Dynamic Scene Models for Incremental, Long-Term, Appearance-Based Localisation

Edward Johns and Guang-Zhong Yang
The Hamlyn Centre, Imperial College London

Abstract— **In this paper we present a new appearance-based mapping system that is able to deal with dynamic elements in a scene. By independently modelling the properties of local features observed in a scene over long periods of time, we show that feature appearances and geometric relationships can be learned more accurately than when representing a location by a single image. We also present a new dataset consisting of a 6 km outdoor path traversed once per month for a period of 5 months, which contains several challenges including short-term and long-term dynamic behaviour, lateral deviations in the path, repetitive scene appearances and strong illumination changes. We show superior performance of the dynamic mapping system compared to state of the art techniques on both ours and existing datasets.**

I. Introduction

State-of-the-art appearance-based localisation techniques typically adopt an image retrieval approach [5], [6], [2]. Here, each location in a topological map is represented by one or more images in a database, and the most similar database image to a new query image returns the most likely location of the robot. Recent developments in dictionary-based image retrieval techniques have introduced scalable systems that can efficiently query databases of the order of millions [7]. Images are typically compared by their distributions of *visual words* [14], with a dictionary of quantised feature descriptors enabling fast feature comparisons. Geometric constraints are either embedded within the dictionary itself [18], [17] or imposed subsequently [12], [15]. In small environments, location recognition can be achieved by using the distribution of visual words alone [5], whereas at a larger scale the geometric verification is necessary [4].

Whilst these methods have yielded promising results in retrieval engines, where the goal is to return all images of the same scene from a database [3], they are perhaps less suitable for place recognition in environments when the appearance of each location may change over time. Typically, a large number of database images are stored for each location to cover all possible states of the scene [13], [19]. However, this results in significant database redundancy due to dynamic scene elements that were present when the database image was captured, but will not appear again in a query image of that same scene, even if captured from precisely the same location. We categorise these dynamic elements into three types:

(i) Short-term dynamics: objects that may temporarily occlude a portion of the scene, such as pedestrians or vechicles.
(ii) Long-term dynamics: structural elements that may appear or disappear over time, such as resulting from building renovations or the seasonal effects on trees.
(iii) Cyclic dynamics: image features that may repeatedly appear and disappear, such as resulting from illumination changes over the course of a day.

In this paper, we deal with all three types of dynamic behaviour with the aim of learning the static properties of a scene that will be repeatable even under strong dynamic conditions.

One further issue with applying image retrieval techniques to location recognition is that, given a feature in a query image, it is necessary to predict the appearance of a matching feature in the database. This is due to the discrete nature of the visual dictionary, and the effect of quantisation often requires significant geometric verification on candidate database images to ensure a correct match [15]. Typically, generic assumptions are made about likely visual word distributions [12][10][7][8] based on a single example, rather than observing the explicit behaviour of each individual scene element over multiple examples.

In this paper, we propose a new framework for appearance-based mapping that applies a more theoretical analysis of both the effect of dynamics on the presence or absence of local features in a scene, and the implicit appearance of the features unavailable from a single feature example. By modelling image features as representations of real-world structural elements and learning the individual properties of each one, unstable features due to dynamic scene elements are filtered out, and stable features from static scene elements are given more importance in the matching process.

Furthermore, we introduce a new long-term dataset acquired along a path of 6 km traversed monthly over a period of 5 months. The dataset exhibits dramatic dynamic behaviour of all three types, together with significant lateral deviations in the path and repetitive scene appearances. Figure 1 demonstrates some sample images from the dataset and examples of difficult cases that are addressed with our proposed approach.

A. Related work

Appearance-based mapping has seen a wide range of applications in the mobile robotics community for loop closure detection in a Simultaneous Localisation and Mapping (SLAM) framework [2]. A probabilistic approach to appearance-based SLAM, denoted FAB-MAP, was presented

(a) Short-term Dynamics

(b) Natural Long-term Dynamics

(c) Man-made Long-term Dynamics

(d) Cyclic Illumination Dynamics

(e) Lateral deviations in Path

(f) Repeating Scenes

Fig. 1. Examples of the challenges with location recognition in our long-term dataset. Our dynamic mapping system addresses all of these challenges and is able to learn appearance-based models of locations that adapt to dynamic elements in the scene.

in [5] whereby a Chow-Liu tree structure models the correlations between observed visual words. Incremental creation of the visual vocabulary was addressed in [1] to fit the dictionary to the observed scenes rather than making prior generic assumptions about visual word distributions. The FAB-MAP method was improved in the PIRV-Nav framework [9] whereby local features tracked between consecutive frames enable filtering out of dynamic bodies and unstable features.

