# Joint Epipolar Tracking (JET):
# Simultaneous optimization of epipolar geometry and feature correspondences

Henry Bradler[1], Matthias Ochs[1], Nolang Fanani[1]
[1]Visual Sensorics & Information Processing Lab
Goethe University, Frankfurt, Germany
{bradler,ochs,fanani,mester}@vsi.cs.uni-frankfurt.de

Rudolf Mester[1,2]
[2]Computer Vision Laboratory, ISY
Linköping University, Sweden
mester@isy.liu.se

## Abstract

*Traditionally, pose estimation is considered as a two step problem. First, feature correspondences are determined by direct comparison of image patches, or by associating feature descriptors. In a second step, the relative pose and the coordinates of corresponding points are estimated, most often by minimizing the reprojection error (RPE). RPE optimization is based on a loss function that is merely aware of the feature pixel positions but not of the underlying image intensities. In this paper, we propose a sparse direct method which introduces a loss function that allows to simultaneously optimize the unscaled relative pose, as well as the set of feature correspondences directly considering the image intensity values. Furthermore, we show how to integrate statistical prior information on the motion into the optimization process. This constructive inclusion of a Bayesian bias term is particularly efficient in application cases with a strongly predictable (short term) dynamic, e.g. in a driving scenario. In our experiments, we demonstrate that the 'JET' algorithm we propose outperforms the classical reprojection error optimization on two synthetic datasets and on the KITTI dataset. The JET algorithm runs in real-time on a single CPU thread.*

## 1. Introduction

The main contribution of this work is the introduction of a joint loss function which is based on the photometric error of all feature correspondences. The correspondences are parameterized by one underlying epipolar geometry. This guarantees all correspondences to be epipolar-conform by construction, and allows to directly optimize the pose based on image intensities. Starting point is the well known Lucas-Kanade tracking method [19] which employs a quadratic photometric loss function (SSD) on a *single* image patch to optimize feature correspondences. Given the
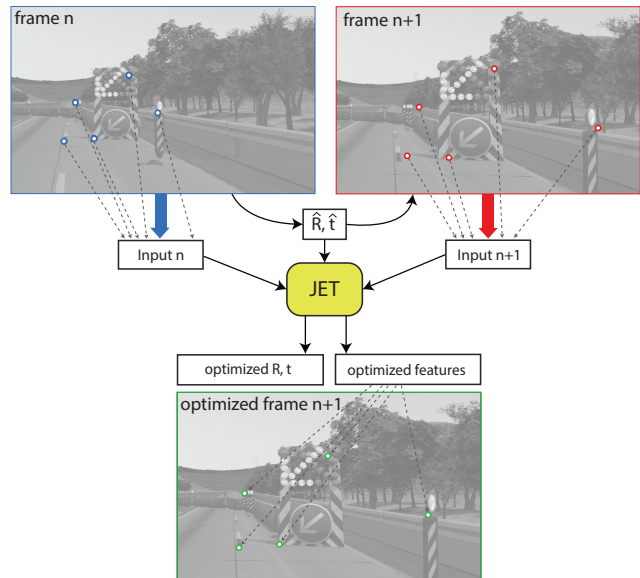


Figure 1: Flow chart of the JET algorithm.

epipolar geometry, the search space can be drastically decreased from a 2D to a 1D search by including an epipolar constraint [23, 24].

We show how to optimize all correspondences simultaneously *and* optimize the epipolar geometry at the same time, given coarse initial values of these entities (the typical situation in many applications). This is achieved by varying the epipolar geometry, and by this the adjusted correspondences of all features, in a way that minimizes the *joint* photometric loss function. We denote the resulting procedure as *Joint Epipolar Tracking* (JET). As this joint optimization is performed directly on the intensity values, JET is a *'direct'* method, in today's terminology. This is in contrast to the widely used minimization of the reprojection error which distills photometric information into geometric information and subsequently disregards the mere image intensities.

We show that JET outperforms the standard minimization of the reprojection error (RPE), when optimizing the relative pose (ego motion of the camera). The comparison was performed on synthetic and real data sets (all publicly available), synthetic data sets in order to have perfect ground truth and real data sets to demonstrate the feasibility under realistic conditions. The synthetic data sets are COnGRATS [5] (driving scenes on a road scene) and RGB-D data from ICL-NUIM [16] (indoor footage of a hand-held camera). As a representative of real data, we utilize the well known KITTI dataset [14, 15], which consists of real driving scenes in urban and highway scenarios. As dense depth information or optical flow ground truth are not available for this data set, we focus on comparing the quality of JET against RPE by regarding the relative pose.

Since we regard monocular image data, the scale of the pose remains undetermined and we only analyze the relative rotation and the relative *unscaled* translation of the motion. For a calibrated camera, these entities will entirely define the epipolar geometry as it is scale independent as well.

## 2. Related Work

Approaches to relative pose or 3D motion estimation can be divided into two basic categories: *feature-based* methods and *direct* methods, with some hybrid approaches existing as well. Feature-based methods are characterized by the extraction and the matching of salient and reproducible features that are tracked over frames. Prominent examples of feature point based optimization methods are [3], [22], and [8]. Usually, these approaches minimize the reprojection error of tracked feature points.

So called 'direct', appearance- or intensity-based methods, on the other hand, operate directly by matching pixel intensities. They propagate the original image information into the optimization scheme, usually using a differential optimization approach and therefore can often provide more accurate estimates of pose and structure. DTAM [21] was among the first real time dense systems. Semi-Dense Visual Odometry for a Monocular Camera [11] and its successor LSD-SLAM [10] as well as SVO [13] are more recent examples. We share the opinion of the authors of [11] who state that the separation into feature detection and tracking versus a state estimation creates an artificial gap between the data and the state sought.

PTAM [17] could also be considered as a hybrid method: A weak motion prior is used to initialize the search for a small number of features using a modified KLT at the highest level of a scale pyramid. The resulting tracks provide a better egomotion prediction which is then used to search for a larger number of KLT tracks at lower pyramid levels and so on until the bottom level is reached.

Stabilizing the estimation of correspondences by integrating a given (or assumed) epipolar relation into the matching process has been used in numerous approaches. For instance, the authors of [2, 23, 25, 27, 28] use epipolar constraints for stabilizing discrete matching, whereas Valgaerts et al. [26] proposed a variational approach to estimate dense optical flow and the epipolar geometry (represented by the fundamental matrix) simultaneously. Other direct methods that also explicitly take into consideration the depth structure of the scene are [4, 18, 20].

