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

Abstract

This paper presents a new framework for visual place recognition that incrementally learns models of each place and offers adaptability to dynamic elements in a scene. Traditional bag-of-features image-retrieval approaches to place recognition treat images in a holistic manner and are typically not capable of dealing with sub-scene dynamics, such as structural changes to a building facade or the rearrangement of furniture in a room. However, by treating local features as observations of real-world landmarks in a scene that are consistently observed, such dynamics can be accurately modelled at a local level, and the spatio-temporal properties of each landmark can be independently updated online. We propose a framework for place recognition that models each scene by sequentially learning landmarks from a set of images, and in the long term adapts the model to dynamic behaviour. Results on both indoor and outdoor datasets show an improvement in recognition performance and efficiency when compared to the traditional bag-of-features image retrieval approach.

1 Introduction

The recognition of a place instance represented in an image has seen a wide range of applications [1, 2, 3, 4]. Typically, the approach for large-scale tasks is the image retrieval challenge [23] where a query image is matched to a database of candidate images. In recent years, efficient matching has been inspired by the Bag-Of-Features (BOF) method [5] where comparisons of histograms of quantised features select candidate images for stronger geometric verification. This has been adopted in a number of closely-related works [11, 15, 17, 21, 22, 23]. In this paper we present a new framework for place recognition that improves both the image retrieval and BOF components of traditional approaches.

1.1 Image Retrieval

Databases for image retrieval often have significant redundancy due to dynamic behaviour influencing an image. We define two types of dynamics: *feature dynamics* and *scene dynamics*. Feature dynamics arise due to the instability of a keypoint when the same real-world point is viewed under different viewpoints or illumination conditions. Scene dynamics arise due to long-term structural changes in a scene, such as renovations of building facades or the rearrangement of furniture in a room, and short-term dynamic

bodies such as pedestrians or cars. As a result of both these types of dynamic behaviour, many features exist in the database that are never matched to by features in a query image.

We propose a model-based recognition framework that compresses database images into a set of *scene models*, each representing a place of interest, whilst maintaining the ability to match images from the full range of viewpoints and illumination conditions expressed in the image database. This is achieved by tracking features across multiple images to form a set of *spatio-temporal landmarks*, each representing a real-world point, and learning the distribution of descriptors across the landmark's constituent features. Greater importance can then be assigned to those landmarks that are more likely to appear and those which are assigned to more discriminative descriptors. Similarly, sets of co-occurring landmarks can be learned that co-occur frequently and with consistent spatial relationships.

Feature dynamics are thus accounted for by eliminating those features which occur infrequently, and scene dynamics can be incorporated into the model online by introducing new landmarks into the database as they begin to appear in subsequent visits to a place, and assigning greater importance to the more recently-observed landmarks. This models the dynamics of a scene at the local level, rather than at the image level as in the case of traditional image retrieval, whereby the entire image must be updated.

1.2 Bag-of-Features

BOF methods have the natural drawback that spatial information is not extracted from an image, requiring a large number of candidate images to be passed to the geometric verification stage. Developments in weak geometric encoding [7, 12] have proven effective for classification tasks, but recognising instances of objects or places typically maintains a division between BOF encoding and geometric verification. More recent works have embedded geometric information in the BOF stage [28] or used inter-feature relationships to reduce the set of candidate matches prior to a RANSAC-based geometric verification [18, 19]. We further improve on this latter idea by using co-occurring landmarks to verify a feature-to-landmark match. By learning which landmarks co-occur most frequently, we show that it is possible to verify each feature-to-landmark match using only a single co-occurring landmark, rather than a more complex search across the full set of candidate matches.

A further issue with BOF is that two features representing the same real-world point can be assigned to two different visual words even under tiny variations in viewpoint or illumination. Whilst this has been addressed by refining visual word dictionaries [18, 19, 20] or query expansion methods [17], the problem will always exist in BOF unless the distribution of each landmark is learned explicitly. In our framework, we do just this by accumulating visual words for each landmark from its associated features.

1.3 Related Work

Model-based place recognition has been successful in small indoor environments [8, 16], but large-scale modelling has not been addressed in this way and typically remains an image retrieval problem. Attempts to improve the efficiency of retrieval have included matching to iconic images of a scene [15], but these still require appearance-based image-to-image matching, and as such feature redundancies remain. The work in this paper is related to our work in [29] whereby scene models are learned from image clusters, and we adapt this to demonstrate incremental learning and adaptation to dynamic scenes.