The common theme among most appearance-based mapping systems is the image retrieval approach, where each location is described by an image. In these systems, any updating of a location to adapt to long-term dynamic scene elements requires either the replacement of the original image, or the accumulation of multiple images per location. This does not address the issue of long-term dynamics, because old images representing an outdated scene appearance may still be falsely matched to. Attempts have been made to reduce the number of accumulated images representing each location [16], however, only the Bag-Of-Words vector is modelled, without any geometric information which is necessary for large-scale environments. In this paper, we show that by learning both feature appearances and geometric relationships over a number of training images, a greater understanding of the expected appearance of a scene is achieved. Note that we do not consider odometry or filtering as in [6], [11], [5] and concentrate solely on the computer vision problem of matching a query image to a location in a database.

II. Framework Overview

Our model which we apply to the world is based on the concepts of *scenes* and *landmarks*. Scene y_i represents discrete location i in a topological map. In our experiments, this map consists of a single path, but is extendable to any arbitrary structure. A landmark represents a real-world point in three-dimensional space, that causes a feature observation in an image of the respective scene. The m^{th} landmark in scene y_i is denoted x_i^m. Landmarks are formed by standard geometric feature matching [15] across adjacent images along a path, where a landmark is recorded in scene y_i if the respective feature is matched in either scene y_{i-1} or scene y_{i+1}. As such, the same real-world point can be represented by several landmarks if that point is observable from many images along the path. Each landmark is represented by a set of visual words; every time the landmark is observed in an image, the visual word for the associated feature is added to this set, if it is not already present.

A. Scene similarity score

Given a query image $\mathcal{Z}$, a scene similarity score is computed for each database scene y_i. This is achieved by accumulating the probability of occurrence of each of y_i's landmarks in the query image, and normalising by the expected number of landmarks that are observed in y_i:

$$s(y_i, \mathcal{Z}) = \frac{\sum_{x_i^m \in \mathcal{X}_i} p(x_i^m|\mathcal{Z})}{\sum_{x_i^m \in \mathcal{X}_i} p(x_i^m|y_i)} \quad (1)$$

Here, $\mathcal{X}_i$ is the set of landmarks in y_i and $p(x_i^m|\mathcal{Z})$ is the probability that landmark x_i^m is observed in the query image $\mathcal{Z}$. $p(x_i^m|y_i)$ is the occurrence probability of landmark x_i^m in scene y_i, computed by dividing the number of times x_i^m is observed at y_i by the number of times y_i has been visited. By relaxing the full joint distribution of landmarks across the entire scene and simply summing probabilities for individual landmarks, the model is flexible and tolerant of dynamic bodies that may occlude portions of a scene. The normalising denominator in Equation 1 enables location recognitions at locations where features are sparse and only few matches may be achieved.

This scene similarity score is akin to the standard image retrieval score where the number of visual word matches is normalised and each weighted by a *tf-idf* factor [14]. However, as shown in the next section, our weighting factor is probabilistic and incorporates local geometry within the image, whereas the *tf-idf* factor only relates to the feature descriptor.

B. Landmark observation probability

We define the *landmark evidence* z_i^m as the set of features in $\mathcal{Z}$ that support the hypothesis that x_i^m exists in the query image. A feature in the query image is added to z_i^m if its visual word is in the set of visual words already represented by x_i^m from the training images. Given this evidence, the probability that this feature does indeed represent x_i^m, is computed as follows:

$$p(x_i^m|z_i^m) = \frac{p(z_i^m|x_i^m)p(x_i^m|y_i)p(y_i)}{\sum_{y_j}\sum_{x_j^m \in \mathcal{X}_j} p(z_j^m|x_j)p(x_j|y_j)p(y_j)} \quad (2)$$

The numerator in Equation 2 is the likelihood that landmark x_i^m is observed in the image, and the denominator is the summation of these likelihoods over all landmarks, to yield a probability. Each landmark x_i^m is therefore evaluated not only by its own evidence, z_i^m, but also by the evidences for all other landmarks.