An important property which typically distinguishes appearance, direct, dense or semi-dense from feature based approaches is that direct methods often use parametric models of the flow field and hence can utilize edges as well as corners. If no explicit motion priors or dynamic models are used, these direct methods generally depend on a high frame rate that ensures moderate displacements, whereas feature-based matching can work even with very large displacements. However, even in this case a photometric 'direct' post-optimization can be performed. JET is a well suited method to do just this.

## 3. Approach

We outline our approach and introduce our notation, starting from plain Lucas-Kanade tracking [19] in section 3.1 and subsequently revisiting epipolar constrained KL tracking in section 3.2. This leads to the presentation of joint epipolar tracking in section 3.3.

### 3.1. General Lucas-Kanade Tracking

The aim of differential direct tracking, often denoted as Lucas-Kanade tracking [19] is to successively determine the corresponding image feature point position $\vec{y}_k = \vec{x}_k + \vec{v}_k$ in image $\mathcal{J}$ for a given feature point $\vec{x}_k$ in another image $\mathcal{I}$. We use the weighted sum of squared differences (WSSD) as loss function for patch comparisons, thus implicitly modeling the image noise as signal-independent, i.i.d. and Gaussian.

$$Q_k \stackrel{def}{=} \sum_{\vec{x}} \mathbf{W}[\vec{x} - \vec{x}_k] \cdot \left( \mathcal{I}[\vec{x}] - \mathcal{J}[\underbrace{\vec{x} + \vec{v}_k^0}_{\vec{y}}] \right)^2 . \quad (1)$$

The non-negative pixel weights and the size of the patches are defined by a normalized kernel $\mathbf{W}$. All points $\vec{x}$ and $\vec{y} = \vec{x} + \vec{v}_k^0$ with a non-zero weight $\mathbf{W}[\vec{x} - \vec{x}_k]$ are taken into account for the patch difference. In a typical scenario, a feature point $\vec{x}_k$ and an initial estimate $\vec{y}_k = \vec{x}_k + \vec{v}_k^0$ for the corresponding feature are given and the task is to optimize this correspondence by minimizing the WSSD

$$Q_k(\vec{v}_k) \stackrel{def}{=} \sum_{\vec{x}} \mathbf{W}[\vec{x} - \vec{x}_k] \cdot (\mathcal{I}[\vec{x}] - \mathcal{J}[\vec{y} + \vec{v}_k])^2 \rightarrow \min$$

$$(2)$$

for a specific realization of the image displacement $\vec{v}_k$. Since this problem cannot be solved directly in closed form, a local first order Taylor approximation of the image difference $\mathcal{I}[\vec{x}] - \mathcal{J}[\vec{y} + \vec{v}_k]$ is usually applied. This yields the approximated weighted sum of squared differences:

$$\widetilde{Q}_k(\vec{v}_k) \overset{def}{=} \sum_{\vec{x}} \mathbf{W}[\vec{x} - \vec{x}_k] \cdot \left( \mathcal{I}[\vec{x}] - \mathcal{J}[\vec{y}] - \vec{v}_k^T \cdot \frac{\partial \mathcal{J}}{\partial \vec{x}}[\vec{y}] \right)^2$$
$$= \vec{v}_k^T \cdot \mathbf{A}_k \cdot \vec{v}_k + 2 \cdot \vec{v}_k^T \cdot \vec{b}_k + c_k, \qquad (3)$$

using the abbreviations

$$\mathbf{A}_k \overset{def}{=} \sum_{\vec{x}} \mathbf{W}[\vec{x} - \vec{x}_k] \cdot \left( \frac{\partial \mathcal{J}}{\partial \vec{x}}[\vec{y}] \right) \cdot \left( \frac{\partial \mathcal{J}}{\partial \vec{x}}[\vec{y}] \right)^T,$$

$$\vec{b}_k \overset{def}{=} \sum_{\vec{x}} -\mathbf{W}[\vec{x} - \vec{x}_k] \cdot \frac{\partial \mathcal{J}}{\partial \vec{x}}[\vec{y}] \cdot (\mathcal{I}[\vec{x}] - \mathcal{J}[\vec{y}]),$$

$$c_k \overset{def}{=} \sum_{\vec{x}} \mathbf{W}[\vec{x} - \vec{x}_k] (\mathcal{I}[\vec{x}] - \mathcal{J}[\vec{y}])^2. \qquad (4)$$

Since the image difference has been linearized, this is an approximation to the 'true' optimization problem, well known in nonlinear optimization theory as the *Gauss-Newton* method. The approximation in equation (3) yields a convex parabolic function which allows to solve for the optimal displacement $\vec{v}_k$. Due to the linearization of the image, this approximation should only be used to improve the feature correspondence which then serves as a new initialization for another step of the incremental optimization process.

### 3.2. Epipolar Constrained Tracking

In *constrained epipolar tracking*, we consider the relative pose given by a rotation matrix $\mathbf{R}$ and a translation vector $\vec{t}$ to be known, and adjust the feature correspondences $\{\vec{x}_k \leftrightarrow \vec{y}_k + \vec{v}_k\}$ to comply with the given epipolar geometry. This yields the epipolar constraint:

$$(\vec{y}_k^T + \vec{v}_k^T, 1) \cdot \mathbf{F} \cdot \begin{pmatrix} \vec{x}_k \\ 1 \end{pmatrix} \overset{!}{=} 0 \qquad (5)$$

$$\Leftrightarrow \quad \vec{v}_k^T \cdot \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \cdot \mathbf{F}}_{\overset{def}{=} \mathbf{F}'} \cdot \begin{pmatrix} \vec{x}_k \\ 1 \end{pmatrix} \overset{!}{=} -(\vec{y}_k^T, 1) \cdot \mathbf{F} \cdot \begin{pmatrix} \vec{x}_k \\ 1 \end{pmatrix}.$$

In equation (5), $\mathbf{F}$ is the fundamental matrix,

$$\mathbf{F} \overset{def}{=} \mathbf{K}^{-T} \cdot \left[\vec{t}\right]_{\times} \cdot \mathbf{R} \cdot \mathbf{K}^{-1}, \qquad (6)$$

which is defined up to a scale factor. $\mathbf{K}$ is the camera matrix holding the intrinsic camera parameters and $\left[\vec{t}\right]_{\times}$ is the skew symmetric matrix of the translation vector. We write $\mathbf{F} = \mathbf{F}(\vec{p})$ to denote that, for a given camera matrix, the fundamental matrix is fully determined by the (unscaled)

motion parameters $\vec{p}$. We use the polar parametrization of the rigid transformation as proposed in [6]:

$$\vec{p} \overset{def}{=} (\theta, \psi, \phi, \alpha, \beta)^T. \qquad (7)$$

These parameters are a minimum representation of the relative unscaled pose. The pitch angle $\theta$, the yaw angle $\psi$ and the roll $\phi$ are the rotational degrees of freedom about the $x$-, $y$- and $z$-axis. The azimuth $\alpha$ and the polar angle $\beta$ represent the unscaled translation $\vec{t}$ in polar coordinates.

