Place Recognition with ConvNet Landmarks: Viewpoint-Robust, Condition-Robust, Training-Free

Niko Sünderhauf, Sareh Shirazi, Adam Jacobson, Feras Dayoub, Edward Pepperell, Ben Upcroft, and Michael Milford
ARC Centre of Excellence for Robotic Vision, Queensland University of Technology, Brisbane QLD 4001, Australia
niko.suenderhauf@roboticvision.org

Abstract—Place recognition has long been an incompletely solved problem in that all approaches involve significant compromises. Current methods address many but never all of the critical challenges of place recognition – viewpoint-invariance, condition-invariance and minimizing training requirements. Here we present an approach that adapts state-of-the-art object proposal techniques to identify potential landmarks within an image for place recognition. We use the astonishing power of convolutional neural network features to identify matching landmark proposals between images to perform place recognition over extreme appearance and viewpoint variations. Our system does not require any form of training, all components are generic enough to be used off-the-shelf. We present a range of challenging experiments in varied viewpoint and environmental conditions. We demonstrate superior performance to current state-of-the-art techniques. Furthermore, by building on existing and widely used recognition frameworks, this approach provides a highly compatible place recognition system with the potential for easy integration of other techniques such as object detection and semantic scene interpretation.

I. INTRODUCTION

Visual place recognition research has been dominated by sophisticated local feature-based techniques such as SIFT and SURF keypoints, hand-crafted global image descriptors such as GIST and bag-of-words approaches. However, as robots operate for longer periods of time in real-world environments, the problem of changing environmental conditions has come to the fore, where conventional recognition approaches fail. To address this problem, a number of techniques have been adopted – matching image sequences [27,29,33,28], creating shadow-invariant images [6,43,25,24], learning salient image regions [26] or learning temporal models that allow the prediction of occurring changes [31]. Recent research has also demonstrated how generic deep learning-based features trained for object recognition can be successfully applied in the domain of place recognition [41,3]. However, all current approaches have introduced at least one significant performance or usability compromise, whether it be a lack of invariance to camera viewpoint changes [27,28], extensive environment-specific training requirements [26], or the lack of appearance change robustness [7]. If visual place recognition is to be truly robust, it must simultaneously address three critical challenges: 1) condition invariance; 2) viewpoint invariance; and 3) generality (no environment-specific training requirements).

In this paper, we present a unified approach that addresses all three of these challenges. We use a state-of-the-art object proposal method to discover potential landmarks in the images. A convolutional network (ConvNet) is then used to extract general purpose features for each of these landmark proposals. We show that the ConvNet features are robust to both appearance and viewpoint change; the first two critical challenges. We also emphasize that landmark proposals require no training and the ConvNet is pre-trained on ImageNet, a generic image database; the third critical challenge. By conducting experiments on a number of datasets we show that our system is training-free in that no task-specific or even site-specific training is required. We also highlight that only single images are required for matching and the system does not require image sequences. We demonstrate the generality of our system on a number of existing datasets and introduce
new challenging place recognition datasets, while comparing to state of the art methods.

The novel contributions of this paper are:

1) A place recognition system that is robust to viewpoint and appearance variation, requiring no environment specific training, and
2) The introduction of new challenging datasets exhibiting extreme viewpoint and appearance variation.

The paper proceeds as follows. Section II provides a brief overview of related work. The method is described in detail in Section III followed by an overview of the four sets of experiments. We present results in Section V before concluding with a discussion and outlining future work.

II. RELATED WORK

The focus of research in place recognition has recently moved from recognizing scenes without significant appearance changes to more challenging, but also more realistic changing environments.

Place Recognition: Methods that address the place recognition problem span from matching sequences of images [27, 17, 40, 33, 29], transforming images to become invariant against common scene changes such as shadows [6, 35, 25, 24, 21], learning how environments change over time and predicting these changes in image space [30, 21, 31], particle filter-based approaches that build up place recognition hypotheses over time [23, 39, 22], or build a map of experiences that cover the different appearances of a place over time [5].

