Beyond Photometric Consistency: Gradient-based Dissimilarity for Improving Visual Odometry and Stereo Matching

Jan Quenzel Radu Alexandru Rosu Thomas Läbe Cyrill Stachniss Sven Behnke

Abstract— **Pose estimation and map building are central ingredients of autonomous robots and typically rely on the registration of sensor data. In this paper, we investigate a new metric for registering images that builds upon on the idea of the photometric error. Our approach combines a gradient orientation-based metric with a magnitude-dependent scaling term. We integrate both into stereo estimation as well as visual odometry systems and show clear benefits for typical disparity and direct image registration tasks when using our proposed metric. Our experimental evaluation indicats that our metric leads to more robust and more accurate estimates of the scene depth as well as camera trajectory. Thus, the metric improves camera pose estimation and in turn the mapping capabilities of mobile robots. We believe that a series of existing visual odometry and visual SLAM systems can benefit from the findings reported in this paper.**

I. Introduction

The ability to estimate the motion of a mobile platform based on onboard sensors is a key capability for mobile robots, autonomous cars, and other intelligent vehicles. Computing the trajectory of a camera is often referred to as visual odometry or VO and several approaches have been presented in this context [1], [2], [3], [4], [5]. VO as well as stereo matching approaches should provide accurate estimates of the relative camera motion and scenes depth under various circumstances. Thus, optimizing such systems towards increased robustness is an important objective for robots operating in the real world.

The gold standard for computing the relative orientation of two images of a calibrated camera is Nister's 5-point algorithm [6]. This approach computes the 5-DoF transformation between two monocular images based on known feature correspondences. It requires at least five corresponding points per image pair. In practice, more points are required to combine the 5-point algorithm with RANSAC followed by a least-squares refinement using only the inliers correspondences. An alternative approach to using explicit feature correspondences are comparisons of the pixel intensity values within the image pair. This approach is also called direct alignment and one often distinguishes semi-dense and dense methods, depending on the amount of compared pixels [5], [7], [8].

Features are often designed to be resilient against changes in the intensity values of the images, for example caused by illumination changes. Often, features are sparsely distributed over the image and their extraction can be a time consuming operation. In contrast to that, the intensity values of each pixel are directly accessible, raw measurements, and can be compared easily. Several direct methods consider the so-called photometric consistency of the image as the objective function to optimize. A key challenge of direct approaches is to achieve robustness because slight variations of the camera exposure, illumination change, vignetting effects, or motion blur directly affect the intensity measurements. In this paper, we address the problem of robustifying the direct alignment of image pairs through a new dis-similarity metric and in this way enable an improved depth estimate and alignment of image sequences.

The main contribution of this paper is a novel metric for direct image alignment and its exploitation in direct visual odometry. We build upon the gradient orientation-based metric proposed by Haber and Modersitzki [9] and improve it through the introduction of a magnitude depending scaling term. We furthermore integrate our metric into four different estimation systems (OpenCV, MeshStereo, DSO and Basalt) to show that our metric leads to improvements and evaluate our system to support our key claims, which are: First, our proposed metric is better suited for stereo disparity estimation than existing approaches. Second, it is also well-suited for direct image alignment. Third, our metric can be integrated into existing VO systems and increase their robustness while running at the frame rate of a typical camera.

II. Related Work

There has been extensive work to improve the robustness of visual odometry and visual SLAM methods towards illumination changes to ensure photometric consistency. Typically, feature-based methods are more resilience towards illumination changes since descriptors are designed to be distinguishable even under severe changes, across different seasons and invariant of camera type. SIFT is the standard choice for Structure-from-Motion [11] but has a significant computational cost. PTAM [12] using FAST [13] features and ORB_SLAM [4] are two prominent examples, which show that feature-based visual SLAM can work well in many scenarios while maintaining real-time performance when exploiting binary descriptor.

Under the assumption of a good initial guess, direct methods can obtain more accurate estimates of the camera

This work has been supported as part of the research group FOR 1505 by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as well as under Germany's Excellence Strategy, EXC-2070 - 390732324 (PhenoRob). Jan Quenzel, Radu Alexandru Rosu and Sven Behnke are with the Autonomous Intelligent Systems Group, University of Bonn, Germany. Thomas Läbe and Cyrill Stachniss are with the Robotics and Photogrammetry Lab, Institute of Geodesy and Geoinformation, University of Bonn, Germany.

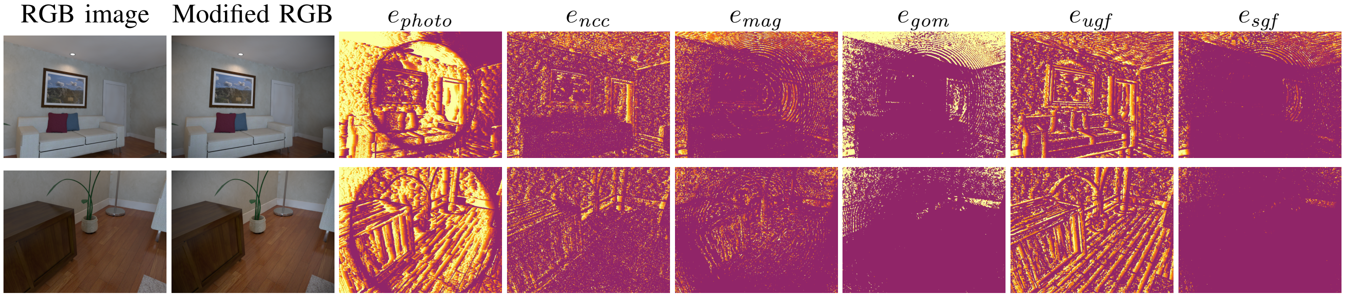


Fig. 1: Matching cost comparison on [10]: Disparity estimation against the same image with slight vignetting and different exposure time results in large disparity errors. The circle in e_{photo} occurs where vignetting and exposure change cancel out.

trajectory than feature-based approaches as they exploit all intensity measurements of the images. For this reason, Dai et al. [14] use features for initialization and to constrain a subsequent dense alignment. A popular approach, e.g. used by Schneider et al. [15], is to extract GoodFeaturesToTrack based on the Shi-Tomasi-Score and use the KLT optical flow tracker operating directly on intensity values. Similar to that, the Basalt system [16] uses locally scaled intensity differences between patches at FAST features within optical flow.

