Direct Visual Odometry in Low Light using Binary Descriptors

Hatem Alismail[1], Michael Kaess[1], Brett Browning[2], and Simon Lucey[1]

***Abstract*— Feature descriptors are powerful tools for photometrically and geometrically invariant image matching. To date, however, their use has been tied to sparse interest point detection, which is susceptible to noise under adverse imaging conditions. In this work, we propose to use binary feature descriptors in a direct tracking framework without relying on sparse interest points. This novel combination of feature descriptors and direct tracking is shown to achieve robust and efficient visual odometry with applications to poorly lit subterranean environments.**

I. Introduction

Visual Odometry (VO) is the problem of estimating the relative pose between two cameras sharing a common field-of-view. Due to its importance, VO has received much attention in the literature [1] as evident by the number of high quality systems available to the community [2], [3], [4]. Current systems using conventional cameras, however, are not equipped to tackle challenging illumination conditions, such as poorly-lit environments. In this work, we devise a novel combination of direct tracking using binary feature descriptors to allow robust and efficient vision-only pose estimation in challenging environments.

Current state-of-the-art algorithms rely on a *feature-based* pipeline [5], where keypoint correspondences are used to obtain an estimate of the camera motion (*e.g.* [6], [3], [7], [8], [9], [10], [11], [12]). Unfortunately, the performance of feature extraction and matching using conventional hardware struggles under challenging imaging conditions, such as motion blur, low light, and repetitive texture [13], [14] thereby reducing the robustness of the system. Examples of such environments include operating at night [13], mapping subterranean mines as shown in Fig. 1 and even sudden illumination changes due to automatic camera controls as shown in Fig. 2. If the feature-based pipeline fails, a vision-only system has little hope of recovery.

An alternative to the feature-based pipeline is to use pixel intensities directly, or what is commonly referred to as *direct methods* [15], [16], which has recently been popularized for RGB-D VO [17], [18], [19], [20], [21] and monocular SLAM from high frame-rate cameras [2], [4]. When the apparent image motion is small, direct methods deliver robust and precise estimates as many measurements could be used to estimate a few degrees-of-freedom [22], [23], [4], [24].

Nonetheless, as pointed out by other researchers [3], the main limitation of direct methods is their reliance on a consistent appearance between the matched pixels, otherwise known as the *brightness constancy* assumption [25], [26] requiring constant irradiance despite varying illumination, which is seldom satisfied in robotic applications.

Due to the complexity of real world illumination conditions, an efficient solution to the problem of appearance change for direct VO is challenging. The most common scheme to mitigating the effects of illumination change is to assume a parametric illumination model to be estimated alongside the camera pose, such as the gain+bias model [17], [27]. This approach is limited by definition and does not address the range of non-global and nonlinear intensity deformations commonly encountered in robotic applications. More sophisticated techniques have been proposed [24], [28], [29], but either impose stringent scene constraints (such as planarity), or heavily rely on dense depth estimates, which are not always available.

In this work, we relax the brightness consistency assumption required by most direct VO algorithms thus allowing them to operate in environments where the appearance between images vary considerably. We achieve this by combining the illumination invariance afforded by binary feature descriptors within a direct alignment framework. This is a challenging problem for two reasons: Firstly, binary illumination-invariant feature descriptors have not been shown to be well-suited for the iterative gradient-based optimization at the heart of direct methods. Secondly, binary descriptors must be matched under a binary-norm such as the Hamming distance, which is unsuitable for gradient-based optimization due to its non-differentiability.

To address these challenges, we propose a novel adaptation of binary descriptors that is experimentally shown to be amenable to gradient-based optimization. More importantly, the proposed adaption preserves the Hamming distance under conventional least-squares as we will show in Section III.

This novel combination of binary feature descriptors in a direct alignment framework is shown to work well in underground mines characterized by non-uniform and poor lighting. The approach is also efficient achieving real-time performance. An open-source implementation of the algorithm is freely available online https://github.com/halismai/bpvo.

II. Background

Direct Visual Odometry: Let the intensity, and depth of a pixel coordinate $\mathbf{p} = (x,\ y)^\top$ at the reference frame be respectively given by $\mathbf{I}(\mathbf{p}) \in \mathbb{R}$ and $\mathbf{D}(\mathbf{p}) \in \mathbb{R}^+$. Upon a rigid-body motion of the camera a new image is obtained

[1]Alismail, Kaess and Lucey are with the Robotics Institute, Carnegie Mellon University, Pittsburgh PA, USA {halismai,kaess,slucey}@cs.cmu.edu. [2]Browning is with Uber Advanced Technologies Center brettbrowning@gmail.com.

Fig. 1: Top row shows an example of commonly encountered low signal-to-noise ratio imagery from an underground mine captured with a conventional camera. The bottom row shows a histogram-equalized version emphasizing the poor quality and the significant motion blur.

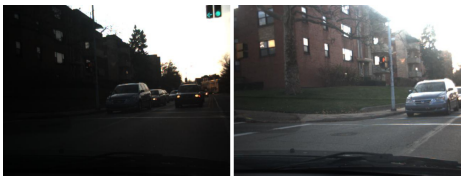
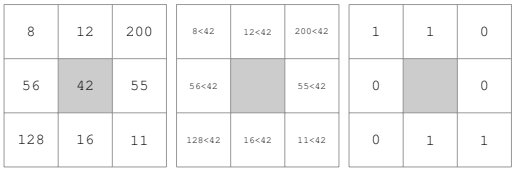


Fig. 2: An example of the nonlinear intensity deformation caused by the automatic camera settings. A common problem with outdoor applications of robot vision.