Using Lagrange multipliers, we solve the approximated problem in equation (3) under the epipolar constraint introduced in equation (5):

$$\widetilde{Q}_k(\vec{v}_k) = \vec{v}_k^T \cdot \mathbf{A}_k \cdot \vec{v}_k + 2 \cdot \vec{v}_k^T \cdot \vec{b}_k + c_k$$
$$+ 2 \cdot \lambda \cdot \left( (\vec{y}_k^T + \vec{v}^T, 1) \cdot \mathbf{F} \cdot \begin{pmatrix} \vec{x}_k \\ 1 \end{pmatrix} \right). \qquad (8)$$

The optimal displacement $\vec{v}_{k,\text{opt}}$ can be computed via the minimization of $\widetilde{Q}_k(\vec{v}_k)$:

$$\frac{1}{2} \cdot \frac{\partial \widetilde{Q}_k}{\partial \vec{v}_k}(\vec{v}_{k,\text{opt}}) = \mathbf{A}_k \cdot \vec{v}_{k,\text{opt}} + \vec{b}_k + \lambda \cdot \mathbf{F}' \cdot \begin{pmatrix} \vec{x}_k \\ 1 \end{pmatrix} \overset{!}{=} \vec{0}. \qquad (9)$$

In combination with equation (5) this yields the linear equation system:

$$\begin{pmatrix} \mathbf{A}_k & \mathbf{F}' \cdot \begin{pmatrix} \vec{x}_k \\ 1 \end{pmatrix} \\ \left( \mathbf{F}' \cdot \begin{pmatrix} \vec{x}_k \\ 1 \end{pmatrix} \right)^T & 0 \end{pmatrix} \cdot \begin{pmatrix} \vec{v}_{k,\text{opt}} \\ \lambda \end{pmatrix} \overset{!}{=}$$
$$\begin{pmatrix} -\vec{b}_k \\ -(\vec{y}_k^T, 1) \cdot \mathbf{F} \cdot \begin{pmatrix} \vec{x}_k \\ 1 \end{pmatrix} \end{pmatrix}. \qquad (10)$$

The matrix on the left hand side of this equation system is symmetric and $\mathbf{F}$ is a function of the motion parameter $\vec{p}$, i.e. there is a closed form solution to the optimal displacement $\vec{v}_{k,\text{opt}}$ showing the following dependency:

$$\vec{v}_{k,\text{opt}} = \vec{f}(\mathbf{A}_k, \vec{b}_k, \vec{x}_k, \vec{y}_k, \mathbf{K}, \vec{p}) = \vec{f}_k(\vec{p}). \qquad (11)$$

For a given epipolar geometry (which is equivalent to a given unscaled relative pose and a calibrated camera) the linear equation system (10) is the extension of the standard Lucas-Kanade equation (see equation (3)) with an epipolar constraint. It can be used to optimize image correspondences if the epipolar geometry is already known in beforehand.

### 3.3. Joint Epipolar Tracking

The present work extends the epipolar constrained tracking in the following sense: We do not only optimize each

feature correspondence $\vec{x}_k \leftrightarrow \vec{y}_k$ individually with respect to a given epipolar geometry, but build a joint loss function which can be optimized with respect to the underlying motion that characterizes the displacements of *all* image points (given that all points obey the same epipolar relation).

Using this approach, we can additionally optimize the motion parameters themselves. We call this method *Joint Epipolar Tracking* (JET). To this end, we perform a reparametrization of the loss function $\widetilde{Q}_k(\vec{v}_k)$ by substituting the functional dependency $\vec{v}_k = \vec{f}_k(\vec{p})$ into it (compare equations (8) and (11) respectively):

$$
\begin{aligned}
\widetilde{Q}_k(\vec{p}) &\stackrel{def}{=} \widetilde{Q}_k(\vec{v} = \vec{f}_k(\vec{p})) \\
&= \vec{f}_k^T(\vec{p}) \cdot \mathbf{A}_k \cdot \vec{f}_k(\vec{p}) + 2 \cdot \vec{f}_k^T(\vec{p}) \cdot \vec{b}_k + c_k \\
&\quad + 2 \cdot \lambda \cdot \underbrace{\left( (\vec{y}_k^T + \vec{f}_k^T(\vec{p}), 1) \cdot \mathbf{F} \cdot \begin{pmatrix} \vec{x}_k \\ 1 \end{pmatrix} \right)}_{0}.
\end{aligned}
\tag{12}
$$

By using this definition of the displacement, the optimization of the loss function is no longer performed with respect to an image displacement $\vec{v}_k$ but with respect to an epipolar geometry which is induced by the relative pose of the camera and the environment. This relative pose is evoking the optical flow in the image domain.

Joining together the loss functions from equation (12) for several feature correspondences $\{\vec{x}_k \leftrightarrow \vec{y}_k\}_k$ and adding a prior term for the motion (expressing a statistical model of 'typical' motion) yields the joint loss function

$$
\widetilde{Q}(\vec{p}) = \underbrace{\frac{1}{N} \sum_{k=1}^{N} \widetilde{Q}_k(\vec{p})}_{\text{image information}} + \underbrace{\xi_Q \cdot (\vec{p} - \hat{\vec{p}})^T \cdot \mathbf{C}_{\vec{p}-\hat{\vec{p}}}^{-1} \cdot (\vec{p} - \hat{\vec{p}})}_{\text{prior term on motion parameters}}.
$$

(13)

The minimization of this function allows to determine the motion parameters, and hence the unscaled relative pose, that best describes the optical flow. In equation (13) the part of the joint loss function that is dependent on the image information has been extended by a second part that incorporates statistical prior knowledge coupled via the coupling constant $\xi_Q$. The prior information is characterized by a prediction of the expected motion $\hat{\vec{p}}$ and a covariance matrix of the prediction residuals $\mathbf{C}_{\vec{p}-\hat{\vec{p}}}$.