A further popular method for motion estimation from camera images is LSD-SLAM [1]. For robustness, the authors use the Huber norm during motion estimation and map creation, while minimizing a variance-weighted photometric error. LSD-SLAM creates in parallel the map for tracking by searching along the epipolar lines minimizing the sum of squared differences. For the stereo version, Engel et al. [17] alternate between estimating a global affine function to model changing brightness and optimizing the relative pose during alignment. As an alternative, Kerl et al. [2] propose to weight the photometric residuals with a t-distribution that better matches the RGB-D sensor characteristics.

Engel et al. [5] furthermore proposed with DSO a sparse-direct approach that further incorporates photometric calibration if available or estimates affine brightness changes with a logarithmic parametrization. They maintain an information filter to jointly estimate all involved variables.

Pascoe et al. [18] proposed to use the Normalized Information Distance (NID) metric for direct monocular SLAM. This works well even for tracking across seasons and under diverse illumination. Yet, the authors report to prefer photometric depth estimation for a stable initialization and only use NID after revisiting. Furthermore, Park et al. [8] presented an evaluation of different direct alignment metrics for visual SLAM. They favored the gradient magnitude due to its accuracy, robustness and speed while the census transform provided more accurate results at a much larger computational cost. In Stereo matching the census transform, e.g. in MeshStereo [19], and the absolute gradient difference combined with the photometric error, e.g. in StereoPatchMatch [20], are common.

In our work[1], we improve the gradient orientation based metric of Haber and Modersitzki [9] by introduction of a magnitude-dependent scaling term to simultaneously matching gradient magnitude and orientation. We apply this to solve direct image alignment for visual odometry as well as semi-dense disparity and depth estimation. We integrated our metric in two stereo matching algorithms as well as two VO systems. Hence, we evaluate and compare the metric against existing approaches on two stereo estimation and VO datasets.

III. Our Method

Our approach provides a new metric for pixel-wise matching and is easy to integrate into existing visual state estimation system. The metric measures the orientation of image gradients while also taking the magnitude into consideration. In the following, we denote sets and matrices with capital letters and vectors with bold lower case letters. We aim to find for a pixel $\mathbf{u}_i$ in the i^{th} image the corresponding pixel $\mathbf{u}_j$ in the j^{th} image that minimizes a dissimilarity measurement $e\left(\mathbf{u}_i, \mathbf{u}_j\right)$. The image coordinates $\mathbf{u} = (u_x, u_y)^\intercal_F$ are defined in the image domain $\Omega \subset \mathbb{R}^2$. For stereo matching, i and j correspond to the left and right image, while in direct image alignment i is often the current frame and j a previous (key-) frame.

A basic error function e_{photo} is photometric consistency

$$e_{photo}\left(\mathbf{u}_i, \mathbf{u}_j\right) = I_i\left(\mathbf{u}_i\right) - I_j\left(\mathbf{u}_j\right), \tag{1}$$

but more robust versions often rely on intensity gradients:

$$e_{gm}\left(\mathbf{u}_i, \mathbf{u}_j\right) = \left(\left\|\nabla I_i\left(\mathbf{u}_i\right)\right\| - \left\|\nabla I_j\left(\mathbf{u}_j\right)\right\|\right), \tag{2}$$

$$\mathbf{e}_{gn}\left(\mathbf{u}_i, \mathbf{u}_j\right) = \nabla I_i\left(\mathbf{u}_i\right) - \nabla I_j\left(\mathbf{u}_j\right). \tag{3}$$

The difference of the gradients $\mathbf{e}_{gn}$ incorporates both, magnitude and orientation. PatchMatch Stereo algorithms [20] typically combine this with the photometric error:

$$e_{pm}(\mathbf{u}_i, \mathbf{u}_j) = (1-\alpha)|e_{photo}(\mathbf{u}_i, \mathbf{u}_j)| + \alpha \left\|e_{gn}(\mathbf{u}_i, \mathbf{u}_j)\right\|_{\ell_1}. \tag{4}$$

[1] An accompanying video is available at https://www.ais.uni-bonn.de/videos/ICRA_2020_Gradient_Dissimilarity.

