogeneous* sky regions *between* the highly textured trees are naturally assigned to the trees (in terms of depth) due to the interpolation property of the variational formulation. Reconstruction accuracy is limited near the epipole (middle) as discussed in Sect. 3.1.1.

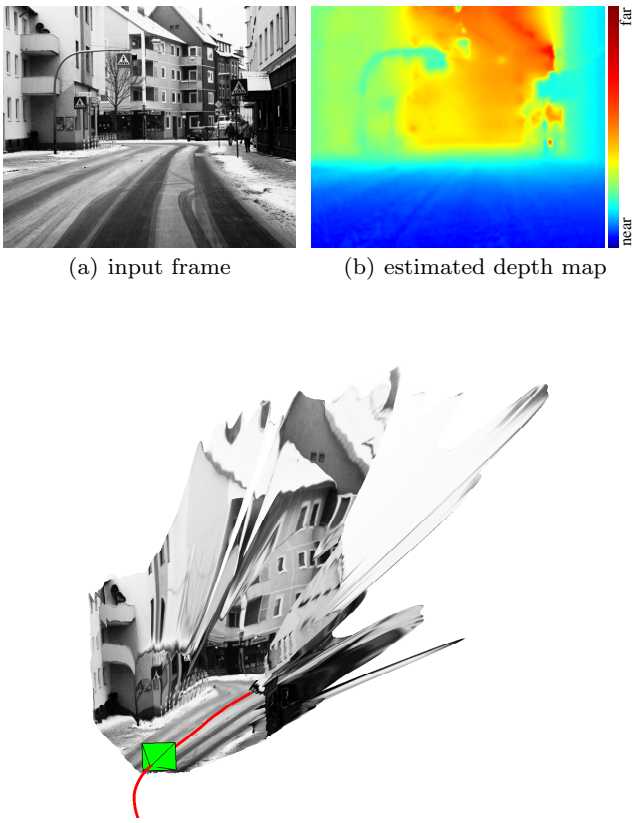
The method provides accurate depth information even during a sharp 90° turn around frame 750. However, after the turn (frame 860), the monocular depth map is distorted in the lower part of the image as the road surface only provides few information relevant for depth estimation. Similar to the situation after the initialization in frame 0, this region is expected to be corrected within the upcoming frames. Depth in the remaining image region is reconstructed with similar accuracy as the reference method.

Synthetic dynamic sequence. The synthetic image sequence *enpeda-2-2* comes with accurate ground truth. Representative results are shown in Fig. 18. Frame 380



(a) frame (b) depth estimated by Geiger et al (2010) (c) depth estimated by Szeliski et al (2008) (d) depth estimated by Rhemann et al (2011)

Fig. 12 **(a)** One original (left) image frame of the *Avenue* sequence and **(b)**–**(d)** depth maps estimated by three state-of-the-art stereo methods, calibrated to meter. Black regions indicate lack of depth information. Only the approach by Geiger et al. is able to accurately reconstruct depth both at low and high distances with many details for this data. See Sect. 5.4.1 for details.



(a) input frame (b) estimated depth map

(c) reconstructed scene and camera track

Fig. 14 *City* sequence: **(a)** input frame, **(b)** estimated depth map, color encoded, **(c)** triangulation of the scene, camera pose (green) and camera track (red). The approach is robust towards (minor) violations of the static scene assumption by moving pedestrians (right).

demonstrates the possible high accuracy of depth reconstruction. The impact of the violation of the static scene assumption is demonstrated in frame 100 and causes only *local* depth map errors.

5.4.3 Statistical Evaluation

Error measure. In order to compare measurements from different image positions x and depth d, we relate them with their expected error $\sigma_g(x,d)$ as defined in (14). Given a depth map $d(x)$ and a reference solution $\overline{d}(x)$, we define the following position-wise error measure:

$$e_d(x) := \frac{f}{p}\frac{|d(x)-\overline{d}(x)|}{\sigma_g(\overline{d}(x),x)} \tag{101}$$

with focal length f and pixel size p as in (100). Furthermore, we define a summarizing error measure as

$$\varepsilon_d := \sqrt{\mathcal{E}\left\{e_d^2(x)\right\}} \tag{102}$$

where $\mathcal{E}\{\cdot\}$ denotes the expectation value over all pixels and all frames of an image sequence.

Remark 1 The proposed measure can be interpreted as the expected error of the involved optical flow estimation process, expressed in pixels units. For the special case of a *stereo line-up*, it approximates the disparity difference $|D-\overline{D}|$ common for evaluating stereo algorithms, e.g. in the Middlebury (Scharstein and Szeliski 2002) and KITTI (Geiger et al 2012) benchmarks.

Proof We consider the reference solution as correct solution, i.e. $\hat{d}=\overline{d}$ in assumption (12). Then the measure

$$\varepsilon_d \overset{(14)}{=} \frac{f}{p}\sigma_u \underbrace{\sqrt{\mathcal{E}\left\{\frac{(d(x)-\overline{d}(x))^2}{\sigma_d^2(\overline{d}(x),x)}\right\}}}_{\approx 1} \approx \frac{f}{p}\sigma_u \tag{103}$$

approximates the expected error of the optical flow measurement. The leading factor scales the displacements from normalized coordinates to more descriptive pixel units.

