Illumination Robust Monocular Direct Visual Odometry for Outdoor Environment Mapping

Xiaolong Wu, Cédric Pradalier

Illumination Robust Monocular Direct Visual Odometry for Outdoor Environment Mapping

Xiaolong Wu[1] and Cédric Pradalier[2]

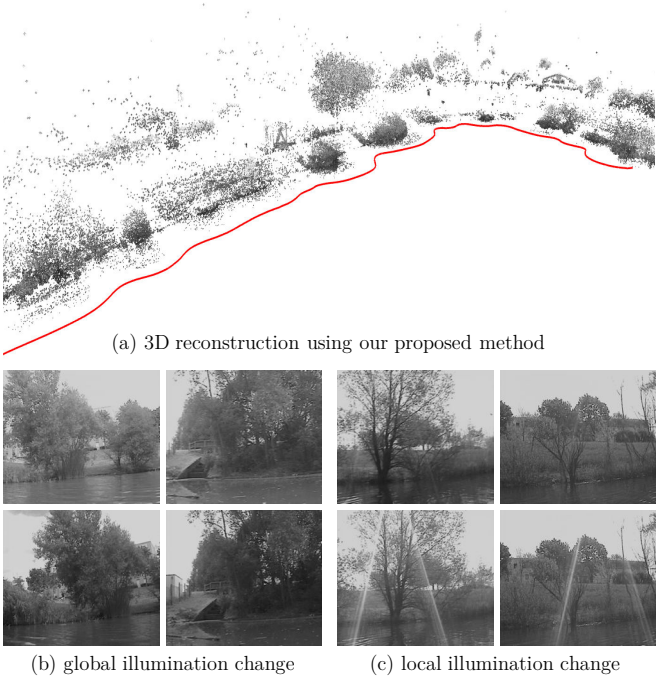
***Abstract*— Vision-based localization and mapping in outdoor environments is still a challenging issue, which requests significant robustness against various unpredictable illumination changes. In this paper, an illumination-robust direct monocular SLAM system that focuses on modeling outdoor scenery is presented. To deal with global and local lighting changes, such as solar flares, the state-of-art illumination invariant photometric costs for RGB-D and stereo SLAM systems are revisited in the context of their monocular counterpart, where the camera motion and scene structure are jointly optimized with a reasonably poor initialization. Based on our analysis, a combined cost is proposed to achieve a high-precision motion estimation with an improved convergence radius. The proposed system is extensively evaluated on the synthetic and real-world datasets regarding accuracy, robustness, and processing time, where our approach outperforms systems with other costs and state-of-art DSO and ORBSLAM2 systems.**

I. Introduction

In recent decades, real-time Visual Odometry (VO) and Simultaneous Localization and Mapping (SLAM) systems have shown their full potentials to assist various robotic applications in outdoor operation - from autonomous driving in the urban scene to environmental monitoring in the natural environment. In large-scale mapping applications, the monocular VO and SLAM systems have gained significant popularity due to the simple calibration procedure and their flexibility introduced by the scale ambiguity, which allows for the seamlessly switching between differently scaled scenes.

Among monocular VO and SLAM systems, the indirect method [1] [2] [3] has dominated the research field for a long time since the self-recognized features can provide considerable robustness to both the photometric noise and geometric distortion in images. However, recent works have shown that the direct method [4] [5] [6] [7] could provide more accurate and robust motion estimation due to their high flexibility of image information usage in comparison of the indirect approaches using certain types of features. In general, the direct VO and SLAM algorithms compare the intensity value of pixels over a local patch across images by making a brightness consistency assumption. However, it is impossible to control the illumination perfectly in real-world applications, especially in outdoor environments, thus resulting in performance degradation.

In our previous works [8], we have observed that one of the primary sources of the tracking failures in outdoor environment stems from the solar glares in acquired images. The affected pixels would either dominate the motion estimation resulting in false trajectory or be treated as outliers resulting in a lack of points in optimization, either of which could devastate the system robustness and enforce a restart. In this work, we propose to treat the solar glares as local and/or global lighting changes and utilizing the illumination-robust method to overcome the described difficulty.



(a) 3D reconstruction using our proposed method

(b) global illumination change (c) local illumination change

Fig. 1. Local and global lighting changes are abundant in Symphony Lake Dataset [9]. The example tracking trajectory and 3D reconstruction using our proposed method is presented in (a), and two pairs of example images with global illumination changes in (b) and two pairs with local+global illumination changes in (c) are illustrated.

Many illumination invariant algorithms have been developed to model the global and local lighting changes or alleviate their adverse effect in motion estimation. However, most of them are merely evaluated for RGB-D or Stereo systems, where the scene structure is usually well initialized. In this paper, we revisit the illumination-invariant costs that have shown excellent performance in indoor evaluations in [10], and we evaluate their performance in the monocular joint optimization framework [7], typically with a poor environmental structure initialization, using the vKITTI dataset [11] with simulated global and/or local lighting changes.

