

On the Performance of ConvNet Features for Place Recognition

Niko Sünderhauf, Sareh Shirazi, Feras Dayoub, Ben Upcroft, and Michael Milford

Abstract—After the incredible success of deep learning in the computer vision domain, there has been much interest in applying Convolutional Network (ConvNet) features in robotic fields such as visual navigation and SLAM. Unfortunately, there are fundamental differences and challenges involved. Computer vision datasets are very different in character to robotic camera data, real-time performance is essential, and performance priorities can be different. This paper comprehensively evaluates and compares the utility of three state-of-the-art ConvNets on the problems of particular relevance to navigation for robots; viewpoint-invariance and condition-invariance, and for the first time enables real-time place recognition performance using ConvNets with large maps by integrating a variety of existing (locality-sensitive hashing) and novel (semantic search space partitioning) optimization techniques. We present extensive experiments on four real world datasets cultivated to evaluate each of the specific challenges in place recognition. The results demonstrate that speed-ups of two orders of magnitude can be achieved with minimal accuracy degradation, enabling real-time performance. We confirm that networks trained for semantic place *categorization* also perform better at (specific) place *recognition* when faced with severe appearance changes and provide a reference for which networks and layers are optimal for different aspects of the place recognition problem.

I. INTRODUCTION

Robots that aim at autonomous long-term operations over extended periods of time, such as days, weeks, or months, are faced with environments that can undergo dramatic changes in their visual appearance over time. Visual place recognition – the ability to recognize a known place in the environment using vision as the main sensor modality – is largely affected by these appearance changes and is therefore an active research field within the robotics community. The recent literature proposes a variety of approaches to address the challenges of this field [1]–[14].

Recent progress in the computer vision and machine learning community has shown that the features generated by Convolutional Networks (ConvNets, see Fig. 1 for examples) outperform other methods in a broad variety of visual recognition, classification and detection tasks [15]. ConvNets have been demonstrated to be versatile and transferable, i.e. even although they were trained on a very specific target task, they can be successfully deployed for solving different problems and often even outperform traditional hand engineered features [16].

Our paper leverages these astounding properties and introduces the first real-time capable ConvNet-based place recognition system. We exploit the hierarchical nature of

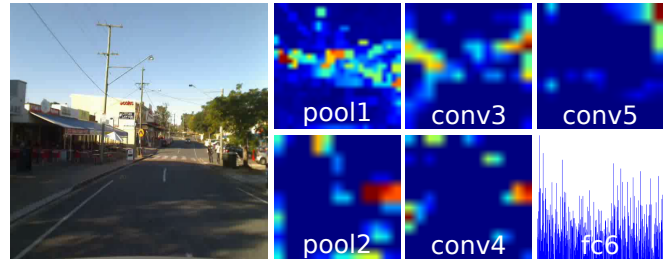


Fig. 1: Convolutional neural networks can extract features that serve as holistic image descriptors for place recognition. We found the features from layer `conv3` to be robust against appearance changes on a variety of datasets. The figure shows an example scene and extracted features from different network layers of the AlexNet network.

ConvNet features and use the semantic information encoded in the higher layers for search space partitioning and the mid-level features for place matching under challenging conditions. Locality-sensitive hashing of these features allows us to perform robust place recognition against 100,000 known places with 3 Hz. We provide a thorough investigation of the utility of the individual layers in the ConvNet hierarchy under severe appearance and viewpoint variations and furthermore compare three state-of-the-art networks for the task of place recognition. We establish the following main results:

- 1) features from the higher layers of the ConvNet hierarchy encode semantic information about a place and can be exploited to significantly reduce place recognition time by partitioning the search space,
- 2) a speed-up of two orders of magnitude can be achieved by approximating the cosine distance between features with the Hamming distance over bitvectors obtained by Locality Sensitive Hashing, compressing the feature data by 99.6% but retaining 95% of place recognition performance;
- 3) when comparing different ConvNets for the task of place *recognition* under severe appearance changes, networks trained for the task of semantic place *categorization* [17] outperform the ConvNet trained for *object* recognition;
- 4) features from the middle layers in the ConvNet hierarchy exhibit robustness against appearance changes induced by the time of day, seasons, or weather conditions; and
- 5) features extracted from the top layers are more robust with respect to viewpoint changes.

In the following we review the related literature before describing the datasets and evaluation protocol used. We analyze the performance of individual ConvNet feature layers for place recognition on several challenging datasets in Section IV and introduce important algorithmic performance improvements in Section V before concluding the paper.

II. RELATED WORK

A. Place Recognition

The focus of research in place recognition has recently moved from recognizing scenes without significant appearance changes [18]–[20] to more demanding, but also more realistic changing environments. Methods that address the place recognition problem span from matching sequences of images [1], [4]–[7], transforming images to become invariant against common scene changes such as shadows [2], [8], [10]–[12], learning how environments change over time and predicting these changes in image space [3], [12], [21], particle filter-based approaches that build up place recognition hypotheses over time [13], [14], [22], or build a map of experiences that cover the different appearances of a place over time [9].

Learning how the appearance of the environment changes generally requires training data with known frame correspondences. [4] builds a database of observed features over the course of a day and night. [3], [21] present an approach that learns systematic scene changes in order to improve performance on a seasonal change dataset. [23] 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.

