Enhancing Direct Camera Tracking with Dense Feature Descriptors

Hatem Alismail*, Brett Browning, and Simon Lucey

The Robotics Institute
Carnegie Mellon University

Abstract. Direct camera tracking is a popular tool for motion estimation. It promises more precise estimates, enhanced robustness as well as denser reconstruction efficiently. However, most direct tracking algorithms rely on the brightness constancy assumption, which is seldom satisfied in the real world. This means that direct tracking is unsuitable when dealing with sudden and arbitrary illumination changes. In this work, we propose a non-parametric approach to address illumination variations in direct tracking. Instead of modeling illumination, or relying on difficult to optimize robust similarity metrics, we propose to directly minimize the squared distance between densely evaluated local feature descriptors. Our approach is shown to perform well in terms of robustness and runtime. The algorithm is evaluated on two direct tracking problems: template tracking and direct visual odometry and using different features descriptors proposed in the literature.

1 Introduction

With the increasing availability of high frame rate cameras, direct tracking is becoming a more popular tool in myriad applications such as visual odometry [1,2], visual SLAM [3,4], augmented and virtual reality [5] and dense reconstruction [6]. Advantages of direct tracking include: (i) increased precision as much of the image could be used to estimate a few degrees of freedom [7], (ii) enhanced tracking robustness in feature-poor environments, where high frequency image content (corners and edges) are not readily available, (iii) improved ability in handling ambiguously textured scenes [8], and (iv) improved running time by exploiting the trivially parallel nature of direct tracking [6].

However, the main limitation of direct tracking is the reliance on the *brightness constancy* assumption [9,10], which is seldom satisfied in the real world. Since the seminal works of Lucas and Kanade [10] and Horn and Schunk [9], researchers have been actively seeking more robust tracking systems [11,12,13,14,15]. Nevertheless, the majority of research efforts have been focused on two ideas: One, is to rely on intrinsically robust objectives, such as maximizing normalized

* Corresponding author halismai@cs.cmu.edu. This manuscript is a pre-print version. The final version will appear in the 13th Asian Conference on Computer Vision (ACCV 16) and will be available on link.springer.com

correlation [16], or the Mutual Information [13], which are inefficient to optimize and more sensitive to the initialization point [17]. The other, is to attempt to model the illumination parameters of the scene as part of the problem formulation [11], which is usually limited by the modeling assumptions.

In this work, we propose the use of **densely evaluated local feature descriptor as a nonparametric means to achieving illumination invariant dense tracking**. We will show that while feature descriptors are inherently discontinuous, they are suitable for gradient-based optimization when used in a multi-channel framework. We will also show that, depending on the feature descriptor, it is possible to tackle challenging illumination conditions without resorting to any illumination modeling assumptions, which are difficult to craft correctly. Finally, we show that the change required to make use of descriptors in current tracking systems is minimal, and the additional computational cost is not a significant barrier.

There exists a multitude of previous work dedicated to evaluating direct tracking. For instance, Baker and Matthews [18] evaluate a range of linearization and optimization strategies along with the effects of parameterization and illumination conditions. Handa *et al*. [4] characterize direct tracking performance in terms of the frame-rate of the camera. Klose *et al*. [19] examine the effect of different linearization and optimization strategies on the precision of RGB-D direct mapping. Zia *et al*. [20] explore the parameter space of direct tracking considering power consumption and frame-rate on desktop and mobile devices. Sun *et al*. [21] evaluate different algorithms and optimization strategies for optical flow estimation. Vogel *et al*. evaluate different data costs for optical flow [22]. Nonetheless, the fundamental question of the quantity being optimized, especially the use feature descriptors in direct tracking, has not yet been fully explored.

Feature descriptors, whether hand crafted [23], or learned [24], have a long and rich history in Computer Vision and have been instrumental to the success of many vision applications such as Structure-from-Motion (SFM) [25], Multi-View Stereo [26] and object recognition [27]. Notwithstanding, their use in direct tracking has been limited and is only beginning to be explored [28,29]. One could argue that this line of investigation has been hampered by the false assumption that feature descriptors, unlike pixel intensities, are non-differentiable due to their discontinuous nature. Hence, the use of feature descriptors in direct tracking has been neglected from the onset.

Among the first application of descriptors in direct tracking is the "distribution fields" work [30,31], which focused on preserving small image details that are usually lost in coarse octaves of the scale-space. Application of classical feature descriptors such as SIFT [32] and HOG [33] to Active Appearance Models have been also explored in the literature demonstrating robust alignment results [28]. The suitability of discrete descriptors for the linearization required by direct tracking has been investigated in recent work [34], where it was shown that if feature coordinates are independent, then gradient estimation of feature channels can be obtained deterministically using finite difference filters. This is advantageous as gradient-based optimization is more efficient and more precise

than discrete optimization [35]. Recent work has applied descriptors to template tracking [36] in an effort to track non-Lambertian surfaces more robustly.

1.1 Contributions

In this work, we propose the use of densely evaluated feature descriptors as a means to *significantly* improve direct tracking robustness under challenging illumination conditions. We show that for dense feature descriptors to be useful for direct tracking, they must be evaluated on a small neighborhood and must have sufficient discrimination power.

We evaluate the use of dense feature descriptors using two direct tracking problems: (i) parametric motion estimation using an Affine motion model, where the warping function is linear; (ii) direct visual odometry, which is more challenging than affine template tracking due to the nonlinear warping function and its dependence on, potentially sparse, depth information.

