Autonomous Robots
https://doi.org/10.1007/s10514-018-9816-4

Autonomous flight with robust visual odometry under dynamic lighting conditions

Pyojin Kim[1] · Hyeonbeom Lee[2] · H. Jin Kim[1]

Received: 29 December 2017 / Accepted: 6 October 2018

Abstract
Sensitivity to light conditions poses a challenge when utilizing visual odometry (VO) for autonomous navigation of small aerial vehicles in various applications. We present an illumination-robust direct visual odometry for a stable autonomous flight of an aerial robot under unpredictable light condition. The proposed stereo VO achieves robustness with respect to the light-changing environment by employing the patch-based affine illumination model to compensate abrupt, irregular illumination changes during direct motion estimation. We furthermore incorporate a motion prior from feature-based stereo visual odometry in the optimization, resulting in higher accuracy and more stable motion estimate. Thorough analyses of convergence rate and linearity index for the feature-based and direct VO methods support the effectiveness of the usage of the motion prior knowledge. We extensively evaluate the proposed algorithm on synthetic and real micro aerial vehicle datasets with ground-truth. Autonomous flight experiments with an aerial robot show that the proposed method successfully estimates 6-DoF pose under significant illumination changes.

Keywords Aerial robotics · Stereo visual odometry · Robustness · Illumination changes

1 Introduction

Autonomous aerial robots that are designed to perform tasks without direct human remote control rely on accurate state information. Due to the limitations of GPS or motion capture system, investigations have been performed to combine multiple sensors such as laser scanner, sonar, barometer in order to localize the aerial robots. Alternatively, vision-based state estimation so-called visual odometry (VO) (Scaramuzza and Fraundorfer 2011) can offer a less expensive solution with up to centimeter-level accuracy without sacrificing too much payload.

Unlike ground vehicle navigation (Nistér et al. 2004; Maimone et al. 2007), however, small autonomous aerial robots pose a challenge in applying VO. VOs for the aerial robots have to compute sufficiently fast and accurate position estimates to maintain active control at high refresh rate and avoid failure. They also should be light enough to run on an onboard computer with limited processing power. Because of these difficulties, VO algorithms for aerial robots are still actively researched with RGB-D camera (Huang et al. 2011; Valenti et al. 2014), stereo camera (Achtelik et al. 2012; Scaramuzza et al. 2014) and a single camera (Forster et al. 2014; von Stumberg et al. 2016).

In this work, we focus on the robustness of VO for an autonomous flight of the aerial robots. Although the accuracy and speed have been main objectives of many VO research (Forster et al. 2017; Mur-Artal et al. 2015), the robustness to external environmental changes has not been addressed much. Among the various environmental factors, light changes in an image, including highlights, shadows caused by the changes of the camera viewing angle, unpredictable changes of a light source, and the automatic exposure control, are inevitable phenomena that the aerial robots must

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10514-018-9816-4) contains supplementary material, which is available to authorized users.

✉ H. Jin Kim
hjinkim@snu.ac.kr

Pyojin Kim
rlavywls@snu.ac.kr

Hyeonbeom Lee
hbeomlee@knu.ac.kr

[1] Department of Mechanical and Aerospace Engineering, Seoul National University, Seoul, South Korea

[2] Department of Electronics Engineering, Kyungpook National University, Daegu, South Korea

Published online: 31 October 2018

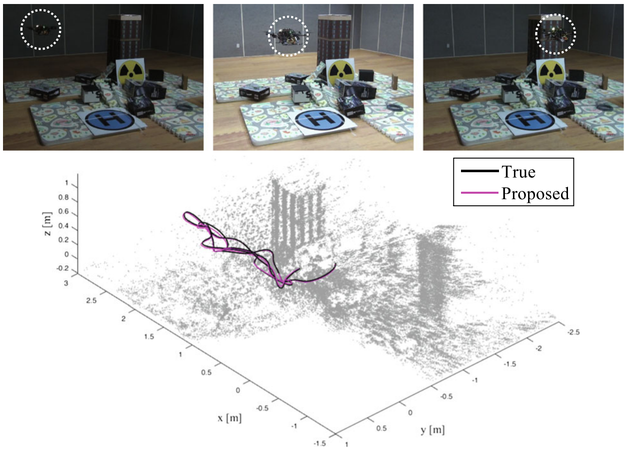


Fig. 1 Top: hexacopter aerial robot used in our autonomous flight experiments with varying light conditions by turning on and off the lights repeatedly. Bottom: the estimated (magenta) and true (black) trajectory with the proposed VO method overlap significantly and the point cloud is consistently recovered despite sudden and severe illumination changes (Color figure online)

deal with in practice. For example, it is well known that direct VO methods (Kerl et al. 2013), which perform state estimation using image brightness values, are very vulnerable to light variations, but the robustification to light changes still remains a challenge.

To address such issue, we propose a robust direct visual odometry algorithm that enables reliable autonomous flight of the aerial robots even in light-changing environments (see Fig. 1). The proposed stereo VO method *simultaneously* estimates the 6-DoF camera pose and the photometric parameters of the affine illumination change model (Jin et al. 2001) for individual patches in an image. Furthermore, we utilize a motion prior from feature-based VO to guide and stabilize direct motion estimation. Extensive evaluations show that the proposed algorithm can achieve a more accurate state estimation than other state-of-the-art VO methods in light-changing environments while maintaining comparable performance in normal light conditions. The contributions of the paper can be summarized as follows.

- We present a novel direct VO algorithm that is robust under challenging lighting environments by including local affine parameters for estimating irregular illumination changes.
- We integrate a motion prior from the feature-based method into the direct approach for stable motion estimation, and analyze its usefulness in terms of convergence property.
- We demonstrate a real-time system enabling autonomous flights of the aerial robot in environments with irregular illumination changes.

Section 2 reviews related literature on direct motion estimation methods, particularly those involving light changes. Section 3 introduces some required notations and the problem. Section 4 provides an overview of the VO pipeline, and Sect. 5 explains the proposed motion estimation algorithm. Section 6 analyzes the convergence property of the proposed method. Section 7 provides evaluation results and demonstrates how the proposed method makes the aerial robot fly autonomously under challenging lighting conditions. We conclude in Sect. 8.

2 Related work

In robotics and computer vision communities, various VO and visual simultaneous localization and mapping (V-SLAM) methods have been researched actively in the last decade. From the vast literature in the visual navigation field, we review related work in terms of illumination changes and implementation on aerial robots.

VO algorithms can be classified into indirect and direct methods depending on the type of visual information (Engel et al. 2018). Indirect methods utilize an intermediate representation to track the camera pose rather than the direct measurements. Feature-based methods, the most widely used indirect methods, show successful 6-DoF camera motion estimation (Geiger et al. 2011; Mur-Artal et al. 2015; Zhang et al. 2017). However, they require enough brightness and textures to extract consistent keypoints from an image (Fang and Scherer 2014). This requirement is not satisfied in varying illumination conditions considered in this paper.

Direct VO methods (Kerl et al. 2013; Engel et al. 2014; Forster et al. 2014) estimate 6-DoF camera motion by minimizing the photometric error between image frames, and they are receiving attention for their improved accuracy and robustness to little texture with the help of hardware progress. They heavily rely on the photo-consistency assumption that a scene point appears with constant brightness intensity across multiple images. Kerl et al. (2013) estimates the RGB-D camera motion accurately with a robust error function which rejects the noise and outliers in the photometric error. In Forster et al. (2014), a semi-direct monocular VO is implemented on the onboard computer of a multirotor, showing precise and fast state estimation results by combining the advantages of feature-based and direct methods, and it is extended to multi-camera systems in Forster et al. (2017). Although these direct VO methods demonstrate impressive levels of accuracy, they have not been fully tested in challenging environments where the photo-consistency assumption does not hold (e.g., abrupt and irregular illumination changes occur).

Only a few direct VO methods give consideration to illumination changes during the direct motion estimation. It is assumed in Klose et al. (2013) that the entire pixels follow the same affine illumination change model (Jin et al. 2001). In order to ignore the illumination changes altogether between image frames, Kerl et al. (2014) estimates a pure albedo image of the texture. In Engel et al. (2015), the modified photometric error based on the affine brightness change is employed. Alismail et al. (2017) addresses the vulnerability of light changes by using a binary descriptor, which is invariant to monotonic changes in intensity. Park et al. (2017) evaluates various direct image alignment methods for their accuracy and robustness under challenging lighting conditions. Recently, the direct sparse model (Engel et al. 2018; Wang et al. 2017) is proposed, and photometric camera calibration is considered explicitly to mitigate photo-consistency assumption (Bergmann et al. 2018). While these methods present superior motion estimation even in light-changing environments, they have not been applied to the autonomous flight of an aerial robot in an environment with severe light changes.

The work which is the most similar to the proposed approach is Krombach et al. (2016) and Forster et al. (2017)), which are the two-stage VO methods combining feature based and direct tracking approaches in a sequential manner. Krombach et al. (2016) extends LSD-SLAM to the stereo camera and employs a feature-based VO to estimate the motion between keyframes. The feasibility analysis of using the feature-based VO as a motion prior has not been addressed in detail while our manuscript provides in-depth analyses of convergence property of direct and feature-based functions. In Forster et al. (2017), the photometric error is minimized at first, and then the estimated camera pose and the position of the observed 3D points are again optimized to reduce the reprojection residuals. SVO requires separate mapping thread additionally to minimize the reprojection residuals, and the minimization procedure of different cost functions is different from the proposed method. Both approaches have not performed sufficient performance evaluation with the aerial robot in a challenging environment where severe lighting changes occur.

Our algorithm builds on our previous work of Kim et al. (2015), which is the patch-based illumination-robust direct visual odometry that estimates not only the 6-DoF camera pose but also the parameters of the affine illumination change model for individual patches. We newly integrate feature-based VO as a motion prior to the proposed direct VO method to guide the optimization by seeding it with an estimate closer to the true solution, resulting in more stable estimates. Importantly, we analyze and compare convergence rate and linearity index of each cost function used in feature-based and direct VO to support the usage of the feature-based VO as the motion prior. We validate the effectiveness and accuracy of our VO algorithm by recovering the 6-DoF camera motion and the photometric parameters of the author-collected dataset where irregular illumination changes exist in the stereo image sequences as well as on manually disturbed sequences of the EuRoC dataset (Burri and Nikolic 2016). Furthermore, we implement the proposed approach on an aerial robot with a stereo camera, achieving stable autonomous 3-D flight in light-changing environments.