B. Convolutional Networks

A commonality between all the aforementioned approaches is that they rely on a fixed set of hand-crafted traditional features or operate on the raw pixel levels. However, a recent trend in computer vision, and especially in the field of object recognition and detection [15], [24], is to exploit learned features using deep convolutional networks (ConvNets). It therefore appears very promising to analyze these features and experimentally investigate their feasibility for the task of place recognition.

Convolutional network is a well-known architecture and was proposed by LeCun et al. in 1989 [25] to recognize hand-written digits. Several research groups have recently shown that ConvNets outperform classical approaches for object classification or detection that are based on hand-crafted features [15], [16], [26]–[28]. The availability of pre-trained network models makes it easy to experiment with such approaches for different tasks: The software packages Overfeat [27] and Caffe [29] provide network architectures pre-trained for a variety of recognition tasks.

[30] was the first to consider ConvNets for place recognition. Our investigation is more thorough, since we cleanly separate the performance contribution of the ConvNet features from the matching strategy, conduct more systematic experiments on various datasets, compare three different ConvNets, and contribute important algorithmic improvements which enable real-time performance.

III. PRELIMINARIES

A. Feature Extraction using a Convolutional Neural Network

For the experiments described in the following section, we deploy the AlexNet ConvNet [26] provided by Caffe [29]. This network was pre-trained on the ImageNet ILSVRC dataset [24] for object recognition. It consists of five convolutional layers followed by three fully connected layers and a soft-max layer. The output of each individual layer can be extracted from the network and used as a holistic image descriptor. Since the third fully connected and the soft-max layer are adopted specifically to the ILSVRC task (they have 1000 output neurons for the 1000 object classes in ILSVRC), we do not use them in the following experiments.

Table I lists the used layers and compares their dimensionality; Fig. 1 displays some exemplar features extracted from different layers. All images are resized to 231×231 pixels to fit the expected input size of the ConvNet.

B. Image Matching and Performance Measures

Place recognition is performed by single-image nearest neighbor search based on the cosine distance of the extracted feature vectors. We analyze the performance in terms of precision-recall curves and F_1 scores¹. In contrast to [30] we explicitly do *not* apply sequence search techniques or other specialized algorithmic approaches to improve the matching performance in order to observe the baseline performance of the ConvNet features under investigation.

C. Datasets used in the Evaluation

We used four different datasets in the experiments presented in Section IV. From the summary of their characteristics in Table II we can see that we carefully selected these

¹A match is considered a *positive* if it passes a ratio test (ratio of the distances of the best over the second best match found in the nearest neighbor search), and a *negative* otherwise. Since every scene in the datasets has a ground truth match (i.e. there are no *true negatives*), every *negative* is a *false negative*. A match is a *true positive* when it is within ± 1.5 frames of the ground truth (depending on the frame rate of the recorded dataset) and a *false positive* otherwise. The running parameter for creating the PR curves is the threshold on the ratio test.

Layer	Dimensions	Layer	Dimensions
pool1	$96 \times 27 \times 27$	conv5	$256 \times 13 \times 13$
pool2	$256 \times 13 \times 13$	fc6	$4096 \times 1 \times 1$
conv3	$384 \times 13 \times 13$	fc7	$4096 \times 1 \times 1$
conv4	$384 \times 13 \times 13$		

TABLE I: The layers from the AlexNet ConvNet used in our evaluation and their output dimensionality.

datasets to form three groups of varying condition changes: Two datasets exhibit severe appearance change but virtually no variation in viewpoints, one shows viewpoint changes but only mild appearance changes, and three others feature both types of variations.

1) *The Nordland Dataset*: The Nordland dataset consists of video footage of a 10 hours long train journey recorded from the perspective of the front cart in four different seasons. Fig. 6 gives an impression of the severe appearance changes between the different seasons. The Nordland dataset is a perfect experimentation dataset since it exhibits no viewpoint variations whatsoever and therefore allows to test algorithms on pure condition changes. See [3] for a more elaborate discussion of this dataset. For our experiments we extracted image frames at a rate of 1 fps and exclude all images that are taken inside tunnels or when the train was stopped. We furthermore exclude images between the timestamps 1:00h-1:30h and 2:47h-3:30h since these form a training dataset we use in parallel work.

2) *The Gardens Point Dataset*: The Gardens Point dataset has been recorded on the Gardens Point Campus of QUT in Brisbane. It consists of three traverses of the environment, two during the day and one during the night. The night images have been extremely contrast enhanced and converted to grayscale in the process. One of the day traverses was recorded keeping on the left side of the walkways, while the other day and the night datasets have been recorded from the right side. This way, the dataset exhibits both appearance and viewpoint changes.

3) *The Campus Human vs. Robot Dataset*: This dataset was recorded in different areas of our campus (outdoor, office, corridor, food court) once by a robot using the Kinect RGB camera and once by a human with a GoPro camera. While the robot footage was recorded during the day, the human traversed the environment during dawn, resulting in significant appearance changes especially in the outdoor scenes. The GoPro images were cropped to contain only the center part of half of the size of the original images.