(a) **(b)** **(c)**

Fig. 3: Local intensity comparisons in a 3×3 neighborhood. In Fig. 3a the center pixel is highlighted and compared to its neighbors as shown in Fig. 3b. The descriptor is obtained by combining the result of each comparison in Fig. 3c into a single scalar [30], [31].

$\mathbf{I}'(\mathbf{p}')$. The goal of conventional direct VO is to estimate an increment of the camera motion parameters $\Delta\boldsymbol{\theta} \in \mathbb{R}^6$ such that the photometric error is minimized

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

where Ω is a subset of pixel coordinates of interest in the reference frame, $\mathbf{w}\left(\cdot\right)$ is a warping function that depends on the parameter vector we seek to estimate, and $\boldsymbol{\theta}$ is an initial estimate. After every iteration, the current estimate of parameters is updated additively. This is the well-known Lucas and Kanade algorithm [15].

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 [32]. Under the IC formulation we seek an update $\Delta\boldsymbol{\theta}$ that satisfies

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

The optimization problem in Eq. (2) is nonlinear irrespective of the form of the warping function, as in general there is no linear relationship between pixel coordinates and their intensities. By equating to zero the derivative of the first-order Taylor expansion of Eq. (2), we arrive at the solution given by the following closed-form (normal equations)

$$\Delta\boldsymbol{\theta} = \left(\mathbf{J}^\top\mathbf{J}\right)^{-1}\mathbf{J}^\top\mathbf{e}, \quad (3)$$

where $\mathbf{J} = \left(\mathbf{g}(\mathbf{p}_1)^\top, \ \ldots, \ \mathbf{g}(\mathbf{p}_m)^\top\right) \in \mathbb{R}^{m\times p}$ is the matrix of first-order partial derivatives of the objective function, or the Jacobian, m is the number of pixels, and $p = |\boldsymbol{\theta}|$ is the number of parameters. Each $\mathbf{g}$ is $\in \mathbb{R}^{1\times p}$ and is given by the chain rule as

$$\mathbf{g}(\mathbf{p})^\top = \nabla\mathbf{I}(\mathbf{p})\frac{\partial\mathbf{w}}{\partial\boldsymbol{\theta}}, \quad (4)$$

where $\nabla\mathbf{I} = (I_x, \ I_y) \in \mathbb{R}^{1\times 2}$ is the image gradient along the x- and y-directions respectively. Finally,

$$e(\mathbf{p}) = \mathbf{I}'(\mathbf{w}(\mathbf{p};\boldsymbol{\theta})) - \mathbf{I}(\mathbf{p}) \quad (5)$$

is the vector of residuals, or the *error image*. Parameters of the motion model are updated via the IC rule given by

$$\mathbf{w}\left(\mathbf{p},\boldsymbol{\theta}\right) \leftarrow \mathbf{w}\left(\mathbf{p},\boldsymbol{\theta}\right) \circ \mathbf{w}\left(\mathbf{p},\Delta\boldsymbol{\theta}\right)^{-1}. \quad (6)$$

We refer the reader to the comprehensive work by Baker and Matthews [32] for a detailed treatment.

Image Warping: Given a rigid body motion $\mathbf{T}(\boldsymbol{\theta}) \in SE(3)$ and a depth value $\mathbf{D}(\mathbf{p})$ in the coordinate frame of the template image, warping to the coordinates of the input image is performed according to:

$$\mathbf{p}' = \pi\left(\mathbf{T}(\boldsymbol{\theta})\pi^{-1}(\mathbf{p};\mathbf{D}(\mathbf{p})\right), \quad (7)$$

where $\pi\left(\cdot\right) : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ denotes the projection onto a camera with a known intrinsic calibration, and $\pi^{-1}\left(\cdot,\cdot\right) : \mathbb{R}^2 \times \mathbb{R} \rightarrow \mathbb{R}^3$ denotes the inverse of this projection given the intrinsic camera parameters and the pixel's depth. Finally, the intensity values corresponding to the warped input image $\mathbf{I}(\mathbf{p}')$ is obtained using bilinear interpolation.

III. Binary Descriptor Constancy

A limitation of direct method is the reliance on the brightness constancy assumption (Eq. (1)), which we address by using a *descriptor constancy* assumption instead. Namely, the parameter update is estimated to satisfy:

$$\Delta\boldsymbol{\theta}^* = \underset{\Delta\boldsymbol{\theta}}{\operatorname{argmin}} \left\| \boldsymbol{\phi}(\mathbf{I}'\left(\mathbf{w}(\mathbf{p};\boldsymbol{\theta}+\Delta\boldsymbol{\theta})\right)) - \boldsymbol{\phi}(\mathbf{I}\left(\mathbf{p}\right)) \right\|^2, \quad (8)$$

where $\phi(\cdot)$ is a robust feature descriptor. The idea of using descriptors in lieu of intensity has been recently explored in optical flow estimation [33], image-based tracking of a known 3D model [34], Active Appearance Models [35], and inter-object category alignment [36], in which results consistently outperform the minimization of the photometric error. To date, however, the idea has not been explored in the context of VO with relatively sparse depth and using binary features.

Prior work [35], [36] relied on sophisticated descriptor such as HOG [37] and SIFT [38]. However, using these descriptor densely in an iterative alignment framework is computationally infeasible for real-time VO with a limited computational budget. Simpler descriptors, such as photometrically normalized image patches [22] or the gradient-constraint [39] are efficient to compute, but do not possess sufficient invariance to radiometric changes in the wild. Furthermore, since reliable depth estimates from stereo are sparse, warping feature descriptors is challenging as it is harder to reason about visibility and occlusions from sparse 3D points.

