Robust Direct Visual Localisation using Normalised Information Distance

Geoffrey Pascoe
gmp@robots.ox.ac.uk
Will Maddern
wm@robots.ox.ac.uk
Paul Newman
pnewman@robots.ox.ac.uk

Mobile Robotics Group
University of Oxford
Oxford, UK

Abstract

We present an information-theoretic approach for direct localisation of a monocular camera within a 3D appearance prior. In contrast to existing direct visual localisation methods based on minimising photometric error, an information-theoretic metric allows us to compare the whole image without relying on individual pixel values, yielding robustness to changes in the appearance of the scene due to lighting, camera motion, occlusions and sensor modality. Using a low-fidelity textured 3D model of the environment, we synthesise virtual images at a candidate pose within the model. We use the Normalised Information Distance (NID) metric to evaluate the appearance match between the camera image and the virtual image, and present a derivation of analytical NID derivatives for the $\mathbb{SE}(3)$ direct localisation problem, along with an efficient GPGPU implementation capable of online processing. We present results showing successful online visual localisation under significant appearance change both in a synthetic indoor environment and outdoors with real-world data from a vehicle-mounted camera.

1 Introduction

Real-time visual localisation is a key technology enabling mobile location applications [33, 35], virtual and augmented reality [8, 17] and robotics [12, 24]. The recent availablity of low-cost GPU hardware and GPGPU programming has enabled a new class of 'direct' visual localisation methods that make use of every pixel from an input image for tracking and matching [28], in contrast to traditional feature-based methods that only use a subset of the input image. The additional information available to direct methods localising against a dense 3D map increases robustness against typical failure modes for feature-based methods, such as motion blur and viewpoint change [10].

For computational reasons these direct methods minimise a cost function based on photometric error on a per-pixel basis, which assumes both the live image and the reference map are embedded in the same space. While methods such as this are reasonable for frame-to-frame tracking or small environments with controlled illumination [28] or for optical flow problems [14], it is not robust to strong changes in illumination intensity, direction and spectrum (e.g. in large outdoor environments or across long periods of time) or differences in

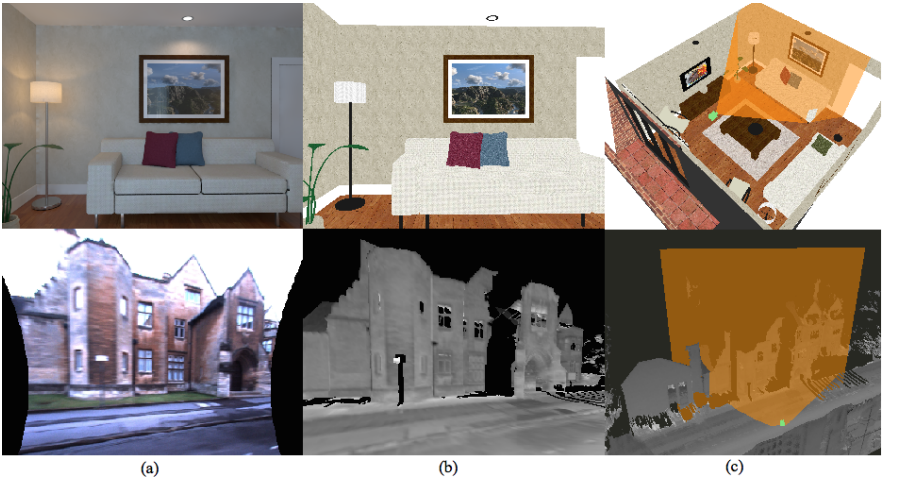


Figure 1: We localise the pose of a camera by registering a rendered image of our prior against the live camera image. Here we show an example from both a synthetic indoor scene and a real-world outdoor scene. (a) Camera image; (b) a render of the prior as used for localisation; and (c) the localised pose and viewing frustrum of the camera within the prior. Our information-theoretic metric is robust to changes in illumination, motion blur, and sensing modality between the live image and prior map.

sensing modality between the image and the map (e.g. if the map is produced by a different camera or LIDAR scanner) where no simple transform exists to map pixel intensity values from the live image to the reference map. The photoconsistency metric therefore limits current direct methods to small-scale environments or short-duration experiments, and although more robust metrics have been proposed for direct localisation against a dense 3D appearance prior [4, 16], these have typically been too computationally expensive for real-time operation.

In this paper we present a direct visual localisation approach based on the Normalised Information Distance (NID) metric, which provides robustness to both viewpoint change and motion blur (in contrast to feature-based methods) as well as illumination change and sensor modality (in contrast to direct photometric methods). Using a low-fidelity 3D appearance prior of the environment, e.g. from a dense reconstruction [27], CAD model or LIDAR scanner as shown in Figure 1, our method is able to localise a camera under a wide range of conditions, including image under/overexposure, outdoor lighting changes, significant occlusions, motion blur, colour space changes, and differences between image and map modality. We present a full derivation of our metric and analytical derivatives for use within a $\mathbb{SE}(3)$ optimisation framework, and illustrate how to leverage GPU rendering and GPGPU programming to accelerate the computation of the metric for online applications. We present localisation results on a synthetic indoor dataset and real-world outdoor data using a LIDAR map as well as computation time analysis demonstrating real-time performance, illustrating the robustness and utility of our method in a wide range of applications in computer vision and robotics.