4) *The St. Lucia Dataset*: The St. Lucia dataset [31] has been recorded from inside a car moving through a suburb in Brisbane at 5 different times during a day, and also on different days over a time of two weeks. It features mild viewpoint variations due to slight changes in the exact route

Dataset	Environment	Variations in	
		Appearance	Viewpoint
Nordland seasons	train journey	severe	none
GP day-right-night-right	campus outdoor	severe	none
GP day-left-night-right	campus outdoor	severe	medium
Campus Human-Robot	campus mixed	medium	medium
St. Lucia	suburban roads	medium	medium
GP day-left-day-right	campus outdoor	minor	medium

TABLE II: The datasets used in the evaluation form three groups of varying severity of appearance and viewpoint changes.

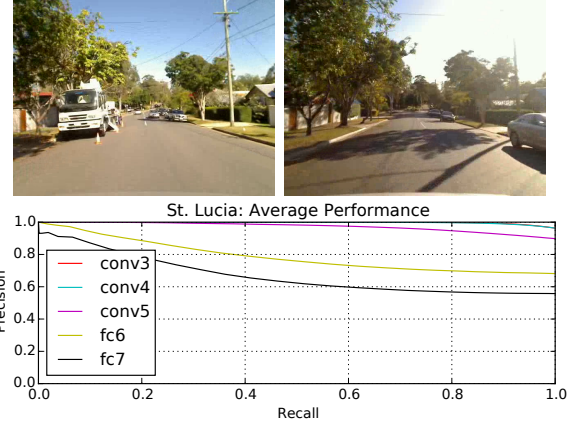


Fig. 2: Top: Two example images from the St. Lucia dataset showing the same place. Bottom: Precision-recall curves averaged over all nine trials (recorded at different times during the day, and several days apart). Layers `conv3` and `conv4` perform almost perfectly (overlapping curves in the upper right corner).

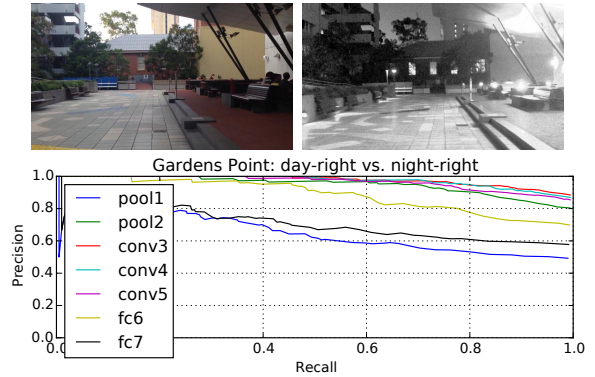


Fig. 3: Extreme appearance changes between day and night images. The nighttime images have been contrast enhanced and converted to grayscale in the process. Despite these changes, layer `conv3` still performs reasonably well.

taken by the car. More significant appearance changes due to the different times of day can be observed, as well as some changes due to dynamic objects such as traffic or cars parked on the street.

IV. LAYER-BY-LAYER STUDIES

This section provides a thorough investigation of the utility of different layers in the ConvNet hierarchy for place recognition and evaluates their individual robustness against the two main challenges in visual place recognition: severe appearance changes and viewpoint variations.

A. Appearance Change Robustness

In a first set of experiments we analyze the robustness of the features from different layers in the ConvNet architecture against appearance changes. We conduct these experiments on the following datasets:

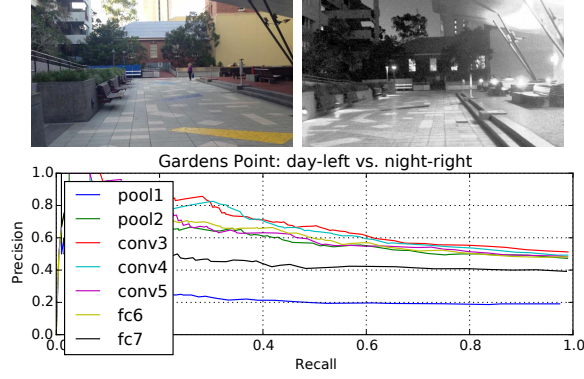


Fig. 4: When combining extreme appearance change with viewpoint changes, place recognition performance begins to deteriorate. This combination is still very challenging for current place recognition systems.

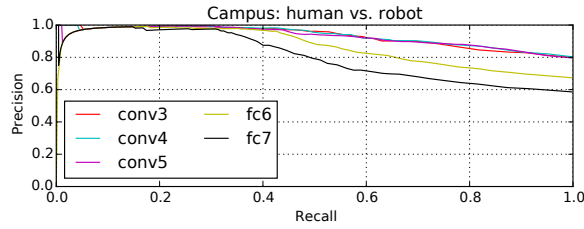


Fig. 5: The Campus dataset was recorded with different cameras by a robot during the day and a human at dawn. The mid-level convolutional layers outperform the fully connected layers $fc6$ and $fc7$.