In this work, we propose a novel adaption of binary descriptors that satisfies the requirements for efficient VO under challenging illumination. Namely, our descriptor has the following properties: (i) Invariance to monotonic changes in intensity, which is important as many robotic applications rely on automatic camera gain and exposure control. (ii) Computational efficiency, even on embedded devices, which is required for real-time VO, and (iii) Suitability for least-squares minimization (*e.g.* Eq. (8)). The last point is important for two reasons. One, solutions to least-squares problems are among the most computationally efficient with a plethora of ready to use software packages. The other, due to the small residual nature of least-squares, only first-order derivatives are required to obtain a good approximation of the Hessian. Hence, a least-squares formulation increases efficiency and avoids numerical errors associated with estimating second-order derivatives that arise when using alignment algorithms based on intrinsically robust objectives [40], [41], [42], [43], [44]. The proposed descriptor is called *Bit-Planes* and is detailed next.

The Bit-Planes Descriptor: The rationale behind binary descriptors is that using relative changes of intensities is more robust than working with the raw values. As with all binary descriptors, we perform local comparisons between the pixel and its immediate neighbors as shown in Fig. 3. We found that a 3×3 neighborhood is sufficient when working with video data and is the most efficient to compute. This step is identical to the Census Transform [31], also known as LBP [30]. Choice of the comparison operator is arbitrary and we will denote it with $\bowtie \in \{>, \geq, <, \leq\}$. Since the binary representation of the descriptor requires only eight comparisons, it is commonly compactly stored as a byte according to

$$\phi_{\text{BYTE}}(\mathbf{x}) = \sum_{i=1}^{8} 2^{i-1} \big[\mathbf{I}(\mathbf{x}) \bowtie \mathbf{I}(\mathbf{x} + \Delta\mathbf{x}_i)\big],$$

where $\{\Delta\mathbf{x}_i\}_{i=1}^{8}$ is the set of the eight displacements that are possible within a 3×3 neighborhood around the center pixel location $\mathbf{x}$.

In order for the descriptor to maintain its morphological invariance to intensity changes it must be matched under a binary norm, such as the Hamming distance, which counts the number of mismatched bits. The reason for this is easy to illustrate with an example. Consider two bit-strings differing at a single bit — which so happens to be at the most significant position — $\mathbf{a} = \{1, 0, 1, 0, 1, 1, 1, 0\}$, and $\mathbf{b} = \{0, 0, 1, 0, 1, 1, 1, 0\}$. The two bit-strings are clearly similar and their distance under the Hamming norm is one. However, if the decimal representation is used and matched under the squared Euclidean norm, their distance becomes $128^2 = 16384$, which does not capture their closeness in the descriptor space. However, it is not possible to use the Hamming distance in least-squares because of its non-differentiability. Approximations are possible using centralized sum of absolute difference [45], but at the cost of reduced photometric invariance.

In our proposed descriptor, we avoid the approximation of the Hamming distance and instead store each bit/coordinate of the descriptor as its own image, namely the proposed descriptor takes the form

$$\phi_{\text{BP}}(\mathbf{x}) = \begin{bmatrix} \mathbf{I}(\mathbf{x}) \bowtie \mathbf{I}(\mathbf{x} + \Delta\mathbf{x}_1) \\ \vdots \\ \mathbf{I}(\mathbf{x}) \bowtie \mathbf{I}(\mathbf{x} + \Delta\mathbf{x}_8) \end{bmatrix} \in \mathbb{R}^8 . \tag{9}$$

Since each coordinate of the 8-vector descriptor is binary, we call the descriptor "Bit-Planes." Using this representation it is now possible to minimize an equivalent form of the Hamming distance using ordinary least-squares.

Bit-Planes implementation details: In order to reduce the sensitivity of the descriptor to noise, the image is smoothed with a Gaussian filter in a 3×3 neighborhood ($\sigma = 0.5$). The effect of this smoothing will be investigated in Section V. Since the operations involved in extracting the descriptor are simple and data parallel, they can be done efficiently with SIMD (Single Instruction Multiple Data) instructions.

Pre-computing descriptors for efficiency: Descriptor constancy as stated in Eq. (8) 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 with sparse depth. An approximation to the descriptor constancy objective in Eq. (8) is to pre-compute the descriptors and minimize the following expression instead:

$$\min_{\Delta\boldsymbol{\theta}} \sum_{\mathbf{p} \in \Omega} \sum_{i=1}^{8} \|\boldsymbol{\Phi}'_i(\mathbf{w}(\mathbf{p}; \boldsymbol{\theta} + \Delta\boldsymbol{\theta})) - \boldsymbol{\Phi}_i(\mathbf{p})\|^2, \tag{10}$$

where $\boldsymbol{\Phi}_i$ indicates the i-th coordinate of the pre-computed descriptor. We found the loss of accuracy when using Eq. (10) instead of Eq. (8) to be insignificant in comparison to the computational savings.

IV. Direct VO using Binary Descriptors

We will use Eq. (10) as our objective function, which we minimize using the IC formulation [32], allowing us to pre-compute the Jacobian of the cost function. The Jacobian is given by

$$\sum_{\mathbf{p}\in\Omega}\sum_{i=1}^{8}\mathbf{g}_i(\mathbf{p};\boldsymbol{\theta})^\top\mathbf{g}_i(\mathbf{p};\boldsymbol{\theta}),\ \text{where} \tag{11}$$

$$\mathbf{g}_i(\mathbf{q};\boldsymbol{\theta}) = \frac{\partial\boldsymbol{\Phi}_i}{\partial\mathbf{p}}\Big|_{\mathbf{p}=\mathbf{q}}\frac{\partial\mathbf{w}}{\partial\boldsymbol{\theta}}\Big|_{\mathbf{p}=\mathbf{q},\boldsymbol{\theta}=\mathbf{0}}\ . \tag{12}$$