2 Related Work

Typical approaches to robust visual localisation in changing environments make use of sparse robust feature descriptors, such as SIFT [22], SURF [2], BRIEF [3] and others. These descriptors use combinations of gradients and histograms of local grayscale patches to provide a degree of invariance to viewpoint and illumination conditions. Systems built on these feature descriptors have demonstrated localisation and mapping at large scales [6, 19, 37, 48], but are not robust to strong viewpoint changes [12], changes in outdoor illumination conditions [13, 24] or motion blur induced by fast camera movement [17].

The recent advent of GPGPU programming has enabled so-called 'direct' methods for visual localisation, where every pixel in the image is used to produce an estimate of the camera position relative to a 3D prior map. Direct approaches have been successfully used for monocular $\mathbb{SE}(2)$ visual odometry at large scale [21, 26], $\mathbb{SE}(3)$ monocular localisation and mapping in indoor environments [28], and visual odometry with stereo [5] and RGB-D [38] sensors. Direct approaches have also been combined with sparse feature approaches to produce 'semi-direct' methods [10, 11]. These methods are able to localise against dense 3D stucture with large viewpoint changes and significant motion blur [28]. However, these methods have not been evaluated in scenarios where areas in the map are revisited after significant illumination change, or where the map is constructed using a different camera or sensing modality.

An effective metric for aligning images from different modalities is mutual information [36], frequently used in medical imaging for aligning X-ray and MRI scans [23, 32, 46]. The advantage of mutual information is that the metric is not a function of the values in each image, but instead a function of their entropies; hence, images from different sensors that are not normalised or even correlated (as in [16]) can be successfully aligned [9]. Mutual-information-based metrics have since been used for visual servoing [7], $\mathbb{SE}(2)$ localisation against a LIDAR map [47], camera-LIDAR calibration [30], and $\mathbb{SE}(3)$ localisation against a coloured pointcloud [40] or textured mesh [4, 31]. However, mutual information metrics are typically computationally expensive, requiring multiple seconds per frame [4] or offline processing [30]. In the following sections we present an efficient approach to mutual-information-based direct $\mathbb{SE}(3)$ localisation, and illustrate its robustness to changes in illumination intensity and spectrum, blur and sensor modality.

3 Problem Formulation

Our stated goal is to find the pose of a camera with respect to a 3D prior appearance model of the environment. We pose this as a minimisation problem, in which we wish to find the camera pose, $\hat{G}_C$, such that:

$$\hat{G}_C = \arg\min_{G_C} \rho\left(I_r, G_C, S\right) \tag{1}$$

where I_r is a reference image captured by the camera, S is the virtual scene (including 3D geometry and texture information) used to generate the synthetic image rendered for a camera at pose G_C, and ρ is a metric that quantifies the match between the reference and synthetic images. A frequently-used metric in the domain of dense tracking and reconstruction is the photoconsistency metric [18], which computes the sum of squared differences between pixel

values in a reference image I_r and a synthetic image I_s as follows:

$$\mathrm{SSD}\left(I_r, G_C, S\right) = \sum_{\mathbf{x} \in I_r} \left\| I_r\left(\mathbf{x}\right) - I_s\left(\mathbf{x}, G_C, S\right) \right\|^2, \tag{2}$$

where $\mathbf{x} = (u, v)^{\mathrm{T}}$ is a pixel location within the image.

Although photoconsistency is efficient to compute and find derivatives for (in order to use in an optimisation framework), as mentioned in [28] it suffers from a number of limitations. Principally, it requires I_s provide a photorealistic rendering of the scene S from location G_C, such that the resulting synthetic image matches the reference image I_r on a pixel-by-pixel basis. A true match would only be possible if the synthetic scene S captured all relevant geometry and material properties of the real-world environment, and the synthetic image rendering function $I_s\left(G_C, S\right)$ accurately modelled scene illumination intensity, spectrum and direction in accordance with the true lighting, which both remain challenging to render for real-time applications. This limitation restricts photoconsistency to applications involving frame-to-frame tracking, where the synthetic image I_s can be derived from a warping of the previous reference image I_r [6, 21, 38], or applications in small indoor environments with controlled illumination where the scene S does not change over time [28].