- 1) the Nordland dataset, using 5 different season pairings that involve the spring or winter season
- 2) the Gardens Point dataset, using all 3 subsets, spanning day and night
- 3) the St. Lucia dataset, using a total of 9 traversals from varying daytimes, over the course of several weeks
- 4) the Campus dataset, using footage recorded by a human and a robot at different times of the day, including dawn

Figures 2-6 show the resulting precision-recall curves for all experiments. Table III summarizes the results further and compares the F_1 scores with other state-of-the-art methods.

The experiments consistently show that the mid-level features from layer `conv3` are more robust against appearance changes than features from any other layer. Both the lower layers (`pool1`, `pool2`) and the higher layers (e.g. `fc6` and `fc7`) in the feature hierarchy lack robustness and exhibit inferior place recognition performance.

We further compared our approach to SeqSLAM [1] a state-of-the-art method for place recognition under extreme appearance changes and found that single image matching using features extracted by layer `conv3` matches or exceeds SeqSLAM’s performance. Previous work [3], [32] established that FAB-MAP [18], [19] is not capable of handling the severe appearance changes of the Nordland and St. Lucia datasets.

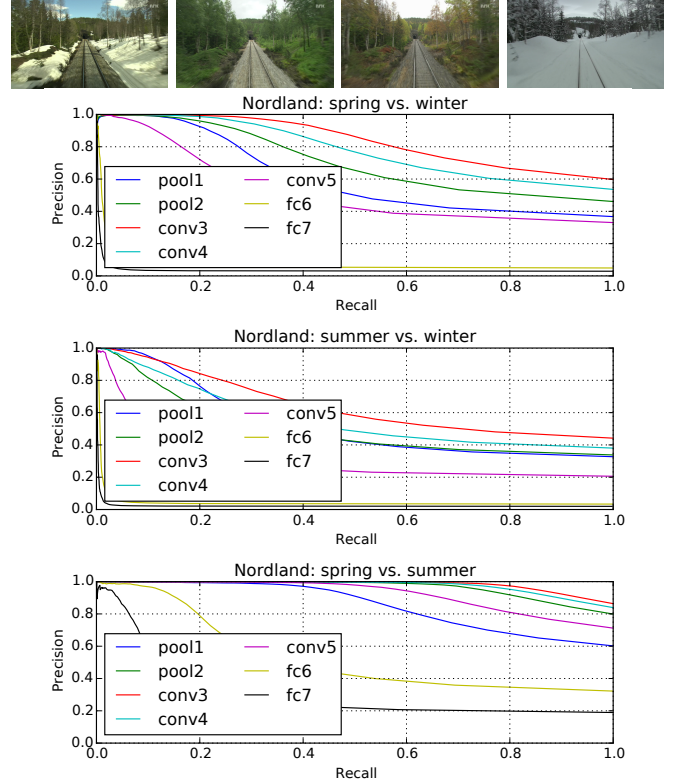


Fig. 6: Place recognition across seasons on the Nordland dataset. Despite the extreme appearance changes `conv3` performs acceptable, given that the image matching is based on a single frame nearest neighbor search only. Notice how `fc6` and `fc7` fail completely.

B. Viewpoint Change Robustness

Viewpoint changes are a second major challenge for visual place recognition systems. While point feature-based methods like FABMAP are less affected, holistic methods such as the one proposed here are prone to error in the presence of viewpoint changes.

In order to quantify the viewpoint robustness of ConvNet features, we perform two experiments:

- 1) We use the Nordland spring dataset to create synthetic viewpoint changes by using shifted image crops.
- 2) We use the Gardens Point dataset (day left vs. day right) to verify the observed effects on a real dataset.

We conduct the first experiment on the Nordland spring dataset using images cropped to half of the width of the original images. We simulate viewpoint changes between two traverses by shifting the images of the second traverse to the right. This results in overlaps between the images of the simulated first and second traverse of between 98% and 20%. For these experiments, the first 5000 images from the spring dataset (excluding tunnels, stoppages and the training dataset mentioned before) were used. Fig. 7 illustrates the results of the experiment with F-scores extracted from precision recall statistics. In a second experiment we used the Gardens Point day-left vs. day-right dataset that exhibits a lateral

Dataset		F ₁ -Scores							SeqSLAM	FAB-MAP
		pool1	pool2	AlexNet Layers			conv5	fc6	fc7	
Nordland	Spring – Winter	0.54	0.63	0.75	0.70	0.50	0.09	0.06	0.80 *	†
	Summer – Winter	0.49	0.51	0.61	0.55	0.34	0.07	0.04	0.64 *	†
	Fall – Winter	0.52	0.60	0.66	0.60	0.38	0.07	0.05	0.63 *	†
	Spring – Summer	0.75	0.89	0.93	0.91	0.83	0.49	0.32	0.86 *	†
	Spring – Fall	0.78	0.90	0.93	0.92	0.84	0.53	0.35	0.88 *	†
Gardens Point	day left – day right	0.61	0.88	0.89	0.88	0.91	0.96	0.96	0.44	—
	day right – night right	0.66	0.89	0.94	0.93	0.92	0.82	0.73	0.44	—
	day left – night right	0.32	0.64	0.68	0.66	0.65	0.65	0.56	0.21	—
St. Lucia	average over 9 trials	—	—	0.98	0.98	0.95	0.81	0.72	—	‡
	worst trial	—	—	0.96	0.96	0.87	0.51	0.38	0.59	‡
Campus	human – robot	—	—	0.89	0.89	0.89	0.80	0.74	0.22	—

TABLE III: Comparison of F_1 -Scores between different layers and other state-of-the-art methods. † We established in [3] that FAB-MAP fails on the Nordland dataset. The maximum measured recall was 0.025 at 0.08 precision. ‡ [32] reported that FAB-MAP is not able to address the challenges of the St. Lucia dataset, creating too many false positive matches. *For SeqSLAM on the Nordland dataset we skipped every second frame (due to performance reasons) and used a sequence length of 5, resulting in an effective sequence length of 10. All other trials with SeqSLAM used a sequence length of 10.