3 Notation and problem statement

We organize the notations using a stereo camera model, but the setup can be transferred to the RGB-D camera model in Kim et al. (2015). The superscripts (l) and (r) denote the left and right camera respectively and k is used to represent the frame index. $I_i^{(l)k}$ is the ith image patch in the left image at time step k. A pixel point is denoted with $\mathbf{x}_{ij}^{(l)k} = \left[x_{ij}^{(l)k}, y_{ij}^{(l)k}\right]^\top$, where the subscript ij represents the pixel index j in the ith image patch. The center point $\mathbf{x}_{ic}^{(l)k}$ of the ith image patch is the detected keypoint in the feature-based VO. The 3D points $\mathbf{X}_{ij}^{(l)k} = \left[X_{ij}^{(l)k}, Y_{ij}^{(l)k}, Z_{ij}^{(l)k}\right]^\top$ expressed in left camera coordinates $\{C^k\}$ are mapped to pixel coordinates $\mathbf{x}_{ij}^{(l)k}$ through the camera projection function $\pi : \mathbb{R}^3 \mapsto \mathbb{R}^2$:

$$\mathbf{x}_{ij}^{(l)k} = \pi\left(\mathbf{X}_{ij}^{(l)k}\right) = \begin{bmatrix} \frac{f \cdot X_{ij}^{(l)k}}{Z_{ij}^{(l)k}} + p_x \\ \frac{f \cdot Y_{ij}^{(l)k}}{Z_{ij}^{(l)k}} + p_y \end{bmatrix} \tag{1}$$

where f, p_x, p_y are the intrinsic calibration parameters of the rectified images. Conversely, we can compute a 3D point $\mathbf{X}_{ij}^{(l)k}$ with the depth value $Z_{ij}^{(l)k}$ and $\mathbf{x}_{ij}^{(l)k}$ through the inverse

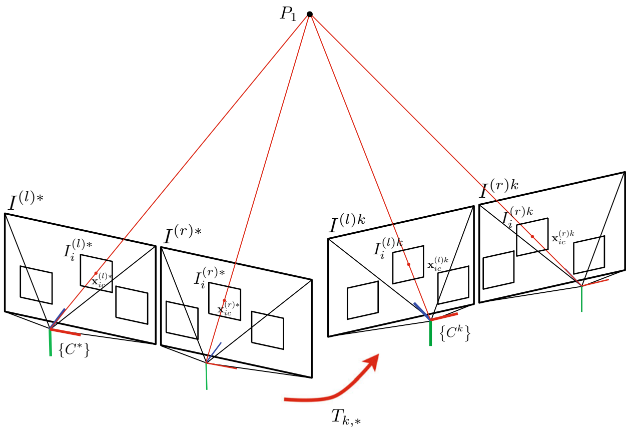


Fig. 2 Stereo camera model and image coordinate systems

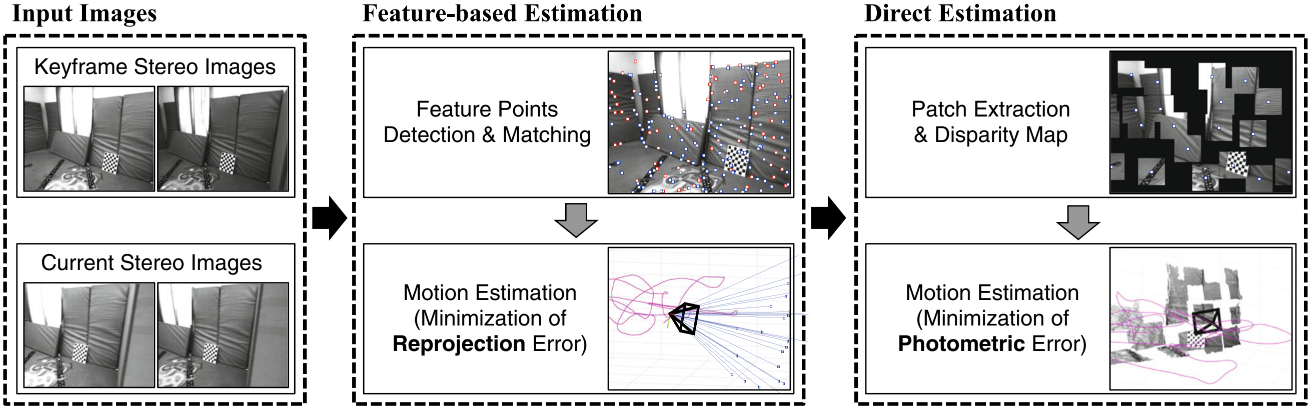

Input Images
Keyframe Stereo Images
Current Stereo Images
Feature-based Estimation
Feature Points Detection & Matching
Motion Estimation (Minimization of Reprojection Error)
Direct Estimation
Patch Extraction & Disparity Map
Motion Estimation (Minimization of Photometric Error)

Fig. 3 Overview of the proposed stereo visual odometry pipeline

projection function $\pi^{-1}: \mathbb{R}^2 \mapsto \mathbb{R}^3$:

$$\mathbf{X}_{ij}^{(l)k} = \pi^{-1}\left(\mathbf{x}_{ij}^{(l)k}, Z_{ij}^{(l)k}\right) = \begin{bmatrix} \frac{x_{ij}^{(l)k} - p_x}{f} Z_{ij}^{(l)k} \\ \frac{y_{ij}^{(l)k} - p_y}{f} Z_{ij}^{(l)k} \\ Z_{ij}^{(l)k} \end{bmatrix} \tag{2}$$

For the proposed direct method, we compute a local dense depth map for each keyframe. We model the relative motion of the left camera frame between $\left\{C^k\right\}$ at time step k and $\{C^*\}$ at keyframe as rigid body transformation $T_{k,*} \in SE(3)$:

$$\tilde{\mathbf{X}}^{(l)k} = T_{k,*}\tilde{\mathbf{X}}^{(l)*} \tag{3}$$

where $\tilde{\mathbf{X}}^{(l)k} = [\mathbf{X}^{(l)k^\top}, 1]^\top$ is the homogeneous form of $\mathbf{X}^{(l)k}$. A minimal representation of Lie group $SE(3)$, i.e., Lie algebra $se(3)$ parameter $\boldsymbol{\xi}$, is used to represent the incremental displacements during the numerical nonlinear optimization. We denote the Lie algebra with a 6 × 1 vector $\boldsymbol{\xi} = \left[\boldsymbol{\nu}^\top, \boldsymbol{\omega}^\top\right]^\top$ where $\boldsymbol{\nu}$ and $\boldsymbol{\omega}$ are infinitesimal translation and rotation in the tangent space of the matrix group $SE(3)$. The exponential map between the Lie algebra $se(3)$ and the rigid body transformation $T \in SE(3)$ can be written as follows:

$$T(\boldsymbol{\xi}) = \exp(\hat{\xi}) \tag{4}$$

where $\hat{\xi}$ is a 4 × 4 twist matrix from the Lie algebra $\boldsymbol{\xi}$ Ma et al. (2012). The above-defined notations and equations are illustrated in Fig. 2.

The problem we want to solve is to estimate the relative motion of the stereo camera $T_{k,*}$ given a sequence of image frames and the corresponding depth maps under arbitrary, abrupt, and partial illumination changes between the time step k and keyframe.

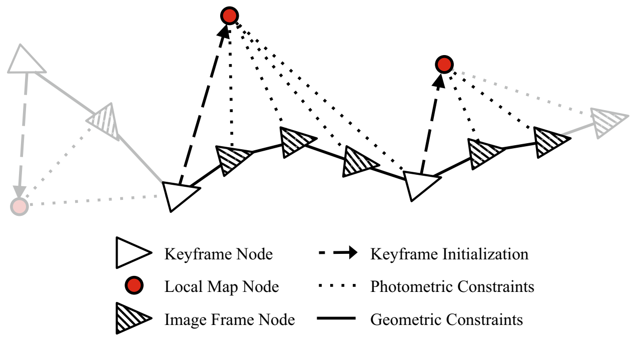


Fig. 4 Topological representation of the proposed algorithm. We estimate an initial camera pose of two consecutive images with the geometric constraints (reprojection error). After initializing the local map at the keyframe, we refine the camera pose and photometric parameters with the photometric constraints (modified photometric error). When the distance between the keyframe and current image frame becomes far enough, we set the current image frame to the keyframe

4 System overview

Figure 3 provides an overview of the proposed stereo VO. The proposed method has two main steps: 1) feature-based VO for estimating initial camera pose as a motion prior; and 2) illumination-robust direct VO for refining the camera pose and photometric parameters to achieve higher accuracy. This sequential VO allows stable and accurate 6-DoF camera tracking in light-changing environments. We obtain the overall trajectory by concatenating the frame-to-keyframe motion estimation illustrated in Fig. 4.

Our feature-based VO method is largely based on Geiger et al. (2011). We detect the salient feature points and obtain feature correspondence. With the matched features, we estimate a camera pose that minimizes the sum of the squared left and right reprojection error using three randomly selected correspondences in a RANSAC scheme (for full details, refer to Geiger et al. (2011)). If the feature-based VO prior fails due to low textured areas or light-changing environments, the proposed method performs the next direct VO approach without motion prior information.

For considering both global and local illumination changes in an image, we generate image patches around the matched feature points used in the feature-based VO. We initialize the camera pose from the feature-based estimation and the model parameters of individual patches following the affine illumination model (Kim et al. 2015). We refine the camera pose and photometric parameters by minimizing the newly proposed photometric error, which is based on the modified photo-consistency assumption explained in Sect. 5.2.1, for compensation of illumination changes between image frames.

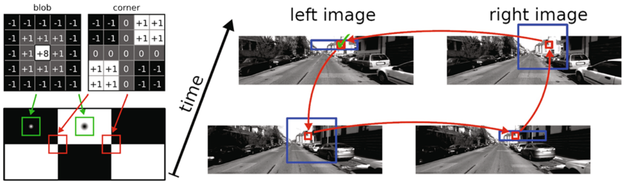
5 Visual odometry pipeline

5.1 Feature-based estimation

5.1.1 Feature detection and matching

We detect the feature points by filtering the left and right images of the two consecutive image frames with the 5 × 5 kernels for finding blobs and corners as shown in the left of Fig. 5. Bucketing Kitt et al. (2010) is utilized to spread them uniformly across the entire image domain and reduce the number of features (for full details, refer to Geiger et al. 2011). We apply non-maximum and non-minimum suppression to the filtered four images for extracting the feature candidates.

The distribution of u- and v-directional image gradient around the feature candidates is employed as a descriptor for feature matching. To measure the similarity between the descriptors, we use the sum of absolute differences (SAD). We solve the feature correspondences by matching features with the two temporally consecutive stereo pairs and perform circular matching illustrated in the right of Fig. 5. We find the best match between the current and previous images within an 11×11 search window. When matching between the left and right images, we additionally utilize the epipolar constraint.



blob

-1	-1	-1	-1	-1
-1	+1	+1	+1	-1
-1	+1	+8	+1	-1
-1	+1	+1	+1	-1
-1	-1	-1	-1	-1

corner

-1	-1	0	+1	+1
-1	-1	0	+1	+1
0	0	0	0	0
+1	+1	0	-1	-1
+1	+1	0	-1	-1

Fig. 5 Kernels for feature detection (left) and circular matching strategy used in solving the feature correspondence (right) (figures courtesy of Andreas Geiger)

5.1.2 Reprojection error minimization

We compute the relative motion of the stereo camera $T_{k,k-1}(\boldsymbol{\xi})$ by minimizing the sum of squared left and right reprojection error given the matched feature points. The reprojection error of the matched features in the left camera can be written as follows:

$$r_i^{(l)}(\boldsymbol{\xi}) = \left\| \mathbf{x}_{ic}^{(l)k} - w^{(l)}(\boldsymbol{\xi}, \mathbf{x}_{ic}^{(l)k-1}) \right\| \tag{5}$$

$$w^{(l)}(\boldsymbol{\xi}, \mathbf{x}_{ic}^{(l)k-1}) = \pi^{(l)}(T_{k,k-1}(\boldsymbol{\xi}) \cdot \pi^{(l)-1}(\mathbf{x}_{ic}^{(l)k-1}, Z_i^{(l)k-1})) \tag{6}$$

where $\boldsymbol{\xi} \in \mathbb{R}^6$ represents the relative motion of the stereo camera and $w^{(l)}(\boldsymbol{\xi}, \mathbf{x}_{ic}^{(l)k-1})$ is the warping function of the left camera, which maps a center point of ith patch $\mathbf{x}_{ic}^{(l)k-1}$ in the previous left image to its pixel coordinate in the current left image frame given the relative camera motion $\boldsymbol{\xi}$. The reprojection error of the features in the right camera can also be written in the same way with the superscript (r) instead of (l) in Eqs. (5) and (6). The objective energy function in the feature-based estimation is the sum of squared left and right reprojection error as follows:

$$\boldsymbol{\xi}^* = \arg\min_{\boldsymbol{\xi}} \sum_{i=1}^{N} \left[\left(r_i^{(l)}(\boldsymbol{\xi}) \right)^2 + \left(r_i^{(r)}(\boldsymbol{\xi}) \right)^2 \right] \tag{7}$$

where N is the number of the matched feature points. We use the relative motion of the camera $\boldsymbol{\xi}$ as a RANSAC model to reject outliers in feature matches. Given all inlier features from the RANSAC, we can obtain the optimal relative motion of the camera with the Gauss–Newton method for solving Eq. (7).