Learning how the appearance of the environment changes generally requires training data with known frame correspondences. [17] builds a database of observed features over the course of a day and night. [30, 31] presents an approach that learns systematic scene changes in order to improve performance on a seasonal change dataset. [26] learns salient regions in images of the same place with different environmental conditions. Beyond the limitation of requiring training data, the generality of these methods is also currently unknown; these methods have only been demonstrated to work in the same environment and on the same or very similar types of environmental change to that encountered in the training datasets.

Although point feature-based methods were shown to be robust against viewpoint changes [7, 8, 38], to the authors' knowledge, significant changes in both viewpoint and environmental conditions have not been addressed in the literature. We show that robustness to variation in both cases can be addressed without site-specific training.

Feature-based Approaches: SIFT [20], SURF [11] and a number of subsequent feature detectors have been demonstrated to display a significant degree of pose invariance but only a limited degree of condition-invariance (illumination, atmospheric conditions, shadows, seasons). Perceptual change as drastic as that illustrated in Fig. 1 has been shown to be challenging for conventional feature detectors [27, 44] and while FAB-MAP [7] is robust with respect to viewpoint changes, it is known to fail in conditions with severe appearance changes [29, 31, 13]. Furthermore, [11, 34] argued that FAB-MAP does not generalize well to new environments without learning a new site-specific vocabulary.

[26] shows that patches and region-based methods within an image can exhibit the same robustness as whole-image techniques while retaining some robustness to scale variation, and thus achieve some of the advantages of both point and whole-image features. However, extensive site-specific training was required. In this research we extend the advantages of region-based methods to address both viewpoint and environmental changes without the requirement for site-specific training.

A commonality between all these approaches is that they rely on a fixed set of hand-crafted traditional features or operate on the raw pixel level. A recent trend in computer vision, and especially in the field of object recognition and detection, is to exploit learned features using deep convolutional networks (ConvNets). The most prominent example of this trend is the annual ImageNet Large Scale Visual Recognition Challenge where for the past two years many of the participants have used ConvNet features [36].

Several research groups have shown that ConvNets outperform classical approaches for object classification or detection that are based on hand-crafted features [19, 37, 10, 12, 35]. The availability of pre-trained network models makes it easy to experiment with such approaches for different tasks: the software packages Overfeat [37] and Decaf [10] or its successor Caffe [16], provide similar network architectures that were pre-trained on the ImageNet ILSVRC dataset [36].

Recent studies have shown that state-of-the-art performance in place recognition can be achieved with networks trained using generic data: [41] demonstrated that ConvNet features representing the whole image outperform current methods for changing environmental conditions. However, whole-image features suffer from sensitivity to viewpoint change. We show that by combining the power of ConvNets and region-based features rather than using whole-image representations, a large degree of robustness to viewpoint change can be achieved.

Consequently in this research we build on the best performing aspects of the state of the art; the recognition performance of ConvNet approaches [41], and the robustness of region-based methods to viewpoint change [25].

III. PROPOSED SYSTEM

In this section we describe the five key components of our proposed place recognition system:

1) landmark proposal extraction from the current image
2) calculation of a ConvNet feature for each proposal
3) projection of the features into a lower dimensional space
4) calculation of a matching score for each previously seen image
5) calculation of the best match

Fig. 2 illustrates our system. The approach has several properties that distinguishes it from previous work:

- The system does not require any task-specific or site-specific training. It uses an off-the-shelf pre-trained con-
volutational network [16] to calculate features and a generic object proposal system to extract landmark proposals from images.