2 Direct Camera Tracking

Let the intensity of a pixel coordinate $\mathbf{p} = (u,\ v)^\top$ in the *reference* image be given by $\mathbf{I}(\mathbf{p}) \in \mathbb{R}$. After camera motion, a new image is obtained $\mathbf{I}'(\mathbf{p}')$. The goal of direct tracking is to estimate an increment of the camera motion parameters $\Delta\boldsymbol{\theta} \in \mathbb{R}^d$ such that the photometric error is minimized

$$\Delta\boldsymbol{\theta}^* = \underset{\Delta\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{\mathbf{p}\in\Omega} \left\|\mathbf{I}'\left(\mathtt{w}(\mathbf{p}; \boldsymbol{\theta} \boxplus \Delta\boldsymbol{\theta})\right) - \mathbf{I}\left(\mathbf{p}\right)\right\|^2, \tag{1}$$

where Ω is a subset of pixel coordinates of interest in the reference frame, $\mathtt{w}\left(\cdot\right)$ is a *warping* function that depends on the parameter vector we seek to estimate, and $\boldsymbol{\theta}$ is an initial estimate of the motion parameters. After every iteration, the current estimate of parameters is updated (*i.e.* $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} \boxplus \Delta\boldsymbol{\theta}$), where $\boxplus$ generalizes the addition operator over the optimization manifold. The process is repeated until convergence, or some termination criteria have been satisfied [10,18].

By (conceptually) interchanging the roles of the template and input images, Baker & Matthews devise a more efficient alignment techniques known as the Inverse Compositional (IC) algorithm [18]. Under the IC formulation we seek an update $\Delta\boldsymbol{\theta}$ that satisfies

$$\Delta\boldsymbol{\theta}^* = \underset{\Delta\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{\mathbf{p}\in\Omega} \left\|\mathbf{I}'\left(\mathtt{w}(\mathbf{p}; \boldsymbol{\theta})\right) - \mathbf{I}\left(\mathtt{w}(\mathbf{p}; \Delta\boldsymbol{\theta})\right)\right\|^2. \tag{2}$$

The optimization problem in Eq. (2) is nonlinear irrespective of the form of the warping function or the parameters, as—in general—there is no linear relationship between pixel coordinates and their intensities. By equating the partial derivatives of the first-order Taylor expansion of Eq. (2) to zero, we reach at solution given by the following closed-form (normal equations) $\Delta\boldsymbol{\theta} = \left(\mathbf{J}^\top\mathbf{J}\right)^{-1}\mathbf{J}^\top\mathbf{e}$,

where $\mathbf{J} = \left(\mathbf{g}(\mathbf{p}_1)^\top, \ \ldots, \ \mathbf{g}(\mathbf{p}_m)^\top\right) \in \mathbb{R}^{m \times d}$ is the matrix of first-order partial derivatives of the objective function, m is the number of pixels, and $d = |\boldsymbol{\theta}|$ is the number of parameters. Each $\mathbf{g}$ is $\in \mathbb{R}^{1 \times d}$ and is given by the chain rule as $\mathbf{g}(\mathbf{p})^\top = \nabla \mathbf{I}(\mathbf{p}) \frac{\partial \mathtt{w}}{\partial \boldsymbol{\theta}}$, where $\nabla \mathbf{I} = (\partial \mathbf{I}/\partial u, \ \partial \mathbf{I}/\partial v) \in \mathbb{R}^{1 \times 2}$ is the image gradient along the u- and v- directions respectively. The quantity $\mathbf{e}(\mathbf{p}) = \mathbf{I}'(\mathtt{w}(\mathbf{p}; \boldsymbol{\theta})) - \mathbf{I}(\mathbf{p})$ is the vector of residuals. Finally, the parameters are updated via the IC rule given by $\mathtt{w}(\mathbf{p}, \boldsymbol{\theta}) \leftarrow \mathtt{w}(\mathbf{p}, \boldsymbol{\theta}) \circ \mathtt{w}(\mathbf{p}, \Delta\boldsymbol{\theta})^{-1}$.

2.1 Direct Tracking with Feature Descriptors

Direct tracking using image intensities (the brightness constraint in Eq. (1)) is known to be sensitive to illumination change. To address this limitation, we propose the use of a *descriptor constancy* assumption. Namely, we seek an update to the parameters such that

$$\Delta\boldsymbol{\theta}^* = \underset{\Delta\boldsymbol{\theta}}{\operatorname{argmin}} \| \boldsymbol{\phi}(\mathbf{I}'(\mathtt{w}(\mathbf{p}; \boldsymbol{\theta} \boxplus \Delta\boldsymbol{\theta}))) - \boldsymbol{\phi}(\mathbf{I}(\mathbf{p})) \|^2, \tag{3}$$

where $\boldsymbol{\phi}(\cdot)$ is a multi-dimensional feature descriptor applied to the reference and the warped input images.

The descriptor constancy objective in Eq. (3) is more complicated than its brightness counterpart in Eq. (1) as feature descriptors are high dimensional and the suitability of their linearization remains unclear. In the sequel, we will show that various descriptors linearize well and are suitable for direct tracking.

2.2 Desiderata

The usual goal of direct tracking is to maximize the precision of the estimated parameters. The linearization required in direct tracking implicitly assumes that we are close enough to the local minima. This fact is usually expressed by assuming small displacements between the input images. In order to maximize precision, it is important to balance the complexity of the descriptor as a function of its sampling density. Namely, descriptors with long range spatial connections such as SIFT [32] and HOG [33], while robust to a range of deformations in the image, they contribute little to tracking precision. This is due to the increased dependencies between pixels contributing to the linear system. We will experimentally validate this hypothesis in the experimental section of this work. Hence, good descriptors for illumination invariant tracking must be: (1) locally limited with respect to their spatial extent, and (2) efficient to compute, which is desired for practical reasons. Both requirements, locality and efficiency, are closely related as most local descriptors are efficient to compute as well.