5.2 Direct estimation

5.2.1 Affine illumination change model

The traditional photo-consistency assumption commonly used in direct visual odometry (Kerl et al. 2013; Forster et al. 2014) denotes that the same 3D points should have the same intensity values across multiple images. Unfortunately, this assumption almost never holds in real-world applications because light variations take place frequently. Thus, we employ the modified photo-consistency assumption which can make up for not only the global but also the local illumination changes between the current and keyframe time steps, proposed in Kim et al. (2015):

$$\lambda_i I_i^{(l)k} + \delta_i = I_i^{(l)*} \tag{8}$$

where λ_i and δ_i denote the photometric parameters for explaining contrast and brightness change of the ith patch in the left image, which will have values close to one and zero, respectively in the normal environments without obvious illumination changes. We generate the image patches around the matched feature points in the feature-based VO, and select them with the planarity test (Kim et al. 2015), which determines whether the selected patches are on the plane in the 3D space or not because 3D points on the same plane undergo the similar illumination changes. If there are few feature points in low-texture environments, we utilize the blob detector like LoG, DoG, SURF to extract some useful plane patches for direct tracking (for full details, refer to Kim et al. 2015). Patch size is one of the user-defined custom parameters, and we create the patches with a size of 91 × 91 pixels. We utilize at most 16 patches spread uniformly across the entire image if there is not enough number of the planar patches. We can compensate both global and local illumination changes because each patch can have different photometric parameters.

5.2.2 Modified photometric error minimization

We simultaneously estimate the camera pose $T_{k,*}(\boldsymbol{\xi})$ and the photometric parameters per patch (e.g., $\{\lambda_1, \delta_1\}, \ldots, \{\lambda_m, \delta_m\}$ where m is the number of patches) by minimizing the sum of squared modified photometric error. The modified photometric error of the jth pixel in the ith patch can be written as follows:

$$r_{ij}(\mathbf{z}) = \lambda_i I_i^{(l)k}(w^{(l)}(\boldsymbol{\xi}, \mathbf{x}_{ij}^{(l)*})) + \delta_i - I_i^{(l)*}(\mathbf{x}_{ij}^{(l)*}) \qquad (9)$$

$$\mathbf{z} := \left[\boldsymbol{\xi}^\top, \lambda_1, \delta_1, \ldots, \lambda_m, \delta_m\right]^\top \in \mathbb{R}^{6+2m} \qquad (10)$$

where $\mathbf{z}$ is the integrated model parameter consisting of the relative motion of the camera and the photometric parameters per patch for the sake of simplicity. To perform the warping in Eq. (9), we generate the local map of the keyframe consisting of the dense depth map from the dense stereo matching method called LIBELAS (Geiger et al. 2010), and the brightness information from the left keyframe image. We fix the dense depth map of the keyframe until the new keyframe is initialized. The optimal model parameter $\mathbf{z}^*$ which minimizes the weighted sum of squared modified photometric error can be obtained by solving the following non-linear weighted least square problem:

$$\mathbf{z}^* = \arg\min_{\mathbf{z}} \sum_{i=1}^{m} \sum_{j=1}^{n} W(r_{ij}) r_{ij}^2(\mathbf{z}) \qquad (11)$$

$$W_\mathrm{t}(r_{ij}) = \frac{\nu + 1}{\nu + \left(\frac{r_{ij}}{\sigma_\mathrm{t}}\right)^2}$$

where n is the number of pixels in each patch and $W(r_{ij})$ is the weighting function from the student t-distribution in order to achieve the robustness against outliers caused by occlusions, dynamic objects, and sensor noise (Kerl 2012). Among the various weight functions like Tukey and Huber, we employ the student t-distribution for its effectiveness in the direct method (for full details, refer to Kerl 2012). We initialize the λ and δ of each patch for the newly created keyframe as one and zero, respectively, and these photometric parameters are continuously updated until the new keyframe is initialized. Our approach utilizes the previous frame's photometric parameters to initialize new optimization in the next subsequent frame. We use the Gauss–Newton algorithm for solving the iteratively re-weighted nonlinear least square (IRLS) problem in Eq. (11). We compute the Jacobian matrix with the efficient second-order minimization (ESM) method (Benhimane and Malis 2004) because it outperforms the other methods such as the forward compositional (FC) and inverse compositional (IC) approaches (Klose et al. 2013; Engel et al. 2014). We employ a coarse-to-fine approach with the image pyramid method for robustness and faster convergence. Note that there should be enough valid pixels in each image patch for accurate and stable correction of light changes.

Our approach might not work well when the feature-based tracking does not give a good initialization for direct tracking, or the cost function in Eq. (11) of the direct tracking is highly nonlinear. Also, the critical part of the proposed algorithm, the compensation for light changes with the photometric parameters, is limited when the images are too dark or too bright. Moreover, there should be enough valid pixels in each image patch for accurate and stable correction of light changes, about 20% of the total number of pixels in each image patch empirically.

5.3 Discussion

The motion prior from the feature-based VO (Sect. 5.1) for the proposed direct VO (Sect. 5.2) seems to be unnecessary and redundant, but we carefully design the proposed stereo visual odometry to solve two critical issues in the direct VO under light-changing environments: more stable motion estimation and higher accuracy.

5.3.1 Stable motion estimation

We occasionally observe in the direct VO that a light-changing environment can lead to "jumps" in the motion estimate (Kerl et al. 2013; Forster et al. 2017). Due to the nature of VO, jumps have a significant impact on the estimated trajectories because VO drift continues to accumulate. We solve this problem by using a motion prior from the feature-based VO whose computation is less intensive than the direct method (Forster et al. 2014) and does not cause noticeable increases in the overall computational time

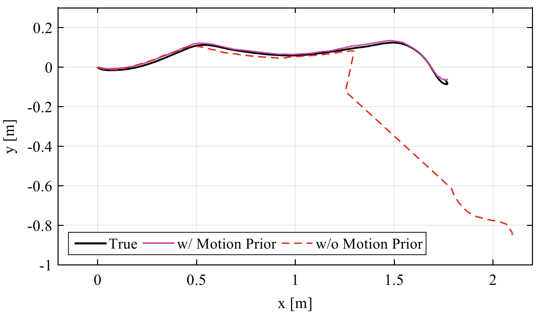


Fig. 6 The motion prior from the feature-based visual odometry (magenta) stabilizes the direct visual odometry significantly. We can observe a jump in the estimated trajectories with the constant motion model (red) (Color figure online)

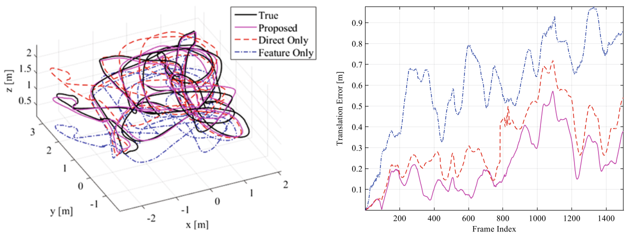


Fig. 7 Our sequential VO shows the best accuracy among other direct only (Kim et al. 2015) or feature only VO methods (Geiger et al. 2011)

as shown in the evaluation section later, resulting in stable motion estimation as illustrated in Fig. 6.

5.3.2 High accuracy

It is well-known that a good initial pose can be beneficial for the direct VO during the nonlinear optimization. The presence of the motion prior knowledge improves not only stability, but also accuracy as shown in Fig. 7. Although both the proposed method and direct only method (Kim et al. 2015) employ the same nonlinear optimization formulation in Eq. (11), we can obtain the more accurate camera pose with the proposed algorithm thanks to the motion prior.

6 Energy function analysis

We analyze the linearity of objective energy function and convergence rate used in feature-based and direct VO. These analyses theoretically support the usage of the feature-based VO as the motion prior.

6.1 Energy function convergence

The objective energy function (cost function) of the feature-based and direct estimation with respect to the camera pose can be written as follows:

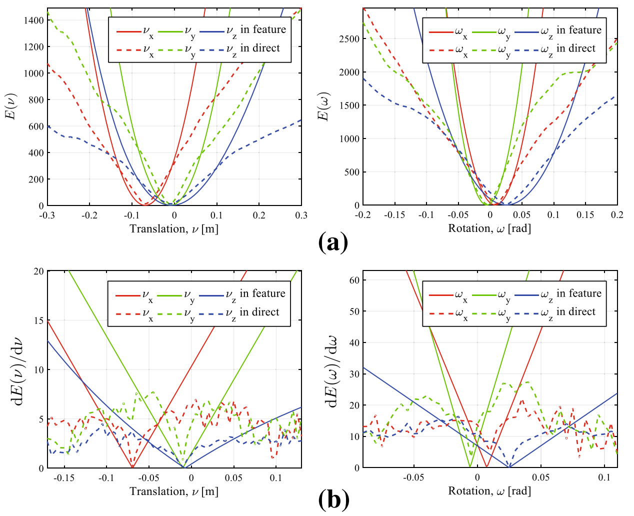


Fig. 8 The tendency of **a** the reprojection and photometric error for transformation with respect to each translational and rotational direction, and **b** the first derivatives of the (**a**)

$$E_f(\boldsymbol{\xi}) = \frac{1}{N}\sum_{i=1}^{N}\left[\left(r_i^{(l)}(\boldsymbol{\xi})\right)^2 + \left(r_i^{(r)}(\boldsymbol{\xi})\right)^2\right] \tag{12}$$

$$E_d(\boldsymbol{\xi}) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} W(r_{ij}) r_{ij}^2(\boldsymbol{\xi}) \tag{13}$$

where the subscripts f and d denote the feature-based and direct estimation, respectively.

For comparing the convergence rate of each cost function, we plot the average reprojection and photometric error in Eqs. (12) and (13) with respect to the 6-DoF camera pose in the vicinity of the true camera pose in Fig. 8a. The range of translation and rotation error is in ± 0.3 m and ± 0.2 radian, respectively. We compute them by warping the feature points and images separately along each degree of freedom in the 6-DoF camera pose (i.e., the Lie algebra parameter $\boldsymbol{\xi} = \left[\nu_x, \nu_y, \nu_z, \omega_x, \omega_y, \omega_z\right]^\top$) while the other parameters in $\boldsymbol{\xi}$ are fixed to the true camera pose. Although the energy functions are highly nonlinear, both error plots have the distinct minimum values at similar places for each axis in (a).