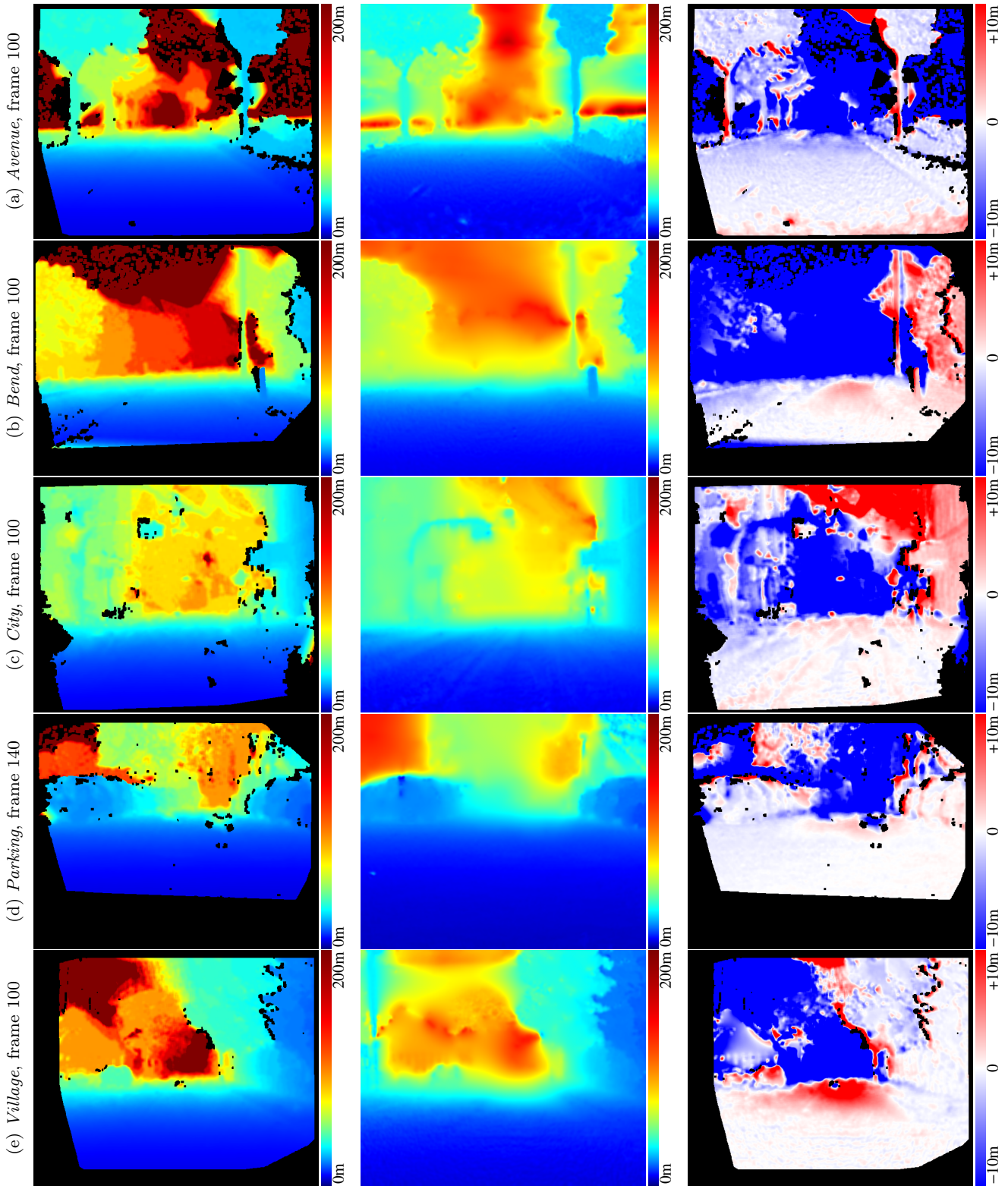


Fig. 15 Comparison between a stereo and our monocular method. **Rows:** one frame of a real static sequences. **Left column:** reference depth map $\overline{d}(x)$ estimated by the stereo approach by Geiger et al (2010). **Center column:** depth map $s \cdot d(x)$ estimated by the proposed monocular method. The unknown global scale s was estimated for each scene as described in Appendix C.1. All depth maps use the same color encoding. **Right column:** point-wise difference $s \cdot d(x) - \overline{d}(x)$, calibrated to meters (clipped). Red and blue indicate over- and underestimation, respectively, of monocular relative to stereo depth. Black pixels lack stereo depth information. The coarse structure is reconstructed correctly and remaining disagreements can be explained by the entirely different camera setups or interpolation schemes, see discussion in Sect. 5.4.2.

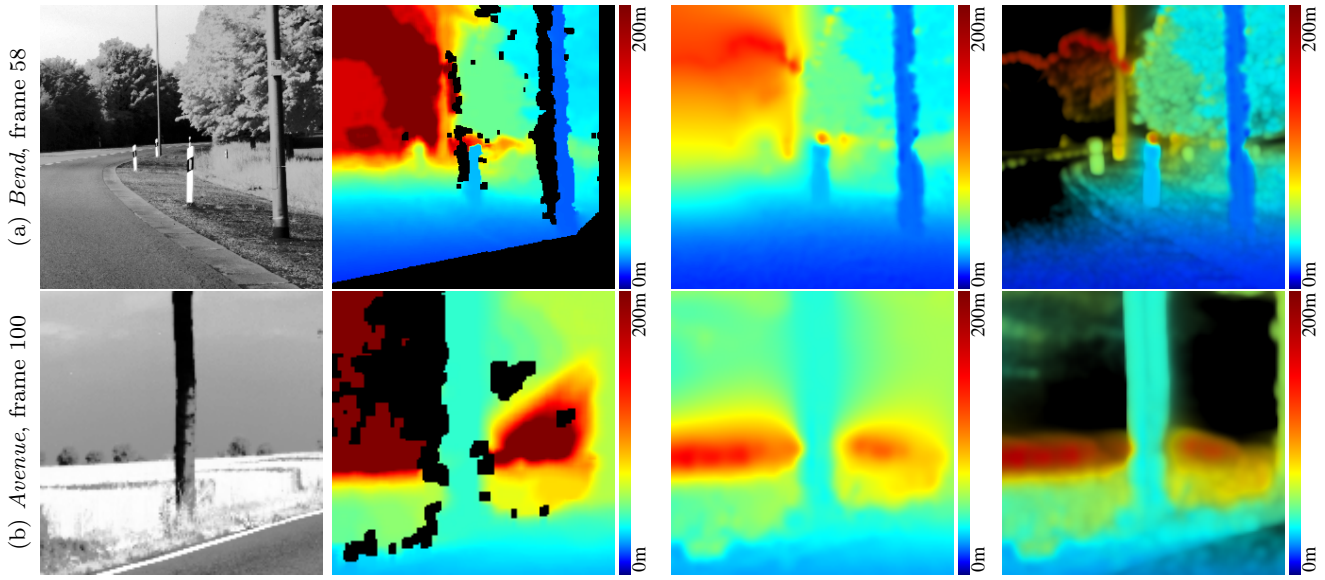


Fig. 16 Detailed comparison between a stereo and our monocular method. **Columns** (left to right): image frame (detail); depth map $\overline{d}$ estimated by the stereo approach by Geiger et al. Black pixels lack stereo depth information; depth map $s \cdot d(x)$ resulting from the monocular method proposed in this work; depth map $s \cdot d(x)$ (color) and estimated variance $\sigma(x)$ (brightness, black = large variance) of the monocular depth map. **(a)**: Stereo and monocular approaches show a similar ability to separate the right lamp post from the background. The nearer the object is to the epipole (left in image) the less accurately it can be resolved by the monocular approach. This inaccuracy is inherent in this camera setup and is annotated accordingly in the covariance map. **(b)**: Stereo and monocular approaches rely on the existence of textured regions to detect distinct depth edges (trunk in front of horizon). In un-textured regions (sky), most stereo methods assume large depth (encoded black) while the proposed monocular method interpolates from neighboring regions, however marks the value as uncertain. See Sect. 5.4.2 for further discussion.

For a stereo setup we have $\sigma_g(d,x) = \sigma_{g,s}(d,x)$ (see (15)) and assuming $d(x) \approx \overline{d}(x)$, the pixel-wise error measure can be reformulated as

$$e_d(x) \overset{(101)}{=} \frac{f}{p} \frac{|d(x) - \overline{d}(x)|}{\sigma_g(\overline{d}(x), x)} \approx \frac{f}{p} \frac{|d(x) - \overline{d}(x)|}{\sqrt{\sigma_g(d(x), x)\sigma_g(\overline{d}(x), x)}}$$

$$\overset{(16)}{=} \frac{fb}{p} \frac{|d(x) - \overline{d}(x)|}{d(x)\overline{d}(x)} \overset{(100)}{=} |D(x) - \overline{D}(x)|$$

which confirms the choice of e_d. Vice versa, the disparity difference can be interpreted as weighting the *depth difference* according to the expected error $\sigma_{g,s}$. □

Thus, ε_d is a sensible measure to assess the quality of the proposed method. In combination with the geometric error model introduced in Sect. 3.1.2, it is possible to predict the expected depth error for other scenarios.