Existing approaches to online learning of dynamic scenes typically adopt an incremental approach to Support Vector Machine (SVM) classification [13, 14]. However, online training of SVMs remains computationally heavy and is not suitable for real-time applications such as robotics. Furthermore, these works are applied to small indoor training sets where discriminative methods are suitable, whereas for large-scale recognition this level of complexity is often not viable.

Direct feature-to-feature matching approaches [26, 27] have been successful on small-scale databases, and more recently this has been speeded up by more selective matching [25], but these methods still require the expensive computation of feature-to-feature descriptor distances. Feature tracking to extract stable features has been applied previously in simple frameworks [9, 24, 25]. However, none of these works learn feature descriptor distributions in a robust probabilistic manner, nor do they exploit the observed spatial relationships between features as they are tracked.

The rest of the paper is organised as follows. Section 2 gives an overview of our scene models. Section 3 provides the methodology for place recognition. Section 4 outlines the method for online learning of dynamic environments. Section 5 demonstrates experimental results and comparisons to the image retrieval method. Section 6 then concludes the work.

2 Scene Models

2.1 Landmark generation

The system is initialised with a training tour to define the set of database scenes. In order to seed the probabilistic framework, a set of 2 initial training images are captured for each defined scene. Local SURF features [6] are extracted, matched and tracked across the training images, using soft quantisation word assignments [18] to form candidate matches, and followed by RANSAC geometric verification [10]. We use a dictionary of 1 million words, structured in a vocabulary tree [11] with 3 levels and 100 branches per level, trained on features across the entire database. For each scene, a set $p_i...p_I$ of landmarks is then formed. Each landmark is assigned a set of visual words $g^p{}_k...g^p{}_K$ accumulated from the tracked features, representing the expected range in feature space that the landmark will occupy under reasonable viewpoints.

2.2 Spatial words

Each landmark p is then assigned a set $q_j...q_J$ of co-occurring landmarks that appear at least once in the same scene as p, and the most frequently co-occurring landmark is denoted q^*. As in [29], we then model the spatial relationship between p and each q as a set of *spatial words* $h^{pq}{}_l...h^{pq}{}_L$ to encompass the expected range of spatial distances, spatial angles and orientation differences between the two landmarks when viewed on an image. Rather than computing the absolute values, these are quantised into a dictionary of spatial words to reduce the effect of overfitting spatial distributions from only a small number of training images, allowing for small inter-landmark displacements that may arise from previously-unobserved viewpoints. Furthermore, when updating landmark properties after recognition, should a false positive feature-to-landmark match occur, then the rate of occurrence of this incorrect spatial word will be small as the scene model is updated from several images, and will thus have less weight in the recognition stage. Spatial distances are quantised into bins with a separation of 10 pixels. Spatial angles and orientation differences are quantised into

discrete bins with 10° separation. The spatial dictionary is thus of size $(400 / 10) \times (360 / 10) \times (360 / 10) = 51840$, where 400 is the diagonal length in pixels of the 320×240 images used in our experiments.

3 Recognition

In order to deal with noisy and dynamic environments, and to sequentially improve the model over time, the system continually updates the scene models as further images are required. This involves a two-stage process, with the recognition stage first performing an image-to-scene match, and then returning its result to the online learning stage. Here, new landmarks that have entered a scene are acquired, and properties of the existing landmarks are updated.

3.1 BOF Filtering

Given a query image, we wish to compute the most likely scene from the database. As in standard BOF image retrieval [5], we first compute the cosine similarity of the word frequency vector between the image and each scene. In our method, the inverse document frequency (*idf*), representing the rate of word occurrence across the entire database, is as standard. However, the term-frequency (*tf*) for the scene vector is computed as an average of occurrences across images previously matched to the scene. In this way, feature dynamics are dampened and those words that occur consistently in the scene have a greater impact. The top k scenes returned from this stage are then passed on for geometric verification.

3.2 Geometric Verification

A score u_x is now computed for each candidate scene x by considering geometrically-verified feature-to-landmark matches, in a similar manner to [29]. An inverted file system is employed similar to traditional BOF [5], which accumulates tentative matches based upon a visual word match between an image feature and a scene landmark. Each match is then verified by considering feature-to-landmark matches for the most frequently co-occurring landmark. Based on this verification, for each landmark p_i in the scene, the likelihood that the landmark has been found in the image is computed as v_i. Finally, u_x is then computed as a weighted sum over all landmarks:

$$u_x = \frac{\sum_{i=0}^{i=I} v_i}{\sum_{i=0}^{i=I} r_i} \tag{1}$$

The term r_i in the denominator is the rate of occurrence of landmark p_i in scene x and exists to normalise the scene score.