Similar to other direct VO algorithms [46], [17] pose parameters are represented using the exponential map, *i.e.* $\boldsymbol{\theta} = [\boldsymbol{\omega},\ \boldsymbol{\nu}]^\top \in \mathbb{R}^6$, such that

$$\mathbf{T}(\boldsymbol{\theta}) := \exp\left(\begin{bmatrix}[\boldsymbol{\omega}]_\times & \boldsymbol{\nu}\\ \mathbf{0}^\top & 0\end{bmatrix}\right) \in SE(3), \tag{13}$$

where $[\boldsymbol{\omega}]_\times$ indicates a 3×3 skew-symmetric matrix. To improve the computational running time of the algorithm, we subsample pixel locations for use in direct VO. A pixel is selected for the optimization if its absolute gradient magnitude is non-zero and is a strict local maxima in a 3×3 neighborhood. The intuition for this procedure is that pixels with a small gradient magnitude contribute little, if any, to the objective function as the term in Eq. (12) vanishes. We compute the pixel saliency map for all eight Bit-Planes coordinates as

$$\mathbf{G} = \sum_{i=1}^{8}\sum_{\mathbf{p}}\left(|\nabla_x\boldsymbol{\Phi}_i(\mathbf{p})| + |\nabla_y\boldsymbol{\Phi}_i(\mathbf{p})|\right). \tag{14}$$

Pixel selection is performed if the image resolution is at least 320×240. For lower resolution images (coarser pyramid levels) we use all pixels with non-zero saliency. The effect of pixel selection on the accuracy of pose estimation depends on the dataset as shown in [47, ch.4] and [48].

Minimizing the objective function is performed using an iteratively re-weighted Gauss-Newton algorithms with the Tukey bi-weight function [49]. The approach is implemented in a coarse-to-fine manner. The number of pyramid octaves is selected such that the smallest image dimension at the coarsest level is at least 40 pixels. Termination conditions are fixed to either a maximum number of iterations (100), or if the relative change in the estimated parameters, or the relative reduction of the objective, fall below 1×10^{-6}.

Finally, we implement a simple keyframing strategy to reduce drift accumulation over time. A keyframe is created if the magnitude of motion exceeds a threshold (data dependent), or if the percentage of "good points" falls below 60%. A point is deemed good if its weight from the M-Estimator is at the top 80-percentile. Points that project outside the image (*i.e.* no longer visible) are assigned zero weight.

V. Experiments & Results

Effect of smoothing: Fig. 4 shows the effect of smoothing the image prior to computing the descriptors. The experiment is performed on synthetic data with translational shifts. Larger kernels wash out image details required to estimate small motions, while no smoothing at all is noise-sensitive. Hence, we use a 3×3 kernel with $\sigma = 0.5$.

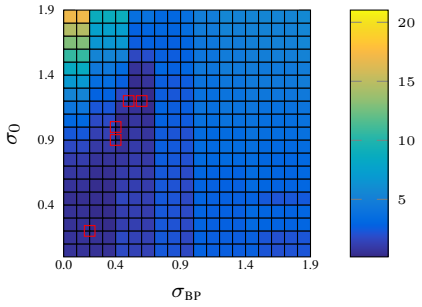


Fig. 4: Error as function of pre-smoothing the image with a Gaussian kernel of standard deviation of σ_0 as well smoothing the Bit-Planes with σ_1. The lowest error is associated with smaller kernels.

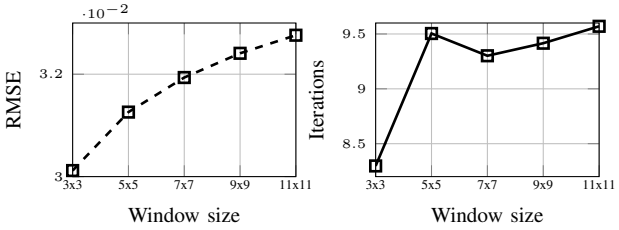


Fig. 5: Effect of window (neighborhood) size used to compute the binary descriptors on the final trajectory estimation accuracy and the mean number of iterations required for convergence.

Effect of binary descriptor window size: Our proposed binary descriptor is implemented using a window size of 3×3, which yields eight channels per pixel. It is possible, however, to use larger window sizes at the expense of increased runtime. In Fig. 5, we evaluate the effects of the window size on the accuracy of the estimated trajectory as well as the number of iterations required for convergence using the Tsukuba dataset. Results indicate that larger window sizes reduce the estimation accuracy as well as increase the mean number of iterations required for convergence.

Comparison with central image gradients: Our proposed binary descriptor can be thought of as thresholded directional image gradients as discussed by Hafner *et al.* [50], who study the robustness of the Census Transform using a continuous representation. In this section, we compare the performance of bit-planes to raw directional gradients. Each channel of the directional/central gradient per pixel in a 3×3 neighborhood is given by:

$$\Phi_{\text{CG}}(\mathbf{x}) = \begin{bmatrix}\mathbf{I}(\mathbf{x}) - \mathbf{I}(\mathbf{x}+\Delta\mathbf{x}_1)\\ \vdots \\ \mathbf{I}(\mathbf{x}) - \mathbf{I}(\mathbf{x}+\Delta\mathbf{x}_8)\end{bmatrix} \in \mathbb{R}^8\ . \tag{15}$$

In contrast to Bit-Planes (Eq. (9)), no thresholding step is performed and the output is a real-valued 8-vector.