C. Feature evidence likelihood

We now consider the computation of $p(z_i^m|x_i^m)$. This is achieved by considering the geometric relationships between z_i^m and the other landmarks in the scene. These relationships are defined by the relative location of features within a regular grid of square cells overlayed on the image. We use a grid of 100-by-75 cells for 640-by-480 pixel images.

The feature evidence likelihood is then computed by considering a simple Bayesian network, with a parent node representing the presence or absence of landmark x_i^m in the query image, and child nodes representing the presence or absence of x_i^m's neighbouring landmarks $x_i^n \in \bar{\mathcal{X}}_i^m$ in scene y_i, where $\bar{\mathcal{X}}_i^m$ is the set of all other landmarks in the scene. Given the parent node, we relax the structure into a naive Bayesian network by assuming neighbouring landmarks to be independent, to avoid over-fitting of the network parameters. In this way, for each landmark x_i^m, we only consider co-occurrence statistics between x_i^m and x_i^n, and not those between each x_i^n.

For a feature that is a candidate match to landmark x_i^m based on visual words, the visual word assigned to the feature is denoted ϕ_i^m, and similarly ϕ_i^n for a feature putatively matching a neighbouring landmark. The spatial relationship between these two features is then denoted as *spatial word* $\delta_{x_i^n x_i^m}$.

The likelihood of the feature evidence is now computed as follows:

$$p(z_i^m|x_i^m) = p(\phi_i^m|x_i^m) \prod_{x_i^n \in \bar{\mathcal{X}}_i^m} p(\phi_i^n|x_i^n)p(\delta_{x_i^m x_i^n}|x_i^n, x_i^m)p(x_i^n|x_i^m) \quad (3)$$

Here, $p(x_i^n|x_i^m)$ is the probability that landmark x_i^n is observed in an image given that x_i^m is observed, i.e. the co-occurrence rate of these two landmarks in the training images. The values of $p(\phi_i^n|x_i^n)$ and $p(\delta_{x_i^m x_i^n}|x_i^n, x_i^m)$ are simply drawn from the normalised frequencies of visual word assignments to x_i^m and spatial word assignments to the pair x_i^m, x_i^n.

All parameters are learned from the training images representing a landmark. The quantisation of images in both feature space and image space thus allows parameters to be stored in a file index relating probabilities for different combinations of visual and spatial words, for rapid calculation of Equation 3.

The value of $p(\phi|x)$ is only stored for a given landmark and visual word combination if it is greater than 0.01, otherwise it is assumed to be zero. In this way, a more efficient inverted file index is employed by linking each visual word to a set of landmarks containing only those landmarks which are likely to be assigned to visual word ϕ. Similarly, spatial words are only linked to landmark co-occurrences if the value of $p(\delta|x_1, x_2)$ is greater than 0.01.

D. Recognition process

The contents of the neighbouring landmark set $\bar{\mathcal{X}}_i^m$ could feasibly incorporate all landmarks in scene y_i other than landmark x_i. Whilst this provides the maximum evidence from which to compute $p(x_i^m|\mathcal{Z})$ and is likely to achieve greatest recognition performance, it would be an unnecessary use of computational time if a confident location recognition can be performed with a smaller set. As such, the recognition process proceeds iteratively, starting with only one neighbouring landmark in the set, and increasing the set by one after each iteration. The process stops when a database location has been found with sufficiently high confidence.

First, we extract peaks in the distribution of scene scores $s(y, \mathcal{Z})$ by use of non-maximal suppression. Scene y_i is retained if, and only if, its score is greater than those for scenes y_{i-1} *and* y_{i+1}. This is to reduce the effect of perceptual aliasing whereby adjacent scenes along a path appear similar. Eliminating non-maximal locations allows computation of a confidence level that the query image depicts scene y; without non-maximal suppression, this confidence may never converge to a sufficient level due to a wide distribution of high scores centred on y.