Evaluation basis. For each frame, the unknown common scale of depth map d and translation h were corrected as described in Appendix C.1. For the calculation of σ_g, the monocular egomotion and the stereo depth map was used if no ground truth was available. Pixels without a reference value or containing distinct dynamic components were excluded. For sequence *enpeda-2-2* ground truth provides pixel-wise scene flow information while for the *Junction* sequence, we completely excluded affected frames 584–620 and 644–670. The proposed method requires up to 30 frames to compensate the weak initialization and we also do not include them in the analysis. In total, more than $638 \cdot 10^6$ depth measurements were considered.

Results and discussion. Figure 19 plots the empirical versus the predicted depth error and confirms the approximate linear relation between them which we assumed in the definition of the error measure (101)–(102).

Table 1 lists the computed error measures ε_d for each of the considered image sequences. For most sequences the error stays below 1 px. A closer analysis in Fig. 20 shows that for the *Junction* sequence major errors are located mostly near the lower image boundary where the road surface does not provide any hints for monocular methods, see also Fig. 17. The same holds for the *Avenue* sequence, where also large differences occur near the horizon, see also Fig. 16(b).

5.5 Egomotion Evaluation

Reference camera motion is obtained by ground truth information or by processing the monocular data with

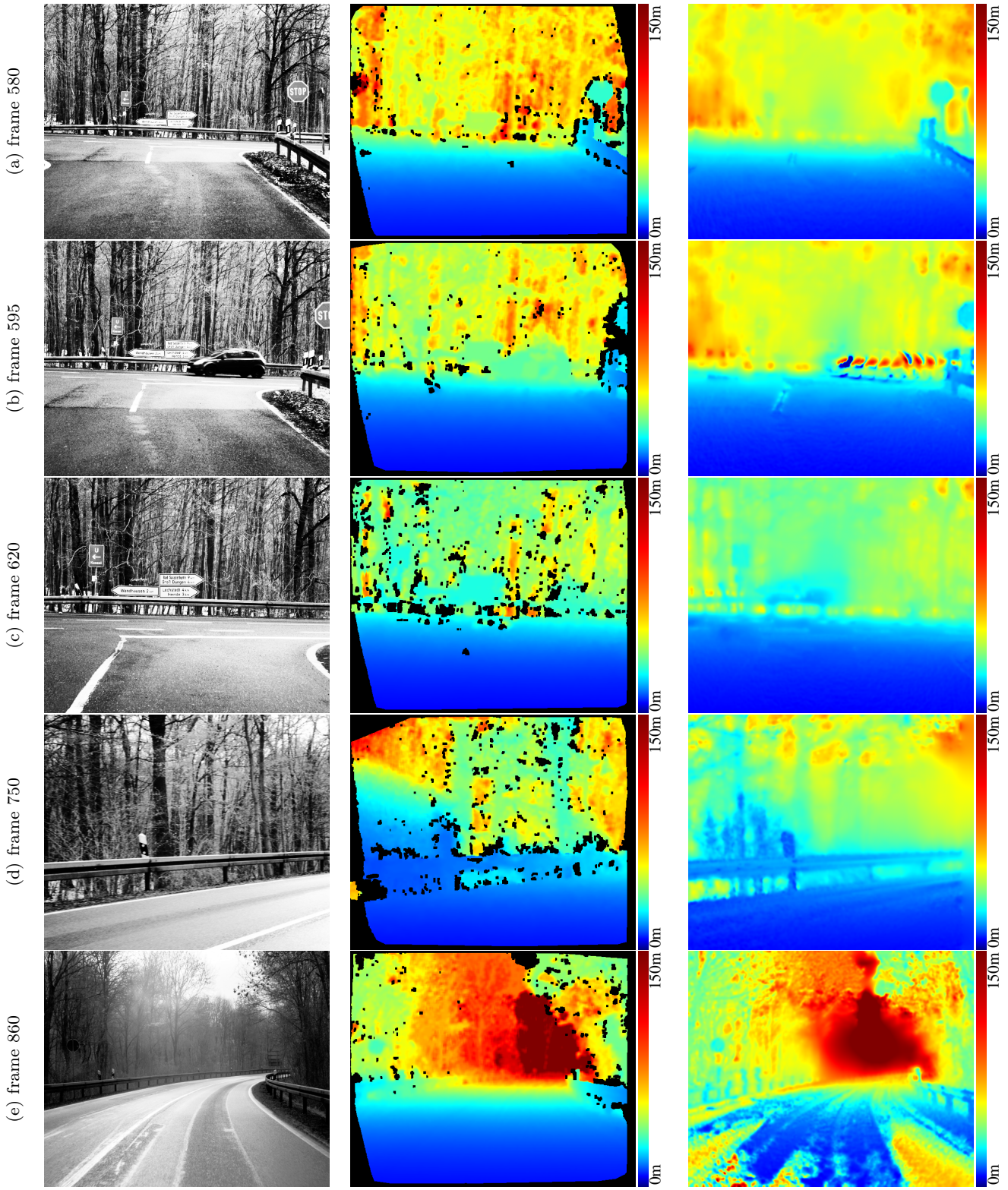


Fig. 17 Five frames from the *Junction* sequence which is challenging due to moving objects and a 90°-turn. **Columns** (left to right): image frame; stereo depth map; calibrated monocular depth map. After **(a)** frame 580 a car crosses the scene. **(b)** The monocular depth map is disturbed as the approach tries to explain the optical flow by a static depth map. Regions outside the dynamic regions are hardly affected, which demonstrates the robustness against model violations. **(c)** The distortions are corrected when the car has left. **(d)** During the following turn, the monocular depth maps shows similar accuracy as the stereo map despite large horizontal displacements. **(e)** Right after the turn, the scene is reconstructed mostly correct, but disturbances in the lower region can be observed. Here, the low-textured road surface does not provide any correspondences *along* the lane which corresponds to depth in this case while the stereo estimation can rely on *horizontal* gradients.