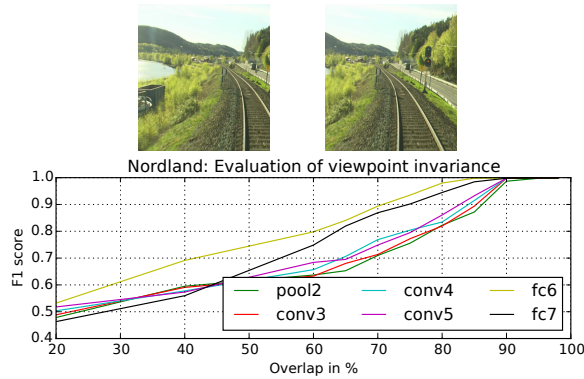


Fig. 7: Top row: Examples for the synthetic viewpoint variation experiments conducted on the Nordland train dataset. We cropped the original images to half of their width and then created shifted versions with varying overlap. Bottom: F_1 scores for different layers and overlap values. $fc6$ performs best, but all layers are invariant to small variations that maintain $\approx 90\%$ overlap.

camera movement of 2-3 meters for medium viewpoint changes between two traversals of the environment. Only minor variations in the appearance of the scenes can be observed, mostly caused by people walking on the campus. Fig. 8 shows the resulting precision recall curves.

Both experiments show that features from layers higher in the ConvNet architecture, especially $fc6$, are more robust to viewpoint changes than features from lower layers. The increased viewpoint robustness of the upper layers can be accounted to the *pooling* layers that are part of the network architecture and perform max-pooling after the first, second, and fifth convolutional layer. The synthetic experiments summarized in Fig. 7 also show that all layers are robust to mild viewpoint changes with more than 90% overlap between scenes. This effect is due to the resampling (resizing) of images before they are passed through the network and the

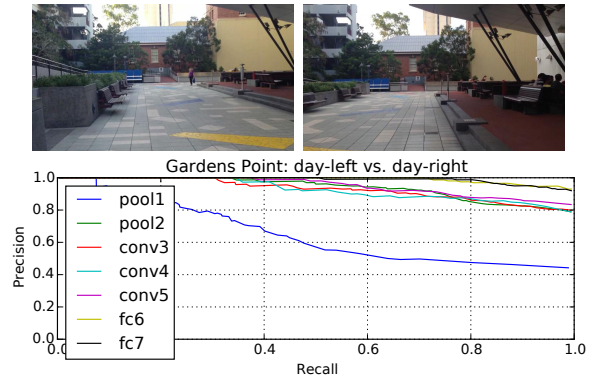


Fig. 8: Testing the viewpoint invariance on a realistic dataset. Two traverses through the same environment were recorded, one on the left and another on the right side of the walkway. The top two images compare the difference in viewpoint from the same place in both sequences. The plot shows that the higher layers $fc6$ and $fc7$ perform best,

convolution and pooling operations that occur in the first layer. In subsequent work [33] we addressed the scenario of viewpoint and appearance changes occurring simultaneously using a number of new dataset.

C. Summary and Discussion

Table IV summarizes our experiments. It is apparent that place recognition based on single image matching performs well even under severe appearance changes when using the mid-level $conv3$ layer as a holistic feature extractor. Features from the higher layers, especially $fc6$, only outperform $conv3$ in situations with viewpoint changes but none or only mild appearance variations.

This intuitively makes sense, since the features from the first convolutional layers resemble very simple shape features [34] that are not discriminative and generic enough to allow

place recognition under severe appearance changes. The layers *higher* in the hierarchy, and especially the fully connected layers, are more semantically meaningful but therefore lose their ability to discriminate between individual places within the same semantic type of scene. For example in the St. Lucia footage, the higher layers would encode the information ‘sub-urban road scene’ equally for all images in the dataset. This enables place *categorization* but is disadvantageous for place *recognition*. The mid-level layers – and especially `conv3` – seem to encode just the right amount of information; they are more informative and more robust to changes than pure low-level pixel or gradient based features, while remaining discriminate enough to identify individual places.

Extracting a ConvNet feature from an image requires approximately 15 ms on a Nvidia Quadro K4000 GPU. The bottleneck of the place recognition system is the nearest neighbor search that is based on the cosine distance between 64,896 dimensional feature vectors – a computationally expensive operation. Our Numpy/Scipy implementation requires 3.5 seconds to find a match among 10,000 previously visited places. As we shall see in the next section, two important algorithmic improvements can be introduced that lead to a speed-up of 2 orders of magnitude.

