NID-SLAM: Robust Monocular SLAM using Normalised Information Distance

Geoffrey Pascoe, Will Maddern, Michael Tanner, Pedro Piniés and Paul Newman
Oxford Robotics Institute
University of Oxford, UK
{gmp,wm,mtanner,ppinies,pnewman}@robots.ox.ac.uk

Abstract

We propose a direct monocular SLAM algorithm based on the Normalised Information Distance (NID) metric. In contrast to current state-of-the-art direct methods based on photometric error minimisation, our information-theoretic NID metric provides robustness to appearance variation due to lighting, weather and structural changes in the scene. We demonstrate successful localisation and mapping across changes in lighting with a synthetic indoor scene, and across changes in weather (direct sun, rain, snow) using real-world data collected from a vehicle-mounted camera. Our approach runs in real-time on a consumer GPU using OpenGL, and provides comparable localisation accuracy to state-of-the-art photometric methods but significantly outperforms both direct and feature-based methods in robustness to appearance changes.

1. Introduction

Real-time monocular simultaneous localisation and mapping (SLAM) is a key technology enabling augmented and virtual reality applications [3]; 3D survey and reconstruction [7]; and robotics, in particular micro aerial vehicle (MAV) navigation [9]. Monocular SLAM approaches typically track a sparse set of visual features matched using descriptors that are robust to limited lighting, scale and viewpoint changes.

By using sparse key-point matching and efficient bundle adjustment, feature-based methods offer computational savings at the cost of reduced accuracy and robustness, since most of the information contained in each image is discarded [6]. Recent methods that directly minimise the photometric error between image and map are designed to address these limitations [32], providing increased accuracy, dense reconstructions and some robustness to viewpoint change and blur [7, 24]. However, the key limitation for these methods is the implicit assumption of *static scene illumination* required for the photometric error metric; this only holds in controlled indoor environments or over short periods of time outdoors. This severely limits applications of photometric visual SLAM methods, since maps can only be used in the lighting conditions in which they were generated.

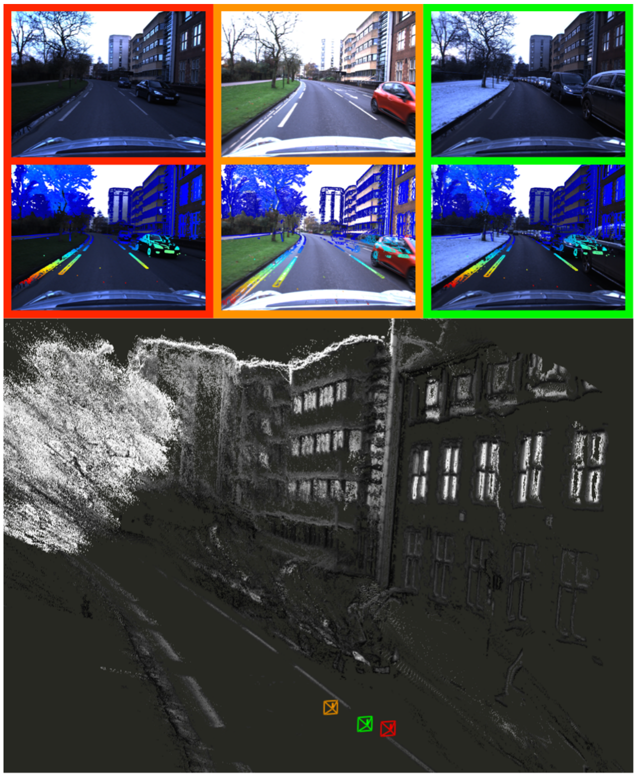

Figure 1. Robust monocular SLAM in changing conditions with NID-SLAM. After traversing an outdoor environment (red) and building key-frame depth maps, we are able to re-localise and refine the map under different lighting (orange) and weather (green) conditions using the robust Normalised Information Distance metric. Depth maps projected into images (center) show the alignment and camera frustrums (bottom) show the tracked $\mathfrak{sim}(3)$ poses.

In this paper we address the challenge of long-term visual SLAM in the presence of outdoor lighting, weather and structural scene changes. We present a monocular SLAM approach based on the Normalised Information Distance (NID) metric, dubbed NID-SLAM, and demonstrate robust localisation and mapping in the presence of appearance change over time as illustrated in Fig. 1. Unlike photometric error, the NID metric is not a function of the intensities of an image but instead a function of their *entropies*; hence, images collected under significantly different lighting, weather and seasonal conditions can be localised relative to a common map and used to update depth maps despite appearance changes. Using both synthetic and real-world data we demonstrate that NID-SLAM provides robustness exceeding that of state-of-the-art feature-based methods, while retaining accuracy comparable with state-of-the-art direct photometric methods. Finally, we present details of our real-time portable OpenGL implementation and address some limitations of the method in very challenging conditions.

1.1. Related Work

Most monocular SLAM approaches are so-called *indirect methods*; they rely on a feature-based front-end to determine sparse keypoints and descriptors (e.g. [17, 28]) to estimate camera pose using a filter-based [5, 13] or optimisation-based [14, 31, 22] back-end. However, indirect methods rely entirely on the feature detector to determine what parts of the image are useful for localisation (often ignoring edges and other geometry that provide useful cues [15]), as well as relying on the feature descriptor to provide robustness to changes in appearance due to scale, viewpoint and illumination [21]. In particular, feature descriptor matching is not robust to outdoor appearance changes caused by strong lighting variation, weather conditions and seasonal changes over longer time periods [12, 34].

Recently a number of *direct methods* have been proposed, which operate on pixel intensities without explicit feature extraction by minimising photometric error between a camera frame and a dense [32, 24] or semi-dense [7, 10] depth map. These methods claim to be more robust to viewpoint changes and motion blur and can offer higher tracking accuracy compared to indirect methods since the entire image is used. More recent results in [6] illustrate the advantages of explicit photometric calibration and exposure/gain compensation for accurate visual odometry (VO), however these methods still rely on the underlying static scene lighting assumptions inherent to photometric approaches. The recent direct VO method in [1] extends photometric error across a set of 'bit-planes' (similar to dense descriptor approaches [16]); this increases robustness to local lighting variations but does not address global changes in scene appearance over longer time periods.

An effective global metric for image alignment under changing conditions is mutual information (MI), often used to align images from multiple modalities [19, 36]. MI-based metrics have been used for camera pose tracking against prior maps [2, 27], and have demonstrated robustness to changing outdoor illumination conditions, structural change, blurred images, and occlusions over long time periods [37, 30, 26]. We believe our approach is the first to incorporate a robust whole-image MI metric into a monocular SLAM framework, providing both robust camera tracking and depth map updates in the presence of lighting, weather and structural scene changes over time.