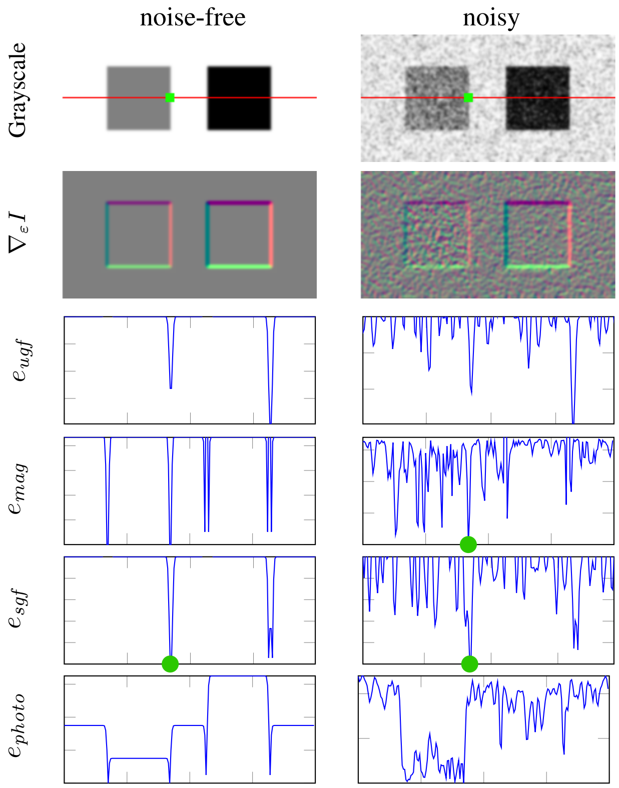


Fig. 2: Error comparison for gradient based metrics on a toy example. The lower boxes show the error between the green reference box and a shifted box along the red horizontal line. e_{ugf} prefers strong edges with same orientation, while e_{mag} does not take the orientation into account and thus generates further local minima. Our e_{sgf} provides the correct minima which are marked with a green circle.

A. Normalized Gradient-based Direct Image Alignment

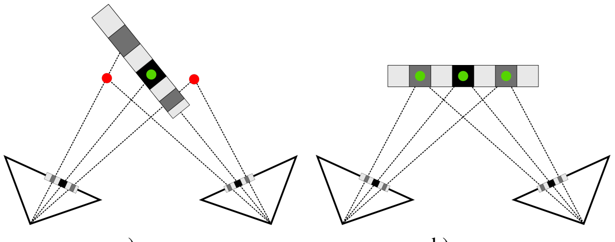
A complementary approach is to align the gradients orientation. The naïve approach may use the costly atan-operation to obtain the orientation angle θ and simply calculate differences. Instead, we follow the approach of [9], [21] to use the dot product and its relation to the cosine as a measure of orientation. If the two vectors $\mathbf{a}, \mathbf{b}$ have unit length, the dot product is equal to the cosine of the angle between the vectors, which is zero for perpendicular vectors, one for same and minus one for opposite orientation. Simply normalizing the gradient by its magnitude is undesirable as noise in low gradient regions will predominate the orientation. Hence, Taylor et al. [21] normalizes the dot product by its magnitude over a window:

$$e_{gom}\left(\mathbf{u}_i, \mathbf{u}_j\right) = 1 - \frac{\sum_{u \in W} \left|\nabla I_i\left(\mathbf{u}_i\right) \cdot \nabla I_j\left(\mathbf{u}_j\right)\right|}{\sum_{u \in W} \left\|\nabla I_i\left(\mathbf{u}_i\right)\right\| \left\|\nabla I_j\left(\mathbf{u}_j\right)\right\|}. \quad (5)$$

Instead we follow [9] and regularize the magnitude by a parameter ε:

$$\varepsilon = \frac{1}{|\Omega|} \sum_{\mathbf{u} \in \Omega} \left\|\nabla I(\mathbf{u})\right\|^2, \quad (6)$$

$$\nabla_\varepsilon I = \frac{\nabla I}{\sqrt{\left\|\nabla I\right\|^2 + \varepsilon}}. \quad (7)$$



a) e_{ugf} b) e_{sgf}

Fig. 3: Association impact: e_{ngf} and e_{ugf} tend to match patches with similar gradient orientation but stronger magnitude. This can cause severe distortions in the 3D reconstruction (left). Associating patches with similar gradient orientation and magnitude using e_{sgf} allows for correct triangulation (right).

This effectively downweighs the gradients magnitude in low gradient regions such that $\|\nabla_\varepsilon I\|$ will be close to zero. We estimate the parameter ε on a per image basis and will use ε and ϑ to make the distinction between different images more visible.

In the context of multi-modal image registration the authors of [9] minimize the per pixel error e_{ngf}:

$$e_{ngf}\left(\mathbf{u}_i, \mathbf{u}_j\right) = 1 - \left[\nabla_\varepsilon I_i(\mathbf{u}_i) \cdot \nabla_\vartheta I_j(\mathbf{u}_j)\right]^2. \quad (8)$$

Squaring the dot product, or taking the absolute value, ensures, that not only gradients with same orientation but also with opposite orientation coincide. This is important for registering CT to MRT data and vice versa where the image gradients may have opposite direction. This error has an important flaw as low gradient pixels prefer to match with higher magnitude ones rather than similar gradients. If the largest magnitude edge is always matched, we would obtain inconsistent depth estimates with high reprojection errors or when successively reducing the search region skew the region and obtain wrong estimates as visualized in Fig. 3.

Since we want to use images from the same sensor type, we can omit the square and only use the following residual:

$$e_{ugf}(\mathbf{u}_i, \mathbf{u}_j) = 1 - \nabla_\vartheta I_j\left(\mathbf{u}_j\right) \cdot \nabla_\varepsilon I_i\left(\mathbf{u}_i\right). \quad (9)$$

The errors e_{ngf} and e_{ugf} are bounded in the interval $[0, 2]$. To ensure the correct behavior for smaller gradients as visualized in Fig. 2, we scale the dot product by the maximum value:

$$e_{sgf}(\mathbf{u}_i, \mathbf{u}_j) = 1 - \frac{\nabla_\vartheta I_j\left(\mathbf{u}_j\right) \cdot \nabla_\varepsilon I_i\left(\mathbf{u}_i\right)}{\max\left(\left\|\nabla_\varepsilon I_i\left(\mathbf{u}_i\right)\right\|^2, \left\|\nabla_\vartheta I_j\left(\mathbf{u}_j\right)\right\|^2, \tau\right)}. \quad (10)$$

The scaling term of SGF thereby increases the number of successfully estimated points in semi-dense depth estimation. Here, τ is a small constant to prevent division by zero.