In Fig. 8b, which shows the derivatives of the two error plots, a notable difference exists between the feature-based and direct estimation. When the camera pose is far from the true camera pose, the slope of error plots of the feature-based method is very steep compared to the direct method. Therefore, the farther the currently estimated camera pose is from the true camera pose, the faster the estimated camera pose approaches the distinct minimum, especially when the feature-based method is applied instead of the direct method. In the vicinity of the valley, however, the slope of the feature-

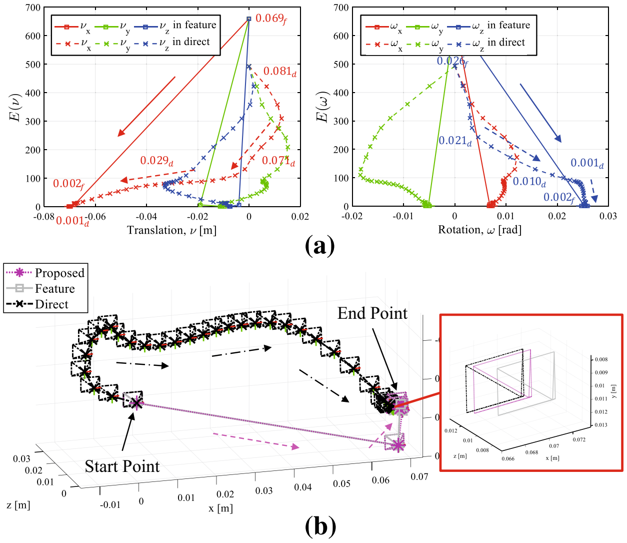


Fig. 9 Convergence history of the feature-based, direct, and the proposed VO approaches with the error distance from the true value. We visualize (**b**) each 6-DoF camera pose corresponding to the square and cross marks in (**a**). The arrows in (**a**) indicate the direction of the convergence of each method

based estimation error flattens gradually as the camera pose approaches the true camera pose. The direct method can be more efficient and accurate than the feature-based method especially when we start the optimization close to the true camera pose.

Figure 9 shows the convergence property of the feature-based, direct, and the proposed VO approaches in the optimization with respect to the 6-DoF camera pose. The numbers represent the error distance of each moment, and the subscript f and d denote the feature-based and direct estimation method, respectively. We write down the error value only for the x-axis (red) in the translation (left) and z-axis (blue) in the rotation (right) for readability. Each square or cross mark refers to the updated 6-DoF camera motion at each iteration during the optimization. In the feature-based estimation, the estimated camera motion rapidly converges to the neighborhood of the true camera pose within only two iterations, and the optimization stops in six iterations. In contrast, the direct estimation converges slowly near the true camera pose, and the total iteration in the optimization is 35, which is about six times more than the feature-based method. Although the direct method is slower especially from a distance, the finally estimated camera motion is a little closer to the true camera pose than the feature-based method due to the characteristics of this gradual approach. The final translation error of the feature-based and direct estimation is 4.6 mm and 3.7 mm, respectively.

We further analyze the convergence property of the proposed method compared to the two kinds of cost functions in Fig. 9b. The proposed method first jumps off the starting point (the origin) into the end point (true camera pose) very closely by performing a feature-based VO method. After that, with a motion prior from the feature-based VO, our two-stage method performs the direct estimation to get closer to the true camera pose. The enlarged figure on the right shows convergence to a similar 3D point from the direct only and the proposed method. The total number of iteration in the direct only method (black) during the optimization is 35 whereas the proposed two-stage method takes six iterations from the feature-based and three iterations from the direct method, showing the effectiveness of the proposed sequential method design.

6.2 Measurement equation linearity

We validate the better convergence property of the feature-based method by analyzing the dimensionless linearity index (Civera et al. 2008), which represents the degree of linearity in the nonlinear measurement equation. The more linear the measurement equation is for the 6-DoF camera motion, the faster the nonlinear optimization converges. The dimensionless linearity index (DLI) of each measurement equation considering the Lie group $SE(3)$ can be written as follows:

$$L = \left| \frac{\left.\frac{\partial^2 h}{\partial \boldsymbol{\xi}^2}\right|_{\boldsymbol{\xi}=\boldsymbol{\xi}_0} \Delta\boldsymbol{\xi}}{\left.\frac{\partial h}{\partial \boldsymbol{\xi}}\right|_{\boldsymbol{\xi}=\boldsymbol{\xi}_0}} \right|$$

$$h_x(\boldsymbol{\xi}) = \Big[\pi(T(\boldsymbol{\xi}) \cdot \pi^{-1}(\mathbf{x}, Z(\mathbf{x})))\Big]_x$$

$$h_I(\boldsymbol{\xi}) = I(\pi(T(\boldsymbol{\xi}) \cdot \pi^{-1}(\mathbf{x}, Z(\mathbf{x})))) \tag{14}$$

where h in Eq. (14) denotes the observation model that h_x is the x component of the warping function in the feature-based method, and h_I is the image intensity observation model in the direct method. We omit the y component of the warping function, h_y, because its linearity index results are symmetric to the h_x. $\boldsymbol{\xi}_0$ is the center point of camera motion used in the first and second derivative, and $\Delta\boldsymbol{\xi} = \left[\Delta\nu^\top, \Delta\omega^\top\right]^\top$ is the transformation of 6-DoF camera motion from the center point $\boldsymbol{\xi}_0$. When $L \approx 0$, the observation model can be considered as a linear model in the interval $\Delta\boldsymbol{\xi}$, and vice versa. Unlike Civera et al. (2008), we newly derive and calculate the DLI of each observation model with respect to the 6-DoF camera motion with the 3D projection model. We provide the more detailed derivation of the DLI and explanations of each component in the Appendix.

We plot the DLI of each estimation method with respect to the translational and rotational transformation from the true camera pose in Fig. 10. The transformation of the translation and rotation from the true camera pose $\Delta\boldsymbol{\xi} = \left[\Delta\nu^\top, \Delta\omega^\top\right]^\top$ is in the range of ± 0.05 m and ± 0.05 radian, respectively. The main factors, which cause nonlinearity in $L_x\ (\Delta\nu)$, are

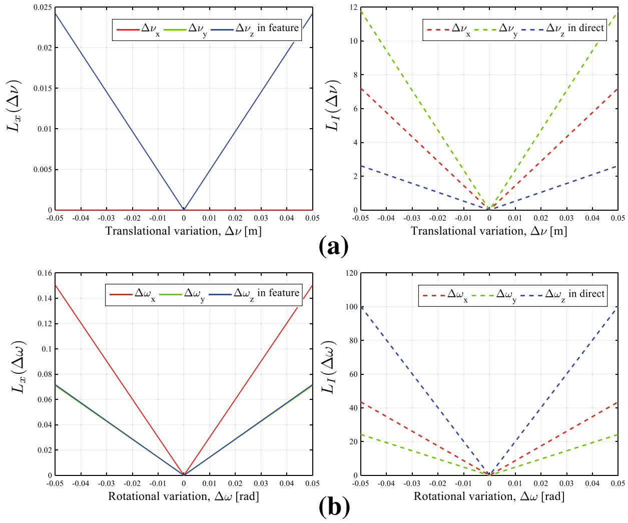


Fig. 10 Dimensionless linearity index (DLI) of the feature-based and direct estimation with respect to the **a** translational and **b** rotational motion of the camera

the depth of the feature point and the camera movement along the forward direction shown in Fig. 10a on the left. In the results of the feature-based method, the DLI in terms of x component in translation is exactly zero, which means that the warping function is linear to the translational movement of the x-axis. Back and forth motion (z component in translational motion) of the camera is nonlinear, and the DLI becomes larger as the camera pose is away from the center point. The direct method, however, has a high degree of nonlinearity for translational motion in all directions, and it is approximately 100 times larger than the feature-based method. In particular, the image intensity function $I(\cdot)$, which maps from the pixel coordinates to the image intensity in Eq. (14), causes such a severe nonlinearity.

Figure 10b shows the similar behavior of the DLI along the rotational motion of the camera. Both estimation methods have the nonlinearity for rotational motion in all directions, and it gets larger as the camera motion drifts farther away from the true camera pose. However, the direct method has approximately 100 times more severe nonlinearity than the feature-based method.

In conclusion, the feature-based method, which minimizes the reprojection error in Eq. (7), converges more quickly to the vicinity of the true camera pose than the direct method as shown in Fig. 9 thanks to the high degree of linearity. However, in the proximity of the true camera pose where the effect of nonlinearity is negligible, the direct method ultimately can approach the true camera pose a little closer. These analyses confirm the effectiveness of our decision to use the feature-based VO as the motion prior, followed by the proposed illumination-robust direct VO.

7 Experimental results

We extensively evaluate the effectiveness of the proposed illumination-robust stereo visual odometry with two different experiments. We test the accuracy of the proposed algorithm under the light-changing environments with the stereo image datasets that include irregular illumination changes. The other VO baselines are applied to these datasets to compare the performance of the motion estimation. Next, we construct a 3-D autonomous flight system with the proposed algorithm for online use. We perform flight experiments with the aerial robot to test the proposed algorithm in terms of accuracy and robustness to light variations, showing the short-term autonomous flight capability.

7.1 Experiments on the datasets

We test the proposed algorithm on the synthetic EuRoC benchmark (Burri and Nikolic 2016) and our own dataset that contains illumination changes through the stereo image sequences. We apply artificial illumination changes to RGB images in the EuRoC benchmark to validate the proposed algorithm under light-changing environments. We collect the stereo image datasets including actual light variations for evaluating the consistency of the proposed algorithm. Since many VO baselines accept only RGB-D input, we convert the stereo camera data to match the desired input format.

We compare the proposed VO method against other VO algorithms: the Dense Visual Odometry (DVO) (Kerl et al. 2013), the Efficient DVO (EDVO) (Klose et al. 2013), and the depth enhanced monocular odometry (DEMO) (Zhang et al. 2017). DVO estimates the camera pose by minimizing the photometric error within the overall images based on the photo-consistency assumption. DVO is a direct VO without the affine illumination change model. EDVO is an advanced direct tracking method which performs per-image brightness correction by considering a global affine illumination. Thus, EDVO estimates the photometric parameters per image whereas the proposed algorithm estimates them per patch. DEMO is one of the state-of-the-art feature-based VO algorithms, which estimates the motion of the camera by utilizing features with and without depth. The proposed method is implemented in Matlab/C++ and runs on a desktop computer with Intel Core i5 3.2 GHz and 8GB memory.

7.1.1 Synthetic EuRoC datasets

The EuRoC micro aerial vehicle (MAV) datasets (Burri and Nikolic 2016) consist of the stereo image pairs at 20 Hz mounted on an AscTec Firefly MAV and a ground-truth position from a motion capture system at 100 Hz. To test the robustness against abrupt and local lighting changes, we modify the intensity values continuously in the stereo images

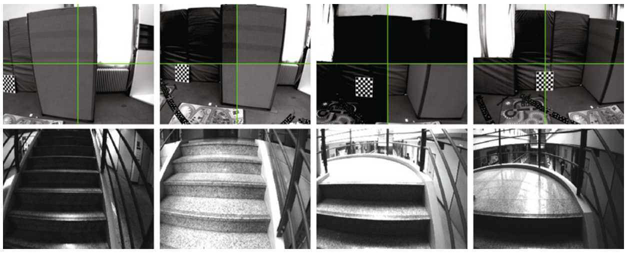


Fig. 11 Extracts from the synthetic EuRoC dataset (top) and author-collected dataset (bottom). Top: the four distinct quadrants showing irregular illumination changes. Bottom: the image brightness of the same place is different because of the automatic exposure control and the variation of the viewpoint

in the four quadrants based on the affine illumination model in the Eq. (8), rather than the same affine illumination in the entire image. Each zone in the four quadrants follows a different affine model to simulate uneven lighting changes in the entire image. We call these stereo image sequences containing the artificial illumination changes as the synthetic EuRoC dataset and extracts are shown in the first row of Fig. 11.