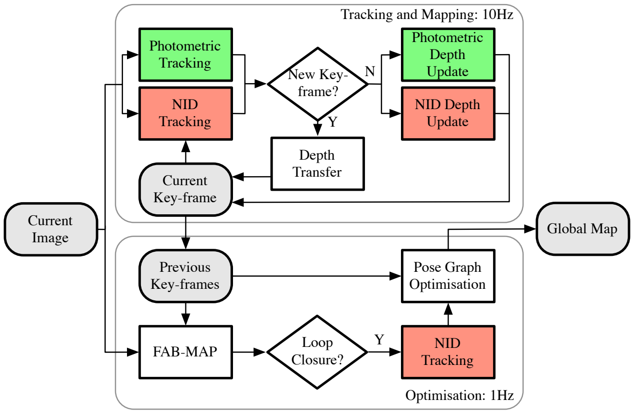

Figure 2. The NID-SLAM pipeline. The key components for photometric tracking and mapping (green) are augmented with robust NID-based methods (red). Loop closures are detected using FAB-MAP [4] and robust NID-based tracking is used to generate constraints to build a consistent global map. We perform tracking and mapping on the GPU at 10Hz using OpenGL, and loop closure detection and optimisation runs in parallel on the CPU at 1Hz.

1.2. Contributions

In this paper we present three novel contributions that form the key components of NID-SLAM as follows:

Robust direct tracking using NID: We present a real-time approach for minimising the NID between a candidate image and a key-frame depth map to recover the $\mathfrak{sim}(3)$ camera pose. In contrast to previous methods we explicitly incorporate depth uncertainty into the NID score.

Multi-resolution tracking using histogram pyramids: We present a novel histogram-pyramid approach for robust coarse-to-fine tracking using NID which increases robustness and the basin of convergence while reducing computation time at smaller scales.

Direct depth map refinement using NID: We present a per-pixel key-frame depth map refinement approach using NID, which allows for map maintenance and depth updates over successive traversals despite appearance changes over time.

2. Direct Monocular SLAM using NID

Fig. 2 provides an overview of the NID-SLAM system, highlighting the novel components in comparison to exist-

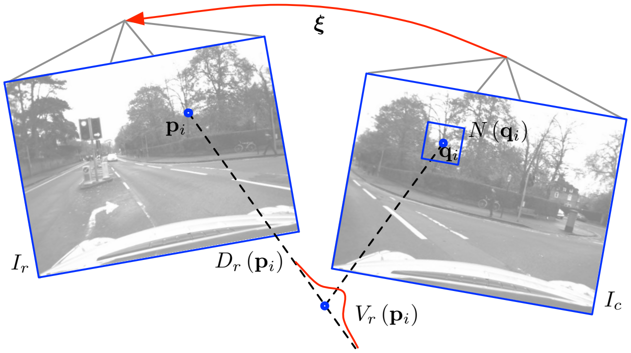

Figure 3. Tracking the current image I_c against a reference image I_r with associated inverse depth map D_r and variance V_r. A point in the reference image $\mathbf{p}_i$ is projected into the current image $\mathbf{q}_i$ using the warping function $\omega(\cdot)$ in Eq. 1, which depends on the relative pose $\boldsymbol{\xi} \in \mathfrak{sim}(3)$. For photometric tracking only the intensities $I_r(\mathbf{p}_i)$ and $I_c(\mathbf{q}_i)$ are needed; for NID tracking the neighbourhood $N(\mathbf{q}_i)$ around point $\mathbf{q}_i$ is also used.

ing photometric monocular SLAM approaches. In this section we detail the components of NID-SLAM, in particular NID-based tracking and depth update.

2.1. Robust Direct NID Tracking

For the following sections we adopt the key-frame and map representations of [7]: a key-frame consists of an image $I : \Omega \to \mathbb{R}^+$, inverse depth map $D : \Omega \to \mathbb{R}^+$ and inverse depth variance $V : \Omega \to \mathbb{R}^+$, where $\Omega \in \mathbb{R}^2$ are normalised pixel coordinates. We select a subdomain of each image $\Omega_D \in \Omega$, where Ω_D are all locations with sufficient gradient to provide meaningful depth estimates. We adopt the notation for a 3D projective warp function ω from [7], which transforms an image point $\mathbf{p}_i \in \Omega_D$ and associated reference inverse depth $D_r(\mathbf{p}_i) \in \mathbb{R}^+$ by the camera pose $\boldsymbol{\xi} \in \mathfrak{sim}(3)$ to yield the new camera frame point $\mathbf{q}_i \in \mathbb{R}^2$, as illustrated in Fig. 3:

$$\mathbf{q}_i = \omega\left(\mathbf{p}_i, D_r(\mathbf{p_i}), \boldsymbol{\xi}\right) \tag{1}$$

Photometric alignment of a current image I_c relative to a reference image I_r is typically performed by solving the following minimisation problem for the relative pose $\boldsymbol{\xi}$:

$$\underset{\boldsymbol{\xi}}{\arg\min} \sum_{\mathbf{p}_i \in \Omega_D} w_i(\boldsymbol{\xi}) \left\| \left(I_r(\mathbf{p}_i) - I_c(\mathbf{q}_i)\right)^2 \right\|_\delta \tag{2}$$

where the image sampling functions $I(\cdot) : \mathbb{R}^2 \to \mathbb{R}^+$ return the scalar intensity value at a subpixel location. The weighting function $w_i(\boldsymbol{\xi}) \in \mathbb{R}^+$ scales the residual based on depth uncertainty, and the robust kernel function $\|\cdot\|_\delta$ reduces the effect of outliers (e.g. Huber norm). However, the photometric error metric is inherently limited to environments where the appearance remains constant over time, which restricts applications to indoor environments with controlled lighting or short time periods outdoors. The more robust NID-based alignment metric is defined as follows:

$$\underset{\boldsymbol{\xi}}{\arg\min} \underset{\mathbf{p}_i \in \Omega_D}{\text{NID}} \left(I_r(\mathbf{p}_i), I_c(\mathbf{q}_i)\right) \tag{3}$$