2.3 Pre-computing descriptors for efficiency

Descriptor constancy as stated in Eq. (3) requires re-computing the descriptors after every iteration of image warping. In addition to the extra computational

cost of repeated applications of the descriptor, it is difficult to warp individual pixel locations if their motion depends on depth, such as in direct visual odometry. The difficulty arises from the lack of a 3D model that could be used to reason about occlusions and discontinuities in the image. An approximation to the descriptor constancy objective in Eq. (3) is to pre-compute the descriptors and minimize the following expression instead (using the IC formulation):

$$\min_{\Delta\boldsymbol{\theta}} \sum_{\mathbf{p}\in\Omega} \sum_{i=1}^{N_c} \|\boldsymbol{\Phi}'_i(\mathtt{w}(\mathbf{p};\boldsymbol{\theta})) - \boldsymbol{\Phi}_i(\mathtt{w}(\mathbf{p};\Delta\boldsymbol{\theta}))\|^2, \tag{4}$$

where $\boldsymbol{\Phi}_i$ indicates the i-th coordinate of the pre-computed descriptor and N_c is the number of channels. This approximation incurs a loss of accuracy especially with nonlinear warps. However, we found that the loss of accuracy induced when using Eq. (4) instead of Eq. (3) to be insignificant in comparison to the computational savings and simplicity of implementation.

The minimization of Eq. (4) is performed similarly to minimizing the sum of squared intensity residuals in Eq. (1). By taking the derivative with respect to parameters of the 1st-order expansion of Eq. (4) we arrive at

$$\sum_{\mathbf{p}\in\Omega} \sum_{i=1}^{N_c} \left(\frac{\partial\boldsymbol{\Phi}_i}{\partial\boldsymbol{\theta}}\right)^\top \left|\boldsymbol{\Phi}'_i(\mathtt{w}(\mathbf{p};\boldsymbol{\theta})) - \boldsymbol{\Phi}_i(\mathbf{p}) - \frac{\partial\boldsymbol{\Phi}_i}{\partial\boldsymbol{\theta}}\Delta\boldsymbol{\theta},\right| \tag{5}$$

where the derivative of the pre-computed descriptor with respect to the parameters is given by the chain-rule per channel i as

$$\frac{\partial\boldsymbol{\Phi}_i}{\partial\boldsymbol{\theta}} = \frac{\partial\boldsymbol{\Phi}}{\partial\mathbf{u}}\frac{\partial\mathtt{w}(\mathbf{u};\boldsymbol{\theta})}{\partial\mathbf{u}}. \tag{6}$$

Subsequently, the Gauss-Newton approximation to the Hessian is given by the summation of the partial derivatives across all pixels and across all channels as

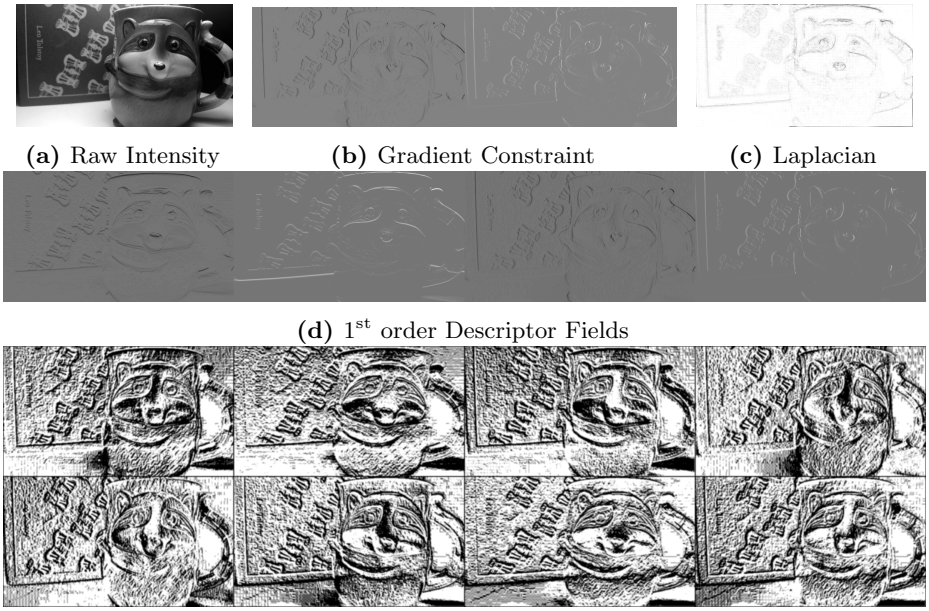
$$\mathbf{H}(\boldsymbol{\theta}) = \sum_{\mathbf{p}\in\Omega} \sum_{i=1}^{N_c} \frac{\partial\boldsymbol{\Phi}_i(\mathbf{p};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}^\top \frac{\partial\boldsymbol{\Phi}_i(\mathbf{p};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}. \tag{7}$$

To this end, we will consider various feature descriptors in the literature that are suitable for high-precision illumination invariant tracking, which we review in the next section.

3 Densely Evaluated Descriptors

We evaluate a number of descriptors suitable for dense tracking as summarized in Table 1 and visualized in Fig. 1. The descriptors are:

– **Raw intensity**: this is the trivial form of a feature descriptor, which uses the raw image intensities. We work with grayscale images, and hence it is a single channel, $\boldsymbol{\Phi}(\mathbf{I}) = \{\mathbf{I}\}$.



(a) Raw Intensity **(b)** Gradient Constraint **(c)** Laplacian

(d) 1st order Descriptor Fields

(e) Bit-Planes (smoothed)

Fig. 1: Visualization of the different descriptors. Best viewed on a screen.

