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

Real-time visual localisation is a key technology enabling mobile location applications [7], virtual and augmented reality [1] and robotics [3]. 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 [6], 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 [2]. In this paper we present a visual localisation system which utilises the entropy-based Normalised Information Distance (NID) metric for image registration.

For computational reasons, existing direct methods generally minimise a cost function based on photometric error on a per-pixel basis (*e.g.* [6]), which assumes both the live image and the reference map are embedded in the same space. Equation (1) defines such a metric based on the sum of squared differences between corresponding pixels in a reference image (I_r) and a synthetic image (I_s).

$$\mathrm{SSD}\left(I_r, I_s\right) = \sum_{\mathbf{x}\in I_r} \|I_r\left(\mathbf{x}\right) - I_s\left(\mathbf{x}\right)\|^2, \tag{1}$$

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 [6] it suffers from a number of limitations. Principally, it requires I_s provide a photorealistic rendering of the scene, such that the resulting synthetic image matches the reference image I_r on a pixel-by-pixel basis. A true match under significant appearance changes would require modelling of the material and illumination properties of the real-world environment, along with the response of the sensor. 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, or applications in small indoor environments with controlled illumination where the scene does not change over time [6].

In this paper we instead make use of the Normalised Information Distance (NID) metric, given by Equation (2) [4].

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

where $\mathrm{H}(I_r, I_s)$ and $\mathrm{MI}(I_r, I_s)$ are the joint entropy and mutual information of the two images respectively, defined as follows:

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

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

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.

As NID is not a function of the actual values of the pixels in the image, but of their distribution, NID provides robustness to illumination change and sensor modality. Unlike mutual information, NID is a true metric, which satisfies the triangle inequality, and is normalised over the number of pixels used in the calculation, thus allowing comparisons between image pairs with differing amounts of overlap.

In order to use NID as a cost function in a gradient-based optimisation framework, we modify the construction of our image histograms to allow the calculation of analytic derivatives. The values in each histogram bin are defined as follows:

$$p_s\left(b\right) = \frac{1}{|I_s|}\sum_{\mathbf{x}\in I_s} \beta_s\left(b, \mathbf{x}\right), \quad p_{r,s}\left(a,b\right) = \frac{1}{|I_s|}\sum_{\mathbf{x}\in I_s} \beta_r\left(a, \mathbf{x}\right)\beta_s\left(b, \mathbf{x}\right) \tag{4}$$

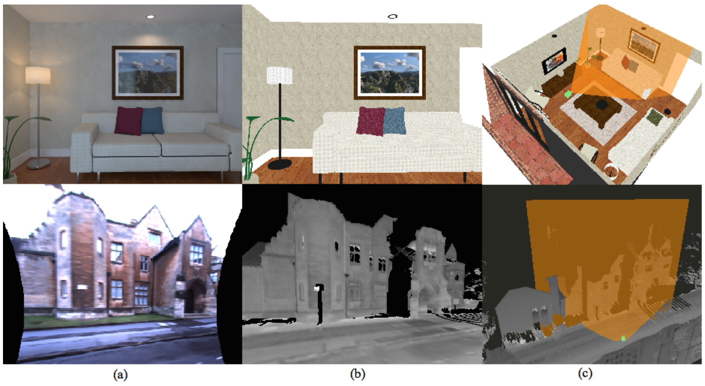

Figure 1: We localise the pose of a camera by registering a rendered image of our prior against the live camera image. Here we show an example from both a synthetic indoor scene and a real-world outdoor scene. (a) Camera image; (b) a render of the prior as used for localisation; and (c) the localised 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.

In a standard histogram, the weighting functions β would take a value of 1 in a single bin, and 0 everywhere else. We instead use the coefficients from a cubic B-spline function as the weighting function, yielding a continuous, twice differentiable function for the value of each histogram bin.

Using a low-fidelity 3D appearance prior of the environment, e.g. from a dense reconstruction [5], 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 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] 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.

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

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

[4] 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.

[5] 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.

[6] 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.

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