To further reduce the number of mathematical operations in above equation, especially the division by the regularized norm, we derived two further combinations of orientation and magnitude:

$$n\left(\mathbf{u}_i, \mathbf{u}_j\right) = \nabla I_j\left(\mathbf{u}_j\right) \cdot \nabla I_i\left(\mathbf{u}_i\right), \tag{11}$$

$$nij(\mathbf{u}_i, \mathbf{u}_j) = \frac{\left\|\nabla_\vartheta I_j\left(\mathbf{u}_j\right)\right\|}{\left\|\nabla_\varepsilon I_i\left(\mathbf{u}_i\right)\right\|}\left\|\nabla I_i\left(\mathbf{u}_i\right)\right\|^2, \tag{12}$$

$$nji(\mathbf{u}_i, \mathbf{u}_j) = \frac{\left\|\nabla_\varepsilon I_i\left(\mathbf{u}_i\right)\right\|}{\left\|\nabla_\vartheta I_j\left(\mathbf{u}_j\right)\right\|}\left\|\nabla I_j\left(\mathbf{u}_j\right)\right\|^2, \tag{13}$$

$$e_{sgf2}(\mathbf{u}_i, \mathbf{u}_j) = \max\left(nij, nji\right) - n\left(\mathbf{u}_i, \mathbf{u}_j\right) \tag{14}$$

$$e_{sgf3}(\mathbf{u}_i, \mathbf{u}_j) = \left\|\nabla I_i\left(\mathbf{u}_i\right)\right\|\left\|\nabla I_j\left(\mathbf{u}_j\right)\right\| - n\left(\mathbf{u}_i, \mathbf{u}_j\right). \tag{15}$$

Given a formulation for the error, we can now formulate stereo matching and direct image alignment. The former aims to find for each pixel $\mathbf{u}_l$ in the left image the corresponding pixel $\mathbf{u}_r$ in the right image that minimizes a dissimilarity measurement $e\left(\mathbf{u}_l, \mathbf{u}_r\right)$:

$$d^*_{\mathbf{u}} = \underset{d \in \mathcal{R}}{\arg\min} \sum_{\mathbf{u}_l \in W} e\left(\mathbf{u}_l, \mathbf{u}_r\left(d\right)\right), \tag{16}$$

$$\mathbf{u}_r\left(d\right) = \mathbf{u}_l - \left(d, 0\right)^\mathsf{T}. \tag{17}$$

Here, the disparity d is defined as the distance along the x-axis of the stereo rectified left and right image pair. For robustness, the error function e is calculated over a patch $W_{\mathbf{u}}$ with window size w centered around the pixel $\mathbf{u}$ rather than a single pixel. In the latter, we seek the transformation T_{cr} that aligns the reference with the current image optimally w.r.t. an error metric e between a reference pixel-patch $\mathcal{N}_{\mathbf{p}_r}$ around $\mathbf{p}_r$ and its projection onto I_c:

$$T_{cr} = \arg\min \sum_{\mathbf{p}_r \in \mathcal{M}} \sum_{\mathbf{p}_k \in \mathcal{N}_{\mathbf{p}_r}} \rho\left(\left\|e\left(\mathbf{p}_i\right)\right\|^2\right). \tag{18}$$

A robust cost function ρ like the Huber norm reduces the effect of outliers. This minimization is typically solved iteratively with the standard Gauss-Newton algorithm.

Hence, the Jacobian for e_{sgf} w.r.t. the pixel $\mathbf{u}_i$ is needed:

$$nn = \nabla_\vartheta I_j\left(\mathbf{u}_j\right) \cdot \nabla_\varepsilon I_i\left(\mathbf{u}_i\right), \tag{19}$$

$$s_1 = nn \begin{cases} -1, & \text{if } {\scriptstyle \|\nabla_\vartheta I_j\|^2 > \|\nabla_\varepsilon I_i\|^2} \\ 1 - \frac{2}{\|\nabla_\varepsilon I_i\|}, & \text{otherwise} \end{cases} \tag{20}$$

$$\frac{\partial e_{sgf}}{\partial \mathbf{u}_i} = -\frac{\left(\nabla_\vartheta I_j + s_1 \nabla_\varepsilon I_i\right)^\mathsf{T}}{\max\left(\left\|\nabla_\varepsilon I_i\right\|, \left\|\nabla_\vartheta I_j\right\|\right)} \frac{\left(\nabla_2\right) I_i}{\left\|\nabla I_i\right\|_\varepsilon}, \tag{21}$$

$$s_2 = \begin{cases} \frac{\|\nabla_\vartheta I_j\|}{\|\nabla_\varepsilon I_i\|}\left(2 - \frac{\|\nabla I_i\|^2}{\|\nabla I_i\|^2 + \varepsilon}\right), & \text{if } {\scriptstyle nij > nji} \\ \frac{\|\nabla_\varepsilon I_i\|}{\|\nabla_\vartheta I_j\|} \frac{\|\nabla I_j\|^2}{\left(\|\nabla I_i\|^2 + \varepsilon\right)}, & \text{otherwise} \end{cases} \tag{22}$$

$$\frac{\partial e_{sgf2}}{\partial \mathbf{u}_i} = \left(s_2 \nabla I_i - \nabla I_j\right)\left(\nabla_2\right) I_i, \tag{23}$$

$$\frac{\partial e_{sgf3}}{\partial \mathbf{u}_i} = \left(\frac{1}{2}\frac{\left\|\nabla I_j\right\|}{\left\|\nabla I_i\right\|}\nabla I_i - \nabla I_j\right)\left(\nabla_2\right) I_i. \tag{24}$$

Here, $\left(\nabla_2\right) I_i$ denotes the hessian of the intensity at pixel $\mathbf{u}_i$.