An alternative metric, Normalised Information Distance (NID) [20], is given by:

$$\mathrm{NID}\left(I_r, I_s\right) = \frac{\mathrm{H}\left(I_r, I_s\right) - \mathrm{MI}\left(I_r; I_s\right)}{\mathrm{H}\left(I_r, I_s\right)}, \quad \mathrm{MI}\left(I_r; I_s\right) = \mathrm{H}\left(I_r\right) + \mathrm{H}\left(I_s\right) - \mathrm{H}\left(I_r, I_s\right) \tag{3}$$

where $\mathrm{H}\left(I_r, I_s\right)$ is the joint entropy of I_r and I_s, $\mathrm{MI}\left(I_r; I_s\right)$ is the mutual information between I_r and I_s, and $\mathrm{H}\left(I_r\right)$ and $\mathrm{H}\left(I_s\right)$ are the marginal entropies of I_r and I_s respectively. Unlike mutual information, the NID is a true metric as it is strictly non-negative and satisfies the triangle inequality [45]. The marginal and joint entropies are defined as follows:

$$\mathrm{H}\left(I_s\right) = -\sum_{b=1}^{n} p_s\left(b\right) \log\left(p_s\left(b\right)\right), \quad \mathrm{H}\left(I_r, I_s\right) = -\sum_{a=1}^{n}\sum_{b=1}^{n} p_{r,s}\left(a,b\right) \log\left(p_{r,s}\left(a,b\right)\right) \tag{4}$$

where $\mathrm{H}\left(I_r\right)$ is defined similarly to $\mathrm{H}\left(I_s\right)$. p_s and $p_{r,s}$ are the marginal and joint discrete distributions of the images I_r and I_s, represented by n-bin discrete histograms where a and b are individual bin indices. The values in each histogram bin are computed as follows:

$$p_s\left(b\right) = \frac{1}{|I_s|} \sum_{\mathbf{x} \in I_s} \beta_s\left(b, \mathbf{x}\right), \quad p_{r,s}\left(a,b\right) = \frac{1}{|I_s|} \sum_{\mathbf{x} \in I_s} \beta_r\left(a, \mathbf{x}\right) \beta_s\left(b, \mathbf{x}\right) \tag{5}$$

The histogram weight functions $\beta_r\left(\right)$ and $\beta_s\left(\right)$ are normalised such that $\sum_{b=1}^{n} \beta\left(b, \mathbf{x}\right) = 1 \; \forall \mathbf{x} \in I$, therefore a change in intensity $I_s\left(\mathbf{x}\right)$ at location $\mathbf{x}$ will update the histogram weight function $\beta_s\left(\right)$ in multiple locations. In the following section we will expand upon the choice of histogram weight function $\beta\left(\right)$ to ensure continuous derivatives from discrete histograms.

Computing the joint distribution is the most computationally demanding process of evaluating the NID between two images, as it requires a bin update for every pixel pair between I_r and I_s. We make use of atomic integer addition in local GPU memory provided by OpenCL [41] to significantly accelerate this process, taking less than 5ms for a 640x480 image.

An important consequence of using an information theoretic metric such as NID to compare images is that the metric is not a function of the (pixel) values contained in the images, but instead a function of their distribution. Therefore, an arbitrary perturbation of the colour space for the reference image I_r will strongly affect a photoconsistency metric such as SSD,

but will have minimal impact on the NID. Figure 2 illustrates some examples of reference image perturbations and the consequences for both metrics.

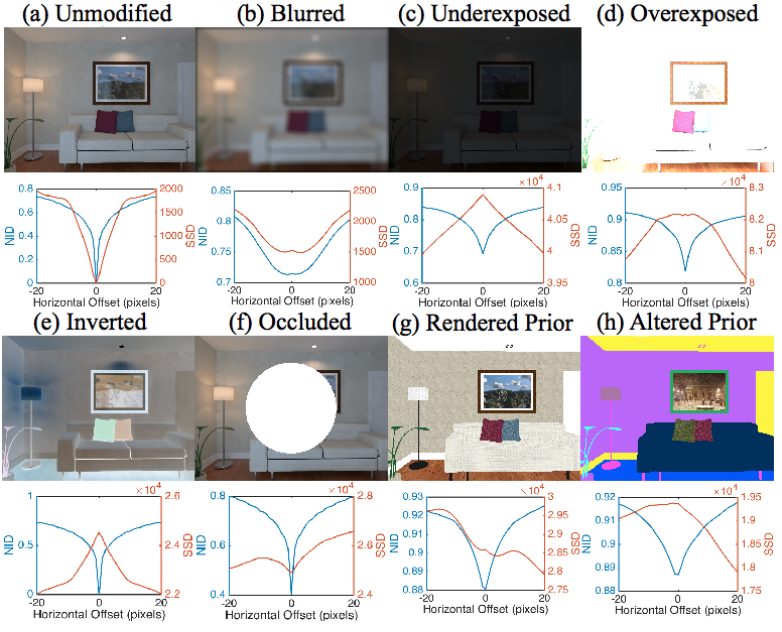


Figure 2: Effect of various image degradations on NID and SSD metric. Each graph shows the NID and SSD compared to the original camera image in (a) with horizontal offsets applied. Whilst NID yields a distinct minima at zero offset under all degradations, SSD is only minimised at zero offset for the simplest of cases.