Unlike mutual information, $\text{NID}(\cdot) : \mathbb{R}^{|\Omega_D|} \times \mathbb{R}^{|\Omega_D|} \to \mathbb{R}^+$ is a true metric bounded by $[0, 1]$ that satisfies the triangle inequality and does not depend on the total information content in a distribution [35]. The NID metric is defined as follows:

$$\text{NID}(I_r, I_c) = \frac{2\text{H}(I_r, I_c) - \text{H}(I_r) - \text{H}(I_c)}{\text{H}(I_r, I_c)} \tag{4}$$

where $\text{H}(I_r, I_c) \in \mathbb{R}^+$ is the joint entropy of the corresponding samples in images I_r and I_c, and $\text{H}(I_r) \in \mathbb{R}^+$ and $\text{H}(I_c) \in \mathbb{R}^+$ are the marginal entropies, defined as follows:

$$\text{H}(I_c) = -\sum_{a=1}^{n} p_c(a) \log\left(p_c(a)\right) \tag{5}$$

$$\text{H}(I_r, I_c) = -\sum_{a=1}^{n}\sum_{b=1}^{n} p_{r,c}(a, b) \log\left(p_{r,c}(a, b)\right) \tag{6}$$

where $\text{H}(I_r)$ is defined similarly to Eq. 5. The marginal $p_c \in \mathbb{R}^n$ and joint $p_{r,c} \in \mathbb{R}^{n \times n}$ distributions are represented by n-bin histograms where a and b are individual bin indices. Since both p_r and p_c can be obtained from $p_{r,c}$ by marginalisation, the primary computation in NID-SLAM is computing the joint distribution $p_{r,c}$ and its derivatives from the set of points $\mathbf{p} \in \Omega_D$ projected from the keyframe into the current image I_c.

We adopt a sampling approach to compute the joint distribution $p_{r,c}$ as illustrated in Fig. 4. In contrast to previous NID-based localisation approaches we explicitly incorporate depth map uncertainty into the pose estimate, in the form of the inverse depth variance $V_r(\mathbf{p}_i)$. The contribution from each sample $\mathbf{p}_i \in \Omega_D$ is added as follows:

$$p_{r,c}(a, b) \leftarrow p_{r,c}(a, b) + \frac{\beta\left(\mathbf{q}_i, N^{(j)}(\mathbf{q}_i)\right)}{kV_r(\mathbf{p}_i)} \tag{7}$$

Here $\beta\left(\mathbf{q}_i, N^{(j)}(\mathbf{q}_i)\right) \in \mathbb{R}^+$ represents a 2D cubic B-spline function that weights the contribution of pixel j in the 4×4 neighbourhood $N(\mathbf{q}_i)$ based on proximity to sub-pixel location $\mathbf{q}_i$. The weights are normalised such that $\sum^j \beta\left(\mathbf{q}_i, N^{(j)}(\mathbf{q}_i)\right) = 1, \ \forall \mathbf{q}_i$. Note that by sampling the j neighbouring pixels $N^{(j)}(\mathbf{q}_i)$ of $\mathbf{q}_i$, we never sample I_r or I_c at sub-pixel locations; in contrast to photometric methods, no interpolation between pixel intensities is required. The cubic B-spline results in a histogram function

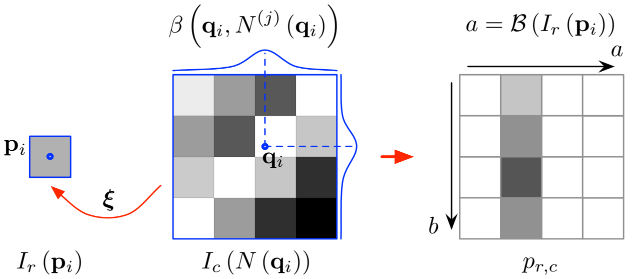

Figure 4. Contribution to the joint distribution $p_{r,c}$ from a single point $\mathbf{p}_i$. Using the relative pose $\boldsymbol{\xi}$ the point is projected into subpixel location $\mathbf{q}_i$ in the current image I_c, and the 4×4 neighbourhood $N(\mathbf{q}_i)$ around location $\mathbf{q}_i$ is retrieved. For each pixel j in the neighbourhood, the B-spline function $\beta(\cdot) \in \mathbb{R}^+$ weights the pairwise contribution to the joint based on proximity to $\mathbf{q}_i$ (shown as blue lines). The function $\mathcal{B}(\cdot) \in \mathbb{N}$ computes the bin index for each pairwise contribution based on the pixel intensity (here a 4-bin histogram is shown). Note that for the reference image I_r the histogram bin $a = \mathcal{B}(I_r(\mathbf{p}_i))$ is constant for all neighbourhood pixels j; hence *at most one column* of the joint $p_{r,c}$ will be updated for each point $\mathbf{p}_i$.

that is C^2 continuous, allowing for its use in a gradient-based optimisation framework. Histogram bin indices (a, b) are computed as follows:

$$a = \mathcal{B}(I_r(\mathbf{p}_i)), b = \mathcal{B}\left(I_c\left(N^{(j)}(\mathbf{q}_i)\right)\right) \tag{8}$$

where $\mathcal{B}(\cdot) : \mathbb{R}^+ \rightarrow \mathbb{N}$ returns the corresponding histogram bin for the intensity value provided by $I(\cdot)$. As the reference image bin index $a = \mathcal{B}(I_r(\mathbf{p}_i))$ is constant for all neighbourhood pixels j, one sample will update at most n bins in the joint $p_{r,c}$, illustrated in Fig. 4. Finally, the constant k normalises the contribution of an individual sample $\mathbf{p}_i$ as follows:

$$k = \frac{1}{|\Omega_D|} \sum_{\mathbf{p}_i \in \Omega_D} \frac{1}{V_r(\mathbf{p}_i)} \tag{9}$$

After computing $p_{r,c}$ (and therefore p_r and p_c), these can be substituted into Eq. 5 and 6 to compute the marginal and joint entropies, which are then substituted into 4 to produce the NID value. By differentiating Eq. 3 with respect to the relative pose parameters $\boldsymbol{\xi}$, we build an optimisation problem that seeks to minimise the NID between an image and key-frame by iteratively updating a relative pose estimate $\boldsymbol{\xi}_k$:

$$\boldsymbol{\xi}_{k+1} = \boldsymbol{\xi}_k - \alpha_k \boldsymbol{\Sigma}_k \left. \frac{\partial \text{NID}(I_r(\mathbf{p}_i), I_c(N(\mathbf{q}_i)))}{\partial \boldsymbol{\xi}} \right|_{\mathbf{p}_i \in \Omega_D} \tag{10}$$

where $\alpha_k \in \mathbb{R}^+$ is a step distance from a robust line search and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{6\times6}$ is an iteratively updated inverse Hessian or covariance approximation computed using the BFGS method [29], available at no extra cost after optimisation.