To measure the accuracy of the proposed and other VO algorithms quantitatively, three types of error metrics are selected: root mean square error (RMSE) of the relative pose error (RPE), absolute trajectory error (ATE) in Sturm et al. (2012), and the final drift error divided by the total traveling distance of a recording platform.

We first evaluate the proposed two-stage design compared to feature only (Geiger et al. 2011) and direct only (Kim et al. 2015) tracking methods while keeping other parts unchanged in Table 1. We measure the root mean squared error (RMSE) of the relative pose error, and present the improved accuracy of the proposed method. Thanks to a good initial pose from the feature-based approach, our direct VO can achieve better performance in terms of RPE value. In particular, the proposed method with motion prior shows more accurate results than the direct only method by preventing it from jumping in the motion estimates. The average RMSE of RPE is 0.151 drift m/s for the proposed method, compared with 0.156, 0.351 drift m/s under feature only and direct only method, respectively.

We present the motion estimation results in Table 2. The smallest error for each dataset is bolded. The proposed algorithm shows better performance in terms of relative pose error for the synthetic EuRoC benchmark. The main reason for the improved results is that the proposed algorithm can cope with non-uniform light variations by integrating the affine illumination model per patch into the direct motion estimation. However, other direct VO algorithms continue to perform the direct motion estimation under the photo-consistency assumption without considering such light changes. Excerpts from the Machine Hall 02, including synthetic image where irregular illumination changes occur and accurate motion tracking result, are shown in Fig. 12. Square patches distributed throughout the image marked red in Fig. 12c help the proposed algorithm to cope with the irregular illumination changes present in the synthetic datasets.

In most cases, the DVO, which is a direct VO without the affine illumination model, greatly loses accuracy due to the light variations. EDVO, which compensates for the global illumination changes, shows good motion estimation results on some datasets: the Vicon Room 2 02 and Machine Hall 03. Since the EDVO performs per-image brightness correction, it cannot effectively deal with the partial light changes. On the other hand, the proposed method performs per-patch illumination correction, and it can handle the irregular illumination changes, resulting in more accurate motion estimation results. If the features are well detected and tracked in the front-end, the DEMO is not sensitive to the light variations unlike the previous direct VO methods. DEMO presents better performance than the direct methods in terms of the final drift error on the Vicon Room 1 03 & 2 03, which have severe light variations. However, high drift error becomes more severe over time in most cases.

The strength of the proposed algorithm becomes clear when analyzing the dataset Machine Hall 02 in detail. During the period from 100 to 300 image index where the irregular illumination changes occur, we can observe that the proposed method maintains the modified photometric error very small whereas the cost values of DVO and EDVO increase noticeably, which are reported in Fig. 13a. The main reason for this difference is that the photo-consistency assumption is

Table 1 Accuracy improvement of the proposed algorithm

Experiment	Proposed	Feature only	Direct only	Length (m)
Vicon Room 1 01	**0.050**	0.061	0.055	57.97
Vicon Room 1 02	**0.050**	0.062	0.105	74.28
Vicon Room 1 03	0.609	**0.575**	1.017	78.70
Machine Hall 01	**0.014**	0.026	0.018	67.53
Machine Hall 02	**0.013**	0.026	0.028	63.00
Machine Hall 03	0.266	0.240	**0.062**	126.87
Machine Hall 04	**0.065**	0.100	1.172	88.39

Table 2 Experimental results on synthetic EuRoC benchmark

Experiment	Relative pose error (m/s)				Absolute trajectory error (m)				Final drift error (%)				Length (m)	# of frame
	Proposed	DVO	EDVO	DEMO	Proposed	DVO	EDVO	DEMO	Proposed	DVO	EDVO	DEMO		
Vicon Room 1 01	**0.050**	0.166	0.059	0.136	**0.289**	1.906	0.668	1.653	**0.353**	3.863	1.543	3.359	57.97	2712
Vicon Room 1 02	**0.050**	0.356	0.088	0.294	**0.253**	6.474	0.975	2.528	**0.510**	9.911	1.714	4.309	74.28	1510
Vicon Room 1 03	**0.609**	1.592	0.771	0.862	6.583	14.346	7.879	**6.488**	7.895	21.400	11.393	**3.881**	78.70	1949
Vicon Room 2 01	**0.035**	0.371	0.089	0.091	**0.888**	5.450	1.475	1.202	4.550	**4.049**	7.300	4.421	35.85	2054
Vicon Room 2 02	**0.063**	0.140	0.079	0.284	1.285	1.485	**1.282**	3.635	2.433	**1.374**	2.300	7.229	83.34	2201
Vicon Room 2 03	**0.359**	1.631	0.734	0.634	**5.086**	16.361	8.888	5.927	12.394	28.150	12.878	**10.014**	86.56	1806
Machine Hall 01	**0.014**	0.040	0.029	0.116	**0.310**	0.770	0.429	1.273	**0.582**	1.182	0.840	2.612	67.53	2582
Machine Hall 02	**0.013**	0.053	0.029	0.101	**0.205**	0.882	0.312	0.953	**0.308**	1.697	1.053	1.750	63.00	2140
Machine Hall 03	0.266	0.600	**0.053**	0.250	4.342	8.781	**1.077**	2.723	5.333	7.799	**1.548**	5.137	126.87	2200
Machine Hall 04	**0.065**	0.365	0.167	0.227	**1.117**	7.591	2.077	5.005	**1.499**	9.536	2.566	6.199	88.39	1533
Machine Hall 05	**0.040**	0.372	0.101	0.147	**0.725**	8.913	0.992	3.260	**0.851**	9.952	1.365	3.468	93.97	1810

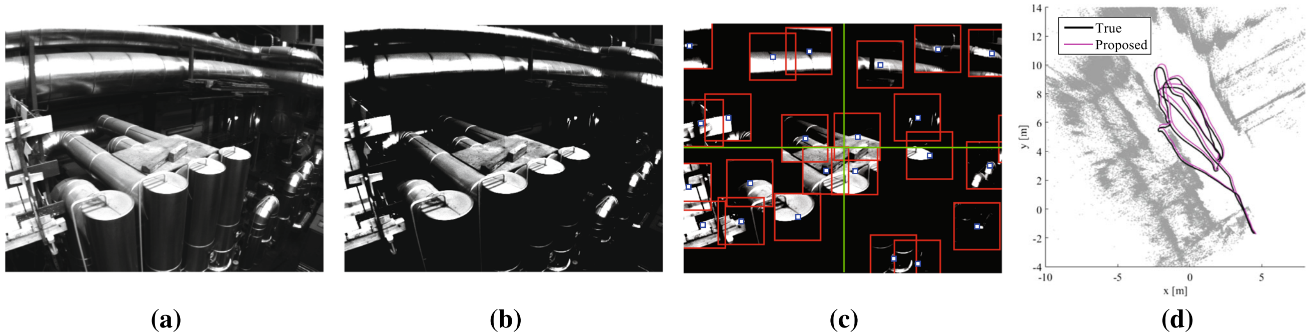


Fig. 12 **a**, **b** Show the original and synthetic images containing artificial lighting changes from the Machine Hall 02 in the EuRoC dataset. Red squares in (**c**) denote the image patches used in the proposed algorithm for compensating irregular illumination changes. **d** Shows the true (black) and estimated trajectories (magenta) and the reconstructed 3D point cloud with the proposed method. No fusion is performed (Color figure online)

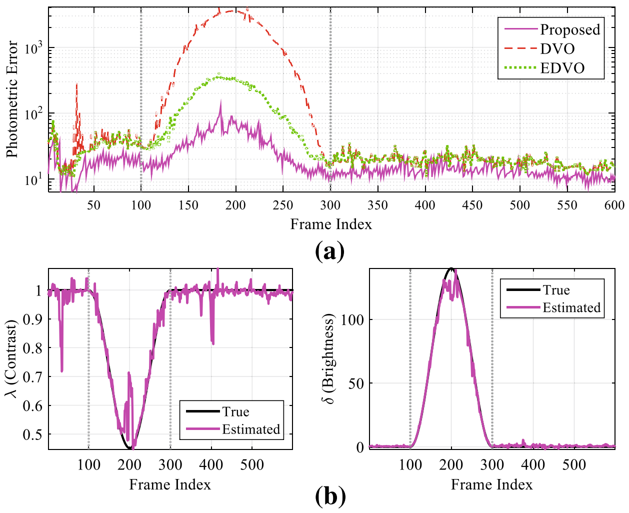


Fig. 13 **a** The photometric error of the three direct VO methods is drawn on a logarithmic scale where photometric disturbances occur between the gray dotted lines. The proposed method shows the lowest growth rate of the photometric error. **b** The true and estimated photometric parameters with the proposed method overlap significantly

severely violated in this period. Although a robust weighting function is employed for removing outliers in DVO or a global affine illumination model is considered in EDVO, these direct VO methods are not effective enough to take into account the sudden, partial lighting variations. The proposed algorithm efficiently handles this kind of illumination changes by using the proposed cost function in Eq. (11), resulting in accurate motion estimates as shown in Fig. 12d.

Figure 13b shows the true and estimated photometric parameters of a randomly selected patch. The proposed method estimates the photometric parameters for contrast and brightness changes correctly, which are used to compensate for the lighting changes in the synthetic EuRoC datasets as shown in Fig. 12b. Some jitters denote that our algorithm compensates for not only the artificial lighting changes we have made, but also the unmodeled and unpredictable light changes from sensor noise. Thanks to accurate estimation of the photometric parameters for each patch, the proposed method properly compensates the partial light variations during the direct motion estimation.

7.1.2 Author-collected datasets

We want to demonstrate that the proposed algorithm also works well in the everyday indoor environments where the actual illumination changes occur due to various reasons such as sunlight entering through windows, automatic exposure control of the camera, etc. We collect stereo image datasets with a handheld VI sensor (Nikolic et al. 2014), capturing a multistoried stairway which includes actual and unknown illumination changes. Figure 11 shows the example images where illumination changes are severe. For evaluating the consistency of the proposed and other VO baselines without the ground truth, we collect the stereo images along the carefully designed movements of the VI sensor in the stairway.

In Fig. 14, the 80 m trajectory going up the stairs from the 1st to the 6th floor of a building is visualized with three different views: top, front, and right side. The top view of the estimated trajectory shows the overlapped, consistent motion estimation result of the proposed method (magenta) while other estimated trajectories gradually diverge from the initially estimated loop. The side and front view of the stairway also support the high consistency of the proposed method compared with other VO methods.

7.2 Experiments on an autonomous aerial robot

We build an aerial robot system capable of flying autonomously in a light-changing environment with only the onboard sensors and computer. In order to evaluate the accuracy of the autonomous flight when integrated with the pro-

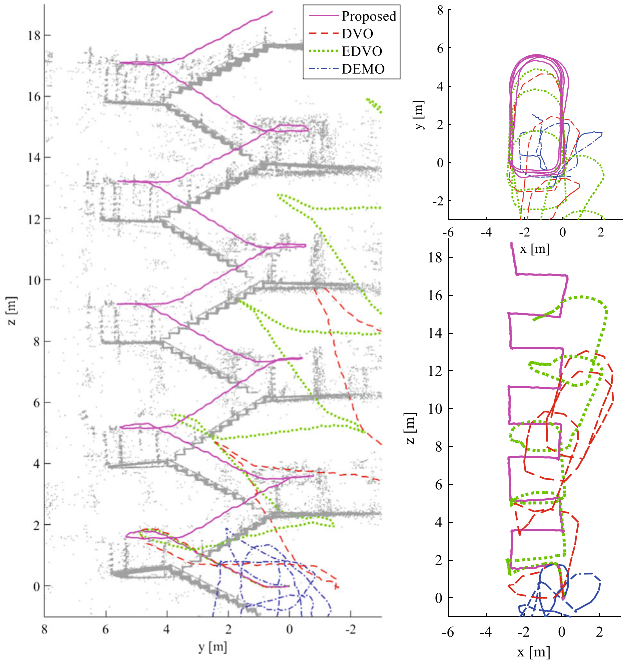


Fig. 14 Comparison of the proposed and other VO methods on the multistoried stairway from the 1st to 6th floor. The figure shows the side (left), front (right down), and top (right up) views of the estimated trajectories in the ascending stairway. The gray dots are reconstructed 3D points used in the feature-based VO in the proposed method, showing the consistently reconstructed stairway

posed method, we perform trajectory tracking experiments. Through the flight experiments under actual light changes, we show the robustness and effectiveness of the proposed algorithm in the autonomous flight of the aerial robot. Although some research on the autonomous flight with VO exists (Forster et al. 2014), there has been no reported result in light-changing environments.

