

Multiple Target Localisation at over 100 FPS

Simon Taylor

<http://mi.eng.cam.ac.uk/~sjt59>

Tom Drummond

<http://mi.eng.cam.ac.uk/~twd20>

Department of Engineering

University of Cambridge

Cambridge, UK

Finding points in different views of a scene which correspond to the same real world locations is a fundamental problem in computer vision, and a vital component of applications such as automated panorama stitching (e.g. [1]), image retrieval (e.g. [4]) and object localisation (e.g. [2]). Many successful approaches to these problems are based on matching features extracted from two images to be matched.

The first stage of all local feature matching schemes is to apply interest region detection to factor out common imaging transformations. A scale-space search for interest regions using methods such as the DoG detector [2] are common. A canonical orientation can be assigned to an interest region, for example by considering the blurred gradient at the centre of the region [1]. The most basic representation of the interest region is obtained by extracting a pixel patch from the canonical frame that has been assigned. As simple patch-matching schemes such as Sum-of-Squared Differences (SSD) do not perform well when subject to the errors introduced by interest region detectors a more complicated matching scheme is usually employed. Two broad approaches have been studied in the literature. The first class of methods is typified by SIFT [2] and perform further processing on the extracted patches to compute a feature descriptor vector which is ideally equal for different views of the same feature. An alternative approach, used in the Ferns method [3], recasts the matching problem as one of classification, and can obtain matches with very little computation on the input patches after classifiers for each database feature have been learnt in an offline training stage.

We propose a method which falls into the matching-by-classification category. Our classifiers are far simpler than those used in the Ferns approach as they employ independent per-pixel distributions rather than the large joint distributions used in Ferns. This means individual features in our method require much less memory and so leads to a more scalable approach.

Our approach is based on simple pixel patches extracted from around interest points. Although SSD-based matching is not robust when subject to small registration errors, registration errors do not affect all pixels equally; samples from the interior of large regions of solid colour in a patch are more robust to registration errors. We employ a training phase to learn a model for the range of patches expected for each feature using independent per-pixel distributions, which we refer to as a Histogrammed Intensity Patch (HIP). This model allows runtime matching to use simple pixel patches whilst providing sufficient viewpoint invariance to handle registration errors from interest point detection. The histograms are quantised to give a small binary representation that can be very efficiently matched at runtime.

In previous work [5] we introduced the approach and used the efficient FAST-9 detector to factor out translation changes. This paper presents a number of significant improvements to the approach. Firstly canonical orientation computation is added to the interest point detection stage, allowing the number of features for a target to be reduced by a factor of around 15. A novel two-pass approach to training accounts for errors related to interest point detection and orientation assignment, and allows us to choose efficient methods for those stages without sacrificing robustness of the overall system. A tree-based matching scheme is introduced to exploit common information in different features and prevent the need for an exhaustive comparison against all database features. Finally a framework for rapid independent multiple-target localisation using HIPs is presented.

A two-stage training approach is used to identify repeatable features in each region and build feature models for them. The first stage is to run interest point detection and orientation assignment on all of the training images in a large set generated by artificially warping a reference target. The position and orientation of each detection and the appearance of the surrounding image region is stored in a structure termed a *subfeature*. The second stage then clusters subfeatures based on position and orientation to identify repeatable features, and builds a Histogrammed Intensity Patch



Figure 1: Four frames from a sequence demonstrating independent multiple object localisation. No frame-to-frame tracking is performed; the objects are localised in each frame. The mean total computation time per frame is 7.46ms using a single core of a 2.4GHz CPU.

model for each feature by combining the appearance information from the subfeatures in each cluster.

Runtime performance considerations led us to select FAST-9 as the interest point detector. Typical approaches to assigning orientation require computationally expensive blurring [1] or histogramming [2] and would add significant computation to the runtime processing. Instead we propose a simple and cheap method based on differences between pixels in the 16-pixel ring used in FAST corner detection. Our training phase clustering of subfeatures naturally compensates for inaccuracies in interest point detection and orientation assignment by adding multiple database features to represent a single cluster if the errors are too large, so we can use an inexpensive and inaccurate orientation assignment scheme without a major degradation in matching robustness.

The Histogrammed Intensity Patch model is stored as a binary representation that permits a very fast dissimilarity score computation. Shared bits between binary database feature representations can be exploited to build a binary search tree such that matches can be found without requiring an exhaustive comparison against all database features.

We present the results of a comparison with state-of-the-art fast localisation schemes [6] and show we achieve better matching robustness in under a quarter of the computation time and requiring 5-10 times less memory. The scalability of the method to support multiple independent target localisation is also discussed.

- [1] M. Brown, R. Szeliski, and S. Winder. Multi-image matching using multi-scale oriented patches. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 510–517, 2005.
- [2] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2:91–110, 2004.
- [3] M. Ozuysal, P. Fua, and V. Lepetit. Fast keypoint recognition in ten lines of code. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2007.
- [4] Cordelia Schmid and Roger Mohr. Local greyvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:530–535, 1997.
- [5] Simon Taylor, Edward Rosten, and Tom Drummond. Robust feature matching in 2.3 μ s. In *IEEE CVPR Workshop on Feature Detectors and Descriptors: The State Of The Art and Beyond*, June 2009.
- [6] Daniel Wagner, Gerhard Reitmayr, Alessandro Mulloni, Tom Drummond, and Dieter Schmalstieg. Pose tracking from natural features on mobile phones. In *Proc. ISMAR 2008*, Cambridge, UK, Sept. 15–18 2008.