A Constant-Time Efficient Stereo SLAM System

Christopher Mei
Gabe Sibley
Mark Cummins
Paul Newman
Ian Reid
{cmei,gsibley,mjc,pnewman,ian}@robots.ox.ac.uk

Department of Engineering Science
University of Oxford
Oxford
OX1 3PJ
UK

Continuous, real-time mapping of an environment using a camera requires a constant-time estimation engine. This rules out optimal global solving such as bundle adjustment. In this article, we investigate the precision that can be achieved with only local estimation of motion and structure provided by a stereo pair. We also discuss the integration of a loop closure and relocalisation mechanism that is essential for working in non-controlled environments where tracking assumptions are often violated. The system is comprised of three key components: (i) a representation of the global environment in terms of a *continuous* sequence of relative locations; (ii) a visual processing front-end that tracks features with sub-pixel accuracy, computes temporal and spatial correspondences, and estimates precise local structure from these features; (iii) a method for loop-closure which is independent of the map geometry, and solves the loop closure and kidnapped robot problem; to produce a *system* capable of mapping long sequences over large distances with high precision, but in constant time processed at 30-45 Hz.

A SLAM system requires a way to represent the map in the environment and several possible representations are possible. In this work, a continuous relative representation (CRR), illustrated in Fig. 1, was used. This approach is beneficial for two reasons. First, it allows constant time state-updates even when loop-closures are detected and relative bundle adjustment (RBA) is applied [4]. Second, optimisation using CRR effectively handles problems inherent in sub-mapping, such as map merging and splitting, data duplication and inconsistency.

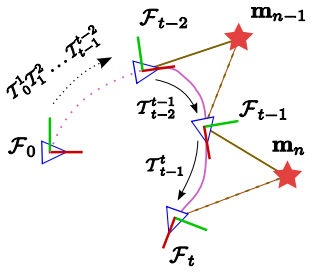


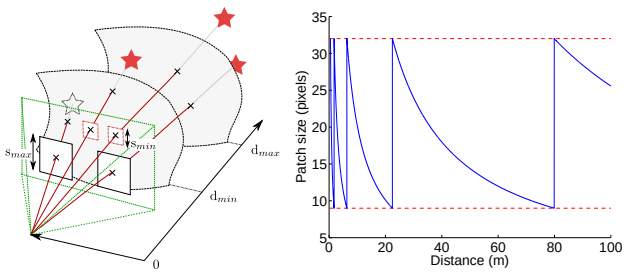
Figure 1: Continuous relative representation (CRR).

The visual front-end combines careful engineered computer vision algorithms [2] for robust motion estimation from a stereo pair with two novel components: "true scale" SIFT and quadtree feature selection.

The idea of "true scale" is to reduce the expensive part of the SIFT algorithm [3] that consists in finding an extremum in scale in a Difference of Gaussians (DoG) pyramid. Landmark descriptors are built corresponding to regions in the world of same physical size. It can be achieved at no extra computational cost as the left-right matching that provides the 3-D location is required in any case for the motion estimation. True scale requires choosing a set of 3-D sizes for different depth ranges called "rings" to ensure the projection size of the 3-D template lies within a given pixel size range (Fig. 2).

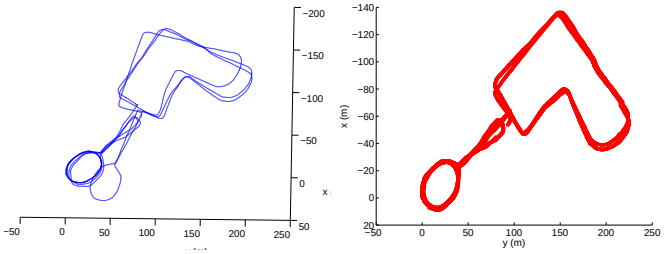
A quadtree representation is used to ensure a even distribution of points in the image. The quadtree contains the number of 3-D features that currently project to that cell (these are done during temporal matching). Potential new features, in order of their Harris distinctiveness scores, are tested against the quadtree to ensure that the number in each cell does not exceed a pre-set maximum percentage. On selecting a potential feature from the left image, its correspondence in the right image is sought by scanline search in our rectified images. This approach greatly improves the conditioning on real sequences.

The "true scale" SIFT provides the information required to ensure efficient relocalisation and loop closure capabilities using the FABMAP bag-of-words representation [1].



(a) (b)

Figure 2: True scale. (a) A fixed 3-D region size between distances d_{min} and d_{max} projects to template sizes ranging from s_{max} to s_{min} with size in $\frac{1}{d}$. For $d > d_{max}$, a bigger 3-D size is used to provide image templates within the same pixel size range. (b) Patch sizes according to distance.



(a) New College without loop closure (b) New College top view with loop closure

Figure 3: Estimated trajectory for the New College sequence (Tab. 1).

	New College
Distance Travelled	2.26 km
Frames Processed	51K
Reprojection Error Min/Avg/Max	0.03 / 0.13 / 1.01 pixels
Accuracy without loop closure	∼15-25m in (x-y) plane, ∼15m in z
Accuracy with loop closure	∼10cm in (x-y) plane, ∼10cm in z

Table 1: Results for the New College sequence.

The experiments demonstrate how a continuous relative representation (CRR) combined with careful engineering (true scale, subpixel minimisation and quadtrees) can provide constant-time precise estimates, efficiency and good robustness. An important aspect is that loop closure using CRR greatly improves the accuracy even without a global relaxation as shown on the New College data set [5] (Tab. 1 and Fig. 3). Furthermore the CRR framework is more than a simple re-parametrisation, it leads to a different cost function that makes it possible to represent trajectories that cannot be embedded in a Euclidean space as in the case of non-observable ego-motion (e.g. if the platform takes a means of transport). This opens up new prospects for mapping algorithms.

[1] 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] R. Hartley and A. Zisserman. *Multiple View geometry in Computer vision*. Cambridge university press, 2000.

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

[4] G. Sibley, C. Mei, I. Reid, and P. Newman. Adaptive relative bundle adjustment. In *Robotics Science and Systems Conference*, 2009.

[5] M. Smith, I. Baldwin, W. Churchill, R. Paul, and P. Newman. The new college vision and laser data set. *The International Journal for Robotics Research*, 28(5):595–599, 2009.