Place Recognition and Online Learning in Dynamic Scenes with Spatio-Temporal Landmarks

Edward Johns and Guang-Zhong Yang
ej09@imperial.ac.uk g.z.yang@imperial.ac.uk
The Hamlyn Centre, Imperial College London

1. Image Retrieval Issues

Traditional methods for place recognition typically adopt an image retrieval approach. Given a query image, a database of existing images is searched in order to find the most similar image(s). The now standard Bag-Of-Features (BOF) [1] technique acquires set of candidate images by quantizing local image features and computing a vector of "visual word" occurrences. Tentative feature matches based on these visual words are then verified geometrically, usually with a RANSAC-based affine estimation of the epipolar geometry between the two views.

One of the issues with this approach is that many features exist in the database that are never matched to. These arise due to unstable keypoints, or dynamic elements in a scene. As a result, many feature matches are considered that are, in fact, very unlikely to appear again in an image, resulting in unnecessary computational costs. A second issue is that quantisation errors can cause the same real-world point, when viewed under different illumination or viewpoint conditions, to be assigned to a different visual word. A third issue is that updating the database over time can be troublesome when false positive matches introduce incorrect images for representing a scene.

2. Landmarks and Scene Models

We address these issues by modelling each scene as a set of spatio-temporal landmarks, each representing a real-world point, and matching a query image to a database scene models, rather than a database of images. These landmarks are formed by tracking features across several images of the same scene. Each landmark therefore corresponds to stable features, and we explicitly learn the set of visual words that are assigned to the feature. Furthermore, we can compute the rate of occurrence of landmarks, together with the expected spatial relationships between landmarks.

In the BOF filtering stage, the vector of visual word frequencies for a scene is computed as an average across all images assigned to a scene. Candidate scenes with similar vectors to a query image are then matched geometrically. Each tentative feature-to-landmark match is verified or discarded by considering the spatial relationship between the most frequently co-occurring landmark and the landmark we are verifying. We describe this spatial relationship in terms of a spatial word reflecting the image distance, image angle and orientation angle between two features. If this spatial word has previously been observed between the two landmarks, then the landmark is verified.

The score given to each candidate scene is a normalised summation of verified landmarks. Each verified landmark is given a weight corresponding to the likelihood that the landmark is in fact present, conditional on the visual words and spatial words detected in the query image. In this way, landmarks with more discriminative visual words, or a more discriminative spatial relationship with its co-occurring landmark, are assigned a greater weight in the scoring function.

3. Online Learning

The ability to treat landmarks independently from the images they come from also enables these above properties to be updated online as further images are acquired for a scene. Additionally, new landmarks that enter the scene from dynamic objects can be incorporated. By tracking features across a history of images, we show that the recognition performance increases as further images are acquired.

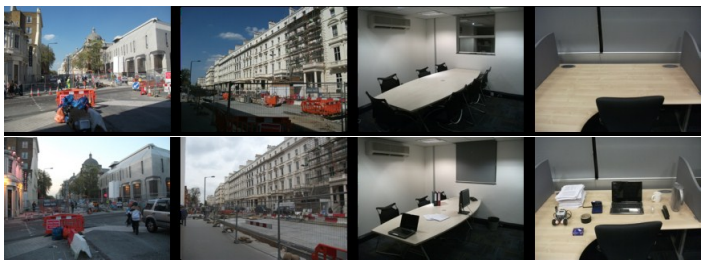
We compare our method to the soft quantisation technique [2] which we adapt for two applications. In application A, each scene is represented by a single image, as standard. In application B, each scene is represented by a pool of all images that have previously been matched to the scene. Several tours of 1.6 km in length were conducted of a busy outdoor environment, with each tour consisting of 1000 discrete scenes. We used the first two tours to initially train our scene models and assign images to the pool in application B of [2].

Figure 2 (a) shows the performance of our method compared with the image retrieval approach. In the long term, our method outperforms the others because the scene models are learned probabilistically and so any false-positive scene matches only change scene properties marginally, rather than falsely assigning an entire image to the pool of images representing a scene.

4. Adapting to Dynamic Environments

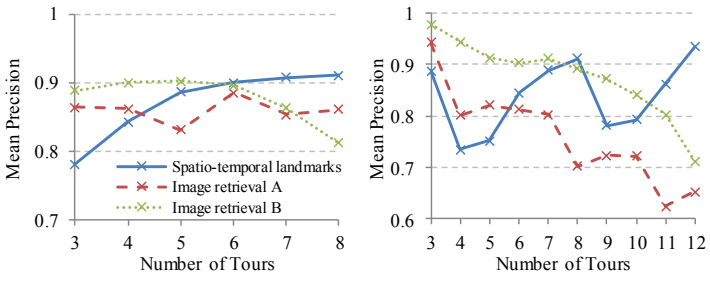
We also show that the expected occurrence probability of each landmark can be treated in a temporal manner, such that those landmarks which have not occurred for some time are given less weight in the scene score function. We rearranged the furniture in the indoor environment to assess the ability of the system to adapt to dynamic objects, by updating existing landmarks, introducing new landmarks and filtering out old landmarks.

Figure 2 (b) shows the performance of our technique under these conditions. Whilst the image retrieval approach decreases in performance as the room is rearranged, our method is able to adapt in the long term and incorporate the new landmarks accordingly.



(a) (b)

Figure 1: Example images from the outdoor dataset, and the indoor dataset, where scene objects were rearranged to examine performance in dynamic environments.



(a) (b)

Figure 2: Mean precision for scene recognition in the outdoor (a) and indoor (b) datasets. For the indoor dataset, scene rearrangements were made after the 3rd and 8th tours of the environment. Image retrieval A stores one image per scene, whereas image retrieval B accumulates images as they are matched.

[1] J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. In *Proc. ICCV*, 2003.

[2] J. Philbin, O. Chum, M. Isard, J. Sivic and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale databases. In *Proc. CVPR*, 2008.