Each v_i is computed by considering the likelihood of landmark p_i conditional on three elements. First, the visual word assigned to the feature tentatively matching the landmark. Second, the visual word assigned to the feature tentatively matching the co-occurring landmark. Third, the spatial word assigned to the geometric relationship between these two features. In this way, v_i can be computed by considering how frequently and

discriminatively these visual words are actually assigned to the respective landmarks, based on all previous instances of the landmarks, and how frequently and discriminatively they are aligned in the manner defined by the spatial word.

Let us define the Boolean variables P and Q^* to represent the presence, 1, or absence, 0, of landmarks p and q^* in the image, respectively. As before, q^* is the landmark that co-occurs most frequently with landmark p. If a tentative feature-to-landmark match exists to both p and q^*, then v is computed for p, otherwise it is set to zero. This is achieved by considering the visual words, g^p and g^{q^*}, assigned to the tentative feature-to-landmark matches to p and q^*, respectively, together with the spatial word h^{pq^*} assigned to the geometric relationship between the two tentative features. v is then computed via a Bayesian calculation as follows:

$$v = \frac{p(g^p, g^{q^*}, h^{pq^*} \mid P=1)p(P=1)}{\sum_{P=0,1} p(g^p, g^{q^*}, h^{pq^*} \mid P)p(P)} \tag{2}$$

Here, $p(P=1)$ is set to r_p, occurrence rate of landmark p in the scene, and hence $p(P=0)$ is set to $1 - r_p$.

Given that landmark p is present, and to account for the fact that p and q^* do not always co-occur, we marginalise over Q^* the probabilities that visual word g^{q^*} and spatial word h^{pq^*} occur together in an image sporadically:

$$\begin{aligned} &p(g^p, g^{q^*}, h^{pq^*} \mid P=1) = \\ &p(g^p \mid P=1) \sum_{Q^*=0,1} p(g^{q^*}, h^{pq^*} \mid Q^*, P=1)p(Q^* \mid P=1) \end{aligned} \tag{3}$$

These probabilities are all computed statistically by considering the visual word distributions of p and q^*, and the spatial word distributions between p and q^*.

For the case when landmark p is not present, we marginalise over Q^* the probability that visual word g^{q^*} occurs sporadically:

$$\begin{aligned} &p(g^p, g^{q^*}, h^{pq^*} \mid P=0) = \\ &p(g^p \mid P=0)p(h^{pq^*} \mid P=0) \sum_{Q^*=0,1} p(g^{q^*} \mid Q^*, P=0)p(Q^* \mid P=0) \end{aligned} \tag{4}$$

For a given landmark, the probability that visual word g^p occurs sporadically in an image is computed by considering its statistical rate of occurrence across the entire database, w_{g^p}, and the average number of features that occur in an image, n:

$$p(g_p \mid P=0) = 1 - (1 - \frac{1}{w_{g^p}})^n \tag{5}$$

In a similar manner, we compute the statistical rate of occurrence, $w_{h^{pq^*}}$, of spatial word h^{pq^*}, as the inverse of the total number of spatial words in the dictionary, leading to:

$$p(h^{pq^*} \mid P=0) = 1 - (1 - \frac{1}{w_{h^{pq^*}}})^n \tag{6}$$

For the co-occurring landmark q^*, we compute as $(w_{g^{q^*}} \mid P=1)$ the sporadic assignment probability of word g^*, conditional on the presence of landmark p, leading to:

$$p(g^{q^*} \mid Q^*=0, P=1) = 1 - (1 - \frac{1}{w_{g^{q^*}} \mid P=1})^n \tag{7}$$

With v computed for each landmark in Eq. 2, we then return to Eq. 1 to compute the score for each scene. The top scene is then returned as the matched scene, and is passed on to the online learning stage.

4 Online Learning

After the recognition stage has output the most likely scene, training then proceeds to update the distribution of landmarks in this scene. First, the visual word distributions and co-occurrence statistics of existing landmarks are updated. Whilst the landmark matching in Section 3.2 is typically precise enough to allow for effective scene recognition, there typically remain some false positive feature-to-landmark matches and a more robust, if less efficient method, is necessary to reliably update the landmark properties. Furthermore, false negative matches are particularly frequent, yet it is important to update these landmarks too.