We use the four different illuminations provided by the Tsukuba dataset to evaluate the effect of thresholding the

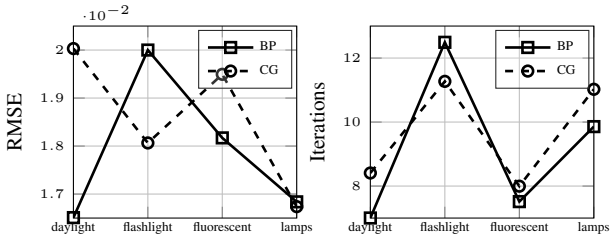


Fig. 6: Comparison between Bit-Planes (BP) and Central Gradients (CG).

TABLE I: Execution time for each major step in the algorithm reported in milliseconds (ms) using image size 640×480. Construction of the pyramid is a common to both raw intensity and Bit-Planes. Descriptor computation for raw intensity amounts to converting the image to floating point. Jacobian pre-computation is required only when creating a new keyframe. The most commonly performed operation is warping, which is not significantly more expensive than warping a single channel of raw intensity. Runtime on the KITTI benchmark with image size 1024×376 is shown in brackets.

	Raw Intensity	Bit-Planes
Pyramid construction	0.31 [0.44]	
Descriptor computation	0.18 [0.28]	4.33 [5.55]
Jacobian pre-computation	3.34 [5.00]	10.47 [13.93]
Descriptor warping	0.35 [0.30]	1.65 [1.74]

central gradients. Referring to Fig. 6, the Bit-Planes binary version is more accurate than raw un-thresholded central gradients except for the "flashlight" dataset. In addition to improved accuracy, the main advantage is a faster run-time and convergence.

Runtime: There are two steps to the algorithm. The first step is pre-computing the Jacobian of the cost function as well as the Bit-Planes descriptor for the reference image. This is required only when a new keyframe is created. We call this step **Jacobians**. The second step is repeated at every iteration and consists of: (i) image warping using the current estimate of pose, (ii) computing the Bit-Planes error, (iii) computing the residuals and estimating their standard deviations, and (iv) building the weighted linear system and solving it. We call this image **Linearization**. The running time for each step is summarized in Table I as a function of image resolution and in comparison to direct VO using raw intensities. A typical number of iterations for a run using stereo computed with block matching is shown in Fig. 7. The bottleneck in the linearization step is computing the median absolute deviation of the residuals, which could be mitigated using histograms [17]. Results are shown in comparison to our implementation of direct VO using raw intensities for a better assessment of the additional computational cost required for direct VO with a descriptor constancy. Finally, we note that due to the compactness of the proposed descriptor, it is possible to obtained additional speed ups using fixed-point arithmetic.

Experiments with synthetic data: We use the "New

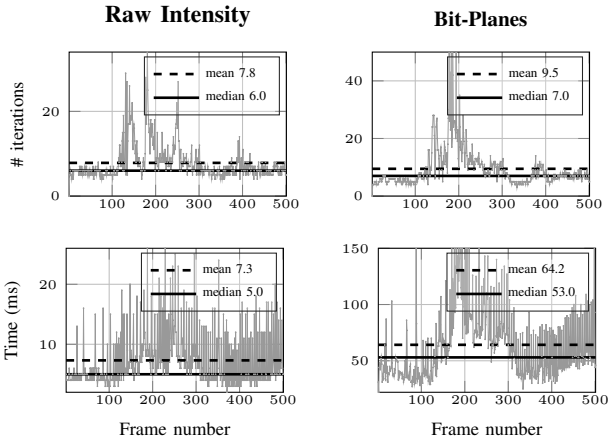


Fig. 7: Number of iterations and runtime on the first 500 frames of the New Tsukuba datasety. On average, the algorithm runs at more than 100 Hz using intensity and 15 Hz using Bit-Planes.

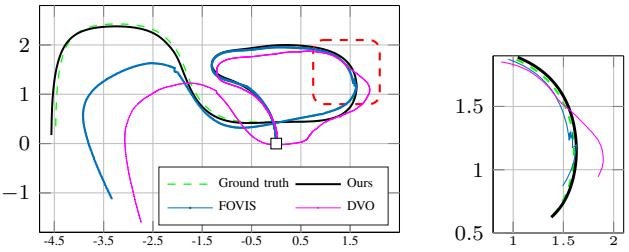


Fig. 8: Evaluation on the Tsukuba sequence [51] using the illumination provided by "lamps" in comparison to other VO algorithms shown in a bird's eye view. The highlighted area is shown with more details on the right. Example images are in shown Fig. 9.

Tsukuba" dataset [51] to compare the performance of our algorithm against two representative frame–frame VO algorithms from the state-of-the-art. The first is FOVIS [11], which we use as a representative of feature-based methods. FOVIS makes use of FAST corners [52] matched with a Zero-mean Normalized Cross Correlation (ZNCC). The second is DVO [46] as a representative of direct methods using the brightness constancy assumption and dense tracking. We use the most challenging illumination for our evaluation (shown in Fig. 9).

Our goal in this experiment is to assess the utility of our proposed descriptor in handling arbitrary changes in illumination. Hence, we initialize all algorithms with the ground truth disparity map. In this manner, any pose estimation errors are caused by failure to extract and/or match features, or failure in minimizing the photometric error under brightness constancy. As shown in Fig. 8 the robustness of our approach exceeds the conventional state-of-the-art methods. Also, as expected, feature-based methods (FOVIS) slightly outperforms direct methods (DVO) due to the challenging illumination conditions.

Evaluation on the KITTI benchmark: The KITTI benchmark [53] presents a challenging dataset for our algorithm,

Fig. 9: Example images from the "lamps" sequence.