- **Gradient constancy**: the image gradient measures the rate of change of intensity and hence it is invariant to additive changes [37]. We found that including the raw image intensities in the optimization with the gradient constraint to work better. The descriptor is composed of three channels and is given by: $\boldsymbol{\phi} = \{\mathbf{I},\ \nabla_u \mathbf{I},\ \nabla_v \mathbf{I}\}$
- **Laplacian**: the Laplacian is based on the 2nd order derivatives of the image and, similar to the gradient constraint, it provides invariance to additive change, but using only a single channel. We found that including the raw intensities to improve results. The descriptor is given by: $\boldsymbol{\Phi} = \left\{\mathbf{I},\ |\nabla^2 \mathbf{I}|\right\}$.
- **Descriptor Fields (DF)** [36] where the idea is to separate the image gradients into different channels based on their sign. After that, a smoothing step is performed. Using first-order image gradients, denoted by DF-1, the descriptor is composed of four channels and is given by:

$$\boldsymbol{\Phi}_{\text{DF-1}}(\mathbf{I}) = \left\{[\nabla_u \mathbf{I}]^+,\ [\nabla_u \mathbf{I}]^-,\ [\nabla_v \mathbf{I}]^+,\ [\nabla_v \mathbf{I}]^-\right\}.$$

  The notation $[\cdot]^+$ indicates selecting the positive part of the gradients, or zero otherwise. The 2nd order DF, denoted by DF-2, includes 2-nd order gradient information and is composed of 10 channels. As show in Fig. 1, DF is real-valued and results in sparse channels. We refer the reader to [36] for additional details.
- **Bit-Planes** [38]: is binary descriptor based on the Census transform [39], where channels are constructed by performing local pixel comparisons. When

evaluated in a 3×3 neighborhood, the descriptor results in eight channels given by: $\boldsymbol{\Phi}(\mathbf{I}) = \{\mathbf{I}(\mathbf{x}) \geq \mathbf{I}(\mathbf{x} + \Delta\mathbf{x}_j)\}_{j=1}^{8}$, where $\mathbf{I}(\mathbf{x} + \Delta\mathbf{x}_j)$ indicates the image sampled at the j-th neighbor location in the neighborhood. We refer the reader to [38] for additional details.

Table 1: Descriptors evaluated in this work

Name	Acronym	channels
Raw Intensity	RI	1
Gradient Constraint	GC	3
Laplacian	LP	2
1^{st} order DF [36]	DF-1	4
2^{nd} order DF [36]	DF-2	10
Bit-Planes [38]	BP	8

4 Evaluation

We experiment with the different feature descriptors summarized in Table 1 on two direct tracking problems. One, is parametric motion estimation using an Affine motion model, which we use to illustrate performance on synthetically controlled illumination variations. The other, is direct visual odometry, which is more challenging as the nonlinear warping function depends on sparse depth.

4.1 Affine Template Alignment

Using the notation introduced in Section 2, we desire to estimate the parameters of motion between a dense descriptor designated as the template $\boldsymbol{\Phi}_{\text{DESC}}$ and a dense descriptor evaluated on an input image $\boldsymbol{\Phi}'_{\text{DESC}}$, where DESC is one of the descriptors in Table 1. Under affine motion $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_6)^\top \in \mathbb{R}^6$, the image coordinates of the two descriptors are related via an affine warp of the form

$$\begin{bmatrix} \mathbf{p}' \\ 1 \end{bmatrix} \equiv \mathtt{w}(\mathbf{p}; \boldsymbol{\theta}) = \mathbf{A}(\boldsymbol{\theta}) \begin{bmatrix} \mathbf{p} \\ 1 \end{bmatrix}, \tag{8}$$

where $\mathbf{A}(\boldsymbol{\theta}) \in \mathbb{R}^{2\times 3}$ represents a 2D affine transform.

Experiments in this section are performed using a set of natural images, where the parameters of the affine transformation are randomly generated to create a synthetically input/moving image with known ground truth.

Performance under ideal conditions While ideal imaging conditions are uncommon outside of controlled imaging applications, it is important to study the effect of any form of image deformation on the system's accuracy. Ideal conditions in this context indicate lack of appearance variations such that brightness constancy assumption is satisfied.

The question we answer in the following experiments is: *How does nonlinear deformations of the image (feature descriptors) affect estimation accuracy under ideal conditions?* Especially under the additional quantization effects caused by descriptors. The answer to question is shown in Fig. 2, where all descriptors are evaluated without additional illumination change.

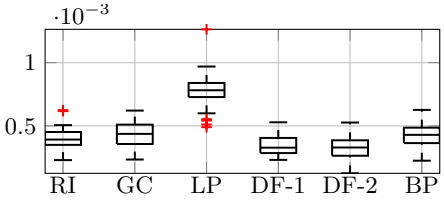


Fig. 2: Performance of dense descriptor under ideal conditions.

4.2 Performance under varying illumination

To generate illumination variations we synthesise the input image from the template using a nonlinear intensity change model of the form

$$\mathbf{I}'(\mathbf{p}) = \texttt{floor}\left(255\left(\frac{\alpha\mathbf{I}(\mathtt{w}(\mathbf{p};\boldsymbol{\theta}))+\beta}{255}\right)^{1+\gamma}\right), \tag{9}$$

where $\boldsymbol{\theta}$ is a randomly generated vector of warp parameters, α and β are respectively multiplicative and additive terms, while $|\gamma| < 1$ is a nonlinear gamma correction term. Parameters controlling the illumination changes are also generated randomly. An example of this illumination change is shown in Fig. 3.

Results are shown in Fig. 5 and Fig. 4 using the end-point RMSE metric [18]. As expected, we observe a large RMSE when using raw intensities. No significant improvement is obtained using the gradient constraint. The Laplacian improves results only slightly. The top performing algorithms are DF, and Bit-Planes.

Nonlocal descriptors Another natural question to ask is whether there is any benefit from using nonlocal feature descriptors in direct tracking? By nonlocal, we mean feature descriptors that make use of nonlocal spatial information in the image, such as a making use of a large neighborhood during the descriptor's computation. Examples include SIFT [32], HOG [33], and BRIEF [40]. As shown

Fig. 3: Example illumination change according to Eq. (9). Cost surfaces for each of the evaluated descriptors for these pair of images are shown in Fig. 4. We observe similar behavior for different image types and different changes in illumination.