Based on the analysis, we propose a combined cost that adaptively weights a global affine model-based cost and

*All authors are with the UMI2958 GeorgiaTech-CNRS, Metz, France. firstname.lastname@georgiatech-metz.fr

a gradient-based cost in different optimization phases to achieve high-precision motion estimation, while preserving a large convergence basin. The proposed system is extensively evaluated regarding accuracy, robustness, and runtime performance using synthetic vKITTI dataset, real-world DEVON Island dataset, and Symphony Lake dataset against various global and local lighting changes, in Fig.1. Besides, a qualitative analysis of the robustness of our proposed method against solar glares is studied to support our claim further.

This paper makes the following contributions: 1) To our knowledge, this is the first paper evaluating illumination-robust costs in the context of a monocular joint optimization framework, considering a large initial inverse depth error; 2) a novel illumination-robust monocular direct VO system is proposed and evaluated against the real-world outdoor lighting changes; 3) the solar glares in the outdoor images are treated as local illumination change, and its adverse effects on motion estimation are proven to be alleviated by implementing global and local illumination invariant costs.

II. RELATED WORK

The so-called direct monocular VO and SLAM has shown exceptional performance in ego-motion estimation and scene structure mapping. As the state-of-art direct VO system, the DSO [7] utilizes a fully direct formulation that jointly optimizes the camera pose and scene structure in a sliding window, while taking advantage of explicit photometric calibration [12] and exposure compensation strategy to further improve the tracking accuracy and robustness. However, this approach makes the static scene lighting assumption, so that is still sensitive to the external lighting changes.

In recent years, a number of illumination-robust VO and SLAM systems have been proposed, implementing various models or descriptors to alleviate the adverse effect of external lighting changes, so as to achieve robust motion estimation in challenging indoor and outdoor environments. To gain robustness against the global illumination changes, either the median value of pixel residuals [13] [14] [15] [16] or an affine brightness transfer function [17] [18] is estimated to compensate the induced adverse effect in the optimization. For local lighting changes, [19] propose to use image gradients, rather than pixel intensities, to formulate the direct energy function, thus gaining local lighting invariance; [20] relies on dense computation of a deliberately designed local descriptor to obtain a clear global minimum in energy function while preserving convergence basin by convolving with a low-pass filter; the methods based on the census transform [21] [22] use a binary descriptor to achieve local illumination invariance during the motion estimation.

A thorough evaluation regarding accuracy and robustness in the context of visual odometry is conducted for all described methods above in [10]. The analysis results suggest that the gradient-based method[19] and census-transform-based method [10] show state-of-the-art performances. However, all the methods and their evaluations are conducted with RGB-D and stereo setup, where the depth information is provided with some precision.

Mutual information (MI) is a global metric which can be used to register images with huge appearance differences. NID-SLAM [23] incorporates a whole-image MI metric, named normalized information distance, into a monocular SLAM framework to achieve illumination-robust motion estimation in real-time on a high-performance GPU.

Illumination invariant imaging provides an alternative solution to our problem. In [24], the illumination invariant image transform is developed to remove the variation from lighting changes in the acquired RGB images. This improves the performance and robustness of localization and mapping.

III. ILLUMINATION ROBUST MONOCULAR DIRECT VISUAL ODOMETRY

In this section, the mathematical representation of the direct formulation is described in Sec. III-A, and the joint optimization algorithm utilized in this paper is presented in Sec. III-B. Most importantly, the state-of-art illumination invariant costs are described as plug-ins of the basic cost function in Sec. III-C, and our proposed cost is then derived in Sec. III-D.

A. Direct Formulation

Consider a reference frame, a gray-scale reference image $I_r : \Omega \to \mathbb{R}$ and an inverse depth map $D_r : \Omega \to \mathbb{R}^+$ are included, where $\Omega \subset \mathbb{R}^2$ is the image domain. A 3D scene point $\boldsymbol{x} = (x,y,z)^T$ of a pixel is parameterized by its inverse depth $d = z^{-1}$ in the reference frame instead of the conventional 3 unknowns. Defining a 3D projective warp function $\pi(\boldsymbol{x}) = (x/z, y/z)^T$, a pixel $\boldsymbol{u} = (u,v)^T \in \Omega$ can be back-projected into 3D world as $\boldsymbol{x} = \pi^{-1}(\boldsymbol{p},d) = \boldsymbol{K}^{-1}\boldsymbol{p}/d$, where $\boldsymbol{p} = (u,v,1)^T$ is the homogeneous coordinate of such pixel and $\boldsymbol{K}$ is the pre-calibrated camera intrinsic matrix.