V. REAL-TIME LARGE-SCALE PLACE RECOGNITION

In contrast to typical computer vision benchmarks where the recognition accuracy is the most important performance metric, robotics applications depend on agile algorithms that can provide a solution under certain soft real-time constraints. The nearest neighbor search is the key limiting factor for large-scale place recognition, as its runtime is proportional to the number of stored previously visited places. In the following we will explore two approaches that will decrease the required search time by two orders of magnitude with only minimal accuracy degradation.

A. Locality Sensitive Hashing for Runtime Improvements

Computing the cosine distance between many 64,896 dimensional `conv3` feature vectors is an expensive operation and is a bottleneck of the ConvNet-based place recognition.

To speed this process up significantly, we propose to use a specialized variant of binary locality-sensitive hashing that preserves the cosine similarity [35]. This hashing method leverages the property that the probability of a random

hyperplane separating two vectors is directly proportional to the angle between these vectors. [36] demonstrates how the cosine distance between two high-dimensional vectors can be closely approximated by the Hamming distance between the respective hashed bit vectors. The more bits the hash contains, the better the approximation.

We implemented this method and compare the place recognition performance achieved with the hashed `conv3` feature vectors of different lengths ($2^8 \dots 2^{13}$ bits) on the Nordland and Gardens Point datasets in Fig. 9. Using 8192 bits retains approximately 95% of the place recognition performance. Hashing the original 64,896 dimensional vectors into 8192 bits corresponds to a data compression of 99.6%. Since the Hamming distance over bit vectors is a computationally cheap operation, the best matching image among 10,000 candidates can be found within 13.4 ms on a standard desktop machine. This corresponds to a speed-up factor of 266 compared to using the cosine distance over the original `conv3` features which required 3570 ms per 10,000 candidates. Calculating the hashes requires 180 ms using a non-optimized Python implementation.

Table V summarizes the required time for the main algorithmic steps. We can see that the hashing enables real-time place recognition using ConvNet features on large scale maps of 100,000 places or more.

B. Search Space Partitioning using Semantic Categorization

We propose a novel approach to exploit the *semantic* information encoded in the *high-level* ConvNet features to partition the search space and constrain the nearest neighbor search to areas of similar semantic place categories such as office, corridor, classroom, or restaurant. This can significantly shrink the search space in *semantically versatile* environments.

To discriminate between semantic place categories we train a nonlinear SVM classifier using the `fc7` layer. The training has to be performed only once on images from the SUN-397 database [37]. When performing place recognition, a separate index for each semantic class c_i is maintained that allows fast access to all stored `conv3` features that were recorded at previously visited places with the probability of belonging to c_i above a threshold θ . The nearest neighbor search can then be limited to places that belong to the same semantic category as the currently observed place.

We tested this approach on the Campus dataset and trained the classifier to discriminate between 11 semantic classes.

Dataset	Variations in		Best Layer
	Appearance	Viewpoint	
Nordland seasons	severe	none	<code>conv3</code>
GP day-right vs night-right	severe	none	<code>conv3</code>
GP day-left vs night-right	severe	medium	<code>conv3</code>
St. Lucia	medium	medium	<code>conv3</code>
Campus	medium	medium	<code>conv3</code>
GP day-left vs day-right	minor	medium	<code>fc6</code>
Nordland synthetic	none	<i>varied</i>	<code>fc6</code>

TABLE IV: The mid-level `conv3` outperforms all other layers in the presence of significant appearance changes.

	original <code>conv3</code>	4096 bits	8192 bits
ConvNet feature	15 ms	15 ms	15 ms
hashing	—	100 ms	180 ms
match 100k candidates	35,700 ms	89 ms	134 ms
frame rate	0.03 Hz	4.9 Hz	3.0 Hz

TABLE V: Runtime comparison between important algorithmic steps for the original features and two different hashes on a desktop machine with a Nvidia Quadro K4000 GPU.

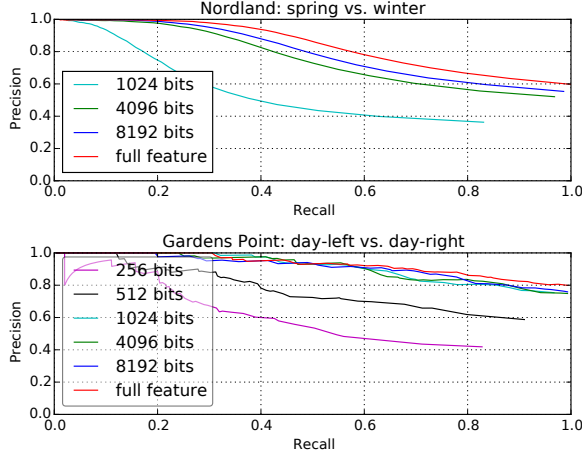


Fig. 9: The cosine distance over the full feature vector of 64896 floating point elements (red), can be closely approximated by the Hamming distance over bit vectors of length 2^{13} (blue) without losing much performance. This corresponds to a compression of 99.6%. The bit vectors are calculated using the cosine similarity preserving Locality Sensitive Hashing method as proposed in [36].