- By using a landmark proposal system (Edge Boxes [45]) that was designed to find arbitrary objects in scenes, we extract recognizable and stable regions in the images that automatically tend to be reliable landmarks.
- Relying on landmark regions rather than the whole image to describe a scene significantly improves the robustness against viewpoint changes or partial occlusions in the scenes.
- ConvNet features have been shown to be more stable against appearance and condition changes than other methods [41]. Since we use these robust features as descriptors for the extracted landmarks, we inherit their robustness against appearance changes such as induced by weather, seasons, or the time of day.
- Since both the landmark proposal and the feature extraction system are used as exchangeable black boxes, any future improvement on these methods by the robotics or computer vision community can be immediately exploited by exchanging the currently used algorithms and network architecture with improved future versions.
- The incorporation of a complete object detection pipeline readily enables future enhancements, such as using the output of the object classifier layer of the ConvNet to discard those proposed landmarks that are likely to contain dynamic objects or scene structures that are otherwise known to be unsuitable as landmarks.

A. Bottom-Up Object Proposals as Region Landmarks

In contrast to previous work we exploit the bottom-up object proposal methods that have been developed in the computer vision community. These methods usually serve as a first step in a general purpose object detection pipeline and extract bounding boxes from an image that are likely to contain an interesting object. R-CNN [12] is a prominent recent example of such a system that extracts approximately 2000 proposal regions per image and passes all of them through a Convolutional Network classifier that determines if an object is present and which of the known classes it belongs to.

To extract landmarks we apply Edge Boxes, an object proposal method developed by Zitnick and Dollár [45]. It has been shown to outperform other recent proposal methods such as BING [4] or Selective Search [42] in the context of object detection and is considered the current state of the art by the computer vision community [15]. In our experiments, we extract 50 or 100 landmark proposals per image.

Edge Boxes mainly relies on the observation that the number of contours that are wholly contained in a bounding box is indicative of the likelihood of the box containing an object. It measures an objectness score by comparing the number of edges within each bounding box with the number of edges passing through it.

B. ConvNet Features as Robust Landmark Descriptors

After landmark proposals have been extracted, we calculate an individual feature vector to describe their appearance. Ideally, this feature vector should be robust against changes in appearance induced by weather, seasonal effects, time of day, and – to a certain degree – occlusions and changes in perspective. While previous work often relied on standard feature descriptors like SIFT or SURF [7] or hand-crafted features to describe landmarks for place recognition, we pass each landmark proposal through a ConvNet to extract a feature vector that is stable under the conditions mentioned above.

We build upon the astonishing results from the computer vision community where ConvNets have been shown to outperform all previous methods in the area of object detection and recognition [19, 37, 10, 12, 35]. In robotics, the first publications that exploited the beneficial properties of these
generic pre-trained features for place recognition appeared very recently [3, 41]. Most available ConvNet frameworks (Overfeat [47], Decaf [10] and its successor Caffe [16]) follow the same principled architecture that was introduced by AlexNet [19]. The network consists of 5 combined layers that perform a convolution, followed by a nonlinear activation function (rectified linear units), and spatial pooling. Three fully connected layers plus a subsequent soft-max layer form the upper parts in the network hierarchy.

[41] has shown that the features of the mid-level features from the 3rd convolutional layer (hereafter called conv3) are highly invariant against the appearance changes that are caused by different weather conditions, seasons or the time of day. They found the cosine distance to be a suitable measure between two of these features. We follow their results and extract a conv3 feature for each of the landmark proposals in an image, utilizing the AlexNet network [19] as implemented by Caffe [16]. It was pre-trained on the ImageNet dataset.

We modified the ConvNet so that only the layers up to conv3 are calculated. This allows us to extract a feature within 15 ms (a speed-up of 6.7×). The landmarks are resized to the expected input size of 231 × 231 × 3 pixels. This procedure has been suggested in the object detection literature [12] and does not seem to degrade the performance.

C. Random Projections for Dimensionality Reduction

The features produced by the conv3 layer of the convolutional network are of size 384 × 13 × 13, i.e. for each proposed landmark we calculate a 64,896 dimensional feature. Calculating the pairwise cosine distances between 50 or 100 of those high-dimensional features per image during the image matching process is an expensive operation.