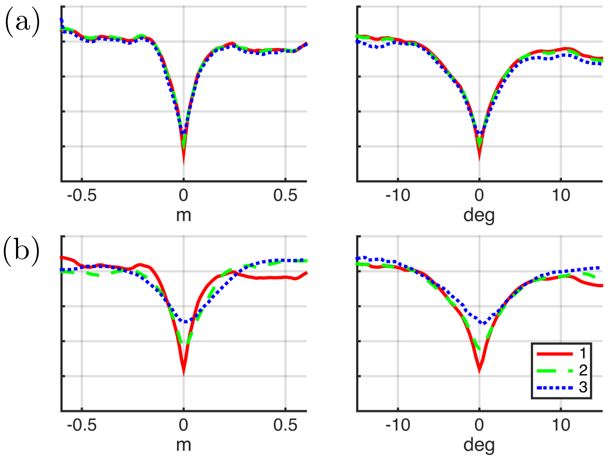

Figure 5. NID variation for 3 pyramid levels with (left) translational and (right) rotational offsets from the ground truth pose, averaged across 10 images from the New Tsukuba dataset. (a) Naïvely downscaling the input image fails to provide any benefit, but the multi-level histogram representation $\mathcal{H}^{(l)}$ (b) yields a smoother cost suface and wider convergence basin at higher pyramid levels, increasing robustness.

We have found NID-based tracking outperforms photometric tracking in both accuracy and robustness for all but the very first key-frame initialised with random depth values.

2.2. Multi-resolution NID Tracking

To increase robustness and the basin of convergence, many direct photometric methods use a coarse-to-fine image pyramid approach [24, 7]. We experimented with a naïve down-scaling approach but found that it did not improve robustness, as illustrated in Fig. 5. Instead, we propose a multi-resolution histogram representation where histograms are averaged instead of pixel intensities.

We build a pyramid of n-channel histogram images, denoted $\mathcal{H}^{(l)}$ for pyramid level l. Each channel a of the base level $\mathcal{H}^{(0)}$ is computed from the input image I as follows:

$$\mathcal{H}^{(0)}(\mathbf{p}_i, a) = \begin{cases} 1, & a = \mathcal{B}(I(\mathbf{p}_i)) \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

Successive levels are produced by down-sampling as follows:

$$\mathcal{H}^{(l+1)}(\mathbf{p}_i, a) = \frac{1}{4} \sum_{j=1}^{4} \mathcal{H}^{(l)}\left(N^{(j)}(2 \cdot \mathbf{p}_i), a\right) \tag{12}$$

where $N(\mathbf{p}_i)$ returns pixels within a 2×2 neighbourhood of $\mathbf{p}_i$. The histogram down-sampling process is illustrated in Fig. 6. The joint distribution update of Eq. 7 is replaced by the multi-resolution histogram form, illustrated in Fig. 7, as follows:

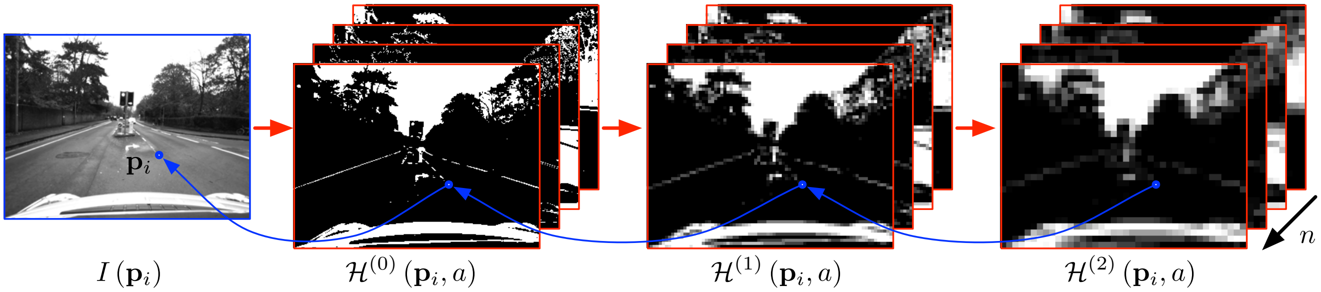
$\mathbf{p}_i$
n
$I\left(\mathbf{p}_i\right)$
$\mathcal{H}^{(0)}\left(\mathbf{p}_i, a\right)$
$\mathcal{H}^{(1)}\left(\mathbf{p}_i, a\right)$
$\mathcal{H}^{(2)}\left(\mathbf{p}_i, a\right)$

Figure 6. Multi-resolution histogram representation. The image I is first converted to a multi-channel binary image $\mathcal{H}^{(0)}$, where the index a selects among the n histogram bins. Each channel of $\mathcal{H}^{(0)}$ is successively down-sampled using 2×2 block averaging, with $\mathcal{H}^{(l)}$ representing the histogram down-sampled l times. By down-sampling the histogram representation instead of building histograms from the down-sampled image I_c, more information is retained at lower image resolutions.

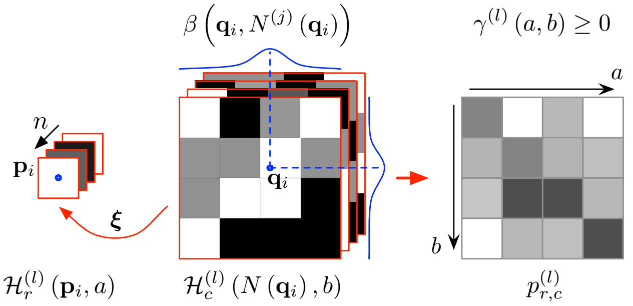
$\beta\left(\mathbf{q}_i, N^{(j)}\left(\mathbf{q}_i\right)\right)$
$\gamma^{(l)}\left(a, b\right) \geq 0$
n
$\mathbf{p}_i$
$\mathbf{q}_i$
$\boldsymbol{\xi}$
a
b
$\mathcal{H}_r^{(l)}\left(\mathbf{p}_i, a\right)$
$\mathcal{H}_c^{(l)}\left(N\left(\mathbf{q}_i\right), b\right)$
$p_{r,c}^{(l)}$