Given a 3D rigid body transformation $\boldsymbol{G} \in SE(3)$ from the reference frame to frame i can be written as:

$$\boldsymbol{G}_{ir} = \begin{bmatrix} \boldsymbol{R}_{ir} & \boldsymbol{t}_{ir} \\ \boldsymbol{0} & 1 \end{bmatrix} \tag{1}$$

where $\boldsymbol{R}_{ir} \in SO(3)$ and $\boldsymbol{t}_{ir} \in \mathbb{R}^3$ are the 3D rigid body rotation and translation from reference frame to frame i, respectively.

To better explain all costs involved in this paper, the pixel-wise direct error E_k of the k^{th} pixel between reference frame and the i^{th} frame can be generally written as:

$$E_k := \sum_{\boldsymbol{p} \in S_{\boldsymbol{p}}} w_{\boldsymbol{p}} \| F_i(\pi(\boldsymbol{p}')) - F_r(\pi(\boldsymbol{p})) \|_\gamma \tag{2}$$

$$\boldsymbol{p}' = \boldsymbol{R}_{ir}\pi^{-1}(\boldsymbol{p}, D_r(\boldsymbol{p})) + \boldsymbol{t}_{ir} \tag{3}$$

where $F(\cdot)$ represents some representation calculated from original Image I, such as intensity, gradient or some descriptors. The set $S_{\boldsymbol{p}}$ is the set of pixels in a pre-defined local patch in the reference frame, $w_{\boldsymbol{p}}$ is the weight assigned for each pixel, $\|\cdot\|_\gamma$ is the Huber norm, and the subscript $k \in S_{pixel}$ is the index of a sampled pixel selected from a selection algorithm.

B. Joint Optimization

A sliding window optimization framework using the Gauss-Newton algorithm, described in [7], is utilized to achieve the real-time motion estimation and 3D structure mapping. The optimization problem, in Eqn. 2, is furthered reduced to solve a nonlinear least-square minimization problem on Lie-manifolds. The corresponding Lie Group component $\boldsymbol{\xi} \in \mathfrak{se}(3)$ is introduced to represent the 6-DoF camera pose, where this element can be mapped to $\boldsymbol{G} \in SE(3)$ through the exponential mapping as:

$$\boldsymbol{G} = exp_{\mathfrak{se}(3)}(\boldsymbol{\xi}) \tag{4}$$

and the update rule in Lie Manifold can be performed through logarithm and exponential mapping as:

$$\boldsymbol{\xi}_{ik} = log_{\mathfrak{se}(3)}(exp_{\mathfrak{se}(3)}(\boldsymbol{\xi}_{ij}) \cdot exp_{\mathfrak{se}(3)}(\boldsymbol{\xi}_{jk})) \tag{5}$$

For monocular VO algorithms, there are typically two phases performing the optimization: tracking phase and reconstruction phase. In the tracking phase, the inter-frame camera pose $\boldsymbol{G}_{ir}$ is estimated given the depth map D_r as a prior. In the reconstruction phase, both the depth map D_r and the camera pose $\boldsymbol{G}_{ir}$ are jointly optimized to improve the overall performance. To boost the tracking robustness over large camera motion, a minimization over an image pyramid, named coarse-to-fine approach in [6], is utilized to achieve a good trade-off between precision and speed.

C. Illumination Robust Cost

Instead of presenting all real-time capable illumination-invariant formulations suitable for the joint optimization framework, we merely describe the intensity-based method as a baseline approach and the ones that achieved excellent performances in [10]. In Sec. III-C.1, the intensity values of the grayscale image is utilized to serve as a baseline method; in Sec. III-C.2, a global affine model is estimated to compensate the global lighting changes; in Sec. III-C.3, the gradient magnitude is used to formulate energy function thus achieving invariance to local illumination change; In Sec. III-C.4, the local descriptor from [22], named Census Transform or bit-plane, is presented to achieve local illumination invariance. Ultimately, an adaptively combined cost, using both an intensity-based energy with a global affine model and gradient-based energy, is proposed to achieve the invariance against global and local lighting changes without compromising convergence robustness.

1) Intensity: For standard implementation [7], we model the unknown function $F(\cdot)$ as an image function mapping 2D pixel coordinate to pixel intensity values obtained by bilinear interpolation, where a mathematical expression can be written as:

$$F_{Int}(\cdot) := I(\cdot) \tag{6}$$

2) Global Affine Model: The global affine model in [17] [18] can be plugged into the basic direct formulation, in Eqn. 2, to compensate for the additive and multiplicative global lighting or exposure changes, where the pixel-wise energy function can be modified as:

$$E_{k,GAff} := \sum_{\boldsymbol{p} \in S_{\boldsymbol{p}}} w_{\boldsymbol{p}} \| I_i(\pi(\boldsymbol{p}')) - \beta_i - \frac{e^{\alpha_i}}{e^{\alpha_r}} I_r(\pi(\boldsymbol{p})) - \beta_r \|_\gamma \tag{7}$$

where $\alpha_{i,r}$ and $\beta_{i,r}$ are global illumination affine model parameters, which are jointly optimized at each iteration. Combined with Huber norm, this *affine* method can work well in an environment without substantial local illumination changes.