These motion prior terms are determined by a linear regression approach on a dataset of motion parameters that are representative for the type of motion to be expected (e.g. restricted car motion, unrestricted motion of a handheld camera). We use a very similar approach as in [6] to determine the parameters of a linear predictor. The difference is that we employ a third order predictor, i.e. the preceding three motion parameter sets are taken into consideration

when evaluating the statistics and performing the dynamic prediction.

Equation (13) can be expressed in vertex form and the optimization of $\widetilde{Q}(\vec{p})$ is represented as the following least squares problem:

$$
\widetilde{Q}(\vec{p}) = \sum_{k=1}^{N} \|\vec{q}_k(\vec{p})\|_2^2 + \text{const.} \quad,
$$

$$
\vec{q}_k(\vec{p}) = \sqrt{\frac{1}{N}} \cdot \begin{pmatrix} \mathbf{A}_k^{1/2} \cdot (\vec{f}_k(\vec{p}) + \mathbf{A}_k^{-1} \cdot \vec{b}_k) \\ \sqrt{\xi_Q} \cdot \mathbf{C}_{\vec{p}-\hat{\vec{p}}}^{-1/2} \cdot (\vec{p} - \hat{\vec{p}}) \end{pmatrix}.
\tag{14}
$$

With an initial estimate $\vec{p}^{(0)}$ of the motion parameters (*e.g.* the prediction $\hat{\vec{p}}$ based on the previous motion parameters), we can now solve this optimization problem using a nonlinear solver like the Ceres solver [1]. The result of this motion optimization $\vec{p}_{\text{opt}}$ is then used to improve the feature correspondences $\vec{x}_k \leftrightarrow \vec{y}_k$ by shifting the corresponding image point to its epipolar line by $\vec{y}_k \rightarrow \vec{y}_k + \vec{f}_k(\vec{p}_{\text{opt}})$ (see equations (10) and (11)). Since the original image difference has been replaced by a linear approximation during the Gauss-Newton approach at the beginning, these improved correspondences and the improved motion parameters serve as an initialization for the second iteration step of this optimization procedure. We continue with this procedure as long as the target loss function

$$
Q(\vec{p}_{\text{opt}}) \stackrel{def}{=} \frac{1}{N} \sum_{k=1}^{N} Q_k(\vec{v}_k = \vec{f}_k(\vec{p}_{\text{opt}}))
\tag{15}
$$

is decreased. Note that the target loss function incorporates the *exact* image difference as introduced in equation (2).

The optimization of the relative pose using JET does not merely minimize the reprojection error[1], but rather than that minimize the photometric error of the feature correspondences by including the full image information encoded in the quantities $\mathbf{A}_k$, $\vec{b}_k$ and $c_k$. Compared to other leading direct methods, such as [9, 10], JET is the most compact formulation of the direct 2-view $n$ points pose optimization problem based on minimizing the photometric error.

## 4. Experiments

We evaluated the JET procedure presented here on synthetic data [5, 16], applying noise to the different input parameters to investigate the stability against noise in our components. As we used synthetic data, we had perfect ground truth for our results to compare against, a situation usually very hard to obtain for real-life driving scenarios, *e.g.* [15, 14].

The aim in our experiment is to optimize the motion parameters and correct the feature correspondences $\{\vec{x}_k \leftrightarrow$

---

[1]Actually the reprojection error is zero, since the feature correspondences are just optimized with respect to the relative pose.

| Dataset | $\boldsymbol{\rho}_{in}$ | | $\boldsymbol{\rho}_{JET}$ | | $\boldsymbol{\rho}_{RPE}$ | | $\boldsymbol{\Omega}_{in}$ | | $\boldsymbol{\Omega}_{JET}$ | | $\boldsymbol{\Omega}_{RPE}$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ |
| Construction | 0.96 | 0.28 | **0.06** | 0.09 | 0.11 | 0.08 | 7.65 | 2.83 | **1.62** | 2.39 | 2.04 | 1.71 |
| Construction* | 0.96 | 0.28 | **0.03** | 0.03 | 0.1 | 0.06 | 7.64 | 2.84 | **0.77** | 0.87 | 0.89 | 0.99 |
| Highway | 0.96 | 0.28 | **0.03** | 0.05 | 0.06 | 0.03 | 7.63 | 2.85 | **0.6** | 1.4 | 1.12 | 0.91 |
| Highway* | 0.96 | 0.28 | **0.02** | 0.02 | 0.05 | 0.05 | 7.65 | 2.85 | **0.25** | 0.47 | 0.23 | 0.18 |
| LivingRoom02 | 4.8 | 1.39 | **0.38** | 0.39 | 0.64 | 0.47 | **12.9** | 5.61 | 14.8 | 14.7 | 20.1 | 14.8 |
| OfficeRoom02 | 4.81 | 1.39 | **0.41** | 0.49 | 0.71 | 0.72 | **12.8** | 5.62 | 18.9 | 20.7 | 23.1 | 18.0 |

Table 1: First two moments of $\rho$ and $\Omega$ from the evaluation on COnGRATS and ICL-NUIM RGB-D dataset. Datasets ending with * indicate the use of prior knowledge. All values are in degrees.

$\vec{y}_k\} \to \{\vec{x}_k \leftrightarrow \vec{y}_{k,\text{opt}}\}$ so that they obey the epipolar geometry induced by the optimized motion parameters $\vec{p}_{\text{opt}}$:

$$\begin{pmatrix} \vec{y}_{k,\text{opt}} \\ 1 \end{pmatrix}^T \cdot \mathbf{F}(\vec{p}_{\text{opt}}) \cdot \begin{pmatrix} \vec{x}_k \\ 1 \end{pmatrix} = 0 \quad, k \in \{1, \dots, N\} \quad (16)$$

We compare the results achieved with JET against the results achieved with a method that minimizes the reprojection error (RPE).