Figure 7. Contribution to the level-l joint distribution $p_{r,c}^{(l)}$ from a single point $\mathbf{p}_i$ using a multi-resolution n-bin histogram representation. Each sample from the reference histogram $\mathcal{H}_r^{(l)}\left(\mathbf{p}_i, a\right)$ and neighbourhood pixel j in current histogram $\mathcal{H}_c^{(l)}\left(N^{(j)}\left(\mathbf{q}_i\right), b\right)$ are indexed by (a, b) to determine the pairwise contribution. The pairwise histogram weighting function $\gamma^{(l)}$ defined in Eq. 14 can be non-zero for any bin index (a, b), and therefore unlike in Fig. 4, *all elements in the joint* may be updated for each point $\mathbf{p}_i$.

$$p_{r,c}^{(l)}\left(a, b\right) \leftarrow p_{r,c}^{(l)}\left(a, b\right) + \frac{\gamma^{(l)}\left(a, b\right) \beta\left(\mathbf{q}_i, N^{(j)}\left(\mathbf{q}_i\right)\right)}{k V_r^{(l)}\left(\mathbf{p}_i\right)} \tag{13}$$

where the pairwise histogram weighting function $\gamma^{(l)}\left(a, b\right) : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{R}^{+}$ is defined as follows:

$$\gamma^{(l)}\left(a, b\right) = \mathcal{H}_r^{(l)}\left(\mathbf{p}_i, a\right) \mathcal{H}_c^{(l)}\left(N^{(j)}\left(\mathbf{q}_i\right), b\right) \tag{14}$$

Importantly, the weighting function $\gamma^{(l)}$ can be non-zero for any combination of bin indices (a, b), and therefore Eq. 13 may update up to n^2 histogram bins in the joint distribution $p_{r,c}^{(l)}$ in contrast to n bins for Eq. 7. This increases computation by a constant factor for each level, but higher pyramid levels quadratically reduce the number of samples which results in an overall computational saving. We also observed an increase in robustness to poor initialisation which justifies the additional computational load as illustrated in Fig. 5. Coarse-to-fine tracking is performed by successively solving Eq. 10 for successive levels l of $p_{r,c}^{(l)}$, using the final pose from each level to initialise tracking for the next.

2.3. NID Depth Map Update

After solving for the estimated camera pose $\hat{\boldsymbol{\xi}}$, direct methods typically refine key-frame depth estimates using small-baseline stereo measurements from the current image I_c [7]. For photometric error this can be performed independently for each pixel $\mathbf{p}_i \in \Omega_D$ using efficient quadratic optimisation. However, when revisiting key-frames after long periods of time, lighting and appearance changes sufficiently to make local photometric depth updates impossible. We propose a global approach to key-frame depth updates using the NID metric to robustly maintain and improve depth estimates across appearance change.

For a camera pose $\boldsymbol{\xi}$ we compute the inverse depth gradient $\nabla_{D_r}\left(\boldsymbol{\xi}\right) \in \mathbb{R}^{|\Omega_D|}$ as follows:

$$\nabla_{D_r}\left(\boldsymbol{\xi}\right) = \left.\frac{\partial \mathrm{NID}\left(I_r\left(\mathbf{p}_i\right), I_c\left(N\left(\mathbf{q}_i\right)\right)\right)}{\partial D_r\left(\mathbf{p}_i\right)}\right|_{\boldsymbol{\xi},\, \mathbf{p}_i \in \Omega_D} \tag{15}$$

The new estimated reference inverse depth map $\hat{D}_r$ is then iteratively updated using the BFGS method, similar to Eq. 10:

$$\hat{D}_r\left(\mathbf{p}\right)_{k+1} = \hat{D}_r\left(\mathbf{p}\right)_{k} - \alpha_{D_k} \boldsymbol{\Sigma}_{D_k} \nabla_{D_r}\left(\boldsymbol{\xi}\right) \tag{16}$$

where $\alpha_{D_k} \in \mathbb{R}^{+}$ is a step distance and $\boldsymbol{\Sigma}_{D_k} \in \mathbb{R}^{|\Omega_D|} \times \mathbb{R}^{|\Omega_D|}$ is the estimated depth covariance after k iterations, which is typically sparse. After optimisation, the inverse depths D_r and inverse depth variances V_r are updated as follows:

$$D_r\left(\mathbf{p}\right) = \frac{\hat{D}_r\left(\mathbf{p}\right)_k \circ V_r\left(\mathbf{p}\right) + D_r\left(\mathbf{p}\right) \circ \mathrm{diag}\left(\boldsymbol{\Sigma}_{D_k}\right)}{V_r\left(\mathbf{p}\right) + \mathrm{diag}\left(\boldsymbol{\Sigma}_{D_k}\right)} \tag{17}$$

$$V_r\left(\mathbf{p}\right) = \left(V_r\left(\mathbf{p}\right)^{-1} + \mathrm{diag}\left(\boldsymbol{\Sigma}_{D_k}\right)^{-1}\right)^{-1} + \mathrm{diag}\left(\sigma_p^2 \mathbf{I}\right) \tag{18}$$

Parameter	Value
Num Histogram Bins (n)	16
Num Histogram Pyramid Levels (l)	3
Min Gradient for Depth	5
Min Frames per Key-frame	5
Min Key-frame Overlap	40%
BFGS Max Iterations Per Level	50
BFGS Max Line Search Iterations	20

Table 1. NID-SLAM Parameters

where $\circ$ is the Hadamard (element-wise) product and σ_p^2 is a process noise term similar to the update in [7] to ensure inverse depth variances do not become overconfident.

In practice we find that the NID depth update is sensitive to the depth initialisation and number of samples; it will fail to converge when initialising a key-frame with random depth values as in [7]. Hence we suggest performing NID depth map updates only when revisiting a previously-initialised key-frame after appearance change; photometric depth updates are currently more robust to poor depth estimates and therefore more effective when initialising new key-frames. Note, however, that we still use NID-based *tracking*, even on the first visit to a keyframe.

2.4. Pose Graph Optimisation

To build a consistent global map from interconnected key-frames we adopt the scale-aware pose graph optimisation approach in [7]. We use FAB-MAP [4] to provide loop closure candidates from the current image to previous key-frames, then perform a multi-resolution NID relative pose optimisation to determine the loop closure constraint. These constraints are optimised with the the key-frame poses to form a global map. Our direct NID tracking is particularly effective at building loop closure constraints in outdoor environments over long time periods, since appearance change is inevitable in the presence of sunlight, shadows and occluding objects.