7.2.1 Experimental setup

We describe the hardware components of the aerial robot and our experimental setup as shown in Fig. 15. The VI sensor captures the stereo images at 752 × 480 pixel resolution at 20 Hz, and is mounted in a front, down-looking position of an AscTec Firefly aerial robot, equipped with an off-the-shelf inertial measurement unit (IMU). The proposed VO algorithm updates the current position at 15 Hz. We obtain the velocity estimates by differentiating the estimated position and gyro from IMU. We integrate the nonlinear sliding mode controller to generate the motor commands used in Lee et al. (2018). All state estimation and control algorithms run on the AscTec Mastermind onboard computer with 2.1 GHz cores and 4 GB memory. For performance comparison only, a Vicon motion capture system is used to obtain the ground truth pose of the aerial robot at 100 Hz. Desired position or trajectory determined by the user or path planning algorithm is sent to the aerial robot from the ground computer with Xbee at 40 Hz.

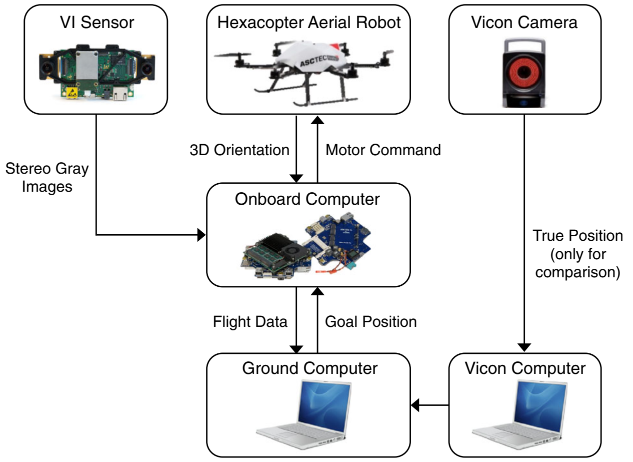


Fig. 15 Schematic diagram of the data flow in our experimental setup. All measured information from the onboard sensors is sent to the onboard computer to perform state estimation. Control inputs are calculated on the onboard computer with the estimated pose and the goal position given by the ground computer. All of the flight data and ground truth pose are sent to the ground computer through WiFi and TCP/IP communication

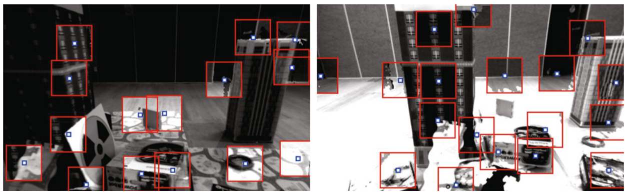


Fig. 16 Example images on an autonomous flight experiment with challenging illumination caused by the on-off of the lights. The image areas in the red patches show successful photometric compensation with the proposed method while other areas undergo severe illumination changes

7.2.2 Autonomous flight with light variations

We evaluate the proposed algorithm in terms of accuracy and robustness through the autonomous flight experiments in an environment where sudden and partial light variations occur frequently as shown in Figs. 1 and 16. While the lights are turned on and off repeatedly and randomly for generating photometric disturbances, we command the aerial robot to follow the given trajectory.

The proposed method allows the aerial robot to fly autonomously along the trajectory even in such a light-changing environment as demonstrated in Fig. 1. The estimated trajectory is qualitatively similar to the ground-truth trajectory, and the average translational RMSE of the proposed method is 0.06 m. The point cloud is also reconstructed consistently with the trajectory estimates. Figure 16 shows that per-patch illumination correction in the proposed method

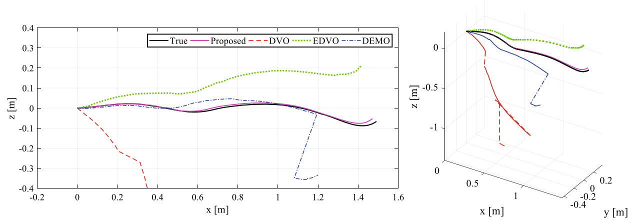


Fig. 17 Comparison of the proposed and other VO methods under severe illumination changes. The figure shows the top (left) and 3-D (right) views of the estimated trajectories

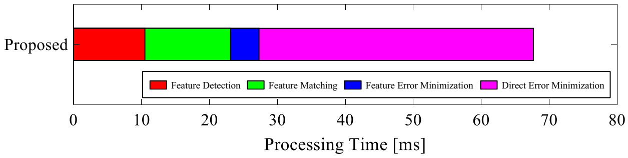


Fig. 18 Runtime evaluation of the proposed algorithm on the aerial robot during the autonomous flight experiments

works successfully during the autonomous flight experiments. If the patch size is too small, we cannot estimate the photometric parameters of each patch stably and correctly. Conversely, if the patch size is too big, it will not be able to respond to local (irregular) lighting changes effectively. The patch size (91 × 91) we have currently used is the result of our experimental evaluations. The autonomous flight experiments validate the accuracy and robustness of the proposed algorithm, which can estimate the accurate 6-DoF pose of the aerial robot using only onboard sensors and computer.

We also compare the proposed method to other VO methods with the stereo images obtained during the autonomous flight experiments under the irregular illumination changes as shown in Fig. 17. The average translational RMSE of the proposed algorithm is 0.06 m, while DVO, EDVO, and DEMO are 0.93, 0.16, and 0.32, respectively. Existing VO approaches (DVO, EDVO, DEMO) cannot cope with the sudden and local lighting changes in the images, showing a significant positional error compared to the ground-truth trajectory. Although EDVO, which compensates for the global illumination changes, shows good motion estimation results among other VO approaches, it cannot effectively deal with the local lighting changes. The proposed method can handle not only the global but also the local illumination changes with different affine models for different patches, showing the benefits of our proposals in challenging light environments.

Figure 18 shows a break-up of the time required to compute the 6-DoF camera motion on the aerial robot. The computation time for the feature-based estimation and direct estimation is about 27.3 ms and 40.4 ms, respectively. Our direct VO approach can achieve stable and robust motion estimation performance without noticeable increases in the overall computational time. The proposed algorithm updates the current position of the aerial robot at 15 Hz, resulting in the stable autonomous flights under light-changing environments.

Figure 19 shows x and y position, the estimated photometric parameter (contrast), sample images, and photometric error of the proposed method under a light-changing environment for about one minute flight. The estimated contrast parameters in the third row closely match the actual lighting conditions observed in the fourth row. When the lights are turned on and off, the photometric parameters of the affine illumination model are changed to compensate for the sudden and irregular illumination changes. Due to such compensation of the lighting changes, the photometric error in the fifth row does not exceed 60, resulting in accurate motion estimation results in the first and second rows. Please refer to the video clips submitted with this paper showing more details about the experiments.[1]

8 Conclusion

We present an illumination-robust direct visual odometry for the autonomous flight of the aerial robot in a light-changing environment. The gain in robustness to irregular illumination changes is due to the fact that the affine illumination model is employed in each image patch and integrated into the direct motion estimation to *simultaneously* estimate the 6-DoF camera motion and the parameters of partial light changes. We further propose to utilize a motion prior from the feature-based visual odometry for stable and accurate motion estimation in a light-changing environment. Detailed analyses with the convergence rate and the degree of linearity of each cost function in feature-based and direct methods support such usage of the motion prior knowledge. The proposed VO algorithm enables the aerial robot to fly autonomously and robustly under changing lighting conditions at the cost of estimating the illumination change model parameters.

The results of this paper have many extensions and applications. Our work only focuses on the autonomous flight of the aerial robots, but the proposed illumination-robust visual odometry can be equally applied to various types of autonomous vehicles such as self-driving cars. Another interesting extension would be to use the proposed VO method to enhance other SLAM algorithms under changing lighting conditions. For example, our initial position estimates could be used as the cornerstone of a full SLAM system under irregular illumination changes. Our approach assumes that light changes will follow the affine illumination model; future work should consider various light change modes that vary more complexly such as bright spots and shadows caused by sunlight coming through the windows.

[1] Video available at https://youtu.be/agOxpphFDfE.

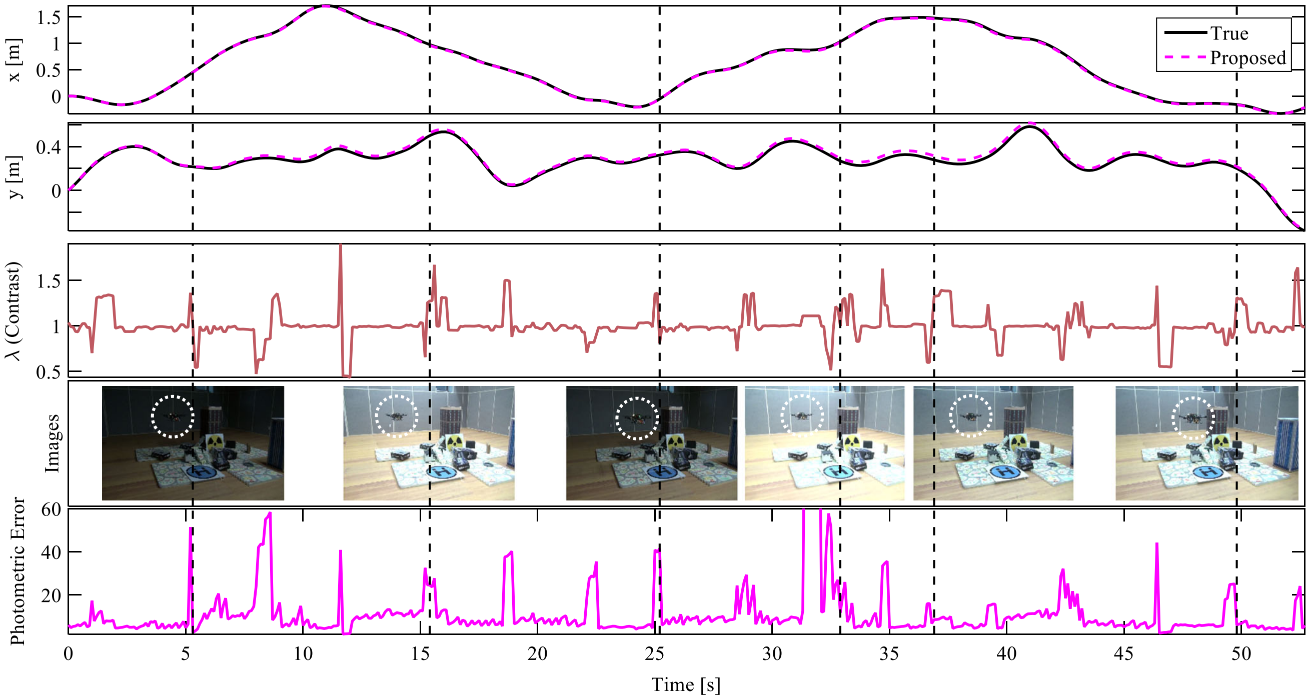


Fig. 19 Flight experiment results in a light-changing environment where lights are turned on and off repeatedly. The dotted vertical lines denote the time instants at which each snapshot is captured. We plot the changes of the estimated contrast in the third row, and all jumps correspond to light changes by switching the lights on and off during the sequence. When the lights dimmed near 5 s, the contrast increases and the illumination changes are compensated, and vice versa when the lights brightened. The estimated photometric parameters of the affine illumination model in the third row show similar behavior to the brightness level of the actual images in the fourth row. Although sudden and severe light changes continue to occur, the photometric error in the fifth row does not exceed 60, resulting in accurate motion tracking in the first and second rows

Acknowledgements This research was supported by the Ministry of Science, ICT, under the Information Technology Research Center (ITRC) program (IITP-2018-2017-0-01637) supervised by the Institute for Information & communications Technology Promotion, and Automation and Systems Research Institute (ASRI), and Samsung Research, Samsung Electronics Co., Ltd.