4 NID Derivatives

In order to incorporate NID into a gradient- or Newton-based optimisation framework, it is necessary to define derivatives of NID for each of the parameters that form the camera pose G_C. Applying the quotient rule to differentiate Equation 3 yields the following:

$$\frac{\partial \mathrm{NID}\left(I_r, I_s\right)}{\partial G_C} = \frac{\frac{\partial \mathrm{H}(I_r, I_s)}{\partial G_C} \mathrm{MI}\left(I_r; I_s\right) - \mathrm{H}\left(I_r, I_s\right) \frac{\partial \mathrm{MI}(I_r; I_s)}{\partial G_C}}{\mathrm{H}\left(I_r, I_s\right)^2} \tag{6}$$

Noting that the reference image I_r does not depend on the camera pose G_C, we can express the mutual information derivative as follows:

$$\frac{\partial \mathrm{H}\left(I_r\right)}{\partial G_C} = 0 \Rightarrow \frac{\partial \mathrm{MI}\left(I_r; I_s\right)}{\partial G_C} = \frac{\partial \mathrm{H}\left(I_s\right)}{\partial G_C} - \frac{\partial \mathrm{H}\left(I_r, I_s\right)}{\partial G_C} \tag{7}$$

The marginal and joint entropy derivatives can be expressed in terms of the marginal and joint distributions as follows:

$$\frac{\partial \mathrm{H}\left(I_s\right)}{\partial G_C} = -\sum_{b=1}^{n} \frac{\partial p_s\left(b\right)}{\partial G_C} \left(1 + \log p_s\left(b\right)\right) \tag{8}$$

$$\frac{\partial \mathrm{H}\left(I_r, I_s\right)}{\partial G_C} = -\sum_{a=1}^{n} \sum_{b=1}^{n} \frac{\partial p_{r,s}\left(a, b\right)}{\partial G_C} \left(1 + \log p_{r,s}\left(a, b\right)\right) \tag{9}$$

The marginal distribution derivative is a straightfoward differentiation of the histogram weighting function $\beta_s()$ for the synthetic image I_s as follows:

$$\frac{\partial p_s(b)}{\partial G_C} = \frac{1}{|I_s|} \sum_{\mathbf{x} \in I_s} \frac{\partial \beta_s(b, \mathbf{x})}{\partial G_C} \tag{10}$$

By noting that the histogram weighting function $\beta_r()$ for the reference image I_r does not depend on camera pose G_C, we can simplify the joint distribution derivative as follows using the product rule:

$$\frac{\partial \beta_r(a, \mathbf{x})}{\partial G_C} = 0 \Rightarrow \frac{\partial p_{r,s}(a,b)}{\partial G_C} = \frac{1}{|I_s|} \sum_{\mathbf{x} \in I_s} \beta_r(a, \mathbf{x}) \frac{\partial \beta_s(b, \mathbf{x})}{\partial G_C} \tag{11}$$

The derivative of the histogram weighting function $\beta_s()$ can be expressed using the chain rule as follows:

$$\frac{\partial \beta_s(b, \mathbf{x})}{\partial G_C} = \frac{\partial \beta_s(b, \mathbf{x})}{\partial I_s(\mathbf{x})} \frac{\partial I_s(\mathbf{x})}{\partial \mathbf{x}} \frac{\partial \mathbf{x}}{\partial G_C} \tag{12}$$

In this expression $\frac{\partial \beta_s(b,\mathbf{x})}{\partial I_s(\mathbf{x})}$ represents the change in the histogram weighting function $\beta_s()$ for all pixels in I_s given a change in pixel value $I_s(\mathbf{x})$ at location $\mathbf{x}$. We follow the approach in [39] and represent the histogram weighting function with cubic B-spline representation [43], yielding C^2 continuous derivatives for each histogram bin, again evaluated using OpenCL atomic operations. The image intensity derivative $\frac{\partial I_s(\mathbf{x})}{\partial \mathbf{x}}$ is represented by a B-spline surface constructed over the synthetic image I_s, approximated using efficient causal and anti-causal Z-transform filters implemented on the GPU [34], requiring approximately the same processing time as the histogram building operation. Finally, $\frac{\partial \mathbf{x}}{\partial G_C}$ is the change in the reprojected location of scene structure S given a change in camera pose G_C. For a minimal Euler angle parameterisation of $G_C = (x, y, z, \alpha, \beta, \gamma)^{\mathrm{T}}$, the structure reprojection derivative is as follows:

$$\frac{\partial \mathbf{x}_i}{\partial G_C} = \begin{bmatrix} -1/z_i & 0 & u_i/z_i & u_i v_i & -\left(1+u_i^2\right) & v_i \\ 0 & -1/z_i & v_i/z_i & 1+v_i & -u_i v_i & -u_i \end{bmatrix} \tag{13}$$

where z_i is the depth associated with the pixel at location $\mathbf{x}_i = (u_i, v_i)^{\mathrm{T}}$, obtained during the rendering process from the z-buffer from an OpenGL [42] rendering operation. Each of the 12 elements in Equation 13 is obtained using a fragment shader during the rendering of I_s, allowing both appearance and derivatives to be computed simultaneously.