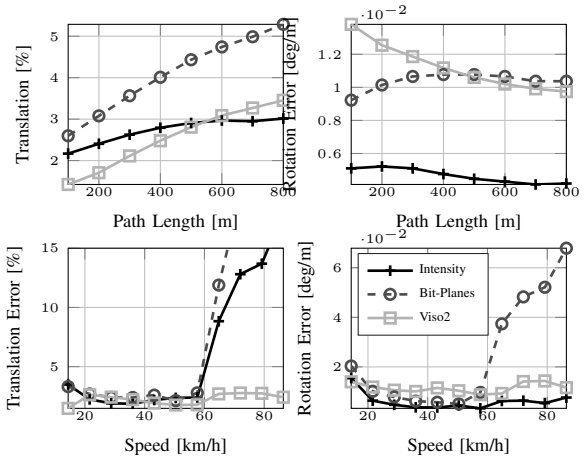


Fig. 10: Performance on the training data of the KITTI benchmark in comparison to VISO2 [12]. The large baseline between consecutive frames presents a challenge to direct methods as can be seen by observing the error as a function of speed.

and all direct methods in general, as the motion between consecutive frames is large. Our algorithm is initialized with disparity estimates obtained using block matching stereo as implemented in the OpenCV library. However, only the left image is used for tracking. Performance of the algorithm is compared against direct tracking using raw intensities and the feature-based algorithm VISO2 [12], which uses both stereo images for VO. Referring to Fig. 10, the algorithm's performance is slightly less accurate than raw intensities. The main limitation, however, is the narrower basin of convergence due to the relatively large camera motion.

We note that the reduced performance with larger camera motions is a limitation of direct methods as they rely on linearization, which assumes small inter-frame displacements. An evaluation of the effect of camera motion on the estimation accuracy is provided in the next section.

Basin of convergence: Direct methods are known to require small inter-frame displacements for the linearization assumption to be valid. In this section, we evaluate the basin of convergence as a function of camera displacements using the variable frame-rate dataset Handa *et al*. [54]. The dataset features the same scene imaged under different camera framerates with an accurate noise model.

Referring to Fig. 11, rotational and translational errors are generally lower at higher framerates. The sudden increase rotational errors at framerates in excess of 120Hz is due to the well known ambiguity of separating the effect of rotation from translation on the apparent flow [10], [55], which can be addressed by using a wider field-of-view lens [56].

Real data from underground mines: We demonstrate

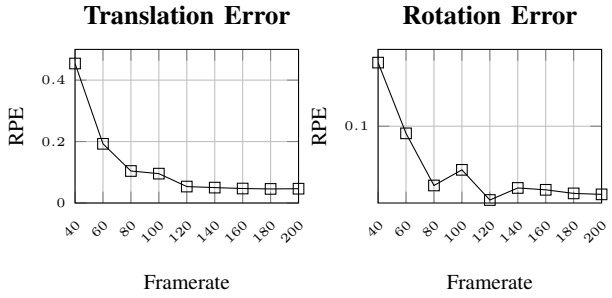


Fig. 11: Relative Pose Error (RPE) (defined in [57]) as a function of the camera frame-rate.

the robustness of our algorithm using data collected in underground mines. Our robot is equipped with a 7cm baseline stereo camera that outputs 1024×544 grayscale images and computes an estimate of disparity using a hardware implementation of SGM [58]. Due to lack of lighting in underground mines, the robot carries its own source of LED lights. However, the LEDs are insufficient to uniformly illuminate the scene due to power constraints. Similar to the previous experiments, disparities are used to initialize the scale and tracking is performed using only a single grayscale image. Examples of the 3D maps generated by our algorithm are shown in Figs. 12 to 14.

In Fig. 13, we show another result from a different underground environment where the stereo 3D points are colorized by height. The large empty areas in the generated map is due to lack of disparity estimates in large portions of the input images. Due to lack of ground-truth we are unable to assess the accuracy of the system quantitatively. But, visual inspection of the created 3D maps indicate minimal drift, which is expected when operating in an open loop fashion. More importantly, the algorithm maintains robustness with limited instances of failure cases as we show next. The performance of the algorithm is also illustrated using the supplementary materials video in comparison to other VO methods.

Failure cases: Most failure cases are due to a complete image wash out. An example is shown in Fig. 15. Theses cases occur when the robot is navigating tight turns where most of the LED power is concentrated very closely to the camera. Addressing such cases using vision-only is a good avenue of future work.

Reconstruction density: Density of the reconstructed point cloud on the tunnel data is shown in Figs. 13 and 14 and on a section of the KITTI data in Fig. 16. The reconstruction is obtained by transforming the selected 3D points into a consistent coordinate system using VO estimates. Denser output is possible by eliminating the pixel selection step and using all pixels with valid disparity estimates. However, denser reconstruction comes at the expense of increased runtime.

Additional evaluation of performance aspects pertaining to 2D parametric image registration problems is available in

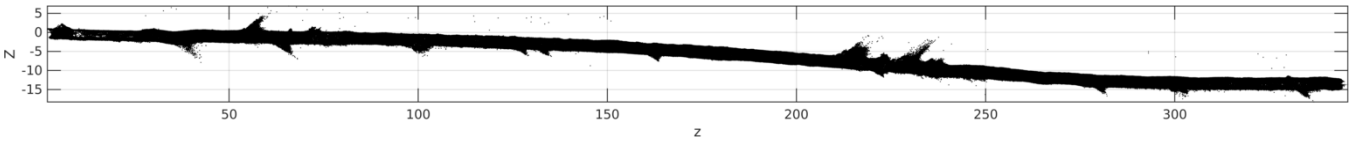


Fig. 12: Dense reconstruction of a long section of ≈ 400 meters from robust VO in a poorly lit underground environments.