TABLE I: Evaluation on Middlebury Stereo 2014 training set [22]

		Orig.	e_{sad}	e_{agm}	e_{pm}	e_{sgf}
StereoBM	mean	7.20	5.80	6.31	4.56	**3.29**
	bad 1	18.36	20.51	21.33	17.19	**12.60**
	bad 2	16.41	17.01	17.79	14.25	**10.36**
	bad 4	14.88	14.19	14.69	11.94	**8.61**
	invalid	40.44	**34.51**	52.69	44.74	45.49
MeshStereo	mean	5.68	11.22	7.85	6.70	**4.17**
	bad 1	**16.87**	46.55	33.45	28.51	20.61
	bad 2	**13.02**	40.25	27.38	23.32	15.94
	bad 4	**10.71**	33.18	22.02	18.78	12.53
	invalid	**0.01**	1.01	0.09	0.08	0.04

TABLE II: Evaluation on KITTI Stereo 2015 training set [23]

		Orig.	e_{sad}	e_{agm}	e_{pm}	e_{sgf}
StereoBM	mean	6.11	3.21	3.17	1.74	**1.61**
	bad 1	19.80	19.79	22.13	15.93	**13.99**
	bad 2	11.60	10.07	11.04	6.87	**5.91**
	bad 4	9.03	6.34	6.73	3.94	**3.41**
	invalid	46.74	**29.57**	53.02	39.33	45.17
MeshStereo	mean	2.03	2.94	2.92	2.07	**2.02**
	bad 1	**27.95**	42.34	33.84	29.60	29.35
	bad 2	**12.00**	25.45	17.32	13.67	13.48
	bad 4	**5.57**	14.01	8.85	6.77	6.67
	invalid	0.07	0.15	0.10	0.08	**0.06**

IV. Evaluation

The first experiment is designed to illustrate the robustness of our metric under small image variations. To underline how even minimal image variations impact the dissimilarity metrics, we used images from the ICL-NUIM "lr kt2" sequence [10] and changed the exposure time and added a vignetting to frames 120 and 808, see Fig. 1 for a visualization. The disparity error is minimal in green regions with ideal disparity being 0 and window size 3. We evaluated $d \in [0, 20)$ for the different metrics. As expected, e_{photo} is large (avg. 8.13 px / 7.76 px), while gradient orientation alone (e_{ugf}) achieves on avg. 4.49 px / 4.78 px. Normalized cross-correlation (e_{ncc}) results in a disparity error of 3.04 px / 2.38 px. The magnitude (e_{mag}) is better suited (2.11 px / 1.40 px) while e_{gom} (2.02 px / 0.49 px) and e_{pm}(1.24 px / 0.18 px) perform best after our metric (1.21 px / 0.18 px) showing the smallest dissimilarity values.

The second experiment is designed to show our metrics suitability for (semi-) dense depth estimation supporting the first claim. For this, we integrated a variety of metrics for cost volume calculation into OpenCVs stereo block matching as well as the more sophisticated MeshStereo algorithm [19]. We evaluate the mean disparity error and report the percentage of bad pixels with 1, 2, and 4 px disparity error. Both algorithms are tested on the training sets of the Middlebury Stereo Benchmark [22] (half size) and the KITTI Stereo Benchmark [23]. We compare our metric e_{sgf} against the sum of absolute differences $e_{sad} = \sum|e_{photo}|$, the absolute difference of gradient magnitude $e_{agm} = |e_{gm}|$, the PatchMatch dissimilarity e_{pm}, and the original imple-

TABLE III: ATE results in meters on EuRoC dataset [24].

		MH1	MH2	MH3	MH4	MH5	V11	V12	V13	V21	V22	Avg
Original	OKVIS	0.085	0.083	0.135	0.143	0.278	0.041	0.956	0.102	0.054	0.063	0.194
	ORB-SLAM2	0.124	0.094	0.253	0.151	0.132	0.090	0.219	0.270	0.149	0.203	0.168
	SVO2	0.093	0.111	0.355	2.444	0.456	0.074	0.174	0.270	0.109	0.158	0.424
	DSO	**0.051**	0.045	0.165	0.164	0.460	0.194	0.151	1.075	0.080	0.098	0.227
	Basalt	0.076	0.045	0.058	0.096	0.141	0.041	0.052	0.073	0.032	**0.046**	0.066
Ours	DSO w/ e_{sgf}	0.071	0.050	0.264	0.235	0.237	0.142	0.178	0.933	0.072	0.086	0.206
	Basalt w/ e_{gm}	0.090	0.044	0.084	**0.091**	0.135	0.049	0.099	0.161	0.030	0.079	0.086
	Basalt w/ $\mathbf{e}_{gn}$	0.076	0.055	0.057	0.112	0.115	**0.039**	**0.042**	0.093	0.037	0.048	0.067
	Basalt w/ e_{sgf}	0.078	0.062	0.080	0.215	0.111	0.043	0.107	0.156	0.037	0.108	0.100
	Basalt w/ e_{sgf2}	0.086	0.065	0.081	0.109	0.148	0.040	0.069	**0.061**	**0.029**	0.058	0.075
	Basalt w/ e_{sgf3}	0.061	**0.042**	**0.065**	0.094	**0.106**	0.041	0.056	0.082	0.034	0.054	**0.063**