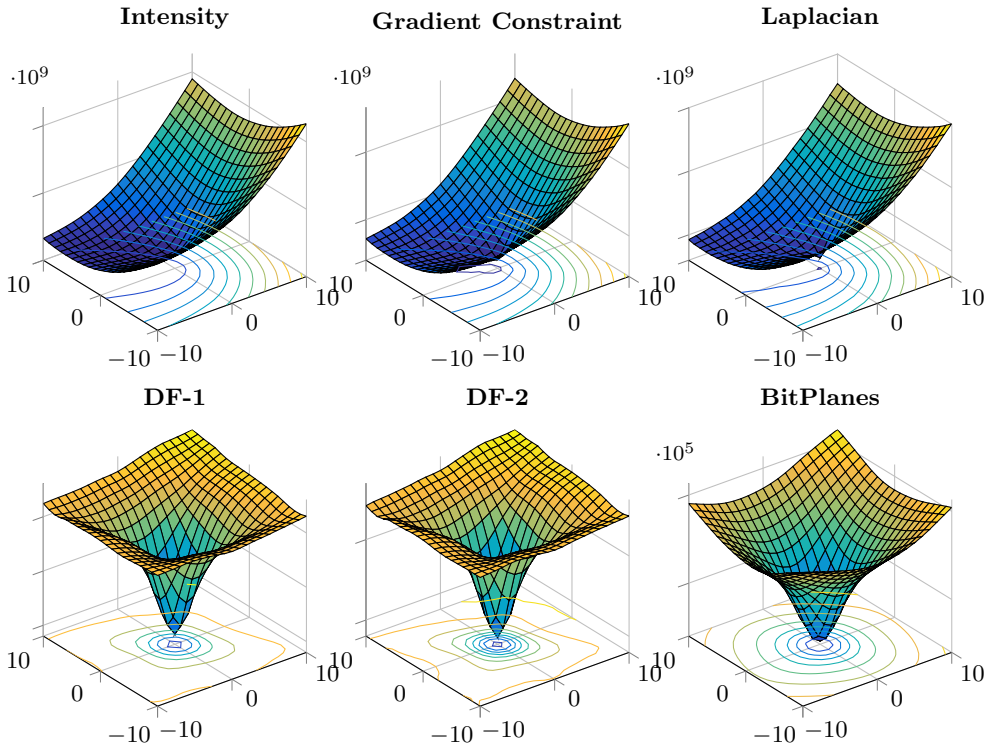


Fig. 4: Cost surfaces for each of the descriptors corresponding to the input pair shown in Fig. 3. The correct minima located at $(0, 0)$. Raw intensity and the gradient constraint fail to capture the correct minima. The Laplacian correctly localizes the minima, albeit a narrow basin of convergence. The feature descriptors at the bottom row correctly identify the minima with an adequate basin of convergence.

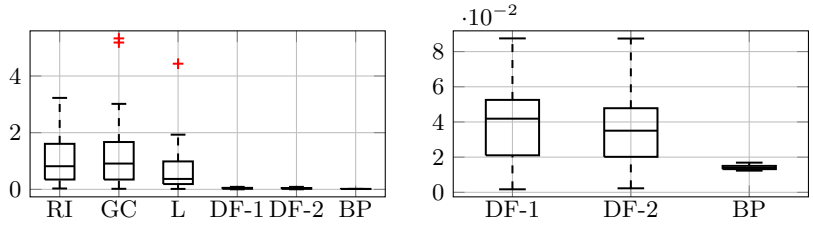
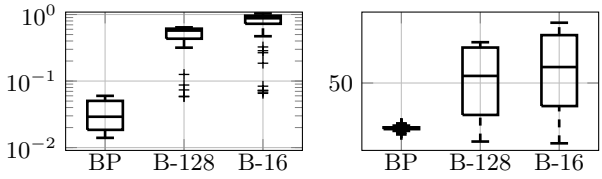


Fig. 5: Accuracy under illumination change. On the left RMSE is shown for all compared descriptors. On the right, we show only the top three for better comparison.

in Fig. 6, the use of nonlocal descriptor appears to hurt performance rather than simpler local ones. In this experiment, we experiment with two possibilities of extracting channels from the BRIEF [40] descriptor. One, is extracting 128 channels similar to [38]. The other, is extracting only 16 channels, where each channel is formed of a single byte. We observed similar degradation in performance using densely evaluated SIFT, and other variations on extracting channels.



(a) Error **(b)** Number of iterations

Fig. 6: Comparison with BRIEF using 128 channels (B-128) and 16 channels (B-16).

4.3 Direct Visual Odometry (VO)

Another popular application of direct tracking is estimating the 6DOF rigid-body motion of a freely moving camera. The warping function takes the form

$$\mathbf{p}' = \pi\left(\mathbf{T}(\boldsymbol{\theta})\pi^{-1}(\mathbf{D}(\mathbf{p}))\right), \tag{10}$$

where $\pi(\cdot) : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ denotes the projection onto a camera with a known intrinsic calibration, and $\pi^{-1}(\cdot,\cdot) : \mathbb{R}^2 \times \mathbb{R} \rightarrow \mathbb{R}^3$ denotes the inverse projection given the camera intrinsic parameters and the pixel's depth $\mathbf{D}(\mathbf{p})$. In our implementation, we parametrize the camera pose with the exponential map $\mathbf{T}(\boldsymbol{\theta}) = \exp(\widehat{\boldsymbol{\theta}}) \in SE(3)$ [41].

The algorithm is implemented in scale space (using 4 octaves) and the solution is obtained by Iteratively Re-Weighted Least-Squares (IRLS) using the

Huber robust weighing function [42]. The maximum number of iterations per pyramid octave is set to 50, but we terminate early if the norm of the estimated parameters vector, or the relative reduction of the objective, fall below a threshold $\tau = 1 \times 10^{-6}$.