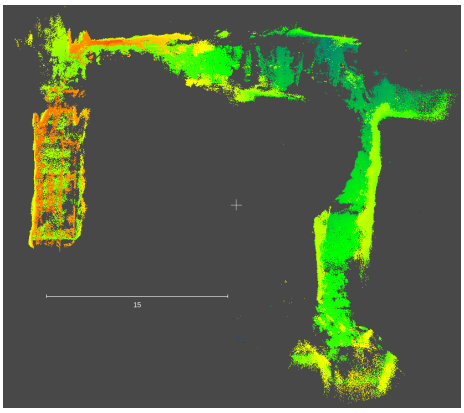


Fig. 13: VO map colorized by height showing the robot transitioning between different levels in the second mine dataset.

Fig. 14: VO map constructed while the robot is navigating a sharp corner. Points are colorized with the intensity values.

our prior work [59].

VI. Conclusions & Future Work

In this work, we presented a VO system capable of operating in challenging environments characterized by poor and non-uniform illumination. The approach is based on direct alignment of binary feature descriptors, where we presented a novel adaptation of the Census Transform, called Bit-Planes, suitable for least-squares optimization. The enhanced robustness as a result of using the binary descriptor constancy proposed in this work, while significant in comparison to the traditional brightness constancy, it requires smaller inter-frame displacements then other direct methods. We plan on addressing this limitation in a future extension.

All in all, by using a binary descriptor constancy, we allow vision-only pose estimation to operate robustly in environments that lack distinctive keypoints and lack the photometric consistency required by direct methods. The approach is simple to implement, and can be readily integrated into existing direct VSLAM algorithms with a small additional computational overhead.

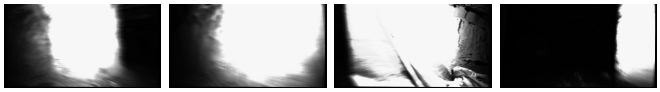


Fig. 15: Illustration of failure cases caused by over saturation.

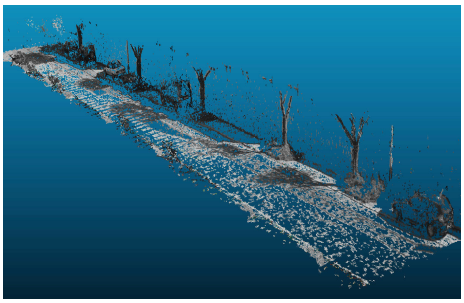


Fig. 16: Reconstruction density on a section of the KITTI dataset.

Acknowledgements

The authors would like to thank the anonymous Robotics and Automation Letters reviewers for their valuable feedback.

References

[1] D. Scaramuzza and F. Fraundorfer, "Visual Odometry [Tutorial]," *Robotics Automation Magazine, IEEE*, vol. 18, Dec 2011.
[2] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-Scale Direct Monocular SLAM," in *ECCV*, 2014.
[3] R. Mur-Artal, J. M. M. Montiel, and J. D. Tards, "Orb-slam: A versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, Oct 2015.
[4] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast Semi-Direct Monocular Visual Odometry," in *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2014.
[5] P. Torr and A. Zisserman, "Feature Based Methods for Structure and Motion Estimation," in *Vision Algorithms: Theory and Practice*. Springer Berlin Heidelberg, 2000, pp. 278–294.
[6] D. Nister, O. Naroditsky, and J. Bergen, "Visual odometry," in *Computer Vision and Pattern Recognition (CVPR)*, June 2004.
[7] H. Badino, A. Yamamoto, and T. Kanade, "Visual odometry by multi-frame feature integration," in *Computer Vision Workshops (ICCVW)*, 2013.
[8] M. Maimone, Y. Cheng, and L. Matthies, "Two years of visual odometry on the Mars Exploration Rovers," *Journal of Field Robotics, Special Issue on Space Robotics*, vol. 24, 2007.
[9] A. Howard, "Real-time stereo visual odometry for autonomous ground vehicles," in *IROS*, 2008.
[10] M. Kaess, K. Ni, and F. Dellaert, "Flow separation for fast and robust stereo odometry," in *IEEE Conf. on Robotics and Automation*, May 2009, pp. 3539–3544.
[11] A. S. Huang, A. Bachrach, P. Henry, M. Krainin, D. Maturana, D. Fox, and N. Roy, "Visual odometry and mapping for autonomous flight using an RGB-D camera," in *International Symposium on Robotics Research (ISRR)*, 2011.
[12] A. Geiger, J. Ziegler, and C. Stiller, "StereoScan: Dense 3d Reconstruction in Real-time," in *Intelligent Vehicles Symposium (IV)*, 2011.
[13] P. Nelson, W. Churchill, I. Posner, and P. Newman, "From dusk till dawn: localisation at night using artificial light sources," in *ICRA*. IEEE, 2015.

[14] M. Milford, E. Vig, W. Scheirer, and D. Cox, "Vision-based Simultaneous Localization and Mapping in Changing Outdoor Environments," *Journal of Field Robotics*, vol. 31, no. 5, 2014.

[15] B. D. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision (DARPA)," in *DARPA Image Understanding Workshop*, 1981.

[16] M. Irani and P. Anandan, "About Direct Methods," in *Vision Algorithms: Theory and Practice*, 2000, pp. 267–277.

[17] S. Klose, P. Heise, and A. Knoll, "Efficient compositional approaches for real-time robust direct visual odometry from RGB-D data," in *IROS*, 2013.

[18] C. Kerl, J. Sturm, and D. Cremers, "Dense visual slam for RGB-D cameras," in *IROS*, 2013.

[19] F. Steinbrucker, J. Sturm, and D. Cremers, "Real-time visual odometry from dense RGB-D images," in *ICCV Workshops*, 2011.

[20] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox, "RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments," *IJRR*, vol. 31, no. 5, 2012.