Appendix

In this section, we derive the analytic form of the dimensionless linearity index (DLI) for each observation model in feature-based and direct estimation in detail. The DLI of each measurement equation in terms of the 6-DoF camera motion can be written as follows:

$$L = \left| \frac{\left.\frac{\partial^2 h}{\partial \boldsymbol{\xi}^2}\right|_{\boldsymbol{\xi}=\boldsymbol{\xi}_0} \Delta\boldsymbol{\xi}}{\left.\frac{\partial h}{\partial \boldsymbol{\xi}}\right|_{\boldsymbol{\xi}=\boldsymbol{\xi}_0}} \right| \tag{14}$$

$$\hat{x} = h_x(\boldsymbol{\xi}) = \left[\pi(T(\boldsymbol{\xi}) \cdot \pi^{-1}(\mathbf{x}, Z(\mathbf{x})))\right]_x$$

$$\hat{I} = h_I(\boldsymbol{\xi}) = I(\pi(T(\boldsymbol{\xi}) \cdot \pi^{-1}(\mathbf{x}, Z(\mathbf{x}))))$$

where the h is the observation model in each estimation method. The division in the Eq. (14) denotes element-wise operation between two vectors with a slight abuse of notation. For the sake of simplicity, we can rewrite the x component of the warping function in the feature-based method as $\hat{x}$, and the pixel intensity of the next image frame in the direct method as $\hat{I}$.

$$T = T(\boldsymbol{\xi}) = \exp(\hat{\xi}) = \exp\left(\sum_{i=1}^{6} \boldsymbol{\xi}_i E_i\right) \tag{4}$$

The 6-DoF camera motion, $T(\boldsymbol{\xi}) \in SE(3)$, is written as T in this section for simplicity. In the Eq. (4), $\boldsymbol{\xi}_i$ is a ith component of the Lie algebra explained in Sect. 3, and E_i is one of the six Lie algebra $se(3)$ bases, each corresponding to either infinitesimal translations or rotations along each axis Blanco (2010).

Feature-based estimation

The first and second-order partial derivatives of $\hat{x}$ with respect to $\boldsymbol{\xi}$, i.e., the Jacobian and Hessian matrices, can be written as follows:

$$J_x = \frac{\partial \hat{x}}{\partial \boldsymbol{\xi}} \in \mathbb{R}^{1\times 6},\ H_x = \frac{\partial^2 \hat{x}}{\partial \boldsymbol{\xi}^2} \in \mathbb{R}^{6\times 6} \tag{15}$$

By applying the chain rule, we can derive analytical Jacobian matrix as follows:

$$J_x = \frac{\partial \hat{x}}{\partial \mathbf{X}'} \cdot \frac{\partial \mathbf{X}'}{\partial T} \cdot \frac{\partial T}{\partial \boldsymbol{\xi}} \tag{16}$$

where $\mathbf{X}' = \left[X', Y', Z'\right]^\top$ is a transformed 3D point from $\mathbf{X} = [X, Y, Z]^\top$ with respect to T. The analytical Jacobian matrix of warping function is:

$$J_x = \left[f_x \frac{1}{Z'} \; 0 \; -f_x \frac{X'}{Z'^2} \; -f_x \frac{X'Y'}{Z'^2} \; f_x \left(1 + \frac{X'^2}{Z'^2}\right) -f_x \frac{Y'}{Z'} \right]$$

For full details of each step, see Blanco (2010) and Kerl (2012).

In the following, the second order partial derivatives of warping function with respect to $\boldsymbol{\xi}$, the Hessian matrix, can be obtained by applying the chain rule same as above:

$$\begin{aligned} H_x &= \frac{\partial}{\partial \boldsymbol{\xi}} \left(\frac{\partial \hat{x}}{\partial \boldsymbol{\xi}} \right) = \frac{\partial}{\partial \boldsymbol{\xi}} \left(\frac{\partial \hat{x}}{\partial T} \cdot \frac{\partial T}{\partial \boldsymbol{\xi}} \right) \\ &= \frac{\partial}{\partial \boldsymbol{\xi}} \left(\frac{\partial \hat{x}}{\partial T} \right) \cdot \frac{\partial T}{\partial \boldsymbol{\xi}} + \frac{\partial \hat{x}}{\partial T} \cdot \frac{\partial^2 T}{\partial \boldsymbol{\xi}^2} \end{aligned} \tag{17}$$

The (i, j)th element of the Hessian matrix H_x can be written as follows:

$$\begin{aligned} H_{x(i,j)} &= \frac{\partial}{\partial T} \left(\frac{\partial \hat{x}}{\partial T} \right) \cdot \frac{\partial T}{\partial \boldsymbol{\xi}_j} \cdot \frac{\partial T}{\partial \boldsymbol{\xi}_i} + \frac{\partial \hat{x}}{\partial T} \cdot \frac{\partial^2 T}{\partial \boldsymbol{\xi}_i \partial \boldsymbol{\xi}_j} \\ &= \left(\frac{\partial T}{\partial \boldsymbol{\xi}_j} \right)^\top \cdot \frac{\partial^2 \hat{x}}{\partial T^2} \cdot \left(\frac{\partial T}{\partial \boldsymbol{\xi}_i} \right) + \frac{\partial \hat{x}}{\partial T} \cdot \frac{\partial^2 T}{\partial \boldsymbol{\xi}_i \partial \boldsymbol{\xi}_j} \end{aligned} \tag{18}$$

where the first and second-order derivatives of T with respect to $\boldsymbol{\xi}_i$ can be computed as follows:

$$\begin{aligned} \frac{\partial T}{\partial \boldsymbol{\xi}_i} &= E_i \cdot T \\ \frac{\partial^2 T}{\partial \boldsymbol{\xi}_i \partial \boldsymbol{\xi}_j} &= \frac{1}{2} \left(E_i E_j + E_j E_i \right) \cdot T \end{aligned}$$

Note that, J_y and H_y, the Jacobian and Hessian matrices of y component of the warping function with respect to the camera motion, are omitted because they are symmetric to J_x and H_x. With the equations derived above, we obtain the Jacobian and Hessian matrices of $\hat{x}$ with respect to the $\boldsymbol{\xi}$, and compute the DLI of the observation model in the feature-based method in terms of the 6-DoF camera pose.

Direct estimation

The first and second-order partial derivatives of $\hat{I}$ with respect to $\boldsymbol{\xi}$, the Jacobian and Hessian matrices of the observation model in the direct method, can be written as follows:

$$J_I = \frac{\partial \hat{I}}{\partial \boldsymbol{\xi}} \in \mathbb{R}^{1 \times 6}, \; H_I = \frac{\partial^2 \hat{I}}{\partial \boldsymbol{\xi}^2} \in \mathbb{R}^{6 \times 6} \tag{19}$$

To obtain the analytical Jacobian and Hessian matrices, we repeat similar calculation procedure again as above. Based on the analytical Jacobian and Hessian matrices of warping function calculated in the previous section and the chain rule, we can easily derive the analytical Jacobian matrix of direct method as follows:

$$J_I = \frac{\partial \hat{I}}{\partial \hat{\mathbf{x}}} \cdot \frac{\partial \hat{\mathbf{x}}}{\partial \boldsymbol{\xi}} = \frac{\partial \hat{I}}{\partial \hat{\mathbf{x}}} \cdot \frac{\partial \hat{\mathbf{x}}}{\partial \mathbf{X}'} \cdot \frac{\partial \mathbf{X}'}{\partial T} \cdot \frac{\partial T}{\partial \boldsymbol{\xi}} \tag{20}$$

where $\hat{\mathbf{x}} = \left[\hat{x}, \hat{y}\right]^\top$ is a 2D pixel point in the image $\hat{I}$. The first term $\frac{\partial \hat{I}}{\partial \hat{\mathbf{x}}}$ denotes the gradient of the image $\hat{I}$ given by the image derivatives in the horizontal and vertical directions and the latter terms are the Jacobian matrix of warping function written in Eq. (16). The analytical Jacobian matrix of the observation model in the direct method is:

$$\begin{aligned} J_I &= \left[\frac{\partial \hat{I}}{\partial \hat{x}} \; \frac{\partial \hat{I}}{\partial \hat{y}} \right] \cdot \begin{bmatrix} \frac{\partial \hat{x}}{\partial \mathbf{X}'} \cdot \frac{\partial \mathbf{X}'}{\partial T} \cdot \frac{\partial T}{\partial \boldsymbol{\xi}} \\ \frac{\partial \hat{y}}{\partial \mathbf{X}'} \cdot \frac{\partial \mathbf{X}'}{\partial T} \cdot \frac{\partial T}{\partial \boldsymbol{\xi}} \end{bmatrix} \\ &= \left[\frac{\partial \hat{I}}{\partial \hat{x}} \; \frac{\partial \hat{I}}{\partial \hat{y}} \right] \cdot \begin{bmatrix} \frac{\partial \hat{x}}{\partial \boldsymbol{\xi}} \\ \frac{\partial \hat{y}}{\partial \boldsymbol{\xi}} \end{bmatrix} \\ &= \left[\nabla \hat{I}_x \; \nabla \hat{I}_y \right] \cdot \begin{bmatrix} J_x \\ J_y \end{bmatrix} \end{aligned}$$

where $\nabla \hat{I}_x$ and $\nabla \hat{I}_y$ are image gradients of $\hat{I}$ along the x and y direction in the image plane.