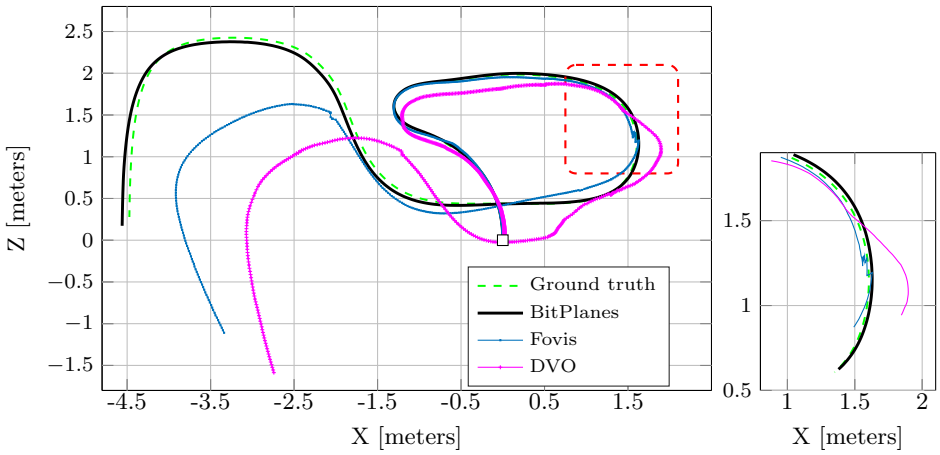
Image gradients required for linearization are implemented with central difference filters, which we found to produce more accurate results than Sobel, or Scharr filters. We also observed an improved accuracy if we selected a subset of pixels at the finest octave (the highest resolution). Pixel selection is implemented as a non maxima suppression of the absolute gradient magnitude across all descriptor channels. For the rest of the pyramid octaves, we use all pixels with non-vanishing gradient information.

The direct approach to VO has been shown to work with stereo [43], mono [3], and RGB-D data [19]. Since our focus is the study of photometric invariance rather than depth estimation, we will evaluate the approach on stereo data and simplify the step of depth estimation using standard stereo algorithms. We evaluate the performance of the different descriptors on the Tsukuba dataset [44,45], which provides a range of illumination as shown in Fig. 7.

Fig. 7: Different illumination condition from the Tsukuba dataset. At the top row from left to right we have: 'fluorescent' (easy), 'lamps' (medium) and 'flashlight' (hard). The bottom row shows the form of appearance change across views.

4.4 Comparison with state-of-the-art

We compare the top performing algorithm from the template tracking section (Bit-Planes) against two state-of-the-art VO algorithms: FOVIS [46], which is a feature-based algorithm, and DVO [1] which is a dense direct tracking method using raw pixel intensities. Qualitative results on the "lamps" dataset are shown in Fig. 8a and Fig. 8b. DVO using intensity only struggles to maintain tracking throughout the dataset. FOVIS's performance is slightly better as features are matched with normalized correlation, which is invariant to affine lighting change. The Bit-Planes based tracking performs the best. Quantitative evaluation for each of the descriptors per image frame is shown in Fig. 9 and summarized for the whole sequence in Table 2.



(a) Bird's eye view. **(b)** Detail.

Fig. 8: Evaluation on the Tsukuba dataset with "lamps" illumination [44]. The figure shows a bird's eye view of the camera path. The highlighted area is shown with more details in Fig. 8b. Example images are in Fig. 7

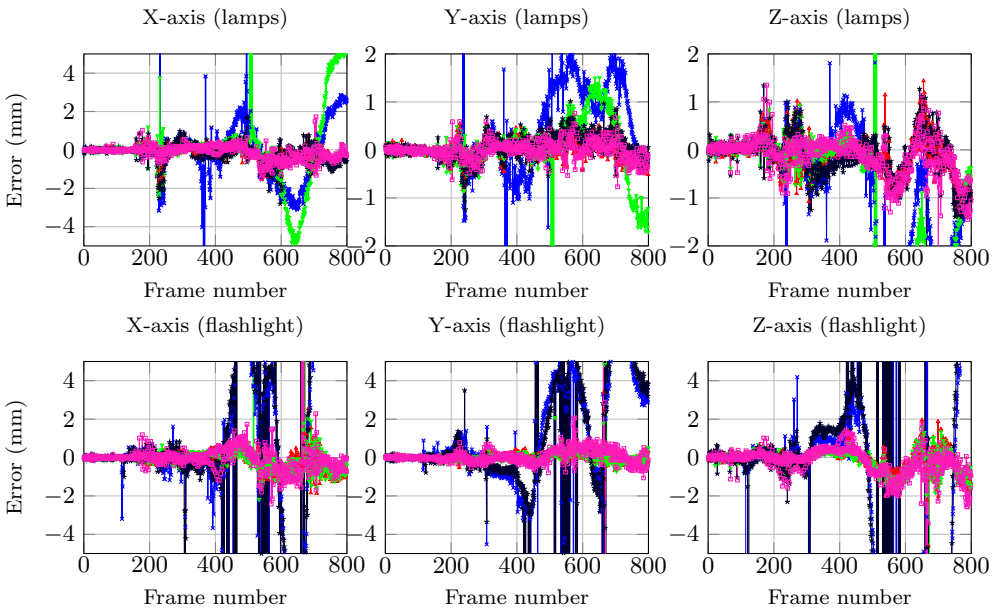


Fig. 9: Trajectory errors using the "lamps" sequence for (RI —•—), (LP —•—), (DF-1 —•—), (DF-2 —•—), and (BP —•—). We truncated the plots for better visualization.

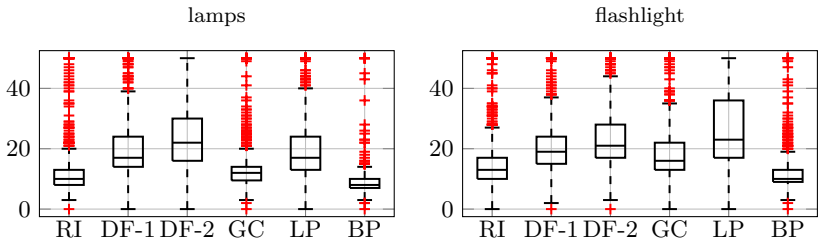


Fig. 10: Number of iterations.