[21] T. Whelan, H. Johannsson, M. Kaess, J. Leonard, and J. McDonald, "Robust real-time visual odometry for dense RGB-D mapping," in *ICRA*, 2013.

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

[23] A. I. Comport, E. Malis, and P. Rives, "Real-time quadrifocal visual odometry," *The International Journal of Robotics Research*, vol. 29, no. 2-3, pp. 245–266, 2010.

[24] G. Silveira and E. Malis, "Real-time Visual Tracking under Arbitrary Illumination Changes," in *Computer Vision and Pattern Recognition*, June 2007, pp. 1–6.

[25] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial intelligence*, vol. 17, no. 1, pp. 185–203, 1981.

[26] M. A. Gennert and S. Negahdaripour, "Relaxing the brightness constancy assumption in computing optical flow," 1987.

[27] J. Engel, J. Stueckler, and D. Cremers, "Large-Scale Direct SLAM with Stereo Cameras," in *IROS*, 2015.

[28] Y. Lu and D. Song, "Robustness to lighting variations: An RGB-D indoor visual odometry using line segments," in *IROS*, 2015.

[29] R. Martins, E. Fernandez-Moral, and P. Rives, "Dense accurate urban mapping from spherical rgb-d images," in *IROS*, Sept 2015, pp. 6259–6264.

[30] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, pp. 51–59, 1996.

[31] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *Computer Vision - ECCV'94*. Springer, 1994, pp. 151–158.

[32] S. Baker and I. Matthews, "Lucas-kanade 20 years on: A unifying framework," *International Journal of Computer Vision*, vol. 56, no. 3, pp. 221–255, 2004.

[33] L. Sevilla-Lara, D. Sun, E. G. Learned-Miller, and M. J. Black, *Optical Flow Estimation with Channel Constancy*, 2014.

[34] A. Crivellaro and V. Lepetit, "Robust 3D Tracking with Descriptor Fields," in *CVPR*, 2014.

[35] E. Antonakos, J. Alabort-i Medina, G. Tzimiropoulos, and S. Zafeiriou, "Feature-Based Lucas-Kanade and Active Appearance Models," *Image Processing, IEEE Transactions on*, vol. 24, no. 9, 2015.

[36] H. Bristow and S. Lucey, "In Defense of Gradient-Based Alignment on Densely Sampled Sparse Features," in *Dense correspondences in computer vision*. Springer, 2014.

[37] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.

[38] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, 2004.

[39] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High Accuracy Optical Flow Estimation Based on a Theory for Warping," in *ECCV*, 2004, vol. 3024.

[40] N. Dowson and R. Bowden, "Mutual Information for Lucas-Kanade Tracking (MILK): An Inverse Compositional Formulation," *PAMI*, vol. 30, no. 1, pp. 180–185, Jan 2008.

[41] M. Irani and P. Anandan, "Robust multi-sensor image alignment," in *ICCV*, 1998, pp. 959–966.

[42] G. Pascoe, W. Maddern, and P. Newman, "Robust direct visual localisation using normalised information distance," 2015.

[43] G. Caron, A. Dame, and E. Marchand, "Direct model based visual tracking and pose estimation using mutual information," *Image and Vision Computing*, 2014.

[44] G. G. Scandaroli, M. Meilland, and R. Richa, "Improving NCC-Based Direct Visual Tracking," in *ECCV*, ser. Lecture Notes in Computer Science, 2012, vol. 7577, pp. 442–455.

[45] C. Vogel, S. Roth, and K. Schindler, "An Evaluation of Data Costs for Optical Flow," in *Pattern Recognition*, 2013.

[46] C. Kerl, J. Sturm, and D. Cremers, "Robust Odometry Estimation for RGB-D Cameras," in *ICRA*, 2013.

[47] H. Alismail, "Direct pose estimation and refinement," Ph.D. dissertation, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, August 2016.

[48] H. Alismail and B. Browning, "Direct Disparity Space: An Algorithm for Robust and Real-time Visual Odometry," Robots Institute, Tech. Rep. CMU-RI-TR-14-20, Oct 2014.

[49] Z. Zhang, "Parameter estimation techniques: A tutorial with application to conic fitting," *Image and vision Computing*, 1997.

[50] D. Hafner, O. Demetz, and J. Weickert, "Why Is the Census Transform Good for Robust Optic Flow Computation?" in *Scale Space and Variational Methods in Computer Vision*, 2013, vol. 7893.

[51] M. Peris, A. Maki, S. Martull, Y. Ohkawa, and K. Fukui, "Towards a simulation driven stereo vision system," in *Pattern Recognition, International Conference on*, Nov 2012.

[52] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *European conference on computer vision*. Springer, 2006, pp. 430–443.

[53] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite," in *CVPR*, 2012.

[54] A. Handa, R. A. Newcombe, A. Angeli, and A. J. Davison, "Real-Time Camera Tracking: When is High Frame-Rate Best?" in *European Conf. on Computer Vision (ECCV)*, 2012, vol. 7578, pp. 222–235.

[55] K. Dannilidis and H.-H. Nagel, "The coupling of rotation and translation in motion estimation of planar surfaces," in *Computer Vision and Pattern Recognition, 1993. Proceedings CVPR'93., 1993 IEEE Computer Society Conference on*. IEEE, 1993, pp. 188–193.

[56] Z. Zhang, H. Rebecq, C. Forster, and D. Scaramuzza, "Benefit of large field-of-view cameras for visual odometry," in *ICRA*, 2016, pp. 801–808.

[57] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, Oct. 2012.

[58] H. Hirschmuller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *CVPR*, 2005.

[59] H. Alismail, B. Browning, and S. Lucey, "Robust tracking in low light and sudden illumination changes," in *3D Vision-3DV 2016, 2016 International Conference on*. IEEE, 2016.