Figure 3(a) shows an example of the derivatives calculated at various offsets from the ground truth pose, along with numeric derivatives calculated via central differences. As can be seen, our analytic derivatives match the numerical approximations very closely.

We found that with scenes that include fine-grained textures, such as that on the couch in Figure 1, the derivatives are dominated by the differences between adjacent pixels in the same texture, and thus don't provide useful information. To obviate this effect, we apply a 5x5 kernel Gaussian blur to our rendered image before calculating NID. Figure 3(b) shows the effect of this blurring on the derivatives for the camera pose shown in Figure 1.

The NID cost function described in Equation (3) is dependent on the number of bins used for the respective histograms. Figure 3(c) shows the NID between a camera image and a render of the prior with varying number of histogram bins. All bin counts produce NID minima at zero offset, with 8 bins producing the clearest minima and 16 bins providing a good trade-off between cost surface steepness and computation time.

5 Evaluation

We evaluate the performance of our algorithm using the *Living Room* dataset by Handa *et al*. [15]. This dataset contains a model of a living room approximately 10m×5.5m×3m in size, as well as photorealistic renders of a camera following a specified trajectory through the room. This provides us with a high quality ground truth for the camera pose at which each of the live images were taken. We preprocessed the data to produce a basic textured mesh of the room without any lighting information, in order to evaluate our system's ability to handle a low-fidelity prior (e.g. from a 3D CAD model) with a different appearance to the camera images. Figure 1 shows an example of a virtual camera image and one of our renders, demonstrating the significant differences in appearance between our low-fidelity prior and the photorealistic live images. The reviewer is encouraged to view the accompanying video submission illustrating the localisation process for different input images and map textures.

5.1 Optimisation Methods

For $\mathbb{SE}(3)$ localisation, our search space for Equation (1) is far too large to explore exhaustively. We thus turn to a gradient-based method to find the minimum of the cost function. Common candidates for an optimisation such as this are the Levenberg-Marquardt (LM) [25] and Broyden-Fletcher-Goldfarb-Shanno (BFGS) [29] algorithms. Both of these algorithms use an approximation to the Hessian, allowing the use of a second-order optimisation without the computational overhead of calculating second derivatives. The LM algorithm, however, is most effective for problems in which the number of residuals is greater than the number of parameters [29]. Given that we have a single residual (NID) and six parameters in $\mathbb{SE}(3)$ space, BFGS would appear to be the more appropriate choice for this problem.

We evaluated both BFGS and LM, using the implementations from Ceres Solver [1]. Each optimiser was tested using each pose of a trajectory through the model, with an initial guess randomly offset from the ground truth pose. The offsets were uniformly distibuted between 0 and 0.6m and between 0 and 10°. We also evaluated BFGS using the-built in numerical differentiation in Ceres Solver, in place of our own analytic derivatives. The results of these evaluations are presented in Table 1. These results show that BFGS with analytic derivatives is superior to the two alternatives for this use case.

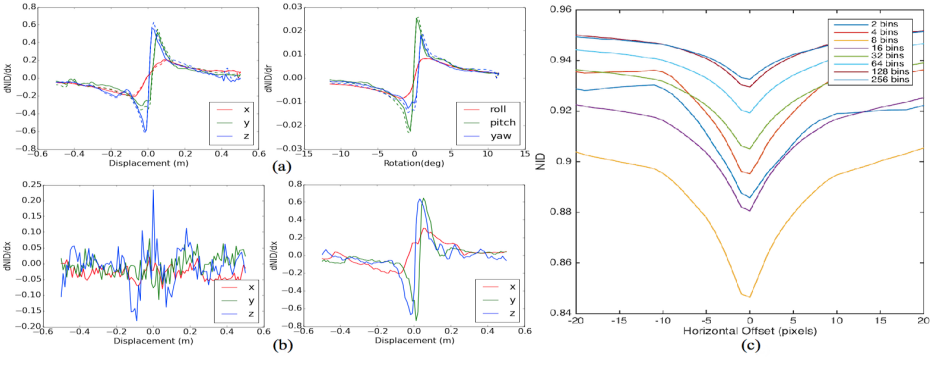


Figure 3: (a) Translational and rotational analytic derivatives (solid lines) compared to numerical approximations (dashed lines). (b) Translational derivatives before (left) and after (right) a Gaussian blur is applied to a rendered image. (c) NID of camera image vs rendered prior for different numbers of histogram bins.

Optimiser	RMSE (m)	p(error < 5cm)	RMSE (°)	p(error < 0.5°)
BFGS (analytic)	0.1692	0.7619	0.2497	0.9263
BFGS (numeric)	0.2553	0.1032	15.5739	0.4444
LM	6.0324	0.1247	38.6402	0.3515

Table 1: RMS errors and success probabilities for random initial offsets.