Table 2: Summary statistics of errors per positional degree of freedom (RMSE in mm). We use the standard right-handed coordinate convention system in vision, where the Z-axis points forward and the Y-axis points downward.

	flahslight			lamps		
	X	Y	Z	X	Y	Z
RI	14.34	8.14	20.94	1.45	1.01	2.16
GC	14.26	7.53	18.86	45.31	30.37	22.76
LP	13.24	6.59	18.16	0.54	0.26	0.46
DF-1	**2.03**	0.45	**0.77**	0.42	0.21	**0.40**
DF-2	2.95	0.83	1.33	2.27	1.09	4.90
BP	2.66	**0.33**	1.08	**0.37**	**0.18**	**0.40**

4.5 Timing information

Experiments in the previous section were carried out using a combination of Matlab and C++. To obtain an accurate comparison of the run time of the different descriptors, we implemented an optimized version in C++. In Table 3 we show the time required to compute the descriptors as a function of image resolution. All experiments were conducted with a single core i7-2640M mobile processor and an 8GB of RAM. When using IC, computing descriptors is only required when creating a new reference frame. The frequency of creating new reference frames (re-initializing the tracker) depends on the application. Based on the photometric invariance performance from the previous evaluation, we conclude that Bit-Planes is the most efficient to compute.

Image warping, however, is required at every iteration of the optimization. Image warping depends on the number of channels as shown in Fig. 11. For instance, there is virtually no difference between warping raw intensities versus the Laplacian as they differ by a single channel and since the algorithm is memory bound. Warping experiments were parallelized on two cores.

Finally, the number of iterations required for convergence for each of the feature descriptor on the two Tsukuba datasets with challenging illumination is shown in Fig. 10. Direct tracking convergence withing 20 iterations during the majority of the time. The “flashlight” illumination is more challenging, and hence all descriptors require additional iterations to converge.

Table 3: Runtime in milliseconds required to compute the descriptors as a function of image resolution

	RI	GC	LP	DF-1	DF-2	BP
80×60	0.00	0.01	0.03	0.09	0.20	0.08
160×120	0.01	0.02	0.10	0.26	0.59	0.20
320×240	0.04	0.11	0.37	0.93	2.09	0.60
640×480	0.17	0.44	1.45	4.04	9.34	3.87
1280×960	0.76	2.56	5.84	17.20	38.65	14.32
2560×1920	2.88	10.15	23.22	68.78	154.29	59.29

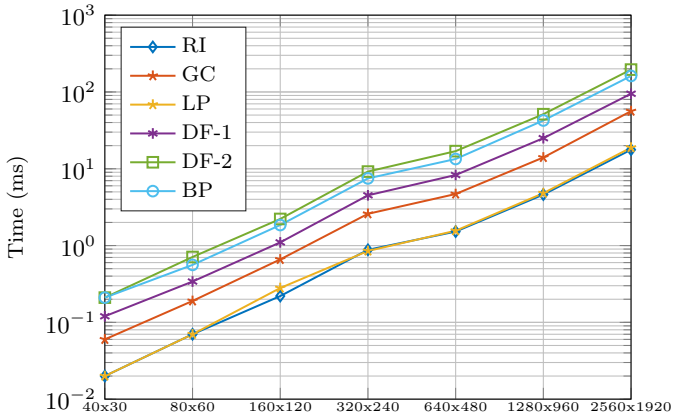


Fig. 11: Image warping running time shown in log scale.

5 Conclusions & Future Work

Locally evaluated dense feature descriptors are a promising avenue for non-parametric illumination invariant dense tracking. We explore various descriptors in comparison to using raw intensities and demonstrate enhanced robustness to arbitrary illumination change. More importantly, we show that local feature descriptor are suitable for the gradient-based minimization required by direct tracking. Our evaluation using two different direct tracking systems (affine template alignment, and direct visual odometry) show that the suitability of linearizing descriptors holds irrespective of the warp. Finally, the algorithmic changes required to allow existing direct tracking system to operate robustly in face of illumination variations are simple to implement without significantly comprising efficiency.

Acknowledgement. We thank the anonymous reviewers for their valuable comments.

References

1. Kerl, C., Sturm, J., Cremers, D.: Dense visual slam for RGB-D cameras. In: Int'l Conf. on Intelligent Robots and Systems. (2013)
2. Comport, A.I., Malis, E., Rives, P.: Real-time quadrifocal visual odometry. The International Journal of Robotics Research **29** (2010) 245–266
3. Engel, J., Schöps, T., Cremers, D.: LSD-SLAM: Large-Scale Direct Monocular SLAM. In: ECCV. (2014)
4. Handa, A., Newcombe, R.A., Angeli, A., Davison, A.J.: Real-Time Camera Tracking: When is High Frame-Rate Best? In: European Conf. on Computer Vision (ECCV). Volume 7578. (2012) 222–235
5. Salas-Moreno, R., Glocken, B., Kelly, P., Davison, A.: Dense planar SLAM. In: Mixed and Augmented Reality (ISMAR), 2014 IEEE International Symposium on. (2014) 157–164
6. Newcombe, R., Lovegrove, S., Davison, A.: DTAM: Dense tracking and mapping in real-time. In: Computer Vision (ICCV), 2011 IEEE International Conference on. (2011) 2320–2327
7. Irani, M., Anandan, P.: About Direct Methods. In: Vision Algorithms: Theory and Practice. (2000) 267–277
8. Forster, C., Pizzoli, M., Scaramuzza, D.: SVO: Fast Semi-Direct Monocular Visual Odometry. In: Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA). (2014)
9. Horn, B.K., Schunck, B.G.: Determining optical flow. Artificial intelligence **17** (1981) 185–203
10. Lucas, B.D., Kanade, T.: An Iterative Image Registration Technique with an Application to Stereo Vision (DARPA). In: Proc. of the 1981 DARPA Image Understanding Workshop. (1981) 121–130
11. Bartoli, A.: Groupwise geometric and photometric direct image registration. IEEE Trans. on Pattern Analysis and Machine Intelligence **30** (2008) 2098–2108
12. Evangelidis, G.D., Psarakis, E.Z.: Parametric image alignment using enhanced correlation coefficient maximization. PAMI **30** (2008)
13. Dowson, N., Bowden, R.: Mutual Information for Lucas-Kanade Tracking (MILK): An Inverse Compositional Formulation. PAMI **30** (2008) 180–185
14. Mller, T., Rabe, C., Rannacher, J., Franke, U., Mester, R.: Illumination-Robust Dense Optical Flow Using Census Signatures. In: Pattern Recognition. Volume 6835 of Lecture Notes in Computer Science. (2011) 236–245
15. Black, M., Anandan, P.: A framework for the robust estimation of optical flow. In: Computer Vision, 1993. Proceedings., Fourth International Conference on. (1993) 231–236
16. Irani, M., Anandan, P.: Robust multi-sensor image alignment. In: Computer Vision, 1998. Sixth International Conference on. (1998) 959–966
17. Nocedal, J., Wright, S.J.: Numerical Optimization. 2nd edn. Springer, New York (2006)
18. Baker, S., Matthews, I.: Lucas-kanade 20 years on: A unifying framework. International Journal of Computer Vision **56** (2004) 221–255
19. Klose, S., Heise, P., Knoll, A.: Efficient compositional approaches for real-time robust direct visual odometry from RGB-D data. In: IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems. (2013)
20. Zia, M.Z., Nardi, L., Jack, A., Vespa, E., Bodin, B., Kelly, P.H.J., Davison, A.J.: Comparative design space exploration of dense and semi-dense SLAM. CoRR **abs/1509.04648** (2015)