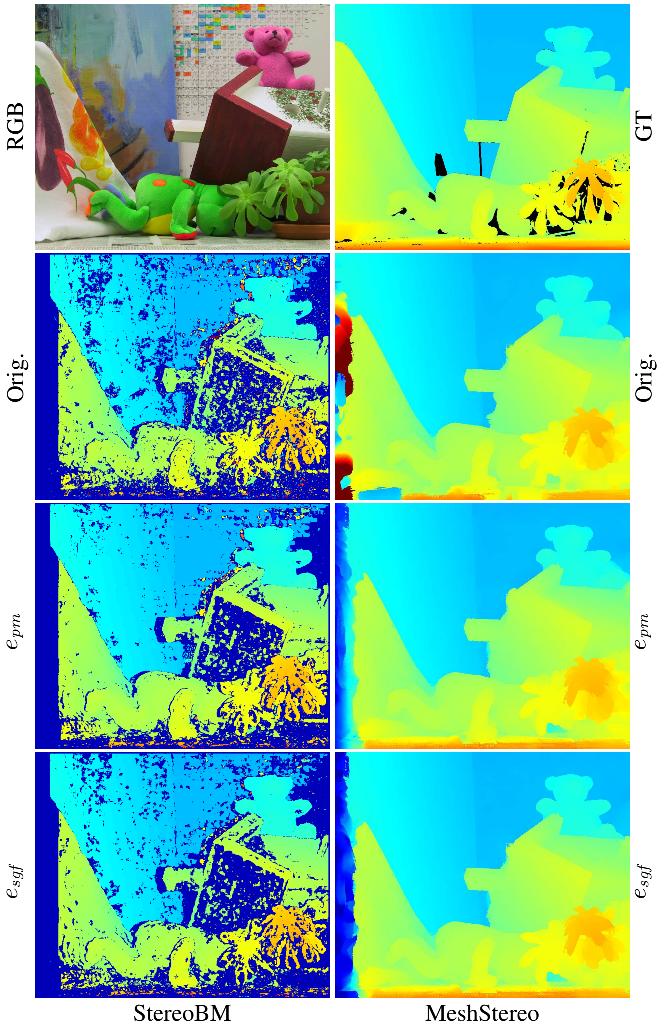


Fig. 4: Disparity comparison on Teddy of the Middlebury Stereo 2014 Benchmark [22] for the original algorithms and the two best metrics.

mentation. The dissimilarity in MeshStereo is calculated with Census-Transform. While OpenCV StereoBM uses e_{sad} too, a different prefilter provided a better result for e_{sad}. All other metrics were evaluated without prefiltering. We omitted $\mathbf{e}_{gn}$ since the results were nearly indistinguishable from e_{pm}. The results are shown in Tab. I and Tab. II. As can be seen, our metric provides in all cases the best mean disparity error. Fig. 5 shows an example on the KITTI Stereo Benchmark. Please note for e_{sgf}, although the bicyclist is not well represented with MeshStereo, it is with StereoBM. Furthermore, in the background less incorrect (too close) disparities are calculated with our metric.

To support our second and third claim, we provide comparisons to a set of state-of-the-art VO and VIO approaches including DSO [5], ORB-SLAM2 [4], OKVIS [25] and SVO2 [7] on the EuRoC dataset. We implemented the different metrics in the optical flow frontend of Basalt and carried out a two-fold cross validation with hyperopt [26] to obtain suitable parameters for each metric. We use the Scharr-Operator [27] on the rotated patches to obtain the intensity gradients. We observed that using finite differences degraded the obtainable precision for this task. For disparity estimation finite differences are sufficient.

In the case of DSO, we also show a modified version which replaces in the depth estimation the original patch similarity metric based on Brightness-Constancy-Assumption e_{photo} with our e_{sgf} term. Fig. 6 shows an example for both on V1_01 of the EuRoC dataset. For a fair comparison we disable the global bundle adjustment of ORB-SLAM2 and use Basalt purely in VIO mode. Furthermore, we evaluate the approaches, if provided, with the tailored parameters for the EuRoC dataset.

We report the mean ATE after alignment using [28] for all the frames which have a pose estimate. We align DSO with a similarity transform and the stereo algorithms with a rigid transform. To achieve a more reliable error estimate we run the algorithms repeatedly for each scenario and average the results. We report also the number of successful trackings for each algorithm out of a total of 250. Tracking is considered failed if the maximum scale error is above $1.5\,\mathrm{m}$ or the median scale error is greater than $0.1\,\mathrm{m}$. Tab. III gathers the final results.

One can see that our modified DSO using the e_{sgf} term for depth estimation performs better than the original DSO, having a lower average ATE. Furthermore, we observed an increase in successful tracking attempts by $10\,\%$ on V1_02 and V1_03 which exhibit strong lighting changes and reduced variance in ATE.

Basalt achieves with all tested metrics excellent results. Presumably e_{sgf} performs worse than our other derived metrics due to the more complex Jacobian, which is more difficult to optimize. Here, the simplifications of e_{sgf2} and e_{sgf3} payoff with e_{sgf3} achieving the best result.