3. Results

We compare the performance of NID-SLAM against two other state-of-the-art monocular SLAM approaches, ORB-SLAM2 [23] and LSD-SLAM [7]. We perform an evaluation using two different datasets, the synthetic indoor New Tsukuba Dataset [20] and sections of the outdoor Oxford RobotCar Dataset [18], illustrated in Fig. 8. Unlike the well-known KITTI dataset [11], the Oxford RobotCar Dataset provides many traversals of the same route at different times. Each dataset includes multiple traversals of the same route under different illumination or weather conditions; the New Tsukuba dataset additionally provides ground-truth trajectory and depth map data for accuracy evaluation. We selected six 500m traversals of the same location in the Oxford RobotCar Dataset to represent outdoor operation in changing conditions.

We performed a total of 16 experiments for the indoor datasets and 36 for the outdoor datasets. Each experiment involved two sequential traversals from two different conditions, where the goal is to successfully localise and track the second traversal against the SLAM map built during the first. To evaluate tracking performance we manually set the first active keyframe for the start of the second traversal to the first keyframe in the map from the first traversal (so that we do not rely on loop closure for the initial tracking). We report the success rate for the second traversal as the percentage of frames successfully tracked against keyframes generated from the first traversal.

Our implementation of NID-SLAM uses OpenGL Shading Language[1] (GLSL) for portability between different platforms; we achieve 10Hz tracking and mapping updates using a single desktop AMD R9 295x2 GPU. For ORB-SLAM2 and LSD-SLAM we use the open source implementations available[2,3], modified to support active keyframe initialisation for multi-session mapping. We also implemented the exposure compensation for LSD-SLAM in [8] to improve performance in outdoor environments. Table 1 lists the parameters used in this evaluation.

3.1. Robust Indoor Tracking

For the indoor New Tsukuba dataset, we use the ground-truth poses to scale the $\mathfrak{sim}(3)$ key-frame transforms to $\mathbb{SE}(3)$ transforms, and report trajectory errors in metric units. Table 2 presents the RMS translational and rotational errors for each method along with the localisation success rate, where a localisation failure is either self-reported by each algorithm (e.g. failed tracking convergence) or a true absolute error of more than 0.5m.

NID-SLAM provides the most reliable tracking estimates for all but two of the experiments where ORB-SLAM provides fractionally higher success rates (e.g. 100% vs 99.3%). Crucially, NID-SLAM typically exceeds 80% success when tracking against a map built under different conditions (exceeding 95% for well-lit traversals), where LSD-SLAM never exceeds 50% and ORB-SLAM varies widely from less than 10% to over 80%. Apart from occasional outliers all three methods provide RMS errors consistently below 100mm and 5°. All three methods failed to localise using the Flashlight traversal against a map built using the Lamp traversal; we attribute this to the low intensity non-uniform illumination in the combination of these two traversals.

[1] `https://www.opengl.org/sdk/docs/man4/`
[2] `https://github.com/raulmur/ORB_SLAM2`
[3] `https://github.com/tum-vision/lsd_slam`

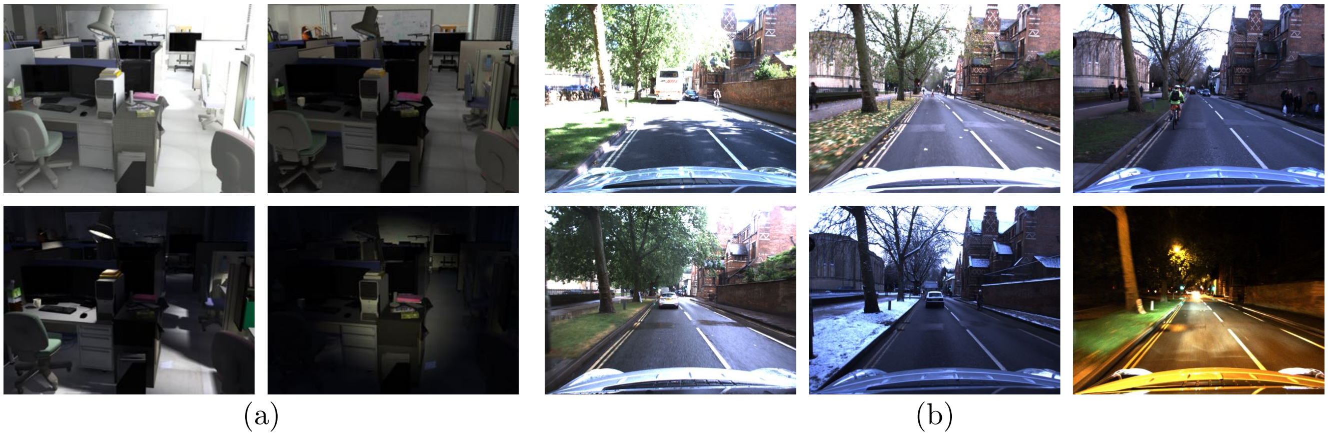

Figure 8. Example images from the indoor New Tsukuba and outdoor Oxford RobotCar datasets used for evaluation. Clockwise from top left: (a) an office environment under daylight, fluorescent, flashlight and lamp illumination; (b) an urban environment in sunny, overcast, dusk, night, snow and rain conditions. The datasets were chosen to provide a range of challenging conditions for monocular SLAM.

Traversal 1 / Traversal 2		Daylight			Fluorescent			Lamps			Flashlight		
		%	RMSE (mm)	RMSE (°)	%	RMSE (mm)	RMSE (°)	%	RMSE (mm)	RMSE (°)	%	RMSE (mm)	RMSE (°)
Daylight	NID	99.3	4.7	0.22	**96.7**	7.8	0.58	**73.9**	67.9	14.24	**74.6**	64.2	2.25
	LSD	96.1	36.4	1.93	48.6	59.8	2.93	26.1	33.2	1.53	25.5	89.5	4.40
	ORB	**100.0**	25.5	1.20	81.4	21.4	0.20	9.5	20.9	0.91	0.1	12.8	0.48
Fluorescent	NID	**95.0**	12.0	1.05	99.7	8.5	0.41	**85.3**	59.4	5.39	**95.8**	27.9	0.80
	LSD	55.5	52.6	2.44	96.5	14.7	0.53	38.6	76.5	3.67	29.7	70.0	3.80
	ORB	85.4	18.4	0.26	**99.7**	15.9	0.64	7.1	291.1	1.51	10.7	125.2	3.39
Lamps	NID	**88.3**	33.8	1.39	**93.6**	25.2	0.95	**93.1**	19.6	0.73	**84.3**	72.8	3.98
	LSD	6.6	90.2	5.01	46.8	93.2	4.30	71.7	8.0	0.05	11.9	72.3	2.80
	ORB	35.4	21.4	0.65	24.6	27.8	0.57	83.1	18.5	0.62	1.6	24.4	0.25
Flashlight	NID	**23.8**	41.2	0.88	**92.2**	38.5	1.72	0.00	N/A	N/A	**92.0**	24.6	1.17
	LSD	15.7	91.0	4.54	27.2	88.4	4.91	0.00	N/A	N/A	88.7	36.4	1.93
	ORB	19.4	17.2	0.29	30.0	24.7	0.25	0.00	N/A	N/A	22.7	50.8	1.28

Table 2. Indoor tracking results on the New Tsukuba dataset.