To avoid the matching process to become too expensive, we apply dimensionality reduction. According to the Johnson-Lindenstrauss-Lemma [18] a set of points in a high-dimensional space can be linearly embedded in a lower dimensional space with high-preserving approximation.

\[(1 - \epsilon)\|u - v\|^2 \leq \|Au - Av\|^2 \leq (1 + \epsilon)\|u - v\|^2\]  (1)

where \(u, v \in \mathbb{R}^n\) are vectors of the original high dimensional space and \(A \in \mathbb{R}^{m \times n}\) is a projection into a lower dimensional space \(\mathbb{R}^m\). We leverage this lemma and apply a Gaussian Random Projection [9] to transfer the original features to a space of much lower dimension. The elements of the projection matrix \(A\) are drawn from a Gaussian distribution. In our experiments, we project the 64,896 original dimensions to 512, 1024, and 4096 dimensions and compare their relative performances.

The Johnson-Lindenstrauss-Lemma is formulated in terms of the euclidean distance between points, but due to the spherical distribution of the features in their high dimensional spaces, we found it to be applicable for the cosine distance too.

D. Image Matching

To determine the similarity between two images \(I_a\) and \(I_b\), we perform matching between all landmark proposals \(I^a_i\) and \(I^b_j\) that were extracted from both images. The landmark matching is performed using a nearest neighbor search based on the cosine distance \(d_{ij}\) of their descriptors (after applying the dimensionality reduction as described above) and applies crosschecking, i.e. only matches are accepted.

As a second step, we score each matched landmark pair \((I^a_i, I^b_j)\) by the similarity of the shape of their bounding boxes. Let \(w_i, h_i, w_j, h_j\) be the width and height of the matched landmark proposals. We then calculate their shape similarity measure \(s_{ij}\):

\[s_{ij} = \exp\left(\frac{1}{2} \left(\frac{|w_i - w_j|}{\max(w_i, w_j)} + \frac{|h_i - h_j|}{\max(h_i, h_j)}\right)\right)\]  (2)

The overall similarity between both images \(I_a\) and \(I_b\) is then calculated as

\[S_{a,b} = \frac{1}{\sqrt{n_a \cdot n_b}} \sum_{ij} 1 - \left(d_{ij} \cdot s_{ij}\right)\]  (3)

where \(d_{ij}\) is the cosine distance between both landmarks and \(n_a, n_b\) are the number of extracted landmarks proposals in both images, including the non-matched ones. In our experiments, \(n_a = n_b = 50\) or 100. The shape similarity score \(s_{ij}\) penalizes false positive matches between landmarks that have a similar conv3 descriptor but are of significantly different shape. Empirically we found this improves overall performance by a small but notable margin. The matched landmarks can still vary in size and aspect ratio, since the appearance-based cosine distance between two landmarks has a bigger influence on the overall similarity score.

To retrieve the best matching database image \(I_a\) for a query image \(I_b\), we search for the database image with the highest similarity score, \(i.e. \arg \max_{a} S_{a,b}\). The matching image is found using the single best match only.

IV. Evaluation and Results

This section describes the conducted experiments and their results. We compare the performance of the proposed system against several state of the art methods using precision-recall plots. Concretely, we compare against the whole image, single ConvNet feature system proposed by [41], the feature-based method FAB-MAP [7] (using OpenFABMAP [14]) and the sequence-based approaches SeqSLAM [27] (using the OpenSeqSLAM implementation [40]) and SMART [33]. FAB-MAP’s vocabulary was trained on the St. Lucia dataset [41]. We found that training a specific vocabulary on the test dataset for each of the individual experiments in the following increased FAB-MAP’s performance slightly, but such an approach would be infeasible in practice.