21. Sun, D., Roth, S., Black, M.: Secrets of optical flow estimation and their principles. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. (2010) 2432–2439
22. Vogel, C., Roth, S., Schindler, K.: An Evaluation of Data Costs for Optical Flow. In Weickert, J., Hein, M., Schiele, B., eds.: Pattern Recognition. Lecture Notes in Computer Science. Springer Berlin Heidelberg (2013)
23. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. Pattern Analysis and Machine Intelligence, IEEE Transactions on **27** (2005) 1615–1630
24. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. (2012) 1097–1105
25. Torr, P., Zisserman, A.: Feature Based Methods for Structure and Motion Estimation. In: Vision Algorithms: Theory and Practice. Springer Berlin Heidelberg (2000) 278–294
26. Furukawa, Y., Hernndez, C.: Multi-view stereo: A tutorial. Foundations and Trends in Computer Graphics and Vision **9** (2015) 1–148
27. Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: Proceedings of the International Conference on Computer Vision. Volume 2. (2003) 1470–1477
28. Antonakos, E., Alabort-i Medina, J., Tzimiropoulos, G., Zafeiriou, S.: Feature-Based Lucas-Kanade and Active Appearance Models. Image Processing, IEEE Transactions on **24** (2015) 2617–2632
29. Bristow, H., Lucey, S.: Regression-based image alignment for general object categories. CoRR **abs/1407.1957** (2014)
30. Sevilla-Lara, L., Sun, D., Learned-Miller, E.G., Black, M.J.: Optical Flow Estimation with Channel Constancy. In: Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I. Springer International Publishing, Cham (2014) 423–438
31. Sevilla-Lara, L., Learned-Miller, E.: Distribution fields for tracking. In: Computer Vision and Pattern Recognition (CVPR), IEEE Conference on. (2012)
32. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision **60** (2004) 91–110
33. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society Conference on. Volume 1. (2005) 886–893 vol. 1
34. Bristow, H., Lucey, S.: In Defense of Gradient-Based Alignment on Densely Sampled Sparse Features. In: Dense correspondences in computer vision. Springer (2014)
35. Liu, C., Yuen, J., Torralba, A.: SIFT Flow: Dense Correspondence across Scenes and Its Applications. IEEE Trans. Pattern Anal. Mach. Intell. **33** (2011) 978–994
36. Crivellaro, A., Lepetit, V.: Robust 3D Tracking with Descriptor Fields. In: Conference on Computer Vision and Pattern Recognition (CVPR). (2014)
37. Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High Accuracy Optical Flow Estimation Based on a Theory for Warping. In: ECCV. Volume 3024 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2004) 25–36
38. Alismail, H., Browning, B., Lucey, S.: Bit-Planes: Dense Subpixel Alignment of Binary Descriptors. CoRR **abs/1602.00307** (2016)
39. Zabih, R., Woodfill, J.: Non-parametric local transforms for computing visual correspondence. In: Computer Vision - ECCV'94. Springer (1994) 151–158

40. Calonder, M., Lepetit, V., Strecha, C., Fua, P.: BRIEF: Binary Robust Independent Elementary Features. In Daniilidis, K., Maragos, P., Paragios, N., eds.: Computer Vision ECCV 2010. Volume 6314 of Lecture Notes in Computer Science. (2010) 778–792
41. Murray, R.M., Li, Z., Sastry, S.S., Sastry, S.S.: A mathematical introduction to robotic manipulation. CRC press (1994)
42. Zhang, Z.: Parameter estimation techniques: A tutorial with application to conic fitting. Image and vision Computing **15** (1997)
43. Engel, J., Stueckler, J., Cremers, D.: Large-Scale Direct SLAM with Stereo Cameras. In: International Conference on Intelligent Robots and Systems (IROS). (2015)
44. Peris, M., Maki, A., Martull, S., Ohkawa, Y., Fukui, K.: Towards a simulation driven stereo vision system. In: Pattern Recognition (ICPR), 2012 21st International Conference on. (2012) 1038–1042
45. Martull, S., Peris, M., Fukui, K.: Realistic CG stereo image dataset with ground truth disparity maps. In: ICPR workshop TrakMark2012. Volume 111. (2012) 117–118
46. Huang, A.S., Bachrach, A., Henry, P., Krainin, M., Maturana, D., Fox, D., Roy, N.: Visual odometry and mapping for autonomous flight using an RGB-D camera. In: International Symposium on Robotics Research (ISRR). (2011) 1–16