5.2 Robust Direct Localisation

We evaluate the robustness of our localisation algorithm by applying various degradations to the prior and the live camera images. We test the following degradations to the camera image as shown in Figure 2: (a) Unmodified input images; (b) Applying a Gaussian blur, with a window of 15×15 and 25×25 pixels around each pixel; (c) Underexposed and (d) overexposed images; (e) Inverting all colours in the image; and (f) Obscuring the image with a randomly placed circle of radius 50, 100, and 150 pixels. In addition, we evaluate reducing the bit depth of the input images to 2 and 4-bit colours.

We also test altering the prior, shown in Figure 2(g), by replacing the textures on all surfaces with different colours, as illustrated in Figure 2(h).

We do not show results for a cost function using photometric error, as this failed to converge in all cases. This is due to the fact that the per-pixel photometric error between the photorealistic renders and the prior does not provide a meaningful signal for localisation, as is shown in Figure 2(g).

Scenario	RMSE (m)	RMSE (°)	p(error < 5cm)	p(error < 0.5°)
Unmodified	0.0410	0.5819	0.9725	0.9817
Gaussian blur (15×15)	0.0441	0.6274	0.9725	0.8073
Gaussian blur (25×25)	0.0589	0.6435	0.9541	0.5505
Underexposed	0.0407	0.5928	0.9462	0.9355
Overexposed	0.0450	0.6301	0.9541	0.9617
Inverted image	0.0400	0.5592	0.9633	0.9725
Circle occlusion (50px rad.)	0.0427	0.6146	0.9905	0.9905
Circle occlusion (100px rad.)	0.0450	0.6260	0.9558	0.9469
Circle occlusion (150px rad.)	0.0526	0.5634	0.9619	0.9429
2-bit colours	0.0395	0.5713	0.9725	0.9725
4-bit colours	0.0419	0.6118	0.9908	0.9908
Altered Prior	0.0397	0.5488	0.9633	0.8899

Table 2: RMS errors from an 882-image trajectory through the living room dataset.

We use an 882-pose trajectory through the room, starting our localiser from the previous ground truth pose for each image. Table 2 shows the RMS errors in position and orientation, and the probabilities of converging to within 5cm and 0.5° respectively for each scenario.

All but two of the scenarios show RMS positional errors less than 5cm. Only the most extreme Gaussian blur and the largest occlusion (which obscures over 20% of the image) exceed 5cm in RMS position error; even these scenarios show RMS errors less than 6cm.

The largest reduction in convergence likelihood comes from the most extreme Gaussian blur, which shows a probability of less than 0.6 of converging with an error of less than 5cm. This is a result of the blurring making accurate localisation difficult, as evidenced by the shallow basin around the NID minima in Figure 2(b).

5.3 Computation Time

Figure 4 shows the total time required to localise from a single image, along with the number of cost function evaluations required for localisation. All timings are shown for images

640×480 pixels in size. We achieve a mean localisation frequency of just over 2Hz, consisting of approximatly 25 cost function evaluations where each evaluation takes approximately 16.75ms, with the largest constituent parts being rendering, application of the B-spline pre-filter, and building of the joint histogram. The largest variation is in rendering time due to differences in scene complexity at different locations.

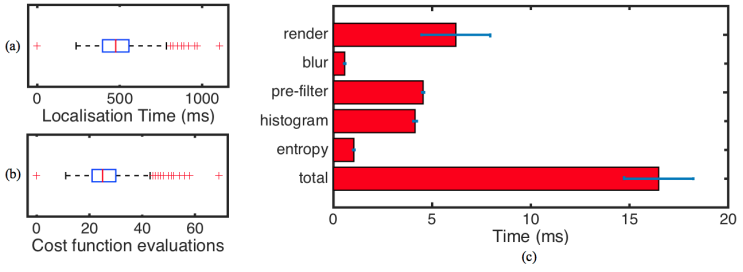
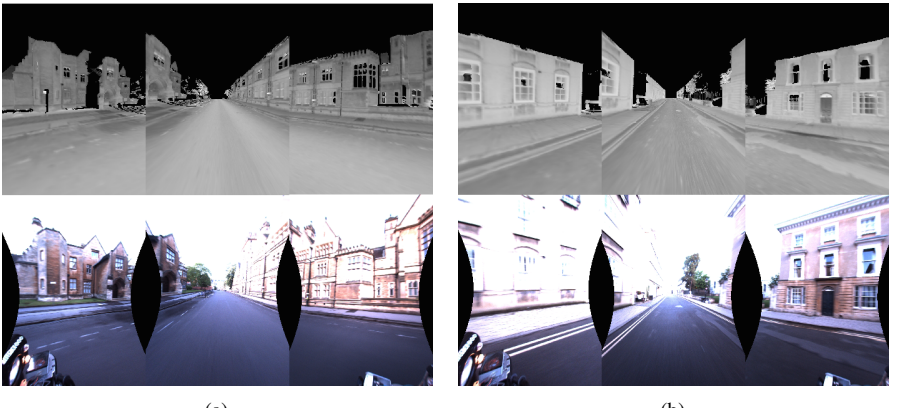


Figure 4: (a) Total time taken for for localisation. (b) Number of cost function evaluations required for localisation. (c) Time taken for different stages of cost function evaluation. All computation is performed with an NVIDIA GTX Titan GPU using OpenGL and OpenCL.