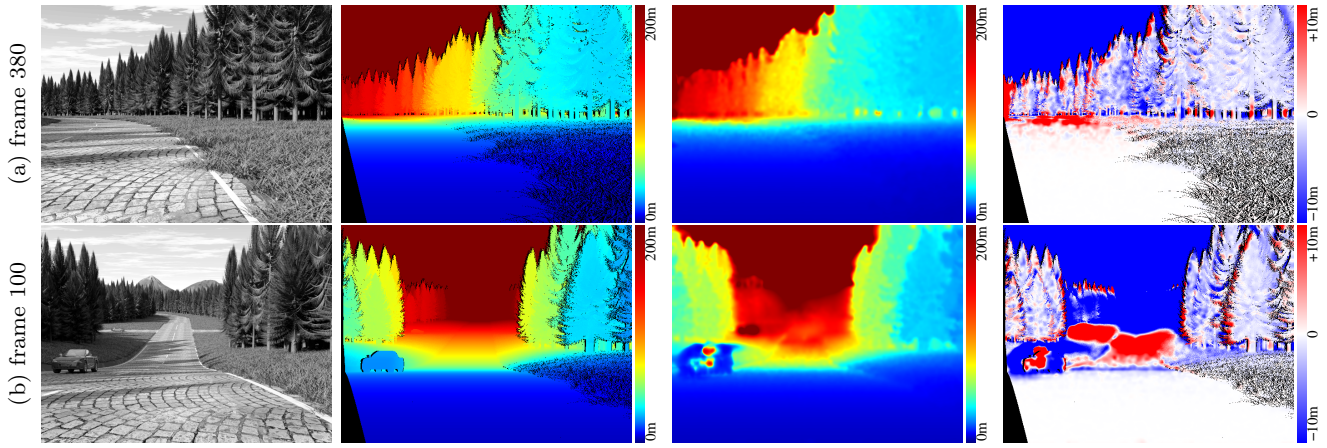


Fig. 18 Synthetic sequence *enpeda-2-2*. **Columns** (left to right): left image frame; ground truth depth map $\overline{d}(x)$; scaled monocular depth map $s \cdot d(x)$; difference $s \cdot d(x) - \overline{d}(x)$. Unknown scale s was estimated as described in Appendix C.1. Black pixels indicate missing reference values. **(a)** Frame 380 shows the potentially high precision of the monocular approach. **(b)** Two moving objects (cars) which clearly violate the static scene assumption and lead to deteriorated depth estimations in this regions. Remaining image regions are not affected, which demonstrates the robustness against model violations. The epipole is near the image center where depth measurement is almost impossible for the monocular setup.

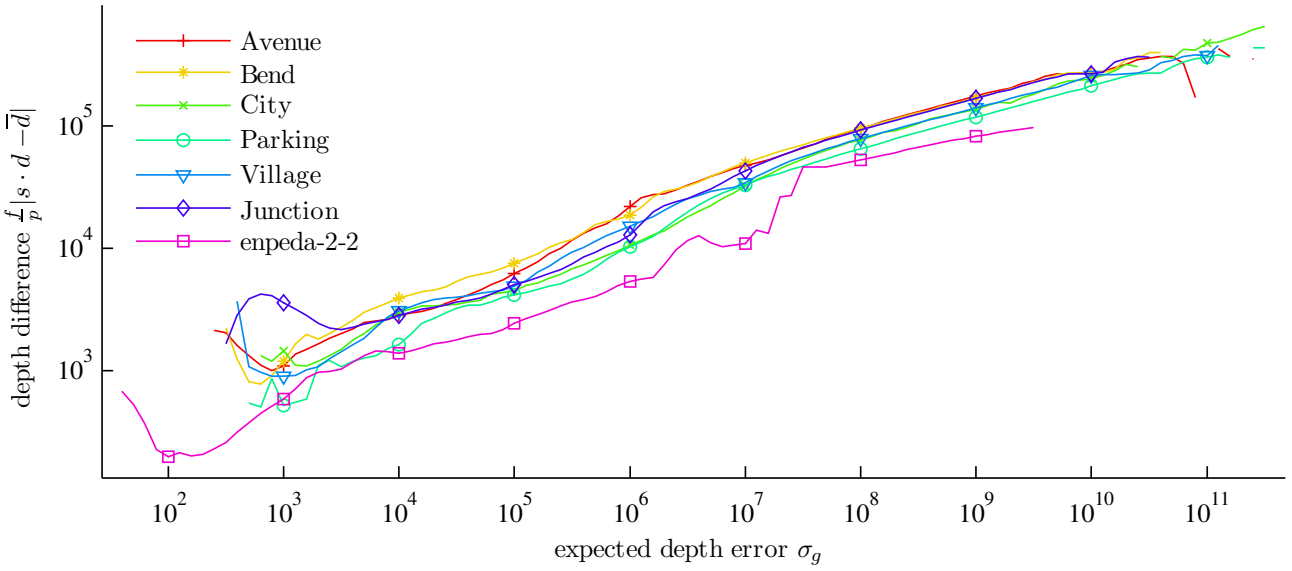


Fig. 19 Log-log-plot of the actual empirical depth error $fp^{-1}|s \cdot d - \overline{d}|$ versus the error prediction σ_g based on the model (14). Samples were gained from most of the pixels and frames of the sequence and summarized by collecting them in σ_g-bins (width 0.1 in $\log_{10}$-scale). The plot confirms the linear relation between the predicted and actual depth error assumed in the definition of the error measure ε_d, see Sect. 5.4.3 for details.

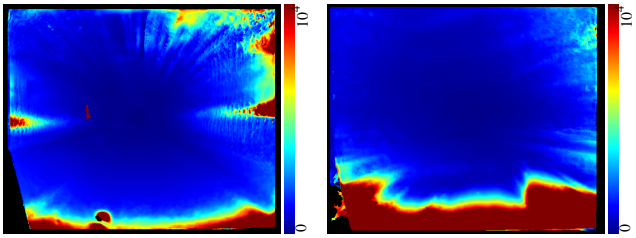
a bundle adjustment method, see Sect. 5.5.1. The integrated camera track and the frame-wise differential motion is evaluated in Sect. 5.5.2 and Sect. 5.5.3, respectively.

5.5.1 Reference Camera Track $(\overline{R}, \overline{h})$

For the synthetic image sequence *enpeda-2-2* we resort to the ground truth camera motion as reference.

To evaluate egomotion estimation of the real image sequences, we measure a reference camera track using the freely available Voodoo Camera Tracker[9] (VCT) which implements a bundle adjustment method and is based on tracking sparse image features. We manually set the internal camera parameters and chose the following parameters: free move mode, Förstner detector, cross-correlation for correspondence analysis, fixed focal length.

[9] http://www.digilab.uni-hannover.de/docs/manual.html, v1.2.0b



(a) *Avenue*: $\varepsilon_d(x)$ (b) *Junction*: $\varepsilon_d(x)$

Fig. 20 Pixel-wise disparity estimation difference $\varepsilon_d(x)$ over all frames for a sequence. The spatial location of large deviations help to identify the source of the error.

scene	ε_d	ε_R	ε_h
Avenue	2.18 px	0.0155°	2.75 %
Bend	0.94 px	0.0143°	1.54 %
City	0.66 px	0.0169°	15.29 %
Parking	0.33 px	0.0107°	9.14 %
Village	0.79 px	0.0171°	3.74 %
Junction	6.11 px	0.0264°	43.04 %
enpeda-2-2	0.98 px	0.0155°	9.15 %

Table 1 The table shows the quantitative evaluation for the considered scenes in comparison to more accurate reference methods (stereo, bundle adjustment) or ground truth. Disparity estimation difference ε_d (in pixel) is crucial for the quality of depth estimation, see Sect. 5.4.3 for details. For assessing the quality of egomotion estimation, we provide the mean difference of the rotational component (ε_R, in degree/frame) and the translational component relative to reference speed (ε_h, in percent/frame), see Sect. 5.5.

Note that bundle adjustment computes the camera poses *jointly* in a batch-processing manner and therefore can be expected to return precise results. In contrast, the proposed monocular approach estimates the trajectory by recursively integrating up camera motion and thus is subject to integration errors.

Unknown global scale. Before comparison, the unknown global scale between the estimated and the reference track is approximated by a least-square match of the camera trajectories. Furthermore, the reference track is normalized to length 1.