We therefore use the method described in Section 2.1 and match features in the new image to features in a defined history of images, consisting of those that have previously been matched to the scene. During this process, new landmarks that have not been accounted for are then added to the scene. These can be landmarks that were previously occluded, or those that have unstable keypoints. In our experiments, because each scene was only visited a small number of times, this history is defined as all of the images accumulated thus far. However, in practice, it would be limited depending on available resources. It should be noted that the "online" is in reference to a sequentially updated model of each scene, but this could also be computed offline, or on a parallel thread. The scene recognition time is dependent upon the number of scenes in the database, whereas the online learning time is dependent upon the number of landmarks in a scene. Thus, for very small databases, online learning may be more time consuming than recognition.

Together with updating visual word distributions and acquiring new landmarks, the expected occurrence rates of landmarks can also be updated after every scene match, based upon temporal data. This is to account for long-term scene dynamics that may introduce landmarks to, or withdraw them from, a scene. Landmarks that are observed more recently are more likely to appear again in the next acquired image, should have a greater impact in both the BOF filtering stage and the geometric verification stage. Those landmarks that have not been observed for a period of time should be gradually filtered out of the system. The occurrence rate, r, of landmark p in a scene, is continually updated to reflect the likely presence of the landmark the next time the scene is observed.

We define t_T as the time period prior to current time t_0 over which a landmark's occurrence rate is evaluated. r_{t_0}, the rate at time t_0, is determined as a weighted average of the rates across images acquired during time $t_0 \ldots t_T$, with exponential weighting giving more importance those more recently-acquired images:

$$r_{t_0} = \frac{\sum_{k=1}^{k=T} r_{t_k} \, e^{\frac{-(t_k - t_0)^2}{2\sigma^2}}}{\sum_{k=1}^{k=T} e^{\frac{-(t_k - t_0)^2}{2\sigma^2}}} \tag{8}$$

Here, r_t is the occurrence rate at time t, and σ is set such that the exponential weighting at time t_T is 0.01. Determining a suitable value for t_T is important in ensuring that the occurrence probability of the landmark is updated appropriately. Assigning t_T to be too small may cause any unusual absences or presences of the landmark in very recent images to have an undesired effect on the occurrence rate. Assigning a value that is too large will include too much historical behaviour and not enough from the more recent images. As such, we compute t_T by ensuring that the standard deviation of the occurrence rate across all images captured within time $t_0 \ldots t_T$ is within an acceptable level:

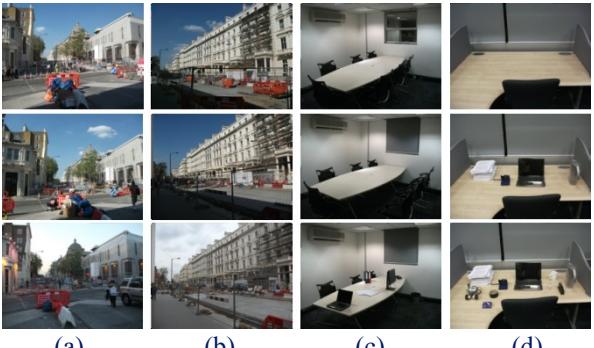
$$t_T = t_k \quad \text{s.t.} \quad \sqrt{\frac{\sum_{k=1}^{T}(r_{t_k} - r_{\text{ave}})^2}{T}} < \sigma_{\max}, \quad T > 3 \tag{9}$$

where r_{ave} is the average rate across all images within the timeframe $t_0 \ldots t_T$. $\sigma_{\max}$ was set at 0.4 and determined heuristically for best results. We require T to be greater than 3 so that there is sufficient data from which to compute a reliable standard deviation.

5 Results

In order to demonstrate the performance of the system across a wide range of scene types, experiments were conducted on both an indoor and an outdoor database. For the outdoor scenes, the first set of training images were captured at discrete locations along a path through a busy town centre, to encapsulate short-term dynamic behaviour of pedestrians and cars, together with longer-term behaviour such as building works and parked cars. The total length of the tour was 1.6 km, encompassing 1000 discrete places. Subsequent tours then followed a similar path and captured ~1000 images per tour, during which recognition and online learning was performed. A total of 8 tours were recorded over a period of a week, with the first 2 defining the initial training set in order to instantiate a set of landmarks for each place.

