Place recognition from disparate views

Rob Frampton
http://www.cs.bris.ac.uk/~frampton
Andrew Calway
http://www.cs.bris.ac.uk/~andrew

Visual Information Laboratory
University of Bristol

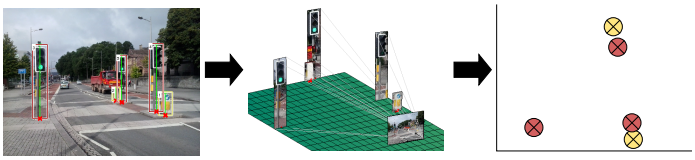

Figure 1: Illustration of our geometric place model

Visual place recognition methods which use image matching techniques have shown success in recent years [1, 2, 3, 6, 8, 9], with some systems operating in real-time on very large databases [4, 7], however their reliance on local features restricts their use to images which are visually similar and which overlap in viewpoint.

In this paper we suggest that a semantic approach to the problem would provide a more meaningful relationship between views of a place and so allow recognition when views are disparate and database coverage is sparse. As initial work towards this goal we present a system which uses detected street objects such as traffic lights and signs as basic features, which are used to generate a 2D geometric place model. We then score the similarity of a pair of models by extracting features which characterise the pair, and use distributions learned from training examples to compute the probability that they depict the same place. We also gain some semantic understanding of the relationship between the two views by estimating the position of each object, the ground plane and the relative pose of the cameras.

Geometry estimation. We use a geometry model inspired in part by the work of Hoiem *et al*. [5], which assumes the relative world height of each object class is known, from which we estimate the depths of the objects as well as the plane they sit on. Ultimately we generate a model like that shown in Figure 1 - essentially a top-down 2-dimensional view of the scene. This is the place model which we use to perform place recognition.

Hypothesis scoring. If the place model is a good approximation of the scene geometry, it should be the case that when we are given two images of the same place, their models look very similar; indeed, if we had perfect measurements, they would be related by just a rotation and translation. With this in mind, we extract a set of features $\mathcal{F}$ from the place models and compute the probability $p(C|\mathcal{F})$, where C is the event that the two images depict the same place. The problem is now treated as a machine learning problem; distributions $p(C|\mathcal{F}_i)$ for each feature are estimated from training data and are used to compute the final probability during testing.

Since we do not know the correspondence of objects between images, we must evaluate each correspondence hypothesis separately. Due to the risk of generating an intractable number of hypotheses, we only consider 5-object correspondences between views and only use the top 10 ranking objects according to the detector confidence scores. We evaluate $p(C|\mathcal{F})$ for all possible 5-object correspondences and use the highest hypothesis probability as the probability that the images depict the same place.

We use five features to score each hypothesis. Two features are based on geometric measurements of the place models. We also compute a rectified ground plane image from the source images and extract edges to quantify their visual similarity, which is used as a feature. The last two features are the residual on the ground plane estimate and the object confidence scores given by the detector. Probabilities are estimates from distributions learned during a training stage, and a simple naive Bayes classifier is used to obtain the final probability.

Experimental results. We evaluated the system by performing a standard place recognition experiment on a dataset of 40 urban locations, with about three viewpoints per location. The system achieved a recognition rate of up to 73.1%. We also observed that some discriminative ability of the system is provided by the different object classes, and whilst this is a legitimate place recognition scenario, we wanted to observe the discriminative ability of the features alone. Thus, we also tested the system on a subset of the dataset with 30 locations, all of which contained the same two object classes, meaning that almost every image was capable of valid object correspondences with every other image. The system achieved a recognition rate of 67.9% in this harder scenario. Figure 2 shows an example of a pair of images correctly identified by our method as representing the same place, as well as the estimated geometry given by the system.

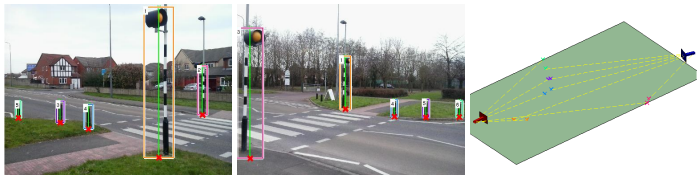

Figure 2: An example of successful output from our system

Due to the lack of similar work against which to compare our result, we also designed a place recognition task for human participants with the aim of providing a benchmark against which to compare our system. Unsurprisingly we found that humans are in general able to perform better (Figure 3), however about a third of participants performed worse than our system - indicating that the problem is not trivial to solve. The experiment also indicated that humans are more likely to use other distinctive semantic cues to discriminate between places rather than the purely geometric approach of our system, which will help to guide future work.

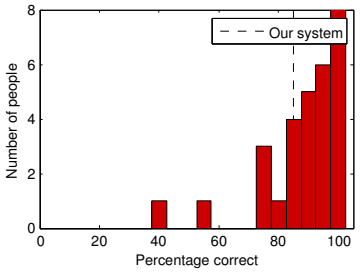

Figure 3: The results of the human experiment for humans and our system.

[1] Adrien Angeli, David Filliat, Stephane Doncieux, and Jean-Arcady Meyer. Fast and incremental method for loop-closure detection using bags of visual words. *IEEE Transactions on Robotics*, 24(5):1027–1037, October 2008.

[2] S. Bazeille and D. Filliat. Combining odometry and visual loop-closure detection for consistent topo-metrical mapping. *RAIRO - Operations Research*, 44(4):365–377, January 2011.

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

[4] Mark Cummins and Paul Newman. Highly scalable appearance-only SLAM - FAB-MAP 2.0. In *Robotics Science and Systems*, pages 1–8, Seattle, USA, 2009.

[5] Derek Hoiem, Alexei a. Efros, and Martial Hebert. Putting objects in perspective. *International Journal of Computer Vision*, 80(1):3–15, April 2008.

[6] K. Konolige, J. Bowman, J. D. Chen, P. Mihelich, M. Calonder, V. Lepetit, and P. Fua. View-based maps. In *Proceedings of Robotics: Science and Systems*, Seattle, USA, June 2009.

[7] M. Labbe and F. Michaud. Appearance-based loop closure detection for online large-scale and long-term operation. *Robotics, IEEE Transactions on*, PP(99):1–12, 2013.

[8] David Nistér and Henrik Stewénius. Scalable recognition with a vocabulary tree. In *CVPR (2)*, pages 2161–2168, 2006.

[9] Grant Schindler, Matthew Brown, and Richard Szeliski. City-scale location recognition. *CVPR*, pages 1–7, June 2007.