## 4.1. Competing method: Optimization of the reprojection error

The competitor *RPE optimization* is a method that minimizes the reprojection error and performs the following steps:

1. Optimize correspondences:
   $\{\vec{x}_k \leftrightarrow \vec{y}_k\} \to \{\vec{x}_k \leftrightarrow \vec{y}_k'\}$ (optional)

2. Minimize the reprojection error:
   $\vec{p} \to \vec{p}_{\text{opt}}$

3. Perform a minimum correction of the correspondences, so that they are in agreement with $\vec{p}_{\text{opt}}$:
   $\{\vec{x}_k \leftrightarrow \vec{y}_k'\} \to \{\vec{x}_k \leftrightarrow \vec{y}_{k,\text{opt}}\}$

The first task is optional and optimizes the feature correspondences using standard Lucas-Kanade tracking as it is implemented in OpenCV [7]. We will run experiments with both, step one enabled and disabled. In the mandatory second step, RPE optimizes the motion parameters by minimizing the reprojection error

$$d_k(\vec{p}) \overset{def}{=} \begin{pmatrix} \vec{y}_k \\ 1 \end{pmatrix}^T \cdot \mathbf{F}(\vec{p}) \cdot \begin{pmatrix} \vec{x}_k \\ 1 \end{pmatrix} \Big/ \left\| \mathbf{F}'(\vec{p}) \begin{pmatrix} \vec{x}_k \\ 1 \end{pmatrix} \right\|_2 \quad (17)$$

for all feature correspondences. $d_k(\vec{p})$ is the distance of the image point $\vec{y}_k$ to the epipolar line specified by the fundamental matrix $\mathbf{F}(\vec{p})$ (see equation (6)) and $\vec{x}_k$. We delegate the optimization of the loss function $R(\vec{p})$ of the RPE method to the *Ceres-Solver* from Google [1].

$$R(\vec{p}) \overset{def}{=} \frac{1}{N} \cdot \sum_{k=1}^{N} d_k^2(\vec{p}) \quad + \quad \xi_R \cdot (\vec{p} - \hat{\vec{p}})^T \cdot \mathbf{C}_{\vec{p}-\hat{\vec{p}}}^{-1} \cdot (\vec{p} - \hat{\vec{p}})$$

$$= \sum_{k=1}^{N} \|\vec{r}_k(\vec{p})\|_2^2 \quad, \text{with}$$

$$\vec{r}_k(\vec{p}) = \sqrt{\frac{1}{N}} \cdot \begin{pmatrix} d_k(\vec{p}) \\ \sqrt{\xi_R} \cdot \mathbf{C}_{\vec{p}-\hat{\vec{p}}}^{-1/2} \cdot (\vec{p} - \hat{\vec{p}}) \end{pmatrix} \quad (18)$$

After having computed the optimized motion parameters $\vec{p}_{\text{opt}}$, we determine the optimized corresponding points $\vec{y}_{k,\text{opt}}$ by projecting all $\vec{y}_k$ to the closest points on their respective epipolar line. For that purpose we introduce the abbreviations

$$\vec{l}_k = \begin{pmatrix} l_{k,0} \\ l_{k,1} \\ l_{k,2} \end{pmatrix} = \mathbf{F}(\vec{p}_{\text{opt}}) \cdot \begin{pmatrix} \vec{x}_k \\ 1 \end{pmatrix} \quad, \quad \vec{y}_k = \begin{pmatrix} y_{k,0} \\ y_{k,1} \end{pmatrix} \quad (19)$$

and obtain for the optimized corresponding point:

$$\vec{y}_{k,\text{opt}} = \frac{1}{(l_{k,0})^2 + (l_{k,1})^2} \cdot$$
$$\begin{pmatrix} y_{k,0} \cdot (l_{k,1})^2 - y_{k,1} \cdot l_{k,0} \cdot l_{k,1} - l_{k,0} \cdot l_{k,1} \\ -y_{k,0} \cdot l_{k,0} \cdot l_{k,1} + y_{k,1} \cdot (l_{k,0})^2 - l_{k,1} \cdot l_{k,2} \end{pmatrix} \cdot \quad (20)$$

## 4.2. Initialization

Both methods were initialized with exactly the same estimated image correspondences and the same estimate of motion parameters. When using synthetic data, it is straightforward to obtain ground truth reference values for the correspondences as well as for the motion parameters. The COnGRATS [5] scenes we used in the evaluation, re-use pose sequences from the KITTI Benchmark. To make a coarse estimate of the variation range of the motion parameters, we checked the statistics of the motion parameters on the KITTI dataset, which covers a wide range of driving scenarios and can be considered as representative for realistic car motion.

If we assume a normal distribution of $\vec{p}_n - \vec{p}_{n-1}$ and use the KITTI motion statistic to find upper bounds for the variances of the translational and rotational degrees of freedom ($\sigma_{\text{rot}}^2 < 10^{-5}$ and $\sigma_{\text{trans}}^2 < 10^{-3}$), we can estimate the $3\,\sigma$ interval to be $3\,\sigma_{\text{rot}} < 10^{-2}$ and $3\,\sigma_{\text{trans}} < 10^{-1}$. More than 99.7% of the motion parameters do not deviate by more than $3\,\sigma_{\text{rot/trans}}$ from their temporal predecessor.

We use these insights to justify a realistic variation range of $\pm 1°$ and $\pm 10°$ for the rotation and translation parameters respectively. These ranges correspond to more than 5 standard deviations $\sigma_{\text{rot/trans}}$. We apply uniformly distributed noise with the just derived intervals to the motion parameters.

A similar consideration for a hand held camera, as it is used in the second synthetic dataset [16], leads to a variation range of $\pm 5°$ and $\pm 20°$ for the rotation and translation parameters, respectively.

For the corresponding image points $\vec{y}_k$, we apply uniform noise to the $x$- and $y$-component of the ground truth value, each with a level of $\pm 5$ pixels.

## 4.3. Evaluation measures

Each experiment gets initialized with an approximation of the pose and with initial image correspondences. To quantify the quality of the input and the output of the methods, the deviation from ground truth is expressed by the following four evaluation measures:

- *Rodrigues angle $\rho$ (rotational error):*
  The rotation parameters $\theta$, $\psi$ and $\phi$ define a rotation matrix $\mathbf{R}$ which is to be compared against the ground truth $\mathbf{R}_{\text{gt}}$ via the relative rotation $\mathbf{R}_{\text{rel}} = \mathbf{R}_{\text{gt}} \cdot \mathbf{R}^T$. According to Rodrigues' formula, $\mathbf{R}_{\text{rel}}$ can be interpreted as a rotation of an angle $\rho$ about some axis $\vec{n}$. The absolute value of the Rodrigues angle $\rho$ serves as a measure for the deviation from the ground truth rotation.