5.5.2 Integrated Track

Figure 21 demonstrates for five of the seven sequences a remarkable agreement of our *monocular online* estimates with VCT. Remaining differences can be explained by a limited sensitivity towards acceleration along camera principal axis, which is crucial for the *Junction* sequence. However, this inaccuracy is correctly reflected by the estimated camera standard deviation, see Fig. 7.

5.5.3 Statistical Evaluation

Error measures. For each frame, we compare the camera motion estimated by our monocular approach (R, h) and the reference method $(\overline{R}, \overline{h})$ and denote their Lie matrix representation (see (1)) by Q and $\overline{Q}$, respectively. Based on the motion difference $Q_e := Q^{-1}\overline{Q}$ with elements (R_e, h_e), we define the frame-wise error measurements

$$e_R := \cos^{-1}((\operatorname{tr} R_e - 1)/2). \tag{104}$$

$$e_h := \frac{\|h_e\|}{\|\overline{h}\|}. \tag{105}$$

The translation measure is normalized w.r.t. $\|\overline{h}\|$ to cancel out the unknown global scale. We define the averages over all frames of a sequence as ε_R and ε_h, respectively.

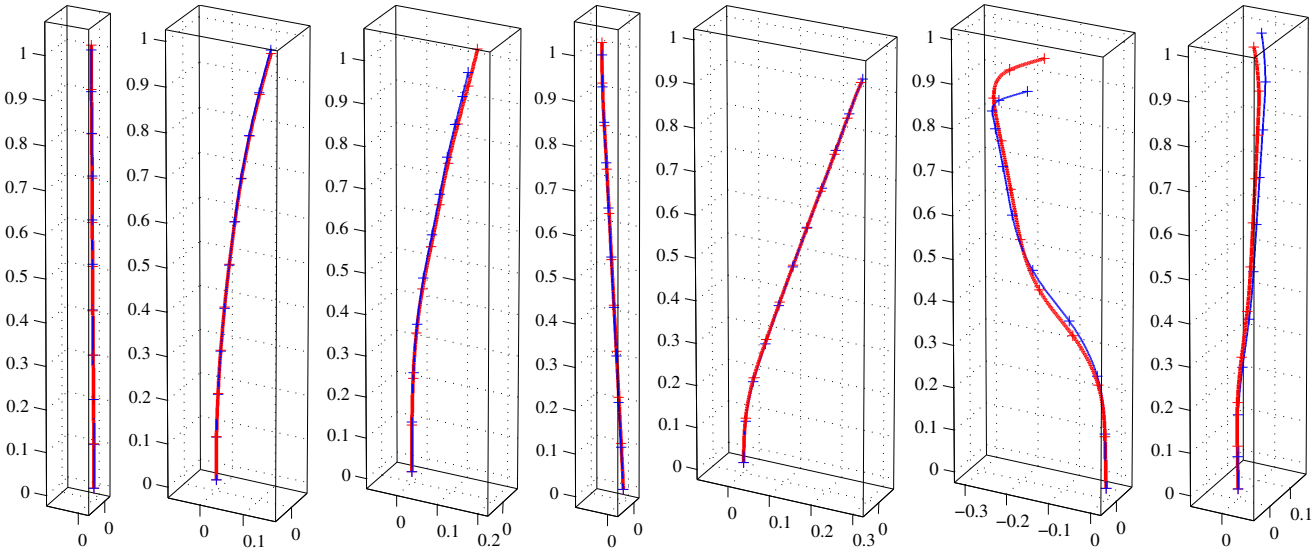
Results and discussion. Table 1 summarizes the egomotion quality for each considered sequence. While the rotational component is very accurate for all cases, the translational measure is high for those sequences which show also large changes in the velocity *along* the camera view, see also Fig. 21. This property is also confirmed by the estimated camera motion uncertainty which is provided for every frame, see Sect. 3.4.1 and Fig. 7.

6 Conclusion and Further Work

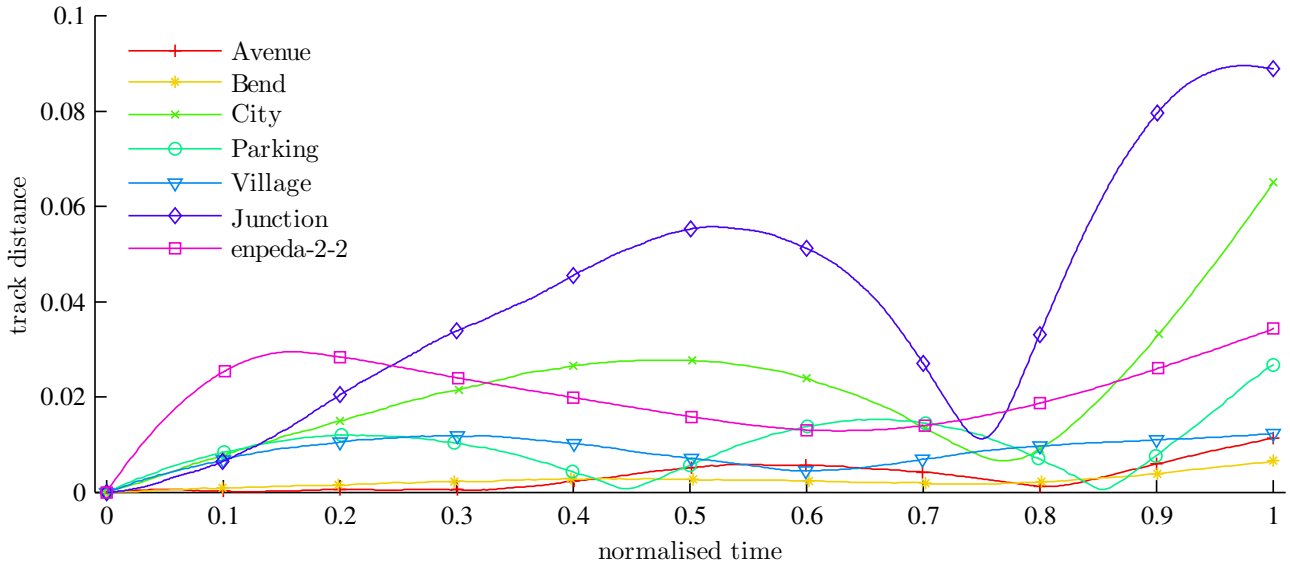
We presented an approach to the estimation of dense scene structure and camera motion from monocular image sequences, taken from a camera positioned inside a fast moving car. The approach optimizes the tradeoff between model expressiveness and computational efficiency. In particular, it works in an online two-frame mode. A sound mathematical framework was presented for the joint update of camera pose and depth map which respects the manifold structure for accuracy, resorts to complete second-order information for efficiency, and guarantees a decrease of the objective function until convergence.

Experiments demonstrated that the reconstruction quality of depth and camera motion is similar to reference methods which can resort to considerably more information. Annotation by uncertainty estimations help to identify components which are inherently inaccurate due to the camera setup or lack of correspondences.