The Hessian matrix used in the direct estimation can be also derived by applying the chain rule and the Eq. (20) as follows:

$$H_I = \frac{\partial}{\partial \boldsymbol{\xi}} \left(\frac{\partial \hat{I}}{\partial \boldsymbol{\xi}} \right) = \frac{\partial}{\partial \boldsymbol{\xi}} \left(\frac{\partial \hat{I}}{\partial \hat{\mathbf{x}}} \cdot \frac{\partial \hat{\mathbf{x}}}{\partial \boldsymbol{\xi}} \right) \tag{21}$$

Each element of the Hessian matrix, an element in the ith row and jth column of H_I, can be written as follows:

$$\begin{aligned}
H_{I(i,j)} &= \frac{\partial}{\partial \boldsymbol{\xi}_j}\left(\frac{\partial \hat{I}}{\partial \boldsymbol{\xi}_i}\right) \\
&= \frac{\partial}{\partial \boldsymbol{\xi}_j}\left(\frac{\partial \hat{I}}{\partial \hat{x}}\cdot\frac{\partial \hat{x}}{\partial \boldsymbol{\xi}_i}+\frac{\partial \hat{I}}{\partial \hat{y}}\cdot\frac{\partial \hat{y}}{\partial \boldsymbol{\xi}_i}\right) \\
&= \frac{\partial}{\partial \boldsymbol{\xi}_j}\left(\frac{\partial \hat{I}}{\partial \hat{x}}\right)\cdot\frac{\partial \hat{x}}{\partial \boldsymbol{\xi}_i}+\frac{\partial \hat{I}}{\partial \hat{x}}\cdot\frac{\partial}{\partial \boldsymbol{\xi}_j}\left(\frac{\partial \hat{x}}{\partial \boldsymbol{\xi}_i}\right) \\
&\quad +\frac{\partial}{\partial \boldsymbol{\xi}_j}\left(\frac{\partial \hat{I}}{\partial \hat{y}}\right)\cdot\frac{\partial \hat{y}}{\partial \boldsymbol{\xi}_i}+\frac{\partial \hat{I}}{\partial \hat{y}}\cdot\frac{\partial}{\partial \boldsymbol{\xi}_j}\left(\frac{\partial \hat{y}}{\partial \boldsymbol{\xi}_i}\right) \\
&= \left(\frac{\partial^2 \hat{I}}{\partial \hat{x}^2}\cdot J_{x(1,j)}+\frac{\partial^2 \hat{I}}{\partial \hat{y}\partial \hat{x}}\cdot J_{y(1,j)}\right)\cdot J_{x(1,i)} \\
&\quad +\frac{\partial \hat{I}}{\partial \hat{x}}\cdot H_{x(i,j)} \\
&\quad +\left(\frac{\partial^2 \hat{I}}{\partial \hat{x}\partial \hat{y}}\cdot J_{x(1,j)}+\frac{\partial^2 \hat{I}}{\partial \hat{y}^2}\cdot J_{y(1,j)}\right)\cdot J_{y(1,i)} \\
&\quad +\frac{\partial \hat{I}}{\partial \hat{y}}\cdot H_{y(i,j)}
\end{aligned} \tag{22}$$

where J_x, J_y and H_x, H_y are the Jacobian and Hessian matrices of warping function derived in the previous section, and $\frac{\partial^2 \hat{I}}{\partial \hat{x}^2}$, $\frac{\partial^2 \hat{I}}{\partial \hat{x}\partial \hat{y}}$, $\frac{\partial^2 \hat{I}}{\partial \hat{y}^2}$ are the second image derivatives in the horizontal and vertical directions. With the above analytical Jacobian and Hessian matrices of the observation model in the direct method, we can compute the DLI of the image intensity observation model with respect to the 6-DoF camera motion.

References

Achtelik, M., Achtelik, M., et al. (2012). Sfly: Swarm of micro flying robots. In: *2012 IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 2649–2650). IEEE.

Alismail, H., Kaess, M., Browning, B., & Lucey, S. (2017). Direct visual odometry in low light using binary descriptors. *IEEE Robotics and Automation Letters*, *2*(2), 444–451.

Benhimane, S., & Malis, E. (2004). Real-time image-based tracking of planes using efficient second-order minimization. In: *2004 IEEE/RSJ international conference on intelligent robots and systems, 2004. IROS 2004. Proceedings* (Vol. 1, pp 943–948). IEEE.

Bergmann, P., Wang, R., & Cremers, D. (2018). Online photometric calibration of auto exposure video for realtime visual odometry and slam. *IEEE Robotics and Automation Letters*, *3*(2), 627–634.

Blanco, J. L. (2010). *A tutorial on se (3) transformation parameterizations and on-manifold optimization* (p. 3). University of Malaga, Tech Rep.

Burri, M., & Nikolic, J. (2016). The EuRoc micro aerial vehicle datasets. *The International Journal of Robotics Research*, *35*, 1157–1163.

Civera, J., Davison, A. J., & Montiel, J. M. (2008). Inverse depth parametrization for monocular slam. *IEEE Transactions on Robotics*, *24*(5), 932–945.

Engel, J., Schöps, T., & Cremers, D. (2014). LSD-SLAM: Large-scale direct monocular slam. In: *European Conference on Computer Vision* (pp. 834–849). Springer.

Engel, J., Stückler, J., & Cremers, D. (2015). Large-scale direct slam with stereo cameras. In: 2015 IEEE/RSJ international conference on intelligent robots and systems (IROS) (pp. 1935–1942). IEEE.

Engel, J., Koltun, V., & Cremers, D. (2018). Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(3), 611–625.

Fang, Z., & Scherer, S. (2014). Experimental study of odometry estimation methods using RGB-D cameras. In: *IROS* (pp. 680–687). IEEE.

Forster, C., Pizzoli, M., & Scaramuzza, D. (2014). SVO: Fast semi-direct monocular visual odometry. In: *2014 IEEE international conference on robotics and automation (ICRA)* (pp. 15–22). IEEE.

Forster, C., Zhang, Z., Gassner, M., Werlberger, M., & Scaramuzza, D. (2017). SVO: Semidirect visual odometry for monocular and multicamera systems. *IEEE Transactions on Robotics*, *33*(2), 249–265.

Geiger, A., Roser, M., & Urtasun, R. (2010). Efficient large-scale stereo matching. In: *Asian conference on computer vision* (pp. 25–38). Springer.

Geiger, A., Ziegler, J., & Stiller, C. (2011). Stereoscan: Dense 3d reconstruction in real-time. In: *2011 IEEE Intelligent Vehicles Symposium (IV)* (pp. 963–968). IEEE.

Huang, A. S., Bachrach, A., Henry, P., Krainin, M., Maturana, D., Fox, D., et al. (2011). Visual odometry and mapping for autonomous flight using an RGB-D camera. In: *ISRR*, pp. 1–16.

Jin, H., Favaro, P., & Soatto, S. (2001). Real-time feature tracking and outlier rejection with changes in illumination. In *ICCV*, pp. 684–689.

Kerl, C. (2012). *Odometry from RGB-D cameras for autonomous quadrocopters*. Master's thesis, Technical University Munich, Germany.

Kerl, C., Sturm, J., & Cremers, D. (2013). Robust odometry estimation for RGB-D cameras. In: 2013 IEEE international conference on robotics and automation (ICRA) (pp. 3748–3754). IEEE.

Kerl, C., Souiai, M., Sturm, J., & Cremers, D. (2014). Towards illumination-invariant 3D reconstruction using ToF RGB-D cameras. In: *3DV*.

Kim, P., Lim, H., & Kim, H. J. (2015). Robust visual odometry to irregular illumination changes with RGB-D camera. In: 2015 IEEE/RSJ international conference on intelligent robots and systems (IROS) (pp. 3688–3694). IEEE.

Kitt, B., Geiger, A., & Lategahn, H. (2010). Visual odometry based on stereo image sequences with ransac-based outlier rejection scheme. In: *2010 IEEE Intelligent Vehicles Symposium (IV)* (pp. 486–492). IEEE.

Klose, S., Heise, P., & Knoll, A. (2013). Efficient compositional approaches for real-time robust direct visual odometry from RGB-D data. In: 2013 IEEE/RSJ international conference on intelligent robots and systems (pp. 1100–1106). IEEE.

Krombach, N., Droeschel, D., & Behnke, S. (2016). Combining feature-based and direct methods for semi-dense real-time stereo visual odometry. In: *International conference on intelligent autonomous systems* (pp. 855–868). Springer.

Lee, H., Kim, H., & Kim, H. J. (2018). Planning and control for collision-free cooperative aerial transportation. *IEEE Transactions on Automation Science and Engineering*, *15*(1), 189–201.

Ma, Y., Soatto, S., Kosecka, J., & Sastry, S. S. (2012). *An invitation to 3-d vision: From images to geometric models* (Vol. 26). Berlin: Springer.

Maimone, M., Cheng, Y., & Matthies, L. (2007). Two years of visual odometry on the mars exploration rovers. *Journal of Field Robotics*, *24*(3), 169–186.

Mur-Artal, R., Montiel, J., & Tardós, J. D. (2015). ORB-SLAM: A versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, *31*(5), 1147–1163.

Nikolic, J., Rehder, J., et al. (2014). A synchronized visual-inertial sensor system with FPGA pre-processing for accurate real-time slam. In: *2014 IEEE international conference on robotics and automation (ICRA)* (pp. 431–437). IEEE.

Nistér, D., Naroditsky, O., & Bergen, J. (2004). Visual odometry. In: *Proceedings of the 2004 IEEE computer society conference on computer vision and pattern recognition, 2004. CVPR 2004* (Vol. 1, pp. I–I). IEEE.

Park, S., Schöps, T., & Pollefeys, M. (2017). Illumination change robustness in direct visual slam. In: 2017 IEEE international conference on robotics and automation (ICRA) (pp. 4523–4530). IEEE.

Scaramuzza, D., Achtelik, M., et al. (2014). Vision-controlled micro flying robots: From system design to autonomous navigation and mapping in GPS-denied environments. *IEEE Robotics & Automation Magazine*, *21*(3), 26–40.

Scaramuzza, D., & Fraundorfer, F. (2011). Visual odometry [tutorial]. *IEEE Robotics & Automation Magazine*, *18*(4), 80–92.

Sturm, J., et al. (2012). A benchmark for the evaluation of RGB-D slam systems. In: *2012 IEEE/RSJ international conference on intelligent robots and systems* (pp. 573–580). IEEE.

Valenti, R. G., et al. (2014). Autonomous quadrotor flight using onboard RGB-D visual odometry. In: *2014 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 5233–5238). IEEE.

von Stumberg, L., Usenko, V., Engel, J., Stückler, J., & Cremers, D. (2016). Autonomous exploration with a low-cost quadrocopter using semi-dense monocular slam. ArXiv preprint arXiv:1609.07835.

Wang, R., Schwörer, M., & Cremers, D. (2017). Stereo DSO: Large-scale direct sparse visual odometry with stereo cameras. In: *International conference on computer vision (ICCV)*, Venice, Italy.

Zhang, J., Kaess, M., & Singh, S. (2017). A real-time method for depth enhanced visual odometry. *Autonomous Robots*, *41*(1), 31–43.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Pyojin Kim received the B.S. degree in Mechanical Engineering from Yonsei University in 2013. He is currently pursuing the M.S. and Ph.D. degrees in the Department of Mechanical and Aerospace Engineering at Seoul National University, Seoul, South Korea. His research interests include 3D computer vision, visual odometry, and visual SLAM.

Hyeonbeom Lee received the B.S. degree in Mechanical and Control Engineering from Handong Global University in 2011, and the M.S. and Ph.D. degrees in Mechanical and Aerospace Engineering from Seoul National University in 2013 and 2017, respectively. From 2017 to 2018, he was a senior researcher in Korea Institute of Machinery and Materials (KIMM). In September 2018, he joined the Department of Electronics Engineering at Kyungpook National University, Deagu, South Korea, as an Assistant Professor. His research interests include autonomous navigation of aerial robots and mobile manipulators.

H. Jin Kim received the B.S. degree from the Korea Advanced Institute of Technology (KAIST) in 1995, and the M.S. and Ph.D. degrees in Mechanical Engineering from University of California, Berkeley (UC Berkeley), in 1999 and 2001, respectively. From 2002 to 2004, she was a Postdoctoral Researcher in Electrical Engineering and Computer Science (EECS), UC Berkeley. In September 2004, she joined the Department of Mechanical and Aerospace Engineering at Seoul National University, Seoul, South Korea, as an Assistant Professor where she is currently a Professor. Her research interests include intelligent control of robotic systems and motion planning.