- *Angle of intersection $\Omega$ (translational error):*
  The translation parameters $\alpha$ and $\beta$ represent the direction of the translation vector. The translation direction is compared to the ground truth via the absolute value of its angle of intersection $\Omega$.

- *RMS distance of corresponding points (positional error):*
  The quality of the point correspondences is characterized by the mean deviation from ground truth:
  $$\text{RMS} \overset{def}{=} \sqrt{\frac{1}{N} \sum_{k=1}^{N} \|\vec{y}_k - \vec{y}_{k,\text{gt}}\|_2^2}.$$

- *Joint weighted sum of squared differences SSD (photometric error):*
  The only measure that is absolute and not relative to the ground truth is the SSD. It is the average squared

gray value difference over all patches of the image correspondences $\vec{x}_k \leftrightarrow \vec{y}_k = \vec{x}_k + \vec{v}_k$:
$$Q \overset{def}{=} \frac{1}{N} \sum_{k=1}^{N} \sum_{\vec{x}} \mathbf{W}[\vec{x} - \vec{x}_k] \cdot (\mathcal{I}[\vec{x}] - \mathcal{J}[\vec{x} + \vec{v}_k])^2.$$

## 4.4. COnGRATS & ICL-NUIM RGB-D dataset

The COnGRATS dataset contains two road scenes of a construction site on a highway ('ConstructionSite') showing maneuvers at low velocities and another highway scene ('Highway') with the car travelling mainly straight ahead at a much higher speed. Both scenes use a setup of the camera similar to KITTI [14] and were generated using the pose information from the KITTI odometry dataset [15]. This enables us to use the extensive motion data in KITTI to generate a statistical model of ego-dynamics to be used as statistical prior. The results are shown in the first and second column of figure 2 and the mean and standard deviation are listed in table 1. The results show that JET, using image information, reduces the rotational error $\rho$ to approximately the half of the value of RPE without using prior knowledge. While using prior knowledge does not seem to have a large impact on the optimization of the rotation of RPE, it does have it for JET. Using the prior, JET is able to nearly halve the rotational error once more, compared to not using a prior. The observations for JET are also true for the translational error $\Omega$: the use of a prior more than halves the error. In contrast to the optimization of the rotation, the translation optimization of RPE also greatly benefits from using the prior, leading to a reduction of the error by more than a half. This behavior becomes very clear when comparing the histograms of $\rho$ and $\Omega$ for the cases with and without prior information (first and second column of figure 2 respectively). JET is the clear winner for the rotation optimization and also dominates the optimization of the translation without using the prior. Enabling the prior leads to a head to head situation for the translational error.

Regarding the SSD, it is very easy to see the influence of the optional Lucas-Kanade tracking for RPE. The value is strongly decreased. However, JET also dominates this area. It achieves SSD values that are clearly below the ground truth value indicating a very good quality of the optimization of the feature correspondences. Nevertheless, on an average the feature correspondences of JET deviate by about 1 pixel off the ground truth position. The reason for this behavior (similar and even better SSD value while still deviating from the ground truth position) can be explained by the use of patch matching and the existence of a locally nonconstant optical flow field (caused by rotation and translation in the direction of the optical axis leading to different scalings). Apart from that, the RMS value of JET is clearly superior to the results of RPE.

The ICL-NUIM RGB-D dataset we evaluated on contains synthetic data of a hand held camera which is carried through a living room ('LivingRoom02') and an office room

| KITTI | Rotation $\rho$ [deg] | | Translation $\Omega$ [deg] | | SSD | |
| Seq No. | RPE | JET | RPE | JET | RPE | JET |
|---|---|---|---|---|---|---|
| 0 | 0.188 | **0.096** | **6.582** | 6.749 | 1209.95 | **180.39** |
| 1 | 0.364 | **0.253** | 8.374 | **7.998** | 1553.32 | **178.64** |
| 2 | 0.154 | **0.061** | 1.703 | **1.502** | 1178.16 | **266.20** |
| 3 | 0.098 | **0.035** | 0.970 | **0.891** | 388.74 | **138.38** |
| 4 | 0.142 | **0.045** | 1.123 | **0.951** | 774.49 | **176.47** |
| 5 | 0.138 | **0.049** | 1.445 | **1.274** | 1120.63 | **212.27** |
| 6 | 0.436 | **0.358** | 11.129 | **10.765** | 1342.75 | **220.85** |
| 7 | 0.262 | **0.152** | 28.411 | **28.070** | 1587.19 | **219.49** |
| 8 | 0.169 | **0.063** | 9.536 | **9.429** | 1188.83 | **206.49** |
| 9 | 0.107 | **0.029** | 0.752 | **0.676** | 687.81 | **243.71** |
| 10 | 0.237 | **0.131** | 1.614 | **1.361** | 1218.97 | **275.59** |

Table 2: Evaluation on the KITTI training dataset.

('OfficeRoom02'). The motion is dominated by strong rotations and involves only slight translation. As the motion is less constrained, compared to vehicle motions, the positive influence of integrating prior knowledge is less pronounced. Therefore, we only present results without using the prior ($\xi_q = 0$, $\xi_R = 0$). They are visualized in the third column of figure 2 and listed in table 1.

In summary, the results of the RGB-D dataset are similar to the ones achieved on COnGRATS. JET is superior to RPE in optimizing the rotation and translation (see histograms in third column of 2). It is dominating the SSD results by achieving SSD values below the ground truth value and it is also clearly superior in optimizing the image point correspondences (RMS). Due to the harder requirements of data from a hand held camera, all results are slightly worse than they were for the COnGRATS dataset. Especially the optimization of the translation direction is very tough (see $\Omega$ in third column of figure 2 and table 1), when only slight magnitudes of the translation can be observed. Already a minor shaking of the hand, as it is simulated in the scenes, can lead to constantly and much pronounced changes in the direction of the translation. Even though the effect of this behavior only has a small influence on the relative pose and the optical flow in the image domain, it has a strong influence when looking at the evaluation of the direction of the translation. This is a limitation of our parametrization: the direction of the translation is almost undetermined due to its vanishing magnitude, and no scale is available due to the use of a mono camera setup.