At each loop of the iterative recognition process, we take the highest and second-highest scores across all locations that are locally maximal, denoted s_1 and s_2 respectively. The level of confidence that the scene scoring s_1 is the correct match, is then defined as:

$$c = \frac{s_1}{s_2} \quad (4)$$

If this confidence is less than threshold c_{min}, then a further neighbouring landmark is included in each set $\bar{\mathcal{X}}_i^m$ for all landmarks, and the process repeats until a sufficient confidence is achieved. Due to the probabilistic nature of Equation 2, if s_1 relates to the true location, then both s_1 will increase and s_2 will decrease as further neighbours are incorporated, allowing for rapid convergence. Neighbouring landmarks are accumulated in $\bar{\mathcal{X}}_i^m$ in order of their co-occurrence rate $p(x_i^n|x_i^m)$, such that those neighbouring landmarks that are more likely to verify the presence of x_i are considered first.

Determining the value of c_{min} is a compromise between efficiency and recognition accuracy. We choose a value of 25 for our experiments, which typically results in between one and five neighbouring landmarks being included in the set $\bar{\mathcal{X}}_i^m$. Increasing c_{min} any further yields little recognition improvement but increases computational time dramatically.

III. Experiments

A. Long-term dataset

In order to test our system's performance against long-term dynamic effects, we introduce a new dataset consisting of GPS-tagged images captured along a 6 km path. Images were manually captured with a standard camera whilst walking along the path, at intervals of around 3 metres. One tour was completed per month over a period of five months to give a total of 6 image sequences. Tours were completed at varying times of day and under varying weather conditions. The route is divided into two sections of roughly equal length, with the first half through a park to assess seasonal effects on the trees and vegetation, and the second half through an urban centre that was undergoing significant structural changes and subject to high short-term dynamic activity from pedestrians and cars. Tours of the path diverge laterally from one another by up to 3 metres. Figure 1 demonstrates some of the dynamic behaviour exhibited by the dataset.

Learning of the parameters in Equation 3 is achieved firstly by detecting tracked landmarks across sequential images in the first tour of the training set. Incremental learning is then subsequently guided by the ground-truth GPS locations of training images, with features from further training tours matched to those from the first tour, to update the landmark properties accordingly. Any new landmarks, which have been detected in later training tours but were previously not detected, are introduced into the map. We do not eliminate landmarks from our map if they do not appear for several training tours. Instead, the value of $p(x|y)$ in Equation 2 is reduced appropriately to reflect a lower likelihood of observing landmark x in scene y.

As an example of how our dynamic system system adapts to dynamic scene elements, Figure 2 shows the same location observed over five months, and the observation likelihoods of landmarks in the as computed after those five months. As each image of the location is captured, the landmarks representing static scene elements dominate and those representing dynamic elements are gradually filtered out.

(a) March (b) April (c) May (d) August

(e) Prior likelihood of landmark observation after 5 months (6 training tours)

Fig. 2. The likelihood of a landmark being observed in a scene is calculated across all available training tours. Dynamic landmarks, such as those caused by moving bodies or tree growth, have a low likelihood, whereas strong structural landmarks, such as those caused by windows, have a high likelihood.

B. Single training tour

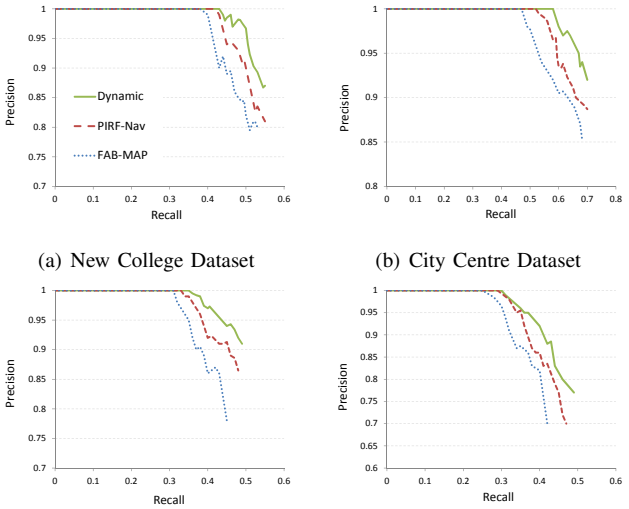
We first evaluated the localisation performance of our system from a single training tour without any incremental learning. Two existing datasets were used together with our own dataset. The New College dataset [5] consists of a 1.9 km path with 2146 images, which several large areas with strong visual repetition. The City Centre dataset [5] consists of a 2 km path with 2474 images, which includes a large number of short-term dynamic bodies such as pedestrians and vechicles