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



(a) Avenue (b) Bend (c) City (d) Parking (e) Village (f) Junction (g) enpeda-2-2



(h) Euclidean distance between tracks over time

Fig. 21 Comparison of the estimated camera tracks to a bundle adjustment method or ground truth (*enpeda-2-2* only). **(a)**–**(g)** Camera tracks estimated by the proposed monocular recursive approach (red, thick line) and the reference track (blue, thin line) for the considered sequences. Markers indicate points equidistant in *time*. The trajectories of the reference solution were normalized to length 1. The monocular tracks were scaled such that they minimize the least-squares Euclidean distance. No rotational fitting was applied. **(h)** Euclidean distances (relative to reference track length) between the trajectory positions over time (normalized to $[0, 1]$). The recursively estimated track agrees well with the reference track for most tracks. Estimated velocities *along* camera view differ and cause divergence of the tracks. This is inevitable for this camera setup and agrees with the estimated camera pose variance, see Fig. 7.

A Geometric Model

A.1 Implicit Epipolar Constraint

Any point pair $(x, x') = (x, x - u(x))$ as defined by (5) is connected via the essential matrix $E := [h]_\times R$ through the constraint (11) (Hartley and Zisserman 2000) as shown here:

$$\begin{pmatrix} x' \\ 1 \end{pmatrix}^\top E \begin{pmatrix} x \\ 1 \end{pmatrix} \tag{106}$$

$$\overset{(4)}{=} \begin{pmatrix} x' \\ 1 \end{pmatrix}^\top ([h]_\times R) d^{-1}(x) R^\top \left(d'(x') \begin{pmatrix} x' \\ 1 \end{pmatrix} - h \right) \tag{107}$$

$$= (d^{-1}(x) d'(x')) \begin{pmatrix} x' \\ 1 \end{pmatrix}^\top [h]_\times \begin{pmatrix} x' \\ 1 \end{pmatrix} = 0 \tag{108}$$

Here we used the equalities $[v]_\times v = 0$ and $v^\top [w]_\times v = 0$ for all $v, w \in \mathbb{R}^3$.

B Optimization on $\mathsf{SE}(3) \times \mathbb{R}^n$

B.1 Manifold $\mathsf{SE}(3)$: Definitions

B.1.1 Mappings Exp *and* Log

For the Lie group SE_3 and the associated Lie algebra se_3, the exponential map $\mathrm{Exp} : \mathsf{se}_3 \mapsto \mathsf{SE}_3$ and its inverse (within the neighborhood of 0) $\mathrm{Log} : \mathsf{SE}_3 \mapsto \mathsf{se}_3$ can be expressed explicitly:

$$\mathrm{Exp}\left(\begin{pmatrix} [\omega]_\times & v \\ 0^\top & 0 \end{pmatrix} \right) = \begin{pmatrix} R(\omega) & P(\omega) v \\ 0^\top & 1 \end{pmatrix} \tag{109}$$

$$R(\omega) = I + \frac{\sin(\|\omega\|)}{\|\omega\|} [\omega]_\times + \frac{1 - \cos(\|\omega\|)}{\|\omega\|^2} [\omega]_\times^2 \tag{110}$$

$$P(\omega) = I + \frac{1 - \cos(\|\omega\|)}{\|\omega\|^2} [\omega]_\times + \frac{\|\omega\| - \sin(\|\omega\|)}{\|\omega\|^3} [\omega]_\times^2 \tag{111}$$

$$\mathrm{Log}\left(\begin{pmatrix} R & h \\ 0^\top & 1 \end{pmatrix} \right) = \begin{pmatrix} [\omega]_\times(R) & P^{-1}(\omega) h \\ 0^\top & 0 \end{pmatrix} \tag{112}$$

$$[\omega]_\times(R) = \begin{cases} 0 & \text{if } \theta(R) = 0 \\ \frac{\theta(R)}{2 \sin \theta(R)} (R - R^\top) & \theta(R) \neq 0 \end{cases} \tag{113}$$

$$\theta(R) = \cos^{-1}\left(\frac{\mathrm{tr}(R) - 1}{2} \right) \tag{114}$$

$$P^{-1}(\omega) = I - \frac{1}{2} [\omega]_\times + \left(1 - \frac{\|\omega\|}{2} \cot \frac{\|\omega\|}{2} \right) \frac{[\omega]_\times^2}{\|\omega\|^2} \tag{115}$$

B.1.2 Christoffel-Symbols Γ_{ij}^k

For SE_3 the Christoffel symbols Γ_{ij}^k, $i, j, k \in \{1, \ldots, 6\}$ are

$$\Gamma_{12}^3 = \Gamma_{23}^1 = \Gamma_{31}^2 = +\frac{1}{2}, \tag{116}$$

$$\Gamma_{13}^2 = \Gamma_{21}^3 = \Gamma_{32}^1 = -\frac{1}{2}, \tag{117}$$

$$\Gamma_{15}^6 = \Gamma_{26}^4 = \Gamma_{34}^5 = +1, \tag{118}$$

$$\Gamma_{16}^5 = \Gamma_{24}^6 = \Gamma_{35}^4 = -1, \tag{119}$$

and zero otherwise, see e.g. Žefran et al (1999).

B.2 Modified Descent Direction

Proposition 3 (Modified descent direction) *Let $\nabla_T f \in T_I \mathrm{G}$ be a gradient of f at $X^{(i)}$ as defined in (55), (57), i.e.*

$$\nabla_T f = L^*_{X^{(i)}} \nabla_{\mathrm{G}} f(X^{(i)}). \tag{120}$$

Then for any positive definite linear operator $B^{(i)}$ on the tangent space, i.e.

$$B : T_I \mathrm{G} \mapsto T_I \mathrm{G}, \quad \langle V, BV \rangle_{\mathrm{G}} > 0 \quad \forall V \in T_I \mathrm{G}, V \neq 0, \tag{121}$$

we obtain a descent direction

$$W := -B^{(i)} \nabla_T f \in T_I \mathrm{G} \tag{122}$$

of f, i.e. it fulfills

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\varphi(t, X^{(i)}, W)) \Big|_{t=0} < 0\,. \tag{123}$$

Proof

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\varphi(t, X^{(i)}, W)) \Big|_{t=0} \tag{124}$$

$$= \left\langle \nabla f(X^{(i)}), \frac{\mathrm{d}}{\mathrm{d}t} \varphi(t, X^{(i)}, W) \right\rangle \Big|_{t=0} \tag{125}$$

$$\overset{(50)}{=} \left\langle \nabla f(X^{(i)}), L_{X^{(i)}} W \right\rangle \tag{126}$$

$$\overset{(52)}{=} \left\langle \nabla_{\mathrm{G}} f(X^{(i)}), L_{X^{(i)}} W \right\rangle_{\mathrm{G}} \tag{127}$$

$$\overset{(122)}{=} - \left\langle \nabla_{\mathrm{G}} f(X^{(i)}), L_{X^{(i)}} B^{(i)} \nabla_T f \right\rangle_{\mathrm{G}} \tag{128}$$

$$= - \left\langle L^*_{X^{(i)}} \nabla_{\mathrm{G}} f(X^{(i)}), B^{(i)} \nabla_T f \right\rangle_{\mathrm{G}} \tag{129}$$

$$\overset{(55)}{=} - \left\langle \nabla_T f, B^{(i)} \nabla_T f \right\rangle_{\mathrm{G}} \overset{(121)}{<} 0\,. \tag{130}$$