Apart from this, the optimization of the unscaled relative pose and the feature correspondences was very successful and largely improved by including the photometric matching information when using JET.

## 4.5. KITTI Dataset

We also performed experiments on the KITTI dataset. Since KITTI does not provide ground truth for image point correspondences (*e.g.* via a dense depth or optical flow map), we cannot use ground truth for the correspondences and apply noise to them to serve as an initializiation. Therefore, we initialize the correspondences by employing propagation based tracking as presented in [12]. Similar to the experiments on the synthetic data, we compare the results of both methods. We use the KITTI ground truth of the pose and compare JET and RPE with respect to the rotational error $\rho$ and the translational error $\Omega$. In order to compare the quality of the feature correspondences of the two methods, we regard the photometric error (SSD).

The results of the experiment on the KITTI dataset are shown in table 2. The table presents the mean values of the Rodrigues angle ($\rho$), the angle of intersection of the translation ($\Omega$), and the SSD for each KITTI sequence that has ground truth available. The results confirm that JET performs clearly better than RPE in matters of rotation optimization. The mean of the rotational error $\rho$ is two to three times lower than the one of RPE. In terms of translation, the results show a head and head situation of RPE and JET with a slight lead of JET. Thus, in summary JET yields a significantly better pose than RPE.

Besides improving the pose, JET also refines the feature correspondences. However, as correspondence ground truth is not available in KITTI, the residual error in feature correspondences after performing JET cannot be determined. However, the feature correspondences from JET possess a much smaller photometric error (SSD) than after RPE optimization as can be seen in table 2.
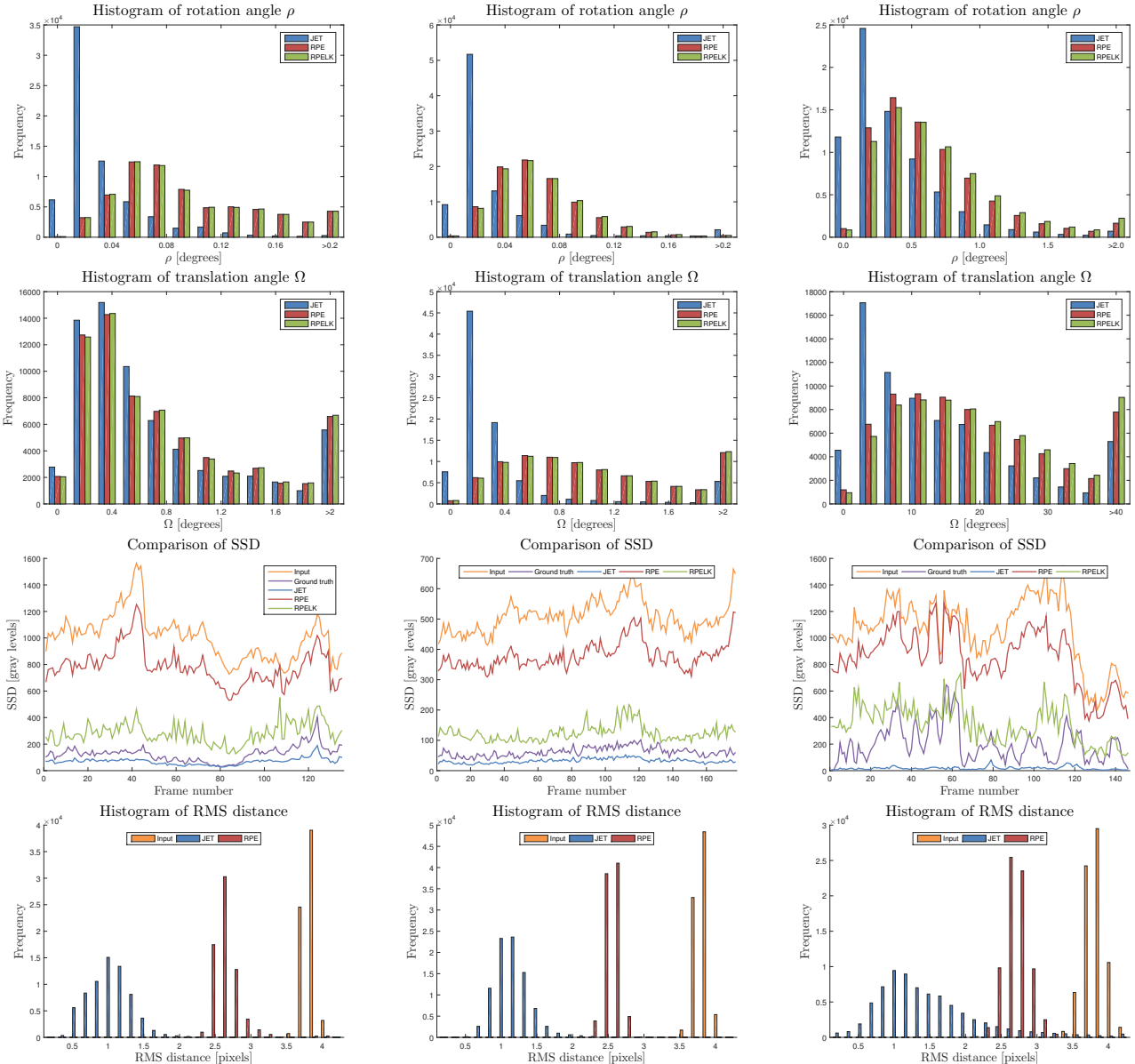
Figure 2: Evaluation on the COnGRATS sequence 'ConstructionSite' using prior knowledge ($\xi_Q = 1$, $\xi_R = 0.5$) (first column) and evaluation on the COnGRATS sequence 'Highway' and on the RGB-D sequence 'LivingRoom02' without using prior knowledge (second and third column). First and second row visualize the distributions of the quality measures $\rho$ (rotational error) and $\Omega$ (translational error) of the relative pose. The third row exhibits the SSD measure (photometric error) over the courses of the sequences and the last row visualizes the distribution of the RMS measures (positional error) of the correspondences.