The indoor dataset was of shorter length, but was generated with the aim of testing the system to more dramatic long-term dynamic behaviour. 200 discrete places were captured over a path length of 200m, with subsequent tours capturing ~200 images per tour. A total of 12 tours were recorded, however, after 3 tours and again after 8 tours, significant structural rearrangements were made to all scenes where possible. Such rearrangements included moving furniture and objects, and opening or closing doors and blinds. See Figure 1 (c) and (d) for examples.



(a) (b) (c) (d)

Figure 1: Images representing places of interest along a tour of the environment. (a) and (b) are outdoor scenes influenced by dynamics such as pedestrians, cars and weather conditions. (c) and (d) are indoor scenes influenced by manually-induced dynamics such as the rearrangement of furniture. These rearrangements took place after the 3rd and 8th tour.

5.1 Recognition

We evaluated our method against recent work in BOF image retrieval [18]. Here, visual word assignments vote for adjacent words in feature space, to reduce the effect of

quantisation. Candidate images from a database are retrieved by computing the similarity between word distribution vectors, and geometric verification prunes out false positive matches. We implemented two applications of this method. In *Application A*, a single image is stored to represent each scene in the database, as standard. In *Application B*, matched images are retained in the pool of images representing the scene, and each image is considered in the image matching. This is the "online learning" equivalent for image retrieval, as features that are introduced into the scene are remembered in the new image. For all the methods, we use a value of $k = 10$ to define the number of images returned from the BOF similarity measure that are passed on for geometric verification. Figure 2 compares the mean precision for our method and the image retrieval method, for both the outdoor and indoor datasets. After the first two tours of the path, which were used to build the scene models, the mean precision was computed for each subsequent tour. Each scene match was then used to update the respective scene model, or add an image to the scene's image pool in Application B of [18].

For the outdoor dataset in Figure 2 (a), our proposed method initially performs poorly compared to the image retrieval approach. This is due to basing the scene models only on a training set of two images per scene. However, as further images are acquired and the model better reflects the landmark occurrence rates and the distributions of visual words and spatial words, our method outperforms the image retrieval method. Application B of [18] initially performs well, but as images from false positive scene matches are added to the scene's image pool, the performance drops dramatically because of the introduction of an entire false positive image. This is not an issue with our method, because updating landmarks based upon false positive scene matches only marginally modifies the expected occurrence rate of landmarks or their visual words and spatial words. As more true positive matches are acquired, the effect of these earlier false positives then becomes negligible.

The performance of our technique under dramatic structural changes can be seen in the results for the indoor dataset in Figure 2 (b). After each restructuring of the environment,

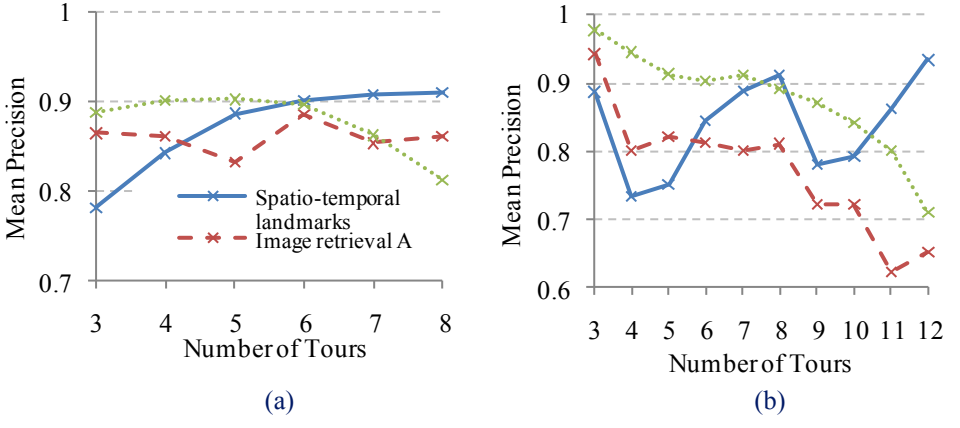


Figure 2: Comparison of the recognition precision of our technique with that of the image retrieval method in (a) outdoor scenes and (b) indoor scenes. Structural changes are manually made to the indoor scenes after 3 images are acquired, and again after 8.

the performance naturally drops across all methods. Application A of [18] continues to drop in performance with each structural change. Because in Application B the system has a recently updated image to represent the scene, its performance does not drop as dramatically after each structural change, but nonetheless its performance drops

sequentially as with the outdoor dataset. However, whilst our method again starts with the poorest performance, it is able to improve its precision after each structural change, resulting in the best performing method in the long term.