3.2. Depth Map Refinement

To evaluate the NID depth map update approach in Section 2.3 we compare the depth map errors before and after the second traversal, computed using the ground truth depth maps provided in the New Tsukuba dataset. Table 3 presents the depth map errors for the indoor evaluation. The leading diagonal lists the median depth error for the first traversal for each condition, while the off-diagonal elements list the depth errors after a second traversal in a different condition.

The NID depth map update successfully reduces depth errors by up to 6% on the second traversal for all of the Daylight, Fluorescent and Lamp conditions. However, traversals involving the Flashlight condition provide between 15% reduced errors and 15% *increased* errors; we attribute this to the highly dynamic illumination conditions during the traversal (since all light in the scene is emitted from the point of view of the camera). Since only a small portion of the scene is lit at a time, there are fewer samples available to update the histogram required for reliable NID-based depth updates.

1 / 2	Daylight	Fluorescent	Lamps	Flashlight
Daylight	60.9	65.2 (-4.25%)	56.5 (-5.99%)	58.0 (-2.84%)
Fluorescent	60.8 (-0.16%)	68.1	57.7 (-3.99%)	67.3 (+12.73%)
Lamps	59.9 (-1.64%)	64.6 (-5.14%)	60.1	67.9 (+13.74%)
Flashlight	52.3 (-14.12%)	68.1 (+0%)	67.2 (+11.8%)	59.7

Table 3. Depth map refinement on the New Tsukuba dataset. All measurements are median depth map errors in units of millimetres; percentages in brackets show the change in errors after the second traversal.

3.3. Robust Outdoor Tracking

For the outdoor RobotCar dataset, no metric ground truth is provided between traversals, but stereo visual odometry for each traversal is available. We generated an approximate key-frame correspondence between datasets based on accumulated metric distance along the route, and classify localisation failure as either self-reported (as above) or more than

2 \ 1	Overcast			Dusk			Snow			Sun			Rain			Night		
	NID	LSD	ORB	NID	LSD	ORB	NID	LSD	ORB	NID	LSD	ORB	NID	LSD	ORB	NID	LSD	ORB
Overcast	**100**	76.28	**100**	**96.6**	0.3	42.6	**85.3**	0.3	12.9	**7.9**	0.0	0.1	**4.3**	0.0	0.0	0.0	0.0	0.0
Dusk	88.2	10.68	**100**	**99.3**	64.1	47.8	**21.8**	18.5	15.2	**21.9**	1.72	0.1	**4.2**	0.37	0.0	0.0	0.0	0.0
Snow	**97.3**	0.0	0.1	**91.7**	4.94	57.8	**99.9**	40.5	45.7	0.1	**0.8**	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Sun	**10.1**	0.2	0.1	0.0	0.0	**0.1**	**11.4**	0.0	0.1	98.7	9.27	**100**	0.0	0.0	0.0	0.0	0.0	0.0
Rain	**6.4**	0.8	0.1	**4.2**	0.0	0.1	0.0	**1.2**	0.1	0.0	**0.1**	**0.1**	**94.1**	5.0	0.0	0.0	0.0	0.0
Night	**1.9**	0.1	0.1	0.1	**0.2**	0.1	**5.7**	0.0	0.0	0.0	0.0	**0.1**	**1.0**	0.1	0.0	0.0	0.0	0.0

Table 4. Outdoor tracking success rates using the Oxford RobotCar Dataset.

3 keyframes from the expected active keyframe (approximately 10m absolute error). Table 4 presents the localisation success rate for each of the outdoor experiments.

The outdoor traversals were significantly more challenging than the indoor experiments; only NID-SLAM successfully generated a map on the first traversal for the first five conditions. Despite the addition of exposure compensation, LSD-SLAM only succeeded in mapping short sections of the Overcast, Dusk and Snow traversals, and ORB-SLAM failed to produce a map for the Rain traversal due to the blurring effects of raindrops on the camera lens. None of the methods could successfully produce SLAM maps using the night traversal.

NID-SLAM again provides the most reliable tracking estimates, with localisation success rates exceeding 80% for all combinations of the first three traversals apart from Dusk against a Snow map, however it struggled to provide better than 10% success for the more challenging traversals (Sun, Rain and Night). ORB-SLAM provided impressive 100% success rates for a small number of traversals, but less than 0.1% success for many of the more challenging conditions. Unsurprisingly, LSD-SLAM provided the least reliable tracking estimates, since the appearance change between traversals (and even between frames in the same traversal) strongly violates the static scene illumination assumption required for photometric error minimisation.

3.4. Limitations

NID-SLAM provides robust and accurate tracking and mapping in the presence of appearance changes in comparison to both ORB-SLAM and LSD-SLAM; however as illustrated above it is not without limitations. The NID metric is less robust to depth errors and relies on well-initialised depth samples. Currently we use photometric tracking for initialisation of the first key-frame and photometric depth update for the first visit to each key-frame to provide well-initialised depth maps for subsequent traversals. As shown in Section 3.3, the combination of darkness, high pixel noise, harsh dynamic lighting and motion blur during the outdoor night dataset caused all approaches to fail to map the first traversal; even the NID metric is not sufficient for these most challenging illumination conditions.

Loop closures provided by FAB-MAP are tolerant to appearance change provided the system has been trained in the appropriate environment, however as reported in [12] even descriptor matching fails under large appearance changes. We seek to replace this stage in the pipeline with either a MI-based approach [25] or a convolutional network approach [33].

Finally, the computational costs of NID-SLAM are higher than both ORB-SLAM and LSD-SLAM due to the use of a dense global metric. Our portable OpenGL implementation currently provides 10Hz updates on a desktop GPU (suitable for robotics or automotive applications); we expect similar computational performance from upcoming mobile graphics processors in the near future.