## 5. Summary and Conclusion

This paper proposed a novel algorithm in the area of feature tracking and frame-to-frame pose estimation, denoted as Joint Epipolar Tracking (JET). The proposed algorithm employs a direct method to simultaneously optimize the epipolar geometry and feature correspondences.

It iteratively solves the minimization problem of the newly introduced joint loss function where additional statistical information about the motion can be included to serve as prior knowledge. The proposed method has been shown to perform better than the competing method of RPE optimization by experiments on several datasets, synthetic and real, such as COnGRATS, ICL-NUIM, and KITTI. It at-

tains real-time performance: approximately 30fps utilizing roughly 400 features with patch size of $9 \times 9$ pixels on a single thread of an Intel Core i7-6700 CPU. On an average, the rotational errors are three times smaller compared to RPE. The translation direction can be improved as well if the translation is sufficiently encoded in the optical flow of the image. Furthermore, the photometric error (SSD) of the feature patches is massively reduced in all cases which suggest a better quality also of the 3D information that can be computed from the point correspondences.

## References

[1] S. Agarwal, K. Mierle, and Others. Ceres solver. http://ceres-solver.org, 2012.

[2] H. Alismail, B. Browning, and S. Lucey. Photometric Bundle Adjustment for Vision-Based SLAM. In *Asian Conference on Computer Vision (ACCV)*, 2016.

[3] H. Badino, A. Yamamoto, and T. Kanade. Visual Odometry by Multi-frame Feature Integration. In *International Conference on Computer Vision (ICCV-W) Workshops*, pages 222–229, 2013.

[4] J. Berger, A. Neufeld, F. Becker, F. Lenzen, and C. Schnoerr. Second Order Minimum Energy Filtering on SE(3) with Nonlinear Measurement Equations. In *International Conference on Scale Space and Variational Methods in Computer Vision (SSVM)*, pages 397–409, 2015.

[5] D. Biedermann, M. Ochs, and R. Mester. COnGRATS: Realistic Simulation of Traffic Sequences for Autonomous Driving. In *Image and Vision Computing New Zealand (IVCNZ)*, 2015.

[6] H. Bradler, B. A. Wiegand, and R. Mester. The Statistics of Driving Sequences - And What We Can Learn from Them. In *International Conference on Computer Vision (ICCV-W) Workshops*, pages 106–114, 2015.

[7] G. Bradski. Opencv. http://opencv.org, 2000.

[8] I. Cvišić and I. Petrović. Stereo odometry based on careful feature selection and tracking. In *European Conference on Mobile Robots (ECMR)*, pages 1–6, 2015.

[9] J. Engel, V. Koltun, and D. Cremers. Direct Sparse Odometry. In *arXiv:1607.02565 [cs.CV]*, 2016.

[10] J. Engel, T. Schöps, and D. Cremers. LSD-SLAM: Large-Scale Direct Monocular SLAM. In *European Conference on Computer Vision (ECCV)*, pages 834–849, 2014.

[11] J. Engel, J. Sturm, and D. Cremers. Semi-dense Visual Odometry for a Monocular Camera. In *International Conference on Computer Vision (ICCV)*, pages 1449–1456, 2013.

[12] N. Fanani, M. Ochs, H. Bradler, and R. Mester. Keypoint trajectory estimation using propagation based tracking. In *Intelligent Vehicles Symposium (IV)*, 2016.

[13] C. Forster, M. Pizzoli, and D. Scaramuzza. SVO: Fast Semi-Direct Monocular Visual Odometry. In *International Conference on Robotics and Automation (ICRA)*, pages 15–22, 2014.

[14] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research (IJRR)*, 32(11):1231–1237, 2013.

[15] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354 – 3361, 2012.

[16] A. Handa, T. Whelan, J. McDonald, and A. Davison. A Benchmark for RGB-D Visual Odometry, 3D Reconstruction and SLAM. In *International Conference on Robotics and Automation (ICRA)*, pages 1524–1531, 2014.

[17] G. Klein and D. Murray. Parallel Tracking and Mapping for Small AR Workspaces. In *International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 225–234, 2007.

[18] F. Lenzen and J. Berger. Solution-Driven Adaptive Total Variation Regularization. In *International Conference on Scale Space and Variational Methods in Computer Vision (SSVM)*, pages 203–215, 2015.

[19] B. D. Lucas and T. Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 674–679, 1981.

[20] A. Neufeld, J. Berger, F. Lenzen, and C. Schnoerr. Estimating Vehicle Ego-Motion and Piecewise Planar Scene Structure from Optical Flow in a Continuous Framework. In *German Conference on Pattern Recognition (GCPR)*, pages 41–52, 2015.

[21] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison. DTAM: Dense tracking and mapping in real-time. In *International Conference on Computer Vision (ICCV)*, pages 2320–2327, 2011.

[22] M. Persson, T. Piccini, M. Felsberg, and R. Mester. Robust stereo visual odometry from monocular techniques. In *Intelligent Vehicles Symposium (IV)*, pages 686–691, 2015.

[23] T. Piccini, M. Persson, K. Nordberg, M. Felsberg, and R. Mester. Good Edgels to Track: Beating the Aperture Problem with Epipolar Geometry. In *European Conference on Computer Vision (ECCV-W) Workshops*, pages 652–664, 2014.

[24] M. Trummer, J. Denzler, and C. Munkelt. KLT Tracking Using Intrinsic and Extrinsic Camera Parameters in Consideration of Uncertainty. In *International Conference on Computer Vision Theory and Applications (VISAPP)*, pages 346–351, 2008.

[25] M. Trummer, C. Munkelt, and J. Denzler. Extending GKLT Tracking - Feature Tracking for Controlled Environments with Integrated Uncertainty Estimation. In *Scandinavian Conference on Image Analysis (SCIA)*, pages 460–469, 2009.

[26] L. Valgaerts, A. Bruhn, and J. Weickert. A Variational Model for the Joint Recovery of the Fundamental Matrix and the Optical Flow. In *German Conference on Pattern Recognition (GCPR)*, pages 314–324, 2008.

[27] C. Vogel, K. Schindler, and S. Roth. 3D Scene Flow Estimation with a Rigid Motion Prior. In *International Conference on Computer Vision (ICCV)*, pages 1291–1298, 2011.

[28] K. Yamaguchi, D. McAllester, and R. Urtasun. Robust Monocular Epipolar Flow Estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1862–1869, 2013.