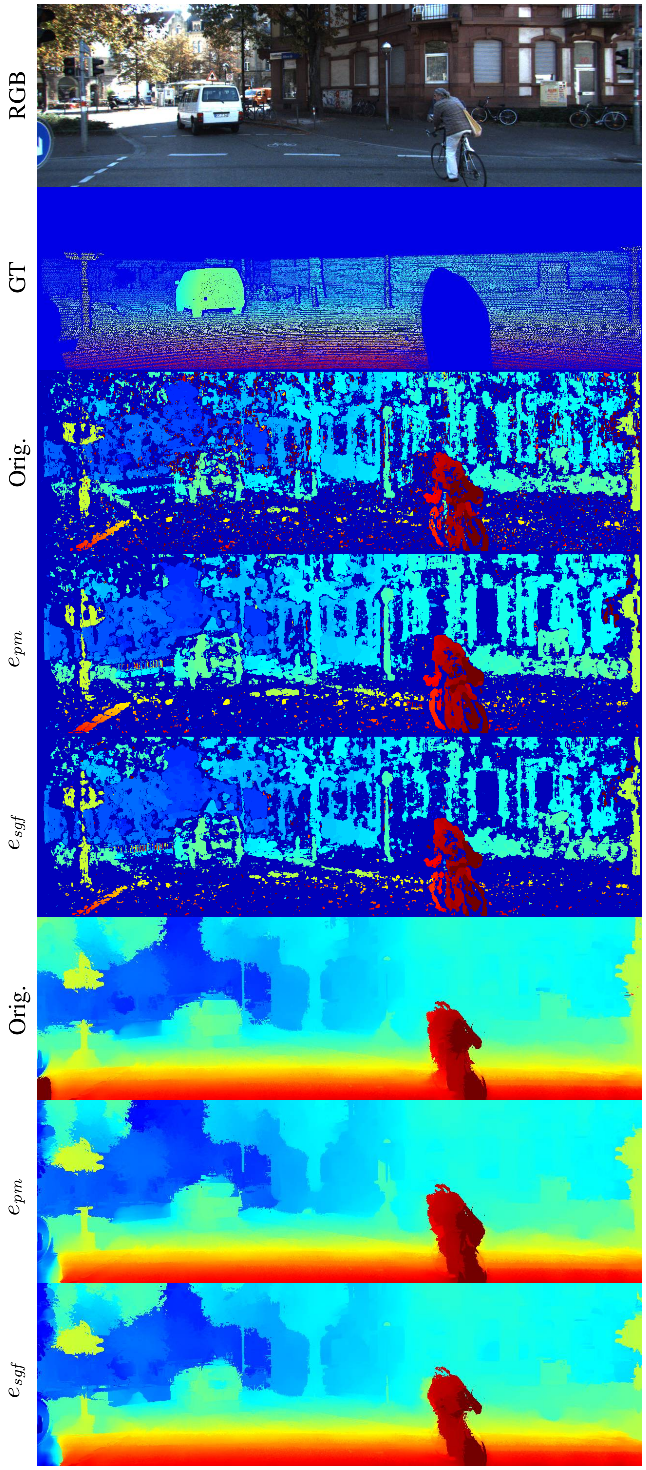
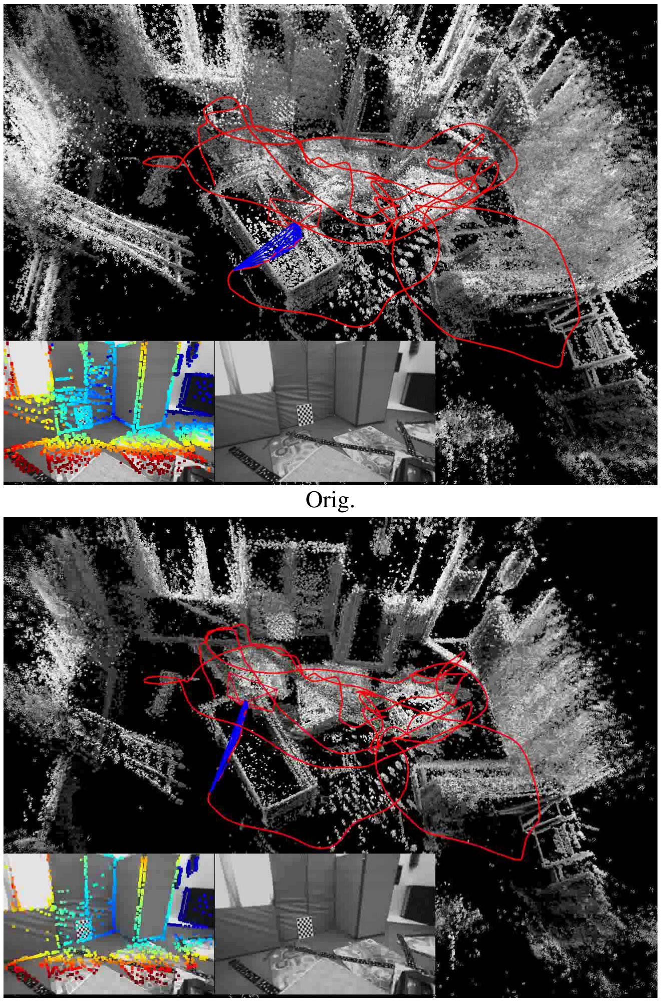


Fig. 5: Disparity comparison on image pair 2 of the KITTI Stereo 2015 Benchmark [23] for the original algorithms and the two best metrics.