5.2 Efficiency

The average processing time required for the matching stage was recorded for the outdoor dataset in the final tour, after the first 7 tours had been processed. Figure 3(a) compares the required processing time of our technique to that of [18], showing a superior efficiency for our framework. Typically, only a small proportion of features in an image are matched across other images of the same scene, and hence our method requires a smaller number of landmarks to be matched to than the number of features in image retrieval methods. Furthermore, geometrically verifying each landmark by only one co-occurring landmark is far more efficient than processing RANSAC-based affine transformation verification as in [18].

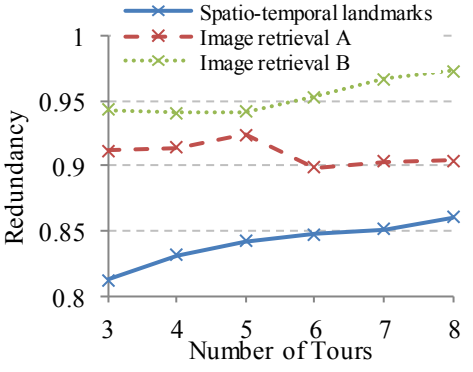
We also investigated the redundancy in our database compared with that of the image retrieval approach. Redundancy is an important concept for memory efficiency and is a measure of how much "waste" the database contains. We define the redundancy of a scene's representation in a database as:

$$\text{scene redundancy} = \frac{\text{number of unmatched database features stored for scene}}{\text{total number of database features stored for scene}} \tag{10}$$

For the computation of scene redundancy above, we substitute the word "features" for "landmarks" to transfer the meaning to our scene models. We account for the storage of landmarks that consist of more than one visual word by incrementing the number of database landmarks accordingly.

Figure 3 (a) shows the database redundancy as the number of tours increases. Naturally, databases for image retreival methods exhibit significant redundancy because most local features are inherently unstable. However, since our framework only stores those features which i) represent static objects or structures and ii) represent the most stable features, the average scene redundancy can be significantly reduced. As further images are acquired, our redundancy increases slightly in our method, but levels out as the acquisition rate of new landmarks reduces. By the time 8 tours of the environment are conducted, most stable features have already been extracted as landmarks, and the remaining landmarks are acquired largely from dynamic objects. We could reduce this redundancy even further by eliminating those landmarks from the database that have not occurred for a period of time, and hence are unlikely to be matched to again, or by limiting the number of stored landmarks to those which have an occurrence rate above a threshold.

Figure 3 (b) shows the mean processing time per recognition, excluding the feature extraction and visual word indexing, and the mean memory requirement per scene, again after 8 tours have been processed. Whilst the BOF filtering stage is typically very similar across all techniques, the geometric verification is significantly more efficient in our method due to the avoidance of an expensive RANSAC-based algorithm. By learning which landmarks co-occur together most frequently, geometric verification of each tentative landmark match can be done with a single co-occuring landmark. An image retrieval approach has no knowledge of co-occurrence rates and has to verify each feature using a number of other occuring features. Application B of [18] has a particularly high redundancy due to false positive scene matches accumulating images in the scene's image pool, which are subsequently never again matched to.



(a)

Method	Mean memory per scene (kB)	Mean processing time for recognition (ms)
Spatio-temporal landmarks	1.29	3.1
Image retrieval A	6.34	8.7
Image retrieval B	50.16	9.8

(b)

Figure 3: Comparison of (a) the database redundancy and (b) the computational requirements of recognition for our technique with that of the traditional image-retrieval approach, for the outdoor database, after 8 tours. The processing time excludes feature extraction and visual word assignment, which is consistent across all techniques.

6 Conclusions

In this paper we have presented a new framework for place recognition and online learning of dynamic changes to scenes. Modelling a place as a distribution of real-world landmarks enables a more robust understanding of the expected distribution of local features in an image, both in terms of descriptors and spatial relationships. By learning which landmarks co-occur most frequently, each landmark can be efficiently geometrically verified by using only a single co-occurring landmark. Furthermore, dynamic elements in a scene can be incorporated online by introducing new landmarks into a scene and filtering out old landmarks. Results show improvements in the long-term recognition precision and efficiency over image-retrieval techniques. Additionally, by storing only stable landmarks in our database, the redundancy in the database, and hence memory requirements, can be dramatically reduced.