A. Place Recognition with Viewpoint Variations

In this experiment we evaluate the robustness to viewpoint variations on the Gardens Point dataset that consists of footage recorded by a pedestrian. It exhibits viewpoint variations that occur from walking on the left or right side of a pathway and mild appearance changes mainly caused by dynamic objects

1http://tinyurl.com/stluciadataset
such as pedestrians. The dataset has been used in a number of place recognition evaluations before (e.g. [41]) and is available online. Fig. 4 shows two example frames along with the extracted and matched landmarks. When comparing to the results obtained by the FAB-MAP [7] and the whole-image ConvNet based method of [41], we see in Fig. 3 that our method outperforms both approaches significantly, coming close to perfect performance.

B. Place Recognition with Viewpoint and Appearance Changes – The Mapillary Dataset

These experiments introduce a new dataset exhibiting significant changes in viewpoint and moderate changes in appearance. Mapillary is a crowdsourced alternative to Google Street View. It is a collaborative photo mapping initiative that allows users to upload sequences of GPS-tagged photos and provides an API interface to download these sequences along with their meta data. Since many roads have been mapped by more than one user, Mapillary is an ideal platform to harvest datasets for place recognition under every-day conditions. We downloaded three sequences of images that exerted significant viewpoint changes and make these available to the community along with ground truth data. For example, the images of the Berlin Kurfürstendamm (201 + 222 frames, 3 km) and Berlin Halenseestraße sequence (157 + 67 frames, 3 km) have been recorded by a bike rider on the bike lane and from the upper deck of a tourist bus, or a dashboard camera in a car respectively. This results in a large variation of viewpoint, as can be seen in Figs. 1, 5, and 6. These figures also illustrate some of the landmark proposals that were correctly matched between the image pairs and demonstrate that the matching process is robust against a reasonable amount of scaling, occlusion, illumination changes, and perspective distortion.

As we can see from the precision recall plots in Fig. 7 our proposed method outperforms the approach of [41] (using a single ConvNet feature over the whole image), FAB-MAP [7] (utilizing the OpenFABMAP implementation [14]) and SMART [33] by a large margin. This underlines the increased robustness against viewpoint changes.

The results also illustrate the effect of different parameters on system performance. Using more landmark proposals (i.e. 100 instead of 50) leads to more accurate place recognition. Performance also does not appear to drop significantly when applying dimensionality reduction using Gaussian Random Projections. The differently colored lines represent the exact cosine distance over all 64,896 dimensions (red), and the reduced feature spaces of 4096 (black), 1024 (green), and 512 (cyan) dimensions. As we can see, the performance drops slightly with reduced dimensionality, but the difference between the exact cosine distance and the reduced spaces of 1024 and 4096 is marginal, especially when using 100 landmarks per image.

Another sequence (Malmö John Ericssons Väg, 221 + 378 frames, 4 km) from Mapillary contains only mild viewpoint changes; both traverses were recorded from the same lane on a road. However, the weather conditions between the two recordings were very different, changing from a bright sunny day to a murky overcast day with wet roads after a rain. Fig. 8 shows two example scenes. In the results illustrated in Fig. 9 we see that when using only 50 landmarks, the performance is worse than the whole-image system of [41], for
Fig. 6: The images in the Berlin Halenseestraße sequence have been recorded by a biker riding on the bike lane (left column) and a dashboard camera in the front of a car (right column). The changes in viewpoint are severe but our proposed method is able to extract landmarks and correctly match them between a large number of scenes.

both the exact cosine distance and the reduced feature space of 1024 dimensions. When the number of landmark proposals is increased to 100 the performance is superior to [41], even when applying dimensionality reduction. Our approach outperforms FAB-MAP [7, 14] for all tested parameters.

C. The Library Robot Indoor Dataset

In this experiment we evaluate our approach on a dataset captured by a service robot in a public library. The robot traversed the library once during the day and a second time during the night. Appearance changes were induced by the different external and internal lighting conditions, while people and moved furniture caused structural changes. In contrast to the previous datasets, this experiment is more realistic since the robot did not revisit all places, i.e. there are true negative non-matching scenes. Furthermore, the robot encounters many weakly textured areas and large parts of the environment suffer from perceptual aliasing. The results and example scenes are depicted in Fig. 10. Our approach again outperforms FAB-MAP in this scenario.