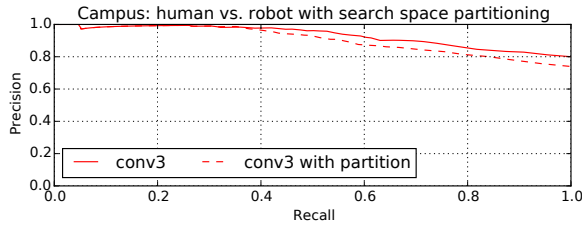


Fig. 10: Partitioning the search space decreases the performance slightly but can significantly reduce the time spent for the nearest neighbor search (76% in this case).

Partitioning the search space decreased the time spent for the nearest neighbor search by a factor of 4. Fig. 10 shows the place recognition performance decreasing slightly when using this technique. However, there is an adjustable trade-off between recognition performance and runtime requirements. Lowering the threshold ($\theta = 10\%$ was used here) allows more candidate matches to be assessed, thus increasing recognition performance at the expense of the runtime.

VI. COMPARING DIFFERENT CONVNETS FOR PLACE RECOGNITION

All experiments described so far used the AlexNet ConvNet architecture [26] as implemented by Caffe [29] that was pre-trained for the task of object recognition on the ILSVRC dataset [24]. While preparing this paper, [17] published the Places205 and Hybrid networks. These two networks have the same principled architecture as AlexNet but have been trained for scene categorization (Places205) or both tasks (Hybrid).

We compare their place recognition performance with

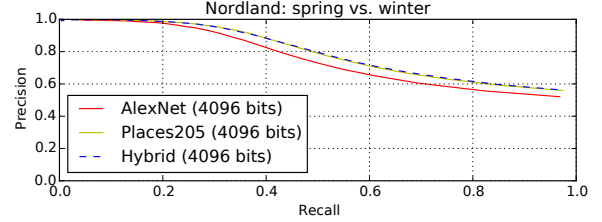


Fig. 11: The Places205 and the Hybrid networks from [17] perform slightly better at place recognition than AlexNet on the Nordland Spring vs. Winter dataset. The hashed conv3 features were used for these experiments.

that of AlexNet, using the hashed conv3 features. Table VI and Fig. 11 summarize the results. Compared to AlexNet, Places205 and the Hybrid network perform slightly better under severe appearance changes. This could be explained by the fact that AlexNet is trained for *object* recognition while the two other networks are specialized for recognizing *scene* categories, i.e. they learned to discriminate places. However, in the presence of viewpoint changes (Gardens Point left vs. right), the results are inconclusive and AlexNet has a slight performance advantage.

VII. CONCLUSIONS

Our paper presented a thorough investigation on the utility of ConvNet features for the important task of visual place recognition in robotics. We presented a novel method to combine the individual strengths of the high-level and mid-level feature layers to partition the search space and recognize places under severe appearance changes. We demonstrated for the first time that large-scale robust place recognition using ConvNet features is possible when applying a specialized binary hashing method. Our comprehensive study on four real world datasets highlighted the individual strengths of mid- and high-level features with respect to the biggest challenges in visual place recognition – appearance and viewpoint changes. A comparison of three state-of-the-art ConvNets revealed slight performance advantages for the networks trained for semantic place categorization.

In subsequent work [33] we applied the insights gained in this paper and extended the holistic approach presented here to a landmark-based scheme that addresses the remaining challenge of combined viewpoint *and* appearance change robustness. In future work we will investigate how training ConvNets specifically for the task of place recognition under changing conditions can improve their performance.

ACKNOWLEDGEMENTS

This research was conducted by the Australian Research Council Centre of Excellence for Robotic Vision (project number CE140100016). We want to thank Arren Glover and Will Maddern for collecting the Gardens Point and St. Lucia datasets.

REFERENCES

- [1] M. Milford and G. F. Wyeth, “SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights.” in *Proc. of Intl. Conf. on Robotics and Automation (ICRA)*, 2012.

Network	Pre-trained for Task	on Dataset	Place Recognition Performance (F-Score) using conv3			
			Nordland spring vs. winter	Gardens Point day-left vs. day-right	day-right vs. night-right	day-left vs. night-right
AlexNet	object recognition	ImageNet ILSVRC [24]	0.68	0.85	0.85	0.48
Places205	place categorization	Places Database [17]	0.71	0.82	0.84	0.46
Hybrid	both of above tasks	Places + ILSVRC [17]	0.71	0.84	0.87	0.44

TABLE VI: The three convolutional networks we compare in this paper. All of them are implemented using Caffe [29]. Notice that none of them is specifically trained for the task of place recognition.