3) Gradient Magnitude: In [19], the gradient magnitudes are utilized, instead of intensities, to formulate unknown function $F(\cdot)$ with a mathematical expression as:

$$F_{Grad}(\cdot) := \|\nabla I(\cdot)\|_2 \tag{8}$$

where $\nabla I(\cdot)$ calculates the gradient vector from the image intensities around a given pixel. The *gradient* method has proven to be robust with the local lighting changes and could achieve state-of-art tracking precision for RGB-D and Stereo applications.

4) Census Transform: The Census Transform [21] [22] is a local binary descriptor that compares a pixel intensity with its neighborhoods and results in one-bit results indicating the neighbors are lighter or darker than this given pixel. The results are then packed to formulate the bit-plane descriptor, which is the $F(\cdot)$ in this work. Considering a $N \times N$ local patch is utilized to calculate the descriptor, where the i^{th} descriptor can be written as:

$$F_{i,CT}(\cdot) := \begin{cases} 1, & \text{if } I(\cdot) > I(N_i(\cdot)) \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

where N_i represent the i^{th} neighborhood of the given pixel. It should be noted that the choice of the comparison operator is flexible, which could be $>$, $\leq$, $<$, or $\geq$, but it must be consistent for all pixels of interest. The gradient of the descriptor is approximated by using its image gradient for Jacobian computation in optimization.

D. Our Proposed Method

From the analysis of tracking accuracy and the convergence radius in Sec. IV-B, we found that the *gradient* method showed maximum robustness against various simulated lighting changes over other methods. However, it has a smaller convergence radius compared with the *affine* approach.

Motivated by descriptor field method [20], we propose to combine the energy functions of *affine*, and *gradient* approaches to obtain a clear global minimum while preserving large convergence basin through an adaptive weighting strategy. Specifically, we adaptively assign weights to *affine* and gradient costs at the different stage of the optimization: more weights put on the *affine* component at the beginning of optimization and more on the gradient component at the end. Mathematically, the proposed pixel-wise energy can be expressed as:

$$E_k := (1 - w_k) E_{k,GAff} + \frac{1}{2\sqrt{2}} w_k E_{k,Grad} \tag{10}$$

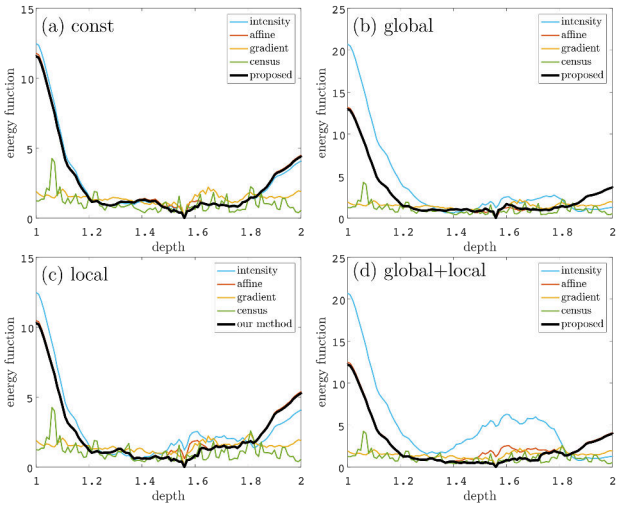


Fig. 2. The pixel-wise energy function of each described method is plotted against the pixel depth in four lighting conditions: (a) no illumination changes, (b) local illumination change, (c) global illumination change, and (d) local and global illumination change.

where the $2\sqrt{2}$ is the factor that scales the gradient magnitude to be the same range with intensity, and the w_k is the adaptive weight that starts from 0 and ends up with 1, which can be defined differently in the tracking phase $w_{k,1}$ and reconstruction phase $w_{k,2}$ as:

$$w_{k,1} := \frac{1}{tr(\mathbf{\Sigma}_\xi)} \quad w_{k,2} := \frac{1}{\sigma^2_{d_k^{-1}}} \tag{11}$$

where $tr(\cdot)$ represents the trace operator, $\mathbf{\Sigma}_\xi$ is the covariance matrix of pose estimates, and $\sigma^2_{d_k^{-1}}$ is the variance of inverse depth of the pixel from joint optimization.

In the tracking phase, the inverse depth of pixel is set constant, and the camera pose is the only parameter to be optimized. In the reconstruction phase, the camera poses are already well-initialized, but the inverse depth estimates could still have large errors. As a result, the uncertainty based weights, $w_{k,1}$ $w_{k,2}$, could force the estimate with a huge amount of uncertainty to have a large convergence basin and guide the determinate estimate to arrive at the global minimum, thus improving the tracking precision and robustness. The formulated pixel-wise energy function is analyzed in four typical lighting conditions in Fig. 2. In the comparison of all other methods, our proposed method holds nice cost functions in all conditions: a clear global minimum and large convergence basin.

IV. EVALUATION

In this section, the dataset utilized in our evaluation is first introduced in Sec. IV-A. The tracking accuracy and convergence radius during tracking phase are quantitatively evaluated and compared for each described methods in Sec. IV-B, and their overall system performance analysis regarding tracking accuracy, robustness, and runtime properties are presented in Sec. IV-C. Besides, a qualitative evaluation is performed to study the effect of solar glare on our proposed system to further support our claims in Sec. IV-D.

A. Dataset

All the evaluations, described in this paper, are performed on both publicly available urban-scene dataset: Virtual KITTI Dataset [11], as well as two challenging natural environment datasets: Devon Island Rover Navigation Dataset [25] and Symphony Lake Dataset [9].

The vKITTI Dataset is a photo-realistic synthetic video dataset, where both the ground truth depth maps and poses are available. We perform the tracking accuracy and convergence radius test using such dataset with the simulated global or/and local illumination changes in Sec. IV-B, and quantitatively compare all described methods concerning tracking accuracy and robustness in Sec. IV-C.

The Devon Island Rover Navigation Dataset, where a rover platform travels ten kilometers at Devon Island, is famous for rocky canyons and sandy terrains, with a forward-looking narrow field of view camera mounted on-board. Unlike vKITTI Dataset collecting data at the high frame rate (10 Hz), Devon Island dataset record images at 2 Hz, which is essentially more difficult for tracking tasks due to the potential large motion variation and the unexpected natural lighting condition changes. The solar glares slightly contaminate this dataset, and our evaluation is mainly conducted using these sequences regarding tracking accuracy and robustness. A pair of Magellan ProMark3 GPS units were utilized to produce differential GPS data for ground truth position. The quantitative comparison of all methods described in this paper, regarding tracking accuracy and robustness, is presented in Sec. IV-C.

The Symphony Lake Dataset consists of millions of natural lakeshore images at a resolution of 704 x 480 pixels captured by an Axis pan-tilt-zoom camera mounted on a boat with an auto-exposure strategy. The images captured in this dataset are abundant with sun glares in Fig. 4, which makes it suitable for our evaluation. However, the ground truth poses are not provided, so that merely the robustness properties can be evaluated in Sec. IV-C. Due to its abundance of solar glares, the qualitative evaluation of the VO performance under different degrees of solar glares are studied to support our claims in Sec. IV-D.

B. Tracking Accuracy and Convergence Radius

The tracking accuracy and the convergence radius, in the tracking phase, are quantitatively evaluated using synthetic vKITTI dataset, where both ground truth pose and depth map are provided. The images are pre-processed to add the global illumination bias or/and solar-glare patterned local illumination changes with the value ranging from -20 to +20. To prepare the inverse depth map as initialization, the ground truth depths are first mapped to inverse depths and normalized to average 1.0, and the corresponding scale is calculated by comparison of average inverse depth before and after normalization. The ground truth pose is then scaled base on this computed scale. A generated Gaussian noise is

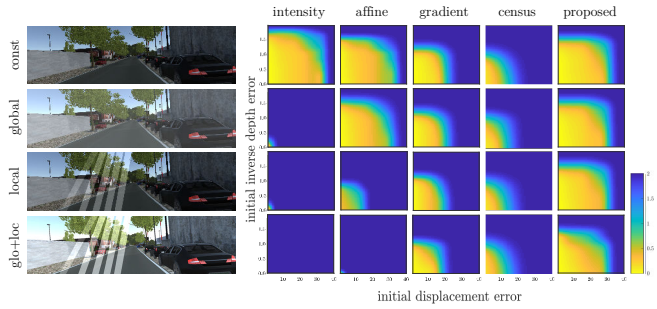


Fig. 3. Left: the example images with simulated illumination changes are provided in four lighting conditions: no lighting changes, global lighting changes, local lighting changes, and global+local lighting changes. Right: the tracking errors are plotted versus initial pose displacement (x-axis) and maximum inverse depth error (y-axis) for each described methods in different conditions.

added to their ground truth values, and the pose initialization is set by moving the camera with specific displacement in a random direction. The absolute value of this metric is meaningless, but it enables the relative comparisons between methods. As the result, there is no absolute metric is defined in Fig. 3.

We examine the tracking accuracy and robustness by merely analyzing the translational drift of pose estimates. A gradient-based pixel selection strategy with 3000 pixels of interest, 5 minimum and 10 maximum iterations are set to all tested systems to facilitate a fair comparison. In Fig. 3, we can observe that (1) the *intensity* method is only working correctly in *const* setting with the largest convergence radius; (2) the *affine* method can provide reasonable tracking precision up to global illumination changes with good convergence basin; (3) both the *gradient* and *census* methods can perform reliable tracking in all kinds of conditions but with a considerably smaller convergence radius, and *gradient* method shows better precision than that of *census*; (4) Our proposed method outperforms all other methods, which exhibits a similar convergence radius to the *affine* method and works correctly for all simulated lighting conditions.

It should be noted that this test is merely optimizing the inter-frame poses rather than the joint optimization of pose and scene structure in the reconstruction phase. This test aims to study and compare the performances of all described methods concerning poor initialization in the tracking phase. A more thorough evaluation of the whole system performance including the joint optimization will be provided in the next subsection.

C. Whole System Performance Evaluation

The whole system performance is assessed regarding the overall tracking accuracy, robustness, and runtime performance. For vKITTI and Devon Island dataset, the ground truth positions are provided, so we choose the translational drift and the average failure rates as metric for our evaluations. For Symphony Lake dataset, where no ground truth poses are available, we therefore merely count the average failure rate. It should be noted that the failure rate is either self-detected as tracking lost event or automatically detected as a substantial abnormal movement - five times larger than average displacement.

To facilitate a fair comparison, all critical parameters are set to be the same: a gradient-based pixel selection strategy, 2000 pixels of interest, 2 minimum and 6 maximum iterations. Since the precision of pose estimation in the initialization phase is lower in average, all pose estimates during initialization are not involved in our final evaluation. We run every test 10 times to calculate the average, and we do not consider loop-closure events. The scales drifts are recovered by comparing estimated camera translations with ground truth translations for every 100 frames.

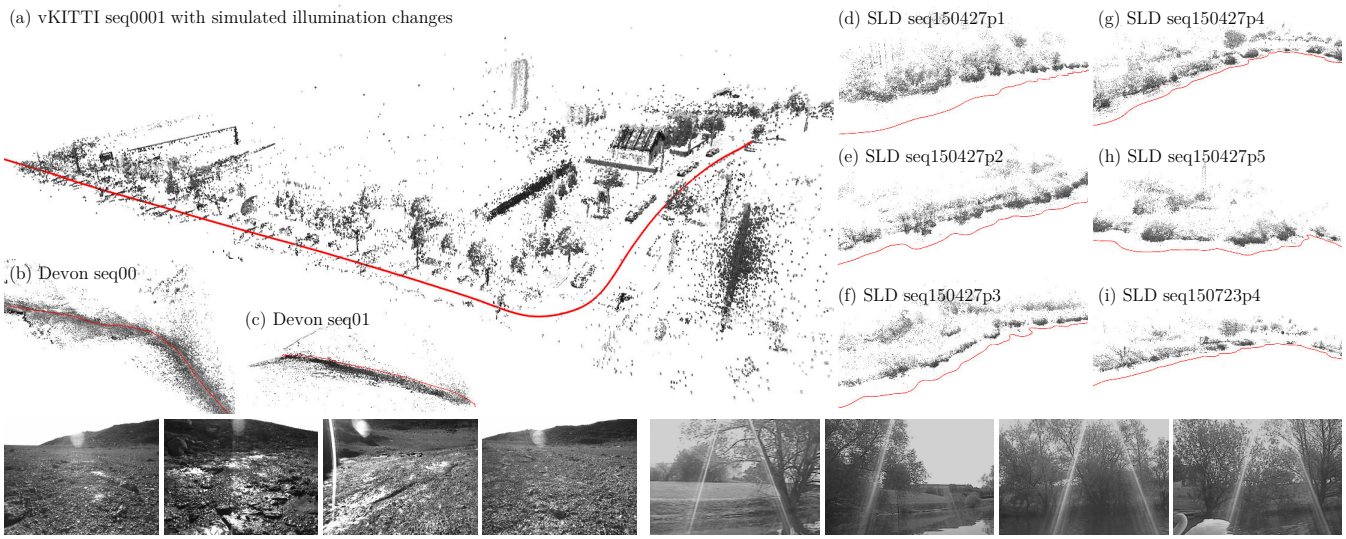
In Fig. 4, the qualitative results generated from our proposed system and the challenging images in the datasets are presented. A quantitative evaluation concerning translatonal drift, failure rate and average runtime between each described methods, as well as state-of-art ORBSLAM2 [3], are presented and compared in Table. I.

Overall, the joint optimization faces a much server robustness issue than that of visual tracking in Sec. IV-B, which is suspected to be the side effect of the joint optimization. The overall robustness of side-looking camera sequence is better than that of forward-looking cameras, however, without a comparison of tracking accuracy we cannot generate more conclusion. Specifically for each method, the *intensity* method achieves a zero failure rate for the original synthetic dataset, but soon faces severe divergence under lighting changes. The *affine* methods show good robustness but less accurate motion estimation in most dataset sequences, while the *census* and *gradient* present attractive tracking accuracy but diverges easily. Our proposed method, combining the sound characteristics of *affine* and *gradient* methods, provides the most reliable tracking estimates in all tested datasets. However, the computational load of our system is higher than that of other methods except for *census*, which is primarily attributed to the doubled residual and Jacobian computational load induced by our combined cost.

It should be noted that the missing entries for the vKITTI dataset indicate that the corresponding method continuously fails at the initialization phase, and the missing run-time performance of *orbslam2* is because it is a GPU implemented algorithm, but all other methods are CPU implemented..

D. Resistance to Solar Glare

In this paper, one of our major claims is that the solar glares can be modeled as global+local illumination changes, and our proposed system can operate in this situation without losing the track. To further support our claim, a qualitative analysis of the effect of solar flares on our motion estimation task is conducted. In Fig. 5, the reconstructed inverse depth map can be seen as an indicator about the tracking performance. Our proposed methods can generate well-spread inverse depth maps even facing strong solar glares, while the *affine* method tends to produce polarized depth map with sun glares nearby (red), which can be seen



(j) Example images from Devan Island Dataset (k) Example images from Symphony Lake Dataset

Fig. 4. The qualitative results of tracking and 3D scene reconstruction are presented utilizing synthetic vKITTI Dataset in (a), real-world Devon Island Dataset in (b) (c), and our Symphony Lake Dataset in (d) - (i). The example images from Devan Island Dataset in (j) and Symphony Lake Dataset in (k) are provided, all of which suffer from the global+local illumination changes induced by solar glare.

TABLE I

WHOLE SYSTEM PERFORMANCE EVALUATION REGARDING TRACKING PRECISION, ROBUSTNESS, AND RUNTIME.

	vKITTI								Devon Island				Symphony Lake				time
	const		*global*		*local*		*glo+loc*		*s00-09*		*s10-19*		*1502*	*1504*	*1507*	*1510*	
	rate	err	rate	err	rate	err	rate	err	rate	err	rate	err	rate	rate	rate	rate	
intensity	**0.0**	**0.37**	-	-	-	-	-	-	15.7	8.32	14.9	7.12	5.2	15.3	10.1	3.1	52
affine	0.3	0.38	**0.3**	0.38	4.5	0.42	-	-	5.9	5.78	6.2	5.66	4.1	13.8	8.2	**1.2**	58
gradient	3.5	**0.37**	3.3	**0.37**	3.4	**0.38**	3.6	0.38	10.9	5.12	11.4	5.27	13.6	12.2	13.5	10.9	71
census	3.4	0.43	3.3	0.43	3.3	0.44	3.2	0.43	10.3	5.13	12.7	**5.21**	12.4	13.2	13.3	13.7	623
proposed	0.3	**0.37**	**0.3**	**0.37**	**0.5**	**0.38**	**0.9**	**0.37**	**4.2**	**5.11**	**3.0**	5.23	**1.2**	**3.1**	**2.1**	1.4	105
orbslam2	0.9	0.38	2.4	0.42	1.8	0.41	5.3	0.55	12.6	6.35	13.1	6.31	7.7	16.7	12.1	5.5	-

Rates are failure rate per sequence, averaged over 10 trials. Err is a drift rate in $[cm/m]$. Processing time in $[ms]$ per frame.

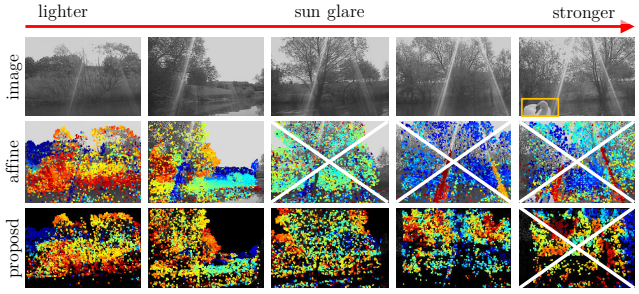


Fig. 5. A collection of images with gradually stronger (left to right) sun glare are chosen to test our system robustness. The reconstructed inverse depth maps using global affine model and our proposed method are presented, where the failure ones are marked with a big 'X'.

as a strong sign of tracking failure when the solar glare gets strong. Our proposed algorithm is tested to be able to reliably reconstructed observed scene even for strong solar glare cases. The only failure case we observed is that the sun glare appears at the same time with a swan close to the camera, which is a significantly more difficult issue.

V. CONCLUSIONS

State-of-art illumination-robust costs are evaluated in the context of a monocular joint optimization framework using a synthetic dataset with simulated light changes. Based on our analysis, a robust monocular VO approach is developed by combining intensity- and gradient-based costs with an adaptive weight. The proposed algorithm is extensively evaluated using real-world datasets regarding tracking accuracy, robustness, and runtime performance, which has shown the superiority of our proposed formulation relative to other methods. We further present a brief evaluation about the mapping performance using images affected by solar glare, which illustrates that the sun glare can be modeled as local illumination changes and its adverse effect on motion estimation can be alleviated by introducing an algorithm robust to such local changes. However, the computational cost of our system is higher than that of other methods, which is primarily attributed to the doubled residual and Jacobian computational load induced by our combined cost.

ACKNOWLEDGMENT

This work has partly been supported by the European Commission under the grant number H2020-ICT-644227-FLOURISH.

REFERENCES

[1] G. Klein and D. Murray, "Parallel tracking and mapping for small ar workspaces," in *Mixed and Augmented Reality, 2007. ISMAR 2007. 6th IEEE and ACM International Symposium on*. IEEE, 2007, pp. 225–234.

[2] H. Strasdat, J. Montiel, and A. J. Davison, "Scale drift-aware large scale monocular slam," *Robotics: Science and Systems VI*, vol. 2, 2010.

[3] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[4] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "Dtam: Dense tracking and mapping in real-time," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2320–2327.

[5] M. Pizzoli, C. Forster, and D. Scaramuzza, "Remode: Probabilistic, monocular dense reconstruction in real time," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. IEEE, 2014, pp. 2609–2616.

[6] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *European Conference on Computer Vision*. Springer, 2014, pp. 834–849.

[7] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 3, pp. 611–625, 2018.

[8] G. Chahine and C. Pradalier, "Survey of monocular slam algorithms in natural environments."

[9] S. Griffith, G. Chahine, and C. Pradalier, "Symphony lake dataset," *The International Journal of Robotics Research*, vol. 36, no. 11, pp. 1151–1158, 2017.

[10] S. Park, T. Schöps, and M. Pollefeys, "Illumination change robustness in direct visual slam," in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE, 2017, pp. 4523–4530.

[11] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4340–4349.

[12] J. Engel, V. Usenko, and D. Cremers, "A photometrically calibrated benchmark for monocular visual odometry," in *arXiv:1607.02555*, July 2016.

[13] M. Meilland, A. Comport, P. Rives, and I. S. A. Méditerranée, "Real-time dense visual tracking under large lighting variations," in *British Machine Vision Conference, University of Dundee*, vol. 29, 2011.

[14] T. Gonçalves and A. I. Comport, "Real-time direct tracking of color images in the presence of illumination variation," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*. IEEE, 2011, pp. 4417–4422.

[15] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart, "Robust visual inertial odometry using a direct ekf-based approach," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*. IEEE, 2015, pp. 298–304.

[16] W. N. Greene, K. Ok, P. Lommel, and N. Roy, "Multi-level mapping: Real-time dense monocular slam," in *Robotics and Automation (ICRA), 2016 IEEE International Conference on*. IEEE, 2016, pp. 833–840.

[17] S. Klose, P. Heise, and A. Knoll, "Efficient compositional approaches for real-time robust direct visual odometry from rgb-d data," in *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*. IEEE, 2013, pp. 1100–1106.

[18] J. Engel, J. Stückler, and D. Cremers, "Large-scale direct slam with stereo cameras," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*. IEEE, 2015, pp. 1935–1942.

[19] A. Dai, M. Nießner, M. Zollhöfer, S. Izadi, and C. Theobalt, "Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 76a, 2017.

[20] A. Crivellaro and V. Lepetit, "Robust 3d tracking with descriptor fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3414–3421.

[21] H. Alismail, B. Browning, and S. Lucey, "Direct visual odometry using bit-planes," *arXiv preprint arXiv:1604.00990*, 2016.

[22] H. Alismail, M. Kaess, B. Browning, and S. Lucey, "Direct visual odometry in low light using binary descriptors," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 444–451, 2017.

[23] G. Pascoe, W. Maddern, M. Tanner, P. Piniés, and P. Newman, "Nid-slam: Robust monocular slam using normalised information distance," in *Conference on Computer Vision and Pattern Recognition*, 2017.

[24] W. Maddern, A. Stewart, C. McManus, B. Upcroft, W. Churchill, and P. Newman, "Illumination invariant imaging: Applications in robust vision-based localisation, mapping and classification for autonomous vehicles," in *Proceedings of the Visual Place Recognition in Changing Environments Workshop, IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China*, vol. 2, 2014, p. 3.

[25] P. Furgale, P. Carle, J. Enright, and T. D. Barfoot, "The devon island rover navigation dataset," *The International Journal of Robotics Research*, vol. 31, no. 6, pp. 707–713, 2012.