Orig.

e_{sgf}

Fig. 6: Resulting map and trajectory (red line) of DSO [5] w/o and with e_{sgf} for depth estimation on V1_01 of the EuRoC dataset [24]. The reduced drift is clearly visible in the sharper edges and an reduction of double walls.

V. Conclusion

In this paper, we proposed a new metric for direct image alignment that is useful for motion and stereo depth estimation. Our metric improves the gradient orientation metric proposed by Haber and Modersitzki [9] and integrates a magnitude-dependent scaling term. This improves the robustness of the image alignment and is beneficiary for stereo matching and visual odometry computation alike. We integrated and evaluated our approach in a multitude of settings showing that the proposed metric is better suited for disparity estimation than existing approaches and well suited for image alignment. Furthermore, our approach is easy to integrate into existing visual systems and thus can make a positive impact on various visual odometry, SLAM, or similar state estimation approaches.

References

[1] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *Proc. of the European Conference on Computer Vision (ECCV)*, 2014.

[2] C. Kerl, J. Sturm, and D. Cremers, "Robust odometry estimation for RGB-D cameras," in *Proc. of the IEEE Int. Conference on Robotics and Automation (ICRA)*, May 2013.

[3] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast semi-direct monocular visual odometry," in *Proc. of the IEEE Int. Conference on Robotics and Automation (ICRA)*, 2014.

[4] R. Mur-Artal and J. 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.

[5] 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.

[6] D. Nistér, "An efficient solution to the five-point relative pose problem," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 6, pp. 756–770, 2004.

[7] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza, "Svo: Semidirect visual odometry for monocular and multicamera systems," *IEEE Trans. on Robotics*, vol. 33, no. 2, pp. 249–265, 2017.

[8] S. Park, T. Schöps, and M. Pollefeys, "Illumination change robustness in direct visual SLAM," in *Proc. of the IEEE Int. Conference on Robotics and Automation (ICRA)*, 2017.

[9] E. Haber and J. Modersitzki, "Intensity gradient based registration and fusion of multi-modal images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2006.

[10] A. Handa, T. Whelan, J. McDonald, and A. Davison, "A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM," in *Proc. of the IEEE Int. Conference on Robotics and Automation (ICRA)*, 2014.

[11] J. Schönberger, M. Pollefeys, and J. Frahm, "Structure-from-Motion revisited," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[12] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *Proc. of the IEEE Int. Symposium on Mixed and Augmented Reality (ISMAR)*, 2007, pp. 225–234.

[13] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Proc. of the European Conference on Computer Vision (ECCV)*, 2006, pp. 430–443.

[14] A. Dai, M. Nießner, M. Zollöfer, S. Izadi, and C. Theobalt, "Bundlefusion: Real-time globally consistent 3D reconstruction using on-the-fly surface re-integration," *ACM Transactions on Graphics*, 2017.

[23] M. Menze, C. Heipke, and A. Geiger, "Joint 3D estimation of vehicles and scene flow," in *ISPRS Workshop on Image Sequence Analysis (ISA)*, 2015.

[15] J. Schneider, F. Schindler, T. Läbe, and W. Förstner, "Bundle adjustment for multi-camera systems with points at infinity," in *Proc. of the Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. (ISPRS)*, 2012.

[16] V. Usenko, N. Demmel, D. Schubert, J. Stückler, and D. Cremers, "Visual-inertial mapping with non-linear factor recovery," *arXiv preprint arXiv:1904.06504*, 2019.

[17] J. Engel, J. Stueckler, and D. Cremers, "Large-scale direct SLAM with stereo cameras," in *Proc. of the IEEE/RSJ Int. Conference on Intelligent Robots and Systems (IROS)*, September 2015.

[18] G. Pascoe, W. Maddern, M. Tanner, P. Piniés, and P. Newman, "NID-SLAM: Robust monocular SLAM using normalised information distance," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[19] C. Zhang, Z. Li, Y. Cheng, R. Cai, H. Chao, and Y. Rui, "MeshStereo: A global stereo model with mesh alignment regularization for view interpolation," in *Proc. of the IEEE Int. Conference on Computer Vision (ICCV)*, 2015, pp. 2057–2065.

[20] M. Bleyer, C. Rhemann, and C. Rother, "PatchMatch Stereo - stereo matching with slanted support windows," in *Proc. of the British Machine Vision Conference (BMVC)*, 2011.

[21] Z. Taylor, J. Nieto, and D. Johnson, "Multi-modal sensor calibration using a gradient orientation measure," *JFR*, vol. 32, no. 5, pp. 675–695, 2015.

[22] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nesic, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth." in *Proc. of the German Conference on Pattern Recognition (GCPR)*, vol. 8753, 2014.

[24] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. Achtelik, and R. Siegwart, "The EuRoC micro aerial vehicle datasets," *The Int. Journal of Robotics Research*, 2016.

[25] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual–inertial odometry using nonlinear optimization," *The Int. Journal of Robotics Research*, vol. 34, no. 3, pp. 314–334, 2015.

[26] J. Bergstra, D. Yamins, and D. D. Cox, "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures," in *Proceedings of the International Conference on Machine Learning*, 2013, pp. I–115–I–123.

[27] H. Scharr, "Optimal filters for extended optical flow," in *International Workshop on Complex Motion (IWCM)*, 2004.

[28] Z. Zhang and D. Scaramuzza, "A tutorial on quantitative trajectory evaluation for visual(-inertial) odometry," in *Proc. of the IEEE/RSJ Int. Conference on Intelligent Robots and Systems (IROS)*, 2018.