- [2] P. Corke, R. Paul, W. Churchill, and P. Newman, "Dealing with shadows: Capturing intrinsic scene appearance for image-based outdoor localisation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2013.
- [3] P. Neubert, N. Sünderhauf, and P. Protzel, "Superpixel-based appearance change prediction for long-term navigation across seasons," *Journal of Robotics and Autonomous Systems*, 2014.
- [4] E. Johns and G.-Z. Yang, "Feature co-occurrence maps: Appearance-based localisation throughout the day," in *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2013.
- [5] N. Sünderhauf, P. Neubert, and P. Protzel, "Are We There Yet? Challenging SeqSLAM on a 3000 km Journey Across All Four Seasons," in *Proceedings of Workshop on Long-Term Autonomy, IEEE International Conference on Robotics and Automation (ICRA)*, 2013.
- [6] E. Pepperell, P. I. Corke, and M. J. Milford, "All-environment visual place recognition with SMART," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2014.
- [7] T. Naseer, L. Spinello, W. Burgard, and C. Stachniss, "Robust visual robot localization across seasons using network flows," in *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, 2014.
- [8] B. Upcroft, C. McManus, W. Churchill, W. Maddern, and P. Newman, "Lighting invariant urban street classification," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2014.
- [9] W. Churchill and P. M. Newman, "Practice makes perfect? Managing and leveraging visual experiences for lifelong navigation," in *Proc. of Intl. Conf. on Robotics and Automation (ICRA)*, 2012.
- [10] C. McManus, W. Churchill, W. Maddern, A. Stewart, and P. Newman, "Shady dealings: Robust, long-term visual localisation using illumination invariance," in *Proc. of IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2014.
- [11] W. Maddern, A. Stewart, C. McManus, B. Upcroft, W. Churchill, and P. Newman, "Illumination invariant imaging: Applications in robust vision-based localisation, mapping and classification for autonomous vehicles," in *Proceedings of the Visual Place Recognition in Changing Environments Workshop, IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2014.
- [12] S. Lowry, M. Milford, and G. Wyeth, "Transforming morning to afternoon using linear regression techniques," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2014.
- [13] C. Stachniss and W. Burgard, "Particle filters for robot navigation," *Foundations and Trends in Robotics*, vol. 3, no. 4, pp. 211–282, 2014.
- [14] S. Lowry, G. Wyeth, and M. Milford, "Towards training-free appearance-based localization: probabilistic models for whole-image descriptors," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2014.
- [15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
- [16] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN Features off-the-shelf: an Astounding Baseline for Recognition," in *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2014.
- [17] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning Deep Features for Scene Recognition using Places Database," *NIPS*, 2014.
- [18] M. Cummins and P. Newman, "FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance," *The International Journal of Robotics Research*, vol. 27, no. 6, pp. 647–665, 2008.
- [19] —, "Appearance-only SLAM at large scale with FAB-MAP 2.0," *The International Journal of Robotics Research*, vol. 30, no. 9, 2011.
- [20] G. Sibley, L. Matthies, and G. Sukhatme, "Sliding Window Filter with Application to Planetary Landing," *J. Field Robotics*, vol. 27, no. 5, 2010.
- [21] P. Neubert, N. Sünderhauf, and P. Protzel, "Appearance Change Prediction for Long-Term Navigation Across Seasons," in *Proceedings of European Conference on Mobile Robotics (ECMR)*, 2013.
- [22] W. Maddern, M. Milford, and G. Wyeth, "Continuous Appearance-based Trajectory SLAM," in *International Conference on Robotics and Automation (ICRA)*, 2011.
- [23] C. McManus, B. Upcroft, and P. Newman, "Scene signatures: Localised and point-less features for localisation," in *Proceedings of Robotics Science and Systems (RSS)*, Berkeley, CA, USA, July 2014.
- [24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, 2014.
- [25] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, 1989.
- [26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* 25, 2012.
- [27] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.
- [28] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," *arXiv preprint arXiv:1310.1531*, 2013.
- [29] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional Architecture for Fast Feature Embedding," in *Proc. of ACM International Conference on Multimedia.*, 2014.
- [30] Z. Chen, O. Lam, A. Jacobson, and M. Milford, "Convolutional Neural Network-based Place Recognition," in *Proceedings of Australasian Conference on Robotics and Automation (ACRA)*, 2014.
- [31] A. Glover, W. Maddern, M. Milford, and G. Wyeth, "FAB-MAP + RatSLAM: Appearance-Based SLAM for Multiple Times of Day," in *Proc. of IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2010.
- [32] A. J. Glover, W. P. Maddern, M. J. Milford, and G. F. Wyeth, "Fab-map+ ratslam: appearance-based slam for multiple times of day," in *Robotics and Automation (ICRA)*, 2010 *IEEE International Conference on*. IEEE, 2010, pp. 3507–3512.
- [33] N. Sünderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, and M. Milford, "Place Recognition with ConvNet Landmarks: Viewpoint-Robust, Condition-Robust, Training-Free," in *Proc. of Robotics: Science and Systems (RSS)*, 2015.
- [34] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision—ECCV 2014*, 2014.
- [35] M. S. Charikar, "Similarity estimation techniques from rounding algorithms," in *Proc. of Annual ACM Symposium on Theory of Computing*, 2002.
- [36] D. Ravichandran, P. Pantel, and E. Hovy, "Randomized algorithms and NLP: using locality sensitive hash function for high speed noun clustering," in *Proc. of the Annual Meeting on Association for Computational Linguistics*, 2005.
- [37] J. Xiao, K. A. Ehinger, J. Hays, A. Torralba, and A. Oliva, "Sun database: Exploring a large collection of scene categories," *International Journal of Computer Vision*, pp. 1–20, 2014.