□

B.3 Function Approximation

Proposition 4 (Quadratic approximation) *The linear equality system (63) can be interpreted as the optimality condition of a quadratic function,*

$$h(W) := f(X^{(i)}) + \langle b, W \rangle_{\mathrm{G}} + \frac{1}{2} \langle W, AW \rangle_{\mathrm{G}}, \tag{131}$$

and $h(tW)$ is a local quadratic approximation of the objective function $f(\varphi(t, X^{(i)}, W))$ at $X^{(i)}$ for small t, i.e.

$$h(tW) \approx f(\varphi(t, X^{(i)}, W)). \tag{132}$$

Proof This can be show by verifying that f and h match at $t = 0$:

Function value:

$$h(tW)|_{t=0} = h(0) = f(X^{(i)}) \tag{133}$$

First derivative:

$$\frac{\mathrm{d}}{\mathrm{d}t} h(tW)|_{t=0} = b = \langle W, \nabla_T f \rangle_{\mathrm{G}} \tag{134}$$

$$\overset{(55)}{=} \left\langle W, L^*_{X^{(i)}} \nabla f(X^{(i)}) \right\rangle_{\mathrm{G}} \tag{135}$$

$$\overset{(124)}{=} \frac{\mathrm{d}}{\mathrm{d}t} f(\varphi(t, X^{(i)}, W)) \Big|_{t=0} \tag{136}$$

Second derivative:

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} h(tW)|_{t=0} \overset{(62)}{=} \langle W, \overline{\nabla}_W \nabla_T f \rangle_{\mathrm{G}} \tag{137}$$

$$\overset{(39)}{=} \Big\langle W, \lim_{t\to 0} t^{-1} \Big(L^*_{\varphi(t,X^{(i)},W)} \nabla f(\varphi(t, X^{(i)}, W)) \tag{138}$$

$$- L^*_{X^{(i)}} \nabla f(X^{(i)}) \Big) \Big\rangle \tag{139}$$

$$= \frac{\mathrm{d}}{\mathrm{d}t} \Big\langle W, L^*_{\varphi(t,X^{(i)},W)} \nabla f(\varphi(t, X^{(i)}, W)) \Big\rangle \Big|_{t=0} \tag{140}$$

$$\overset{(124)}{=} \frac{\mathrm{d}^2}{\mathrm{d}t^2} f(\varphi(t, X^{(i)}, W)) \Big|_{t=0} \tag{141}$$

□

B.4 Schur Complement

Given a linear equality system

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix} \tag{142}$$

with $A \in \mathbb{R}^{n\times n}$, $B, C^\top \in \mathbb{R}^{n\times m}$, $D \in \mathbb{R}^{m\times m}$, $x, a \in \mathbb{R}^n$ and $y, b \in \mathbb{R}^m$. The Schur complement of an invertible D is defined as $S := A - BD^{-1}C$.

Then the solution to (142) can be determined by first solving

$$Sx = a - BD^{-1}b \tag{143}$$

w.r.t. x and by then substituting x in $Bx + Dy = b$ and solving for y:

$$Dy = b - Bx\,. \tag{144}$$

This can be easily verified using back-substitution.

Furthermore, the *symmetric* matrix $M := \left(\begin{smallmatrix} A & B \\ B^\top & D \end{smallmatrix}\right)$ is positive definite if and only if $S \succ 0$ and $D \succ 0$, see e.g. Golub and Loan (1996).

C Evaluation

C.1 Depth Map Scale Estimation

Given a depth map $sd(x)$ with unknown global scale s and a reference $\overline{d}(x)$, we robustly estimate the unknown scale s as

$$s = \operatorname{median} \left\{ \frac{\overline{d}(x)}{d(x)} \,\middle|\, x \in \Omega, w(x) \le p_{10} \right\}, \tag{145}$$

where p_{10} denotes the 10th percentile of the weights

$$w(x) := \sigma_d(x) \sigma_{\overline{d}}(x) \tag{146}$$

and σ_d and $\sigma_{\overline{d}}$ are estimators for the expected error of d and $\overline{d}$, respectively.

For the monocular method, we choose $\sigma_d(x)$ as the local variance of the monocular depth map, see Sect. 3.4.1. If $\overline{d}$ is provided by a stereo method, we approximate the inaccuracy by $\sigma_{\overline{d}}(x) = b^{-1}\overline{d}^2(x)$, see Sect. 3.1.2 for a motivation.

For ground truth as available for the synthetic *enpeda-2-2* data set we assume a unit $\sigma_{\overline{d}}(x)$. Missing values are marked by setting $\sigma_{\overline{d}}(x) = \infty$.

References

Absil PA, Mahony R, Sepulchre R (2008) Optimization Algorithms on Matrix Manifolds. Princeton University Press

Bagnato L, Frossard P, Vandergheynst P (2011) A variational framework for structure from motion in omnidirectional image sequences. J Math Imaging Vis 41(3):182–193

Bain A, Crisan D (2009) Fundamentals of Stochastic Filtering. Springer

Baker S, Matthews I (2004) Lucas-Kanade 20 Years On: A Unifying Framework. IJCV 56(3):221–255

Becker F, Lenzen F, Kappes JH, Schnörr C (2011) Variational Recursive Joint Estimation of Dense Scene Structure and Camera Motion from Monocular High Speed Traffic Sequences. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp 1692–1699

Bonnans JF, Gilbert JC, Lemaréchal C, Sagastizábal C (2003) Numerical Optimization. Springer Verlag

Bredies K, Kunisch K, Pock T (2010) Total Generalized Variation. SIAM J Imaging Sciences 3(3):492–526

Brox T, Bruhn A, Papenberg N, Weickert J (2004) High accuracy optical flow estimation based on a theory for warping. In: Pajdla T, Matas J (eds) European Conference on Computer Vision (ECCV), Springer, Prague, Czech Republic, LNCS, vol 3024, pp 25–36

Bruhn A, Weickert J, Schnörr C (2005) Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods. IJCV 61:211–231

Comport A, Malis E, Rives P (2007) Accurate Quadri-focal Tracking for Robust 3D Visual Odometry. In: IEEE International Conference on Robotics and Automation, ICRA'07, Rome, Italy

Enzweiler M, Gavrila D (2009) Monocular pedestrian detection: Survey and experiments. PAMI 31(12):2179–2195

Fleet D, Weiss Y (2006) Optical Flow Estimation, Springer, pp 239–257

Geiger A, Roser M, Urtasun R (2010) Efficient large-scale stereo matching. In: Asian Conference on Computer Vision, Queenstown, New Zealand

Geiger A, Lenz P, Urtasun R (2012) Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In: Computer Vision and Pattern Recognition (CVPR), Providence, USA

Gerónimo D, López A, Sappa A, Graf T (2010) Survey of pedestrian detection for advanced driver assistance systems. PAMI 32(7):1239–1258

Golub GH, Loan CFV (1996) Matrix Computations, 3rd edn. The Johns Hopkins University Press

Graber G, Pock T, Bischof H (2011) Online 3D reconstruction using Convex Optimization. In: 1st Workshop on Live Dense Reconstruction From Moving Cameras, ICCV 2011, pp 708–711

Hadsell R, Sermanet P, Ben J, Erkan A, Scoffier M, Kavukcuoglu K, Muller U, LeCun Y (2009) Learning long-range vision for autonomous off-road driving. J Field Robot 26:120–144

Hartley R, Zisserman A (2000) Multiple View Geometry in Computer Vision. Cambridge Univ. Press

Helmke U, Hüper K, Lee P, Moore J (2007) Essential Matrix Estimation Using Gauss-Newton Iterations on a Manifold. Int J Comp Vision 74(2):117–136

Hirschmüller H (2008) Stereo processing by semiglobal matching and mutual information. IEEE Trans Pattern Anal Mach Intell 30(2):328–341

Irani M, Anandan P, Cohen M (2002) Direct recovery of planar-parallax from multiple frames. TPAMI

24(11):1528–1534
Jordan M, Ghahramani Z, Jaakkola T, Saul L (1999) An Introduction to Variational Methods for Graphical Models. Mach Learning 37:183–233
Klein G, Murray D (2007) Parallel tracking and mapping for small AR workspaces. In: Proc. Sixth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR'07), Nara, Japan
Konolige K, Agrawal M (2008) FrameSLAM: From Bundle Adjustment to Real-Time Visual Mapping. IEEE Trans Robotics 24(5):1066–1077
Lee DC, Hebert M, Kanade T (2009) Geometric reasoning for single image structure recovery. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)
Lenzen F, Becker F, Lellmann J (2013) Adaptive second-order total variation: An approach aware of slope discontinuities. In: Proceedings of the 4th International Conference on Scale Space and Variational Methods in Computer Vision (SSVM) 2013, Springer, LNCS, in press
Lin WY, Cheong LF, Tan P, Dong G, Liu S (2011) Simultaneous Camera Pose and Correspondence Estimation with Motion Coherence. International Journal of Computer Vision pp 1–17
Liu B, Gould S, Koller D (2010) Single image depth estimation from predicted semantic labels. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp 1253 –1260
Meister S, Kondermann D, Jähne B (2012) An Outdoor Stereo Camera System for the Generation of Real-World Benchmark Datasets with Ground Truth. SPIE Optical Engineering 51(2)
Mester R (2011) Recursive Live Dense Reconstruction: Some Comments on Established and Imaginable New Approaches. In: 1st Workshop on Live Dense Reconstruction From Moving Cameras, ICCV 2011, pp 712–714
Mouragnona E, Lhuilliera M, Dhomea M, Dekeyserb F, Sayd P (2009) Generic and real-time structure from motion using local bundle adjustment. Image Vis Comp 27(8):1178–1193
Newcombe RA, Davison AJ (2010) Live dense reconstruction with a single moving camera. In: CVPR
Newcombe RA, Lovegrove SJ, Davison AJ (2011) DTAM: Dense Tracking and Mapping in Real-Time. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp 2320–2327
Nister D, Naroditsky O, Bergen J (2004) Visual odometry. In: Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, vol 1, pp 652 – 659
Pennec X (2006) Intrinsic Statistics on Riemannian Manifolds: Basic Tools for Geometric Measurements. J Math Imag Vision 25(1):127–154
Rabe C, Müller T, Wedel A, Franke U (2010) Dense, Robust, and Accurate Motion Field Estimation from Stereo Image Sequences in Real-Time. In: Daniilidis K, Maragos P, Paragios N (eds) Proceedings of the 11th European Conference on Computer Vision, Springer, Lecture Notes in Computer Science, vol 6314, pp 582–595
Rasmussen C, Williams C (2006) Gaussian Processes for Machine Learning. MIT Press
Rhemann C, Hosni A, Bleyer M, Rother C, Gelautz M (2011) Fast Cost-Volume Filtering for Visual Correspondence and Beyond. Computer Vision and Pattern Recognition, IEEE Computer Society Conference on pp 3017–3024
Rudin LI, Osher S, Fatemi E (1992) Nonlinear total variation based noise removal algorithms. Phys D 60(1–4):259–268
Saxena A, Chung SH, Ng AY (2008) 3-D Depth Reconstruction from a Single Still Image. IJCV 76:53–69
Scharstein D, Szeliski R (2002) A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int J Comput Vision 47(1-3):7–42
Sheikh Y, Hakeem A, Shah M (2007) On the direct estimation of the fundamental matrix. In: CVPR, IEEE Computer Society
Stühmer J, Gumhold S, Cremers D (2010) Parallel Generalized Thresholding Scheme for Live Dense Geometry from a Handheld Camera. In: Doucet A, De Freitas N, Gordon N (eds) CVGPU
Sturm P, Triggs B (1996) A factorization based algorithm for multi-image projective structure and motion. In: ECCV, Springer, Cambridge, England, pp 709–720
Szeliski R, Zabih R, Scharstein D, Veksler O, Kolmogorov V, Agarwala A, Tappen M, Rother C (2008) A Comparative Study of Energy Minimization Methods for Markov Random Fields with Smoothness-Based Priors. TPAMI 30:1068–1080
Tierney L, Kadane JB (1986) Accurate Approximations for Posterior Moments and Marginal Densities. Journal of the American Statistical Association 81(393):82–86
Triggs B, McLauchlan PF, Hartley RI, Fitzgibbon AW (2000) Bundle Adjustment – A Modern Synthesis, vol 1883. Springer-Verlag
Valgaerts L, Bruhn A, Zimmer H, Weickert J, Stoll C, Theobalt C (2010) Joint estimation of motion, structure and geometry from stereo sequences. In: Proceedings of the 11th European Conference on Computer Vision, Springer-Verlag, Berlin, Heidelberg, ECCV 2010, pp 568–581
Valgaerts L, Bruhn A, Mainberger M, Weickert J (2012) Dense versus sparse approaches for estimating the fundamental matrix. International Journal of Computer Vision 96(2):212–234
Vaudrey T, Rabe C, Klette R, Milburn J (2008) Differences between stereo and motion behavior on synthetic and real-world stereo sequences. In: 23rd International Conference of Image and Vision Computing New Zealand (IVCNZ '08), pp 1–6
Žefran M, Kumar V, Croke C (1999) Metrics and Connections for Rigid-Body Kinematics. The International Journal of Robotics Research 18(2):242–1–242–16
Wedel A, Rabe C, Vaudrey T, Brox T, Franke U, Cremers D (2008) Efficient dense scene flow from sparse or dense stereo data. In: ECCV, LNCS, vol 3021
Weishaupt A, Bagnato L, Vandergheynst P (2010) Fast Structure from Motion for Planar Image Sequences. In: EUSIPCO
Wendel A, Maurer M, Graber G, Pock T, Bischof H (2012) Dense reconstruction on-the-fly. In: CVPR, IEEE, pp 1450–1457
Wojek C, Roth S, Schindler K, Schiele B (2010) Monocular 3d scene modeling and inference: Understanding multi-object traffic scenes. In: ECCV, LNCS, vol 6314, pp 467–481
Yamaguchi K, Hazan T, McAllester D, Urtasun R (2012) Continuous Markov Random Fields for Robust Stereo Estimation. In: ECCV 2012