5.4 Outdoor Results

Our system is designed to handle situations in which the appearance of the prior does not exactly match the appearance at localisation time. One application domain in which this is extremely important is outdoor localisation for mobile robots. Here we present results from testing our system using outdoor data, using images acquired from a vehicle-mounted camera and a map provided by a 3D LIDAR survey coloured by near-infrared (905nm) reflectance.

We found that the system works well in downtown areas where there is significant structure close to the road. Table 3 shows the results from localisation along a 140m route in a downtown area. Figure 5 shows examples of locations along this route where localisation was successful. These results illustrate the viability of our system in an outdoor environment, and its ability to handle data from different sensor modalities.



(a) (b)

Figure 5: Examples of places with successful localisation against a laser reflectance prior.

RMSE (m)	p(error < 5cm)	RMSE (°)	p(error < 0.5°)
0.1035	0.7291	0.6749	0.9016

Table 3: RMS error and probability of obtaining an error less than 5cm and 5°along a 140m stretch of road in the centre of a city.

6 Conclusions

In this paper, we have presented a direct visual localisation method that is robust to significant changes in the appearance of the scene. We use an information-theoretic metric for whole-image comparisons, removing a need for exact colour correspondence between pixels. Through the use of a quasi-Newton optimisation method, we are able to minimise our cost function without an exhaustive search, and by exploiting GPUs for much of the computation, we demonstrate localisation at a rate of approximately 2Hz.

We have shown our system's ability to localise to within 5cm of the true pose in a synthetic indoor environment, even with significant degradations to the prior map and camera images. Furthermore, we have presented results indicating our system's ability to be used across sensor modalities in a large outdoor environment. Further work is required to evaluate the system in larger-scale outdoor environments, along with investigating approximations to the full cost function presented here in order to improve on the 2Hz localisation rate.

Acknowledgements

The authors would like to thank Alex Stewart for his assistance and advice on this work. Geoffrey Pascoe is supported by a Rhodes Scholarship. Will Maddern is supported by EPSRC Grant EP/J012017/1. Paul Newman is supported by an EPSRC Leadership Fellowship, EPSRC Grant EP/I005021/1. We would also like to thank our reviewers for their helpful comments.

References

[1] Sameer Agarwal, Keir Mierle, and Others. Ceres Solver. URL http://ceres-solver.org.

[2] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Computer vision–ECCV 2006*, pages 404–417. Springer, 2006.

[3] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. Brief: Binary robust independent elementary features. In *Computer Vision–ECCV 2010*, pages 778–792. Springer, 2010.

[4] Guillaume Caron, Amaury Dame, and Eric Marchand. Direct model based visual tracking and pose estimation using mutual information. *Image and Vision Computing*, 32 (1):54–63, 2014.

[5] Andrew I Comport, Ezio Malis, and Patrick Rives. Real-time quadrifocal visual odometry. *The International Journal of Robotics Research*, 29(2-3):245–266, 2010.

[6] Mark Cummins and Paul Newman. Appearance-only slam at large scale with fab-map 2.0. *The International Journal of Robotics Research*, 30(9):1100–1123, 2011.

[7] Amaury Dame and Eric Marchand. Mutual information-based visual servoing. *Robotics, IEEE Transactions on*, 27(5):958–969, 2011.

[8] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):1052–1067, 2007.

[9] Nicholas Dowson and Richard Bowden. A unifying framework for mutual information methods for use in non-linear optimisation. In *Computer Vision–ECCV 2006*, pages 365–378. Springer, 2006.

[10] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *Computer Vision–ECCV 2014*, pages 834–849. Springer, 2014.

[11] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. Svo: Fast semi-direct monocular visual odometry. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 15–22. IEEE, 2014.

[12] Paul Furgale and Timothy D Barfoot. Visual teach and repeat for long-range rover autonomy. *Journal of Field Robotics*, 27(5):534–560, 2010.

[13] Arren J Glover, William P Maddern, Michael J Milford, and Gordon Fraser 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.

[14] Simon Hadfield and Richard Bowden. Scene Flow Estimation using Intelligent Cost Functions. In *Proceedings of the British Conference on Machine Vision*, Nottingham, UK, 2014.

[15] A Handa, T Whelan, J B McDonald, and A J Davison. A Benchmark for RGB-D Visual Odometry, 3D Reconstruction and SLAM. In *IEEE Intl. Conf. on Robotics and Automation, ICRA*, Hong Kong, China, May 2014.

[16] Michal Irani and P Anandan. Robust multi-sensor image alignment. In *Computer Vision, 1998. Sixth International Conference on*, pages 959–966. IEEE, 1998.

[17] Georg Klein and David 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.

[18] Reinhard Koch. Dynamic 3-d scene analysis through synthesis feedback control. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 15(6):556–568, 1993.

[19] Kurt Konolige and Motilal Agrawal. Frameslam: From bundle adjustment to real-time visual mapping. *Robotics, IEEE Transactions on*, 24(5):1066–1077, 2008.

[20] Ming Li, Xin Chen, Xin Li, Bin Ma, and P M B Vitanyi. The similarity metric. *Information Theory, IEEE Transactions on*, 50(12):3250–3264, 2004. ISSN 0018-9448. doi: 10.1109/TIT.2004.838101.

[21] Steven Lovegrove, Andrew J Davison, and Javier Ibanez-Guzmán. Accurate visual odometry from a rear parking camera. In *Intelligent Vehicles Symposium (IV), 2011 IEEE*, pages 788–793. IEEE, 2011.

[22] David G Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

[23] Frederik Maes, Dirk Vandermeulen, and Paul Suetens. Medical image registration using mutual information. *Proceedings of the IEEE*, 91(10):1699–1722, 2003.

[24] Larry Matthies, Mark Maimone, Andrew Johnson, Yang Cheng, Reg Willson, Carlos Villalpando, Steve Goldberg, Andres Huertas, Andrew Stein, and Anelia Angelova. Computer vision on mars. *International Journal of Computer Vision*, 75(1):67–92, 2007.

[25] Jorge J Moré. The levenberg-marquardt algorithm: implementation and theory. In *Numerical analysis*, pages 105–116. Springer, 1978.

[26] Ashley Napier, Gabe Sibley, and Paul Newman. Real-time bounded-error pose estimation for road vehicles using vision. In *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, pages 1141–1146. IEEE, 2010.

[27] Richard A Newcombe and Andrew J Davison. Live dense reconstruction with a single moving camera. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1498–1505. IEEE, 2010.

[28] Richard A Newcombe, Steven J Lovegrove, and Andrew J Davison. Dtam: Dense tracking and mapping in real-time. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2320–2327. IEEE, 2011.

[29] J Nocedal and S J Wright. *Numerical Optimization*, volume 43. 1999. ISBN 0387987932. doi: 10.1002/lsm.21040.

[30] Gaurav Pandey, James R McBride, Silvio Savarese, and Ryan M Eustice. Automatic extrinsic calibration of vision and lidar by maximizing mutual information. *Journal of Field Robotics*, 2014.

[31] Geoffrey Pascoe, Will Maddern, Alexander D. Stewart, and Paul Newman. FARLAP: Fast Robust Localisation using Appearance Priors. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Seattle, WA, USA, May 2015.

[32] Josien PW Pluim, JB Antoine Maintz, and Max A Viergever. Mutual-information-based registration of medical images: a survey. *Medical Imaging, IEEE Transactions on*, 22(8):986–1004, 2003.

[33] Duncan P Robertson and Roberto Cipolla. An image-based system for urban navigation. In *BMVC*, pages 1–10, 2004.

[34] Daniel Ruijters and Philippe Thévenaz. Gpu prefilter for accurate cubic b-spline interpolation. *The Computer Journal*, page bxq086, 2010.

[35] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Fast image-based localization using direct 2d-to-3d matching. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 667–674. IEEE, 2011.

[36] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[37] Gabe Sibley, Christopher Mei, Ian Reid, and Paul Newman. Vast-scale outdoor navigation using adaptive relative bundle adjustment. *The International Journal of Robotics Research*, 2010.

[38] Frank Steinbrucker, Jürgen Sturm, and Daniel Cremers. Real-time visual odometry from dense rgb-d images. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 719–722. IEEE, 2011.

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

[40] Alexander D Stewart and Paul Newman. Laps-localisation using appearance of prior structure: 6-dof monocular camera localisation using prior pointclouds. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 2625–2632. IEEE, 2012.

[41] The Khronos Group Inc. OpenCL - The open standard for parallel programming of heterogeneous systems, . URL http://www.khronos.org/opencl/. accessed 11/05/2015.

[42] The Khronos Group Inc. OpenGL - The Industry's Foundation for High Performance Graphics, . URL http://www.khronos.org/opengl/. accessed 11/05/2015.

[43] Michael Unser. Splines: A perfect fit for signal and image processing. *Signal Processing Magazine, IEEE*, 16(6):22–38, 1999.

[44] Christoffer Valgren and Achim J Lilienthal. Sift, surf and seasons: Long-term outdoor localization using local features. In *EMCR*, 2007.

[45] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*, 11:2837–2854, 2010.

[46] Paul Viola and William M Wells III. Alignment by maximization of mutual information. *International journal of computer vision*, 24(2):137–154, 1997.

[47] Ryan W Wolcott and Ryan M Eustice. Visual localization within lidar maps for automated urban driving. In *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*, pages 176–183. IEEE, 2014.

[48] Julius Ziegler, Henning Lategahn, Markus Schreiber, Christoph G Keller, Carsten Knoppel, Jochen Hipp, Martin Haueis, and Christoph Stiller. Video based localization for bertha. In *Intelligent Vehicles Symposium Proceedings, 2014 IEEE*, pages 1231–1238. IEEE, 2014.