Figure 3 shows the precision-recall performance of our technique (labelled "Dynamic") on all three datasets compared to two competing techniques: FAB-MAP 2.0 [4] and PIRF-Nav[9]. The FAB-MAP system uses the standard Bag-Of-Words model together with a Chow-Liu tree structure to learn co-occurrence statistics of visual words, followed by geometric verification using epipolar geometry. PIRF-Nav adopts a similar approach but also employs local tracking of features to eliminate those which are unstable and likely to be due to dynamic bodies. We use only the image matching stages of FAB-MAP and PIRF-Nav for a fair test, without the filtering component. As such, these experiments address the global localisation problem. However, such a filtering system is simple to append by adjusting the value of $p(y)$ in Equation 2 accordingly.

For the results on our long-term dataset, the first tour

was used as the database tour and subsequent tours of 1 month and 5 months later were used for testing. A correct localisation was recorded if the highest scoring location in Equation 1 was within 10m of the location of the query image.

These results show the benefit of our system in its base form, without any incremental learning, using only the probabilistic framework and scene similarity score to query the database. Even with minimal training data, the incorporation of a discriminative probabilistic model in our system proves to be more powerful than the weaker methods of FAB-MAP and PIRF-Nav which are based on image-to-image matching, rather than matching to learned scene models as in our system.



(a) New College Dataset (b) City Centre Dataset

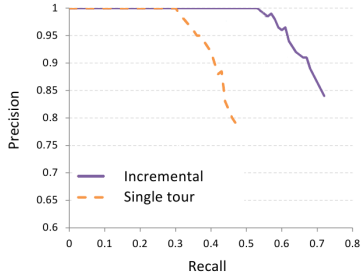
(c) Long-term Dataset after 1 month (d) Long-term Dataset after 5 months

Fig. 3. Recognition performance of our system compared with FAB-MAP 2.0 [4] and PIRF-Nav [9], on three datasets. (c) and (d) show the performance on our long-term dataset using the first tour as the reference tour, and tours after 1 and 5 months respectively as the testing tours.

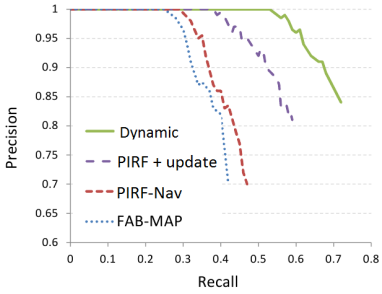
C. Incremental Learning and Dynamic scenes

The final experiments were carried out to incorporate the learning of parameters in Equations 2 and 3, as dynamic elements affect the appearance of the dataset over time. Figure 4 (a) shows the performance of our framework with and without the incremental parameter learning. The full system with incremental learning far outperforms the competing methods in Figure 4 (b). Here, we include an additional competing method in [9] whereby the latest image of each scene is stored rather than the image from the first tour, and we denote this methodd "PIRF-Nav + update". This is the image retrieval approach to dealing with long-term dynamics without having to store multiple images per location, but as can be seen from the results, the ability of our system to deal with each landmark's individual landmark properties still yields superior results.

Our overall system takes, on average, 237 ms to perform a localisation, which is comparable to the competing techniques despite our system having a much deeper model representing each location. The efficiency comes from the discretisation of image space, allowing for an inverted index to address co-occurring landmark pairs of a particular spatial geometry, and also from only evaluating small subsets of neighbouring landmarks in Section IIB, rather than computing expensive geometric transformations over the entire image, as is required in [4].



(a) Effect of incremental learning over a period of 5 months



(b) Comparison with competing methods

Fig. 4. Precision-recall performance for (a) two implementations of our system and (b) a comparison with competing methods.

Figure 5 shows sample localisation attempts with two different challenges. In a), our system trained over time filters out long-term dynamic elements in the scene, such as vegetation, and enables a successful recognition, whereas FAB-MAP is confused by dynamic features that have changed over time. In b), our system localises much more accurately whereas FAB-MAP returns a location much further along the path. This demonstrates how discriminative probabilistic models of feature appearances and geometric distributions can help to make fine distinctions between locations of similar appearance.