D. Quantifying Viewpoint Robustness

a) Real World Scenes: To better quantify the robustness to viewpoint changes, we systematically translate a camera in a complex scene with both close and distant objects. We use the image from the original position as the reference image and move the camera sideways in 10 cm increments under changed illumination conditions. The images from the changed conditions are used as query images, i.e. we attempt to match them with the original image. To make the experiment more significant, we repeat it in 8 different scenes and add 678 unrelated indoor scenes to the dataset. Fig. 11 plots the average accuracy over the sideways camera displacement. Our approach outperforms FAB-MAP [7, 14] under this combination of appearance and viewpoint change.

b) Simulated Viewpoint Changes: For this experiment we use 2289 images of the spring and winter season of the Nordland train dataset (see [31] for an elaborate discussion of the dataset) and crop them to half of their original width. We simulate viewpoint changes between two traverses by shifting
Despite different weather conditions (sunny vs. overcast) the ConvNet landmarks allow for successful place recognition.

The results of this experiment are illustrated in Fig. 12. Our proposed method based on landmark proposals and robust ConvNet features (solid lines) again clearly outperforms the whole-image based method (dashed lines) except for the idealized case of 100% overlap. However, such a scenario is not realistic for real-world applications. Fig. 12 illustrates how the system can pick the same objects as landmark proposals across changing environmental conditions, and match them between images despite the changes in viewpoint and appearance.

E. Runtime Considerations

Like other state-of-the-art methods [26] the system in its current stage is not capable of processing images in real time. Finding landmark proposals using the Matlab implementation of Edge Boxes [45] takes around 1.8 seconds per image on a standard desktop machine. In addition, calculating a single ConvNet feature up to layer conv3 requires approximately 15 ms using Caffe on a NVIDIA Quadro K4000 GPU.

In future work we will adapt the system so that only one forward pass through the ConvNet is performed per image, instead of one individual pass for each of the 100 landmarks. This will result in a $100 \times$ speed-up of the feature extraction.

F. Improving the Absolute Performance by Sequence Search

While single image matching performance can serve as a good evaluation of performance, in many practical robotic applications it is feasible to exploit the inherent temporal information available to a navigating robot. Consequently we replace the simple nearest neighbor search with a state-of-the-art sequence search technique from [33], while keeping the preceding parts of our approach. Performance improves significantly even for short sequence search lengths of 6 images. Fig. 13 summarizes these results and compares the vanilla single-image precision recall curves from above with the results obtained by the addition of the sequence search.

V. CONCLUSION AND DISCUSSION

We have presented a novel place recognition system that builds on state-of-the-art object detection methods and con-
when the landmark is partially occluded or changes its size and appearance changes and can be successfully matched even for sizes in a place recognition context; they are stable under environment-specific training, the results have also revealed further insights: Mid-level ConvNet features appear to be highly suitable as descriptors for landmarks of various sizes in a place recognition context; they are stable under viewpoint, our method extracts and successfully matches the landmark regions illustrated by the colored boxes. Bottom: Precision recall plot showing the proposed region-based method (solid) outperforms the whole image based method (dashed) and SeqSLAM (dotted) significantly for different values of overlap between the query and database images.

The current system has no explicit or implicit metricity. Following on the success in metric localization of landmarks presented in [26], we will investigate whether the camera position can be estimated relative to observed landmarks. Other geometry-within-the-image techniques like geometric verification [8, 28] may also improve performance. In subsequent work we replaced the Gaussian Random Projection by a binary locality-sensitive hashing method that compresses the original 64,896 dimensional $\text{conv3}$ feature vectors to merely 8192 bits. First results indicate this approach regains more than 95% of place recognition performance while achieving a 250× speed-up for the feature matching.

Acknowledgements This research was conducted by the Australian Research Council Centre of Excellence for Robotic Vision (project number CE140100016).
REFERENCES