References

[1] M. Cummins and P. Newman. FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance. In *International Journal of Robotics Research*, 27(6), 2008.

[2] E. Johns and G.-Z. Yang. Global Localization in a Dense Continuous Topological Map. In *Proc. ICRA*, 2011.

[3] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz and R. Szeliski. Building Rome in a Day. In *Proc. ICCV*, 2009.

[4] K. Heath, N. Gelfand, M. Ovsjanikov, M. Aanjaneya and L. J. Guibas. Image webs: Computing and exploiting connectivity in image collections. In *Proc. CVPR*, 2010.

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

[6] H. Bay, A. Ess T. Tuytelaars and L. Van Gool. SURF: Speeded Up Robust. In *Computer Vision and Image Understanding* 110(3), 2008.

[7] Y. Cao, C. Wang, Z. Li, L. Zhang and L. Zhang. Spatial-Bag-of-Features. In *Proc. CVPR*, 2010.

[8] K. Ni and K. Kannan. Epitomic Location Recognition. In *Proc. CVPR*, 2008.

[9] E. Arnaud, E. Delponte, F. Odone and A. Verri. Trains of keypoints for 3D object recognition. In *Proc. ICPR*, 2006..

[10] R. Hartley and A. Zisserman. Multiple View Geometry in Computer Vision. Cambridge University Press, 2003.

[11] D. Nister and H. Stewenius. Scalable Recognition with a Vocabulary Tree. In *Proc. CVPR*, 2006.

[12] M. Marszalek, C. Schmid. Spatial Weighting for Bag-of-Features. In *Proc. CVPR*, 2006.

[13] F. Orabona, L. Jie and B. Caputo. Online-Batch Strongly Convex Multi Kernel Learning. In *Proc. CVPR*, 2010.

[14] J. Luo, A. Pronobis, B. Caputo and P. Jensfelt. Incremental Learning for Place Recognition in Dynamic Environments. In *Proc. IROS*, 2007.

[15] X. Li, C. Wu, C. Zach, S. Lazebnik and J.-M. Frahm. Modeling and Recognition of Landmark Image Collections Using Iconic Scene Graphs. In *Proc. ECCV*, 2008.

[16] A. Pronobis and B. Caputo. Confidence-based cue integration for visual place recognition. In *Proc. IROS*, 2007.

[17] O. Chum, J. Philbin, J. Sivic, M. Isard and A. Zisserman. Total Recall: Automatic Query Expansion with a Generative Feature Model for Object Recognition. In *Proc. ICCV*, 2007.

[18] 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.

[19] H. Jegou, M. Douze and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proc. ECCV*, 2008.

[20] A. Mikulik, M. Perdoc, O. Chum and J. Matas. Learning a Fine Vocabulary. In *Proc. ECCV*, 2010.

[21] Y.-T. Zheng, M. Zhao, Y. Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T.-S. Chua and H. Neven. Tour the World: building a web-scale landmark recognition engine. In *Proc. CVPR*, 2009.

[22] G. Schindler, M. Brown and R. Szeliski. City-Scale Location Recognition. In *Proc. CVPR*, 2007.

[23] N. Vasconcelos. On the complexity of probabilistic image retrieval. In *Proc. ICCV*, 2001.

[24] E. Johns and G.-Z.-Yang. Scene Association for Mobile Robot Navigation. In *Proc. IROS*, 2009.

[25] Y. Li, N. Snavely and D. P. Huttenlocher. Location Recognition using Proritized Feature Matching. In *Proc. ECCV*, 2010.

[26] S. Se, D. Lowe and J. Little. Vision-based mobile robot localization and mapping using scale-invariant features. In *Proc. ICRA*, 2001.

[27] F. Lik and J. Kosecka. Probabilistic location recognition using reduced feature set. In *Proc. ICRA*, 2006.

[28] Y. Zhang, Z. Jia and T. Chen. Image Retrieval with Geometry-Preserving Visual Phrases. In *Proc. CVPR*, 2011.

[29] E. Johns and G.-Z. Yang. From Images to Scenes: Compressing an Image Cluster into a Single Scene Model for Place Recognition. In *Proc. ICCV*, 2011.