Finally, figure 6 demonstrates the effect of incremental learning and adaptation to dynamic environments by considering the distribution of successful localisations over a period of time. Due to the lack of rigid structural bodies, the park section of the route suffers from poor recognition performance with only 1 month of learning (1 training tour of the route). However, after 5 months the system is able to give greater importance to more reliable landmarks and achieve many more successful location recognitions.

IV. Conclusions

In this paper we have presented a new framework for appearance-based place recognition and localisation with the

(a) Loop closure with long-term dynamics

(b) Loop closure with repeating scene appearance

Fig. 5. Sample location recognition attempts with a query image captured 5 months after the training image

application of long-term learning of a topological map. We have showed that by learning individual landmark properties, rather than learning entire image properties, a greater understanding of feature appearances can be gained. Furthermore, allowing each landmark to exhibit independent dynamic behaviour allows our system to adapt to long-term dynamic effects in the environment.

REFERENCES

[1] A. Angeli, S. Doncieux, J.-A. Meyer, and D. Filliat. Incremental vision-based topological slam. In *IROS*, 2008.

[2] A. Angeli, D. Filliat, S. Doncieux, and J.-A. Meyer. Visual topological slam and global localization. In *ICRA*, 2009.

[3] Ondrej Chum, Andrej Mikulík, Michal Perdoch, and Jiri Matas. Total recall ii: Query expansion revisited. In *Proc. CVPR*, pages 889–896, 2011.

[4] M. Cummins and P. Newman. Highly scalable appearance-only slam - fab-map 2.0. In *Robotics: Science and Systems*, 2009.

[5] Mark Cummins and Paul Newman. Fab-map: Probabilistic localization and mapping in the space of appearance. *IJRR*, 27:647–665, June 2008.

[6] A. Glover, W. Maddern, M. Milford, and G. Wyeth. Fab-map + ratslam : appearance-based slam for multiple times of day. In *Proc. ICRA*, 2010.

[7] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Improving bag-of-features for large scale image search. *International Journal of Computer Vision*, 87(3):316–336, feb 2010.

[8] E. Johns and G.-Z. Yang. Feature co-occurrence maps: Appearance-based localisation throughout the day. In *Proc. ICRA*, 2013.

[9] A. Kawewong, N. Tongprasit, S. Tungruamsub, and O. Hasegawa. Online and incremental appearance-based slam in highly dynamic environments. *Trans. IJRR*, 2011.

[10] A. Mikulk and M. Perdoch. Learning a fine vocabulary. In *Proc. ECCV*, 2010.

[11] Michael J. Milford. Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights. In *Proc. ICRA*, 2012.

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

(a) 1 Month (1 training tour)

(b) 3 Months (3 training tours)

(c) 5 months (5 training tours)

Fig. 6. Correct localisation distributions after learning for 1 month, 3 months and 5 months. Red indicates a successful localisation, yellow indicates that the best location match was a false positive.

[13] Rahul Raguram, Changchang Wu, Jan-Michael Frahm, and Svetlana Lazebnik. Modeling and recognition of landmark image collections using iconic scene graphs. *Trans. IJCV*, 95(3):213–239, 2011.

[14] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. pages 1470–1477, 2003.

[15] G. Tolias and Y. Avrithis. Speeded-up, relaxed spatial matching. In *Proc. ICCV*, 2011.

[16] M. Milford W. Maddern and G. Wyeth. Towards persistent localization and mapping with a continuous appearance-based topology. In *Proc. IROS*, 2012.

[17] X. Wang, M. Yang, T. Cour, S. Zhu, K. Yu, and T.X. Han. Contextual weighting for vocabulary tree based image retrieval. In *Proc. ICCV*, pages 209–216, 2011.

[18] Y. Zhang, Z. Jia, and T. Chen. Image retrieval with geometry-preserving visual phrases. In *Proc. CVPR*, pages 809 – 816, 2011.

[19] Yan-Tao Zheng, Ming Zhao, Yang Song, Hartwig Adam, Ulrich Buddemeier, Alessandro Bissacco, Fernando Brucher, Tat-Seng Chua, and Hartmut Neven. Tour the world: building a web-scale landmark recognition engine. In *Proc. CVPR*, 2009.