4. Conclusions

We have presented a robust monocular SLAM approach based on Normalised Information Distance, which we call NID-SLAM. In contrast to existing feature-based and direct photometric methods, NID-SLAM uses a global appearance metric to solve for camera pose relative to a key-frame depth map even in the presence of significant scene changes due to illumination, weather, and occluding objects. We presented three primary contributions: (1) a NID-based tracking approach that explicitly incorporates depth uncertainties into the estimated pose solution; (2) a multi-resolution histogram representation for NID-based tracking that increases the convergence basin for pose estimation; and (3) a NID-based depth map update method, which allows for long-term map refinement and maintenance despite appearance change. Our approach provides tracking and mapping accuracy rivalling state-of-the-art feature-based and direct photometric methods, and significantly outperforms these methods in robustness to appearance changes in both indoor and outdoor environments. We hope NID-SLAM unlocks a wide range of AR/VR and robotics applications that require robust and accurate long-term visual SLAM capabilities.

References

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

[2] G. Caron, A. Dame, and E. Marchand. Direct model based visual tracking and pose estimation using mutual information. *Image and Vision Computing*, 32(1):54–63, 2014. 2

[3] H. Chen, A. S. Lee, M. Swift, and J. C. Tang. 3D Collaboration Method over HoloLens and Skype End Points. In *Proceedings of the 3rd International Workshop on Immersive Media Experiences*, pages 27–30. ACM, 2015. 1

[4] M. Cummins and P. Newman. FAB-MAP: Probabilistic localization and mapping in the space of appearance. *The International Journal of Robotics Research*, 27(6):647–665, 2008. 2, 6

[5] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse. MonoSLAM: Real-time single camera SLAM. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1052–1067, 2007. 2

[6] J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. *arXiv preprint arXiv:1607.02565*, 2016. 1, 2

[7] J. Engel, T. Schöps, and D. Cremers. LSD-SLAM: Large-scale direct monocular SLAM. In *European Conference on Computer Vision*, pages 834–849. Springer, 2014. 1, 2, 3, 4, 5, 6

[8] J. Engel, J. Stückler, and D. Cremers. Large-scale direct slam with stereo cameras. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pages 1935–1942. IEEE, 2015. 6

[9] M. Faessler, F. Fontana, C. Forster, E. Mueggler, M. Pizzoli, and D. Scaramuzza. Autonomous, vision-based flight and live dense 3D mapping with a quadrotor micro aerial vehicle. *Journal of Field Robotics*, 1, 2015. 1

[10] C. Forster, M. Pizzoli, and D. Scaramuzza. SVO: Fast semi-direct monocular visual odometry. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 15–22. IEEE, 2014. 2

[11] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, page 0278364913491297, 2013. 6

[12] A. J. Glover, W. P. Maddern, M. J. Milford, and G. F. Wyeth. FAB-MAP+ RatSLAM: Appearance-based SLAM for multiple times of day. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pages 3507–3512. IEEE, 2010. 2, 8

[13] H. Jin, P. Favaro, and S. Soatto. Real-time 3D motion and structure of point features: a front-end system for vision-based control and interaction. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pages 778–779. IEEE, 2000. 2

[14] G. Klein and D. Murray. Parallel tracking and mapping for small AR workspaces. In *Mixed and Augmented Reality, 2007. ISMAR 2007. 6th IEEE and ACM International Symposium on*, pages 225–234. IEEE, 2007. 2

[15] G. Klein and D. Murray. Improving the agility of keyframe-based SLAM. In *European Conference on Computer Vision*, pages 802–815. Springer, 2008. 2

[16] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):978–994, 2011. 2

[17] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 2

[18] W. Maddern, G. Pascoe, C. Linegar, and P. Newman. 1 Year, 1000 km: The Oxford RobotCar dataset. *The International Journal of Robotics Research*, 36(1):3–15, 2017. 6

[19] F. Maes, D. Vandermeulen, and P. Suetens. Medical image registration using mutual information. *Proceedings of the IEEE*, 91(10):1699–1722, 2003. 2

[20] S. Martull, M. Peris, and K. Fukui. Realistic CG stereo image dataset with ground truth disparity maps. In *ICPR Workshop: TrakMark2012*, volume 111, pages 117–118, 2012. 6

[21] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005. 2

[22] R. Mur-Artal, J. Montiel, and J. D. Tardós. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015. 2

[23] R. Mur-Artal and J. D. Tardós. ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras. *arXiv preprint arXiv:1610.06475*, 2016. 6

[24] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison. DTAM: Dense tracking and mapping in real-time. In *International Conference on Computer Vision*, pages 2320–2327. IEEE, 2011. 1, 2, 4

[25] G. Pandey, J. R. McBride, S. Savarese, and R. M. Eustice. Toward mutual information based place recognition. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3185–3192. IEEE, 2014. 8

[26] G. Pascoe, W. Maddern, and P. Newman. Direct visual localisation and calibration for road vehicles in changing city environments. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 9–16, 2015. 2

[27] G. Pascoe, W. Maddern, and P. Newman. Robust direct visual localisation using normalised information distance. In *British Machine Vision Conference (BMVC), Swansea, Wales*, volume 3, page 4, 2015. 2

[28] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. ORB: An efficient alternative to SIFT or SURF. In *International Conference on Computer Vision*, pages 2564–2571. IEEE, 2011. 2

[29] D. F. Shanno. On the Broyden-Fletcher-Goldfarb-Shanno method. *Journal of Optimization Theory and Applications*, 46(1):87–94, 1985. 4

[30] A. Stewart. *Localisation using the Appearance of Prior Structure*. PhD thesis, University of Oxford, Oxford, United Kingdom, 2014. 2

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

[32] J. Stühmer, S. Gumhold, and D. Cremers. Real-time dense geometry from a handheld camera. In *Joint Pattern Recognition Symposium*, pages 11–20. Springer, 2010. 1, 2

[33] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford. On the performance of convnet features for place recognition. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pages 4297–4304. IEEE, 2015. 8

[34] C. Valgren and A. J. Lilienthal. SIFT, SURF & seasons: Appearance-based long-term localization in outdoor environments. *Robotics and Autonomous Systems*, 58(2):149–156, 2010. 2

[35] N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct):2837–2854, 2010. 3

[36] P. Viola and W. M. Wells III. Alignment by maximization of mutual information. *International Journal of Computer Vision*, 24(2):137–154, 1997. 2

[37] R. W. Wolcott and R. M. Eustice. Visual localization within LIDAR maps for automated urban driving. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 176–183. IEEE, 2014. 2