Towards Vision-Based Pose- and Condition-Invariant Place Recognition along Routes*

Edward Pepperell, Peter Corke and Michael Milford
Australian Centre for Robotic Vision
Queensland University of Technology
e.pepperell@qut.edu.au

Abstract

Vision-based place recognition involves recognising familiar places despite changes in environmental conditions or camera viewpoint (pose). Existing training-free methods exhibit excellent invariance to either of these challenges, but not both simultaneously. In this paper, we present a technique for condition-invariant place recognition across large lateral platform pose variance for vehicles or robots travelling along routes. Our approach combines sideways facing cameras with a new multi-scale image comparison technique that generates synthetic views for input into the condition-invariant Sequence Matching Across Route Traversals (SMART) algorithm. We evaluate the system's performance on multi-lane roads in two different environments across day-night cycles. In the extreme case of day-night place recognition across the entire width of a four-lane-plus-median-strip highway, we demonstrate performance of up to 44% recall at 100% precision, where current state-of-the-art fails.

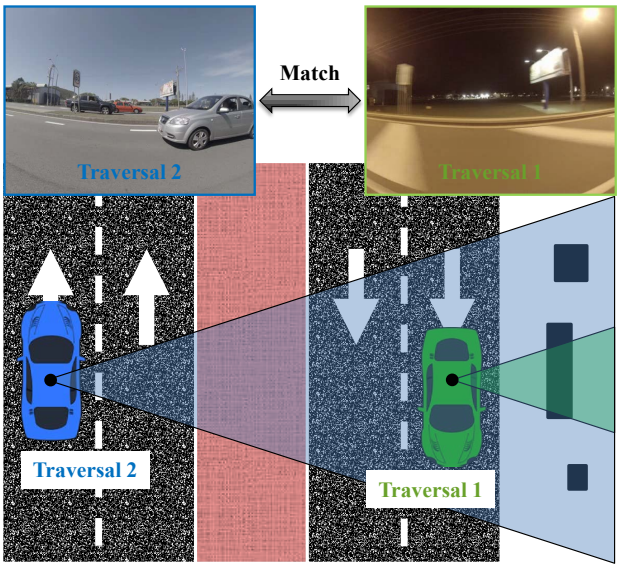

Figure 1: By applying the SMART algorithm over views of multiple scales, the same place can be correctly identified across significant changes in lateral pose and illumination. The example frame pair above is a match (as the centre of an image sequence) correctly identified by our algorithm under the worst case pose change of 3 lanes and a median strip on a divided highway road, where the 2 traverses were performed in opposite directions. Note that this technique uses bilateral cameras, but for clarity, only the east camera views are shown.

1 Introduction

Visual place recognition systems have gained increasing traction due to their many benefits over other sensing modalities, including their low cost, small size, low-power requirements and ability to work in GPS-denied areas such as tunnels. However, a place's visual appearance significantly varies depending on environmental conditions and camera pose. While existing systems cope well with significant camera pose [Cummins and Newman, 2009; Davison *et al.*, 2007; Klein and Murray, 2007; Konolige and Agrawal, 2008] or condition variance [Milford and Wyeth, 2012; Pepperell *et al.*, 2014; Milford *et al.*, 2013; Sünderhauf *et al.*, 2013a; Sünderhauf *et al.*, 2013b], overcoming both simultaneously is an ongoing challenge. Typical approaches to this problem have taken pose-invariant systems and attempted to endow them with invariance to changing environmental conditions. In this paper, we pursue an alternative approach by adding pose invariance to the condition-invariant algorithm, Sequence Matching Across Route Traversals (SMART) [Pepperell *et al.*, 2014].

We present a multiple-scale image comparison method for matching places along a route with lateral camera shift. Using sideways imagery, e.g. by combining information from 2 sideways-facing cameras (Figure 1) – or sideways regions of an omnidirectional sensor – the lateral pose change problem is transformed into one of forward camera translation, which can be emulated using varied image scales. We evaluate the performance of the new multi-scale SMART algorithm against its predecessor on 2 challenging road vehicle datasets in highway

*This work was supported by an Australian Research Council DECRA grant (DE120100995) awarded to M. Milford.

and suburban environments. The evaluations involve performing place recognition across varying degrees of lateral camera shifts up to the complete width of a 4-lane road with a median strip, using images obtained from car journeys in opposing directions across a day-night cycle.

The paper proceeds as follows: Section 2 provides an overview of relevant prior work. In Section 3, we briefly summarise the SMART algorithm and describe the new multi-scale image comparison approach in Section 4. We detail our evaluation methodology in Section 5 and present the results in Section 6. Section 7 discusses the results and suggests improvements to be pursued in future work.

2 Background

A typical goal of place recognition systems is to work in large-scale environments, which is attainable by the current state-of-the-art. One such example is FAB-MAP, which successfully localised within a 1000 km journey [Cummins and Newman, 2009]. This method uses Speeded Up Robust Features (SURF) [Bay *et al.*, 2006] for image comparison, which are effective in coping with changes in scale and rotation (and hence camera pose), but perform poorly across significant changes in illumination [Milford and Wyeth, 2012; Valgren and Lilienthal, 2007]. Other notable methods using feature descriptors include MonoSLAM [Davison *et al.*, 2007] and FrameSLAM [Konolige and Agrawal, 2008]. However, condition invariance is essential for long-term operation.

Several condition invariant solutions use training to dynamically model or predict changes in appearance [Sünderhauf *et al.*, 2013b; Lowry *et al.*, 2014; Johns and Yang, 2013a; Churchill and Newman, 2012; Johns and Yang, 2014], or to build a database of scenes at different times of day with which to match features on query traverses [Johns and Yang, 2013b]. In the context of place recognition, these methods are primarily limited by the overheads in acquiring training data, and currently have limited applicability to previously unseen types of environmental change. Many also lack pose invariance. More recent approaches attempt to learn invariant place-dependent features [McManus *et al.*, 2014b], but again require training. Alternative, training-free approaches have made significant progress towards producing and using "illumination invariant" images for place recognition through knowledge of spectral response to convert colour images to greyscale forms absent of shadows [Corke *et al.*, 2013; McManus *et al.*, 2014a; Maddern *et al.*, 2014].

Exploiting temporal information, Sequence SLAM (SeqSLAM) [Milford and Wyeth, 2012] addresses changing environmental conditions by combining simple whole-image matching techniques with search methods for finding coherent sequences of image matches, rather than globally-best single frame matches. SeqSLAM demonstrated place recognition in the presence of significant motion blur [Milford *et al.*, 2013], on very long journeys [Sünderhauf *et al.*, 2013a] and using non-linear [Johns and Yang, 2013b; Naseer *et al.*, 2014] and particle filter-based search techniques [Liu and Zhang, 2013]. SeqSLAM's successor, SMART, introduced odometry-corrected search, variable offset image matching and sky removal to improve place recognition accuracy and enable the use of significantly shorter (10 m vs. 320 m) image sequences [Pepperell *et al.*, 2014]. However, SeqSLAM and SMART have very limited pose invariance, and neither has demonstrated successful localisation across multiple lanes of lateral pose change.

In this paper, we seek to bridge the gap between pose-invariant, condition-sensitive place recognition algorithms and the condition-invariant, but pose-sensitive performance of the SeqSLAM and SMART algorithms.

3 The SMART Algorithm

This section briefly summarises the SMART algorithm, which facilitates condition-invariant image matching in the combined system. For full details and mathematical formulae, please refer to [Pepperell *et al.*, 2013].

3.1 Image Comparison

SMART uses normalised Sum of Absolute Differences (SAD) applied across corresponding pixels to calculate a difference score between a pair of frames. SMART already provides a slight degree of pose invariance by comparing each query frame to each database frame (template) over a range of offsets up to horizontal and vertical maxima (in the range of $\pm x_{max}$ and $\pm y_{max}$), such that the SAD score of the overlapping region is minimised. Prior to comparison, classification is performed to automatically remove (by pixel blackening) the sky regions of daytime frames to improve matching between traverses with dissimilar sky conditions, such as clear to cloudy or day to night. Frames are then pre-processed into low-resolution, patch-normalised greyscale images, as shown in Figure 2. The multi-scale image comparison method presented here builds on this technique.

3.2 Odometry-Corrected Local Sequence Searching

SMART performs place recognition by searching for coherent sequences of image matches, rather than assessing individual frame pair differences. Odometry is used to normalise the spacing between video frames, such that sequences can be searched as straight-lines

Figure 2: Pre-processing stages of an input image. Each input frame (a) is sky blackened (if daytime) and converted to greyscale (b), then resolution-reduced and patch-normalised (c) prior to image matching.

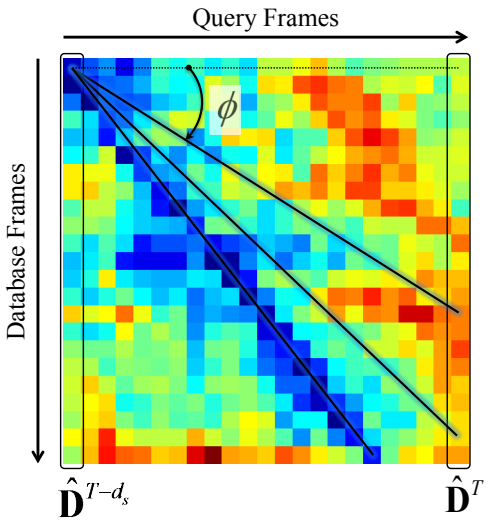

Figure 3: Difference matrix, $\mathbf{M}$. For each image difference vector (column in $\mathbf{M}$), the search algorithm finds the lowest-cost (dark blue) straight-line sequence through the matrix as the starting point (row in $\mathbf{M}$) and search angle, ϕ, are varied. With accurate odometry as provided by a car's onboard system, video frames are equally-spaced and the search can be constrained to $\phi = 45°$ lines. For clarity, only one sequence starting point is shown.

through a matrix of pairwise frame difference scores. As processed, low-resolution query and database images are compared, an image difference matrix, $\mathbf{M}$, is formed. Each element of $\mathbf{M}$ is normalised within a neighbourhood (l) in its column by subtracting the local mean and dividing by the local standard deviation to normalise scores in the local matching cohort. Coherent sequences are then found within this matrix by searching for the lowest-cost linear paths of fixed length, as shown in Figure 3.

4 Multiple-Scale Image Comparison

Our proposed technique rescales images to generate approximate equivalent views at all possible lateral vehicle shifts in both directions. Consider first, a single camera, providing 1 image per place for each

Figure 4: Example of cropping and resampling rectangular regions from an image to produce 4 different scale views of a scene with the same image size. This process is applied to all database and query frames.

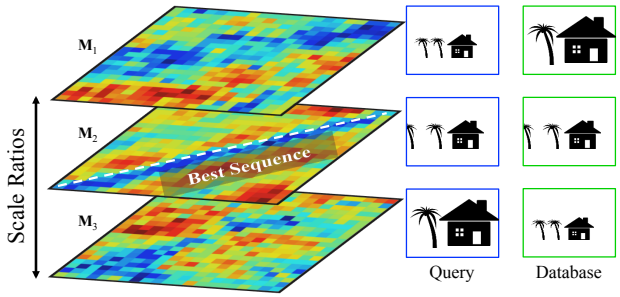

Figure 5: Sequences are independently searched through the combined image difference matrices at each scale ratio, with the lowest-cost overall sequence chosen. For clarity, only 3 scale ratios are shown.

traverse of an environment (as in Figure 1). Scaled images are produced by cropping out rectangular regions of varying size above the horizon (Figure 4) and resampling them to a common image size. The largest (least zoomed) scaled image for each query frame is then compared to all scales of each database frame and vice versa. This process creates a difference matrix for each *scale ratio*, giving $2N - 1$ matrices for N image scales – e.g. 4 scales per image gives 7 scale ratios: 3 scaling the query views to approximate database views to the east, 3 scaling the query views to approximate database views to the west, and the 1:1 scale ratio representing no relative shift.

Each scale ratio matrix, $\mathbf{M}_n$, is then independently searched for sequences (Figure 5), with the lowest-cost sequence across rows and scales chosen for each difference vector. Parallel searching of the scale ratio matrices in this manner adds another dimension to the search process shown in Figure 3.

We apply the above process to a *pair* of sideways views – in this case, from 2 cameras – providing more information about a scene, and the ability to localise

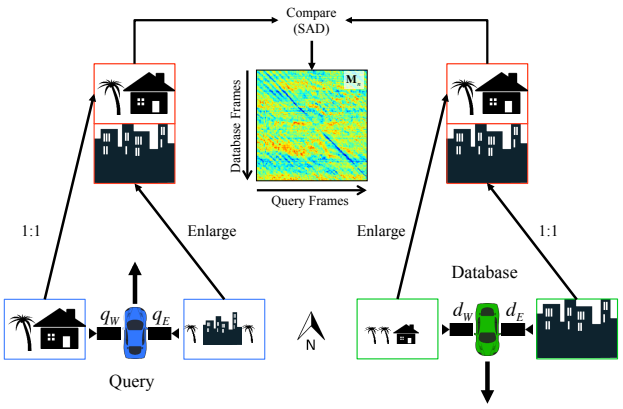

Figure 6: The corresponding camera views from the query and database traverses are scaled such that one database view is enlarged relative to the corresponding query view, and vice versa. The scaled views are then concatenated into a single image for each traverse and compared to form a difference matrix. This figure shows 1 scale ratio, appropriate for the case where the query traverse is closest to western structure and the database traverse is closest to eastern structure.

in opposite directions. Images from each camera pair along a traverse are independently cropped and scaled as above, then concatenated together into a single image before being compared to form a difference matrix for each scale ratio (Figure 6). Note that in each instance, one database camera view is enlarged relative to one query camera view, and the other is reduced.

To enable place recognition across traverses in opposite directions, the left view on one traverse must be compared to the right view of the other traverse, and vice versa. The order of one image sequence relative to the other also reverses. This situation can either be accounted for by searching in both directions (hence doubling compute), or by adding further pose information, e.g. from a compass, to detect changes in polarity between traverses.

In selecting the image regions for comparison, we make 2 assumptions – (1) that the vehicle movement is planar, such that the camera has a fixed, repeatable horizon; and (2) that surrounding structure is at an approximately fixed distance from the road edge (such that perspective distortion is not significant) over the length of a sequence. However, the first assumption could be eliminated by using a gravity signal from an accelerometer to dynamically find the horizon.

Although the second assumption appears restrictive, our results demonstrate good performance using such an approximation. No assumptions are made as to the lateral position (or lane in the case of road routes) of the platform on either traverse, other than that it remains approximately consistent through each individual sequence. The range of scales determines the maximum amount of lateral pose shift.

5 Evaluation

In this section, we describe the experimental setup, image scaling arrangements, vehicular datasets, ground truth measures and studies performed. Each experiment includes a control case as a baseline; using SMART without the new multi-scaling algorithm ("vanilla" SMART).

5.1 Experimental Setup

Video footage was recorded using GoPro Hero3 Black Edition cameras, mounted on each side of a road vehicle, facing sideways as in Figure 1. Odometry information was collected via the vehicle's OBDII port using an OBDPro USB Scantool and a laptop computer. All original videos were recorded with PROTUNE^{TM}enabled (for better low-light exposure) at a resolution of 1920 by 1440 pixels at 24 frames per second. The videos were then converted to a resolution of 480 by 360 prior to cropping and scaling. The cameras were used uncalibrated with no image rectification performed.

Figure 7: Aerial maps of the highway (top) and suburban (bottom) datasets, showing the traversed paths in red. Imagery © 2014 Cnes/Spot Image, DigitalGlobe, Landsat, Map data © 2014 Google.

5.2 Datasets

We evaluated our method on 2 road datasets[1] acquired in Queensland, Australia (Figure 7), each consisting of multiple passes across day and night on approximately 4 km, 4-lane, bidirectional road sections, divided by median strips. The first dataset was collected along the Gold Coast Highway in Palm Beach ("highway") and the second was collected along Christine Avenue in Robina and nearby suburbs ("suburban"). We refer to each lane numerically from 1 to 4, going from east to west on the highway dataset and going from south to north on the suburban dataset.

5.3 Image Cropping and Scaling

All frames of the video footage had regions extracted above the horizon as in Figure 4, which were then linearly downsampled to 64 by 32 pixels.

Scaling ratios were calculated by assuming a 10-metre distance ($D = 10$) from the outside lanes to kerbside structure and using discrete steps based on a standard lane width of 3.3 m to proportionally scale the image region of interest. Note that these quantities are arbitrary approximations and the

[1] https://wiki.qut.edu.au/display/cyphy/Datasets

algorithm was found to be insensitive to significant real-world variations in these distances (see Figures 8 and 9).

Table 1 shows the resultant 7 scale ratios for our camera setup, comprising 4 scaled views from each traverse. Each cropping coordinate is taken with the origin as the top-left corner of the 480 by 360 pixel frame. For generic place recognition with unknown lateral shift ranges, scales ratios from 1:1 to 1:D could be selected using a step size chosen to suit available computational resources.

Table 1: Cropping Parameters for Multiple Scales

Video 1, 2 Structure Distances (m)	Video 1 (East) Video 2 (West) (x, y, w, h)	Video 1 (West) Video 2 (East) (x, y, w, h)	Effective Scale Ratio
13, 23	30, 70, 420, 120	121, 122, 237, 68	1.77:1
13, 20	30, 70, 420, 120	104, 112, 273, 78	1.54:1
13, 16	30, 70, 420, 120	69, 93, 341, 98	1.23:1
13, 13	30, 70, 420, 120	30, 70, 420, 120	1:1
16, 13	69, 93, 341, 98	30, 70, 420, 120	1:1.23
20, 13	104, 112, 273, 78	30, 70, 420, 120	1:1.54
23, 13	121, 122, 237, 68	30, 70, 420, 120	1:1.77

5.4 Ground Truth

Ground truth was calculated by manually synchronising the start and end points of the traversal videos for each tested scenario and interpolating in-between frame correspondences using odometry, verified by manual inspection. The ground truth can be considered correct to within approximately 3 metres.

5.5 Studies and Parameters

We evaluated the multi-scale method against vanilla SMART across 1, 2 and 3 lanes of lateral pose shift on the highway dataset (Table 3), varying which traverse lane acted as the database and which acted as the query. To verify these results, the 1- and 3-lane shift tests were repeated on the suburban dataset with the same parameters, as shown in Table 4. All tests were done across a day-night cycle for illumination variance.

The SMART parameters used in all studies are shown in Table 2, which were chosen based on previously successful values [Pepperell *et al.*, 2013] and applied consistently across both datasets. Sequence lengths of $d_s = 30$ m and $d_s = 100$ m were used to be comparable to a previous study showing typical and perfect results with repeated pose [Pepperell *et al.*, 2014], in addition to tests with $d_s = 200$ m to evaluate whether increased sequence lengths can compensate for degraded image matching performance.

6 Results

In this section, we present example sequence matches, precision-recall curves, ground truth plots and failure cases. A video accompaniment to this paper details the methodology and presents illustrative results.

Table 2: List of SMART Parameters

Parameter	Value	Description
R_x, R_y	64, 32	Reduced image size
f_{dist}	1 m	Frame distance
P	8×8 pixels	Patch size
ϕ	45°	Trajectory search angle
l	80 templates	Neighbourhood length
x_{max}, y_{max}	0, 2	Maximum shift offsets
d_s	30, 100, 200 m	Sequence length, varied

6.1 Precision-Recall Performance and Dataset Coverage

Tables 3 and 4 show the scenarios tested and summarise the recall results at 100% precision on the highway and suburban datasets, respectively. For comparison, recall figures attained by the vanilla SMART algorithm without the new multi-scaling approach are shown in parentheses – except where 100% precision was never reached (denoted by '—').

Table 3: Highway Dataset Results

Scenario	Best Recall at 100% Precision with (without) Multi-Scaling at Sequence Length (m)		
	30	100	200
Lane 1 (Night) to Lane 2 (Day)	40% (6%)	87% (61%)	93% (84%)
Lane 1 (Night) to Lane 4 (Day)	1% (—)	24% (—)	44% (—)
Lane 3 (Day) to Lane 1 (Night)	1% (—)	12% (—)	57% (—)

Table 4: Suburban Dataset Results

Scenario	Best Recall at 100% Precision with (without) Multi-Scaling at Sequence Length (m)		
	30	100	200
Lane 1 (Night) to Lane 2 (Day)	49% (9%)	94% (24%)	95% (57%)
Lane 1 (Night) to Lane 4 (Day)	1% (—)	7% (—)	14% (—)

Precision-recall curves were generated by varying the sequence cost threshold and comparing reported matches to ground truth. Precision refers to the proportion of returned frame pairs that were correct (within 20 metres of ground truth – similar to past studies [Milford and Wyeth, 2012; Pepperell *et al.*, 2013]), and recall is the proportion of total correct frame pairs that were returned. Recall performance at 100% precision is of particular interest in a place recognition system [Cummins and Newman, 2008]. As the first and last $d_s/2$ frames do not form part of complete sequences, it is only possible to approach, but not reach, 100% recall (e.g. maximum of 95% recall for a sequence length of 200 in a database of 4000 frames).

Figures 10 and 11 show the precision-recall performance plots and example ground truth coverage (at 100% precision) for the highway and suburban datasets, respectively. Performance was significantly better in all cases with the new multi-scale algorithm – vanilla SMART performed significantly worse under 1 lane of lateral pose variance, and unusably poorly (in the order of 1% precision and recall) at larger pose variances. The multi-scale algorithm attained

Figure 8: Overhead satellite image (top) and example frame pairs (middle, bottom) equally-spaced along a 200 m sequence of the highway dataset across lanes 1 and 4 from the viewpoint of the east-facing cameras. Manually-marked visual key-points are used to show corresponding features, though these are not used by the algorithm. Overhead imagery © 2014 Nearmap Ltd.

Figure 9: Overhead satellite image (top) and example frame pairs (middle, bottom) equally-spaced along a 200 m sequence of the suburban dataset across lanes 1 and 4 from the viewpoint of the south-facing cameras. Manually-marked visual key-points are used to show corresponding features, though these are not used by the algorithm. Overhead imagery © 2014 Nearmap Ltd.

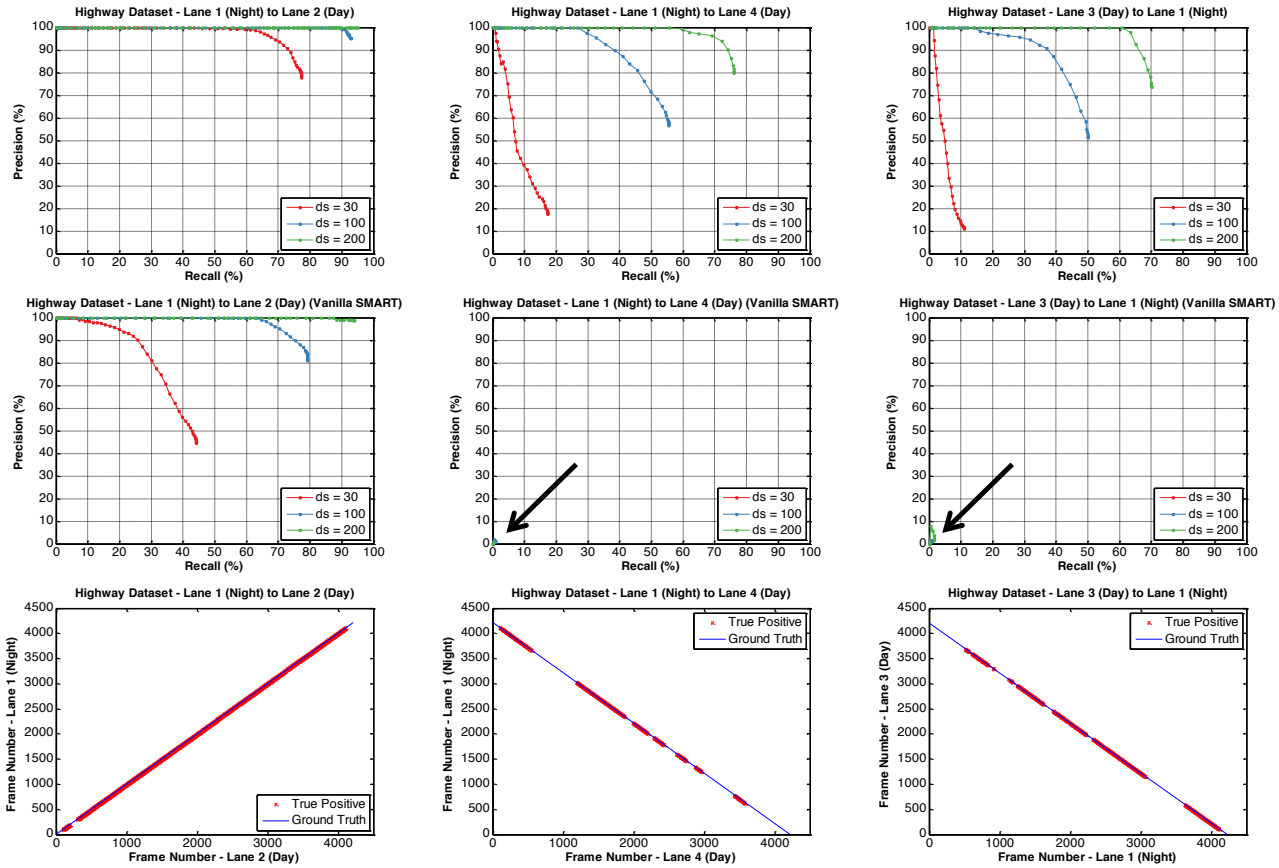

Figure 10: Precision-Recall curves for each highway dataset scenario, showing performance with sequence lengths of 30, 100 and 200 metres with the new multi-scale method (top row), as compared to vanilla SMART (middle row). No usable place recognition performance was attained without multi-scaling for lateral pose changes greater than 1 lane. The corresponding ground truth coverage plots for the new method at a sequence length of 200 m are also shown (bottom row). Note the inverted slopes for traverses in opposite directions.

significant matching coverage over all tests, with the exception of the second scenario on the suburban dataset – see Section 6.2 for a discussion of failure cases.

Example frame matches along a 200 m sequence for the lane 1 (night) to lane 4 (day) scenario in each of the highway and suburban datasets are shown in Figures 8 and 9, respectively.

Another statistic of interest is the distribution of automatically chosen scale ratios. Figure 12 shows the distribution of scale ratios amongst the matched sequences for the second highway dataset scenario, for both all possible matches and matches accepted at 100% precision. The majority (99%) of matched sequences at perfect precision were from the scale ratio representing 3 lanes of lateral pose shift, as expected.

6.2 Failure Cases

Performance was worst on the scenarios with the largest change in lateral pose. Recall performance – and hence coverage – was lowest on the suburban dataset, which can be partly attributed to it having prevalent occlusions on road dividers. Figure 13 shows example scenes from this dataset where matching was impossible due to lack of line of sight across the road. Section 7 discusses possible mitigation strategies.

6.3 Computation and Storage

Our evaluation consisted of exhaustive image comparisons and sequence searches through all database frames across all spatial scales, effectively emulating random localisation within the tested environments. This approach results in a linear growth in computational load with the size of the database. The calculation below provides the approximate computational requirements per frame queried on one of our datasets at 1-metre separation when travelling at 60 km/h:

Whole Image Comparisons: 4000 images × 2 cameras × 7 scales × 2048 pixels/image × 5 vertical offsets × 17 fps = $\mathbf{9.75 \times 10^9}$ **comparisons/s**

Sequence Searches: 200 elements/sequence × 4000 sequences × 7 scales × 17 fps = $\mathbf{9.5 \times 10^7}$ **comparisons/s**

Thus, even with an efficient algorithm; performing one comparison per clock cycle, exhaustive real-time computation is currently not feasible on a typical CPU in 2014 (though could be with a GPU). We propose further use of odometry to constrain the comparisons and search to smaller local neighbourhoods within the database (after an initial slower global localisation). Likewise, while we demonstrated our method working successfully with no knowledge of the database or query traverse lane positions, acquiring some

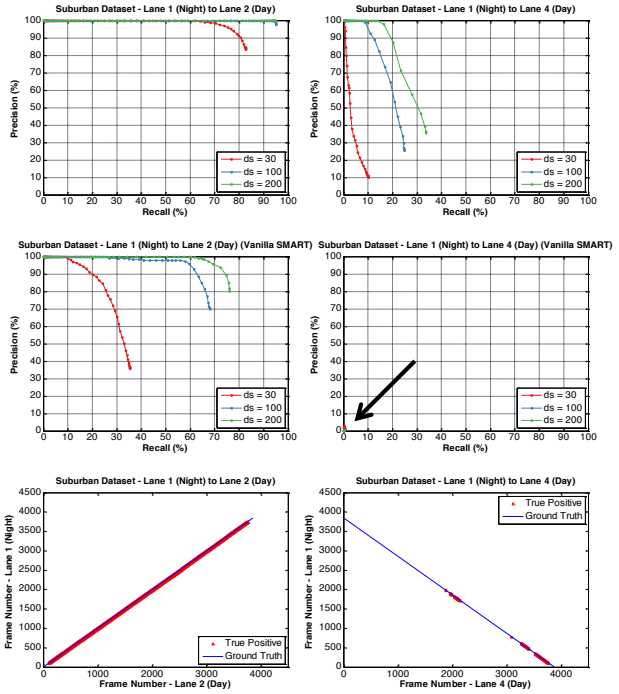

Figure 11: Precision-Recall curves for each suburban dataset scenario, showing performance with sequence lengths of 30, 100 and 200 metres with the new multi-scale method (top row), as compared to vanilla SMART (middle row). No usable place recognition performance was attained without multi-scaling for lateral pose changes greater than 1 lane. The corresponding ground truth coverage plots for the new method at a sequence length of 200 m are also shown (bottom row). Note the inverted slope for traverses in opposite directions.

knowledge of lane positioning (especially in the database case, which can be pre-labelled) would further reduce the scale search space and hence computational requirements.

Storage requirements also grow linearly with the size of the database. Prior calculations regarding the storage of images in large-scale environments showed that the road networks of a medium country could be represented by a few gigabytes of very low-resolution templates [Milford, 2013]. Projecting these calculations to allow for 2 camera views at 7 scales each at a higher resolution of 64 by 32 pixels, an environment of this size has storage requirements in the order of a few terabytes – a feasible ask for a hard drive in 2014.

7 Discussion and Future Work

We presented a technique for achieving laterally pose-invariant place recognition along a route, by using a multi-scale image comparison method that generates scaled views based on possible lateral camera shifts. Combined with the SMART algorithm, the system successfully performed place recognition across day-night cycles, but would also work across weather and seasonal variations where sequence-based approaches have previously been shown effective.

The results demonstrated that short sequences

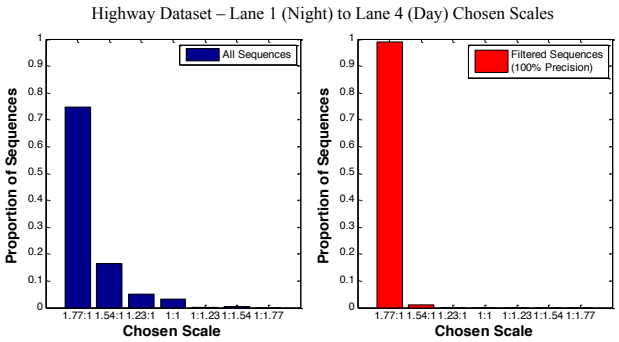

Figure 12: Histograms of scale ratios automatically chosen for the highway dataset for 200-metre sequences in the lane 1 (night) to lane 4 (day) scenario, showing the distribution of scale ratios chosen across all sequences (left) and after filtering to 100% precision (right).

previously demonstrated effective with SMART over repeated pose remain effective with the multi-scale approach in achieving high recall over modest lateral pose change (e.g. 1 lane). However, as the lateral shift is increased, matching performance weakens, as (1) the larger degree of image scaling exaggerates perspective differences; (2) occlusions by other vehicles become more frequent; and (3) in the case of travelling in opposite directions, matching becomes difficult or impossible in sections where the road is separated by foliage or tall concrete barriers. We suspect that using a tall, roof-mounted omnidirectional camera for a higher vantage would somewhat ameliorate the second problem. Additionally, employing a technique to automatically change the sequence length (such as with a particle filter [Liu and Zhang, 2013]) would enable high recall in difficult route sections, while retaining faster, short-sequence localisation in less

Figure 13: Example failure cases as 2 separate night-day frame pairs (top row, bottom row) at the same location from the suburban dataset in the lane 1 (night) to lane 4 (day) scenario, looking south. The algorithm failed to correctly match sequences around these frames due to occlusions from trees and bushes on the median strip and roundabout (false negatives).

challenging sections. Street imagery databases such as Google Street View often collect imagery along both sides of large highways; offering another potential solution.

Our multi-scale approach localised successfully without making assumptions about the lane of travel in either traverse. Road networks could be mapped into databases without any constraints on their acquisition, potentially making crowd-sourcing feasible (and enabling usage of existing variable lane databases such as Google Street View). The traverse direction polarity – whether a vehicle is travelling in the same or opposite direction to the database route – can easily be determined with a compass. Combining a panoramic camera with a compass could enable localisation along arbitrary, "squiggly" route traverses.

The naive assumption of world features being at fixed distances from the camera worked effectively in combination with long sequences at correctly recognising familiar scenes. A potential avenue for future work would be to leverage 3D reconstruction or scene understanding techniques. As the frame spacing is known from odometry, tracking features (or patches) between adjacent frames would enable depth to be inferred and used to scale the camera views accordingly; eliminating the need for brute-force scale ratio searching. Accurate depth estimation is contingent on having line of sight to common scene structure between visits to a given place, and would likely be challenged by occluding obstacles.

We believe that the approach of starting with condition-invariant techniques and endowing them with pose invariance is an under investigated one. Through the research presented here and in future, we hope to stimulate others to pursue this approach; leveraging their expertise in sophisticated techniques for pose-invariant place recognition.

Acknowledgments

The authors thank Richard Pepperell for his assistance in collecting the road datasets used in this paper. Additionally, we acknowledge computing resources provided by the HPC and Research Support Group, Queensland University of Technology, Brisbane, Australia.

References

[Bay *et al.*, 2006] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded Up Robust Features. In *Computer Vision ECCV 2006*, pages 404–417, 2006.

[Churchill and Newman, 2012] Winston Churchill and Paul M Newman. Practice makes perfect? Managing and leveraging visual experiences for lifelong navigation. In *IEEE International Conference on Robotics and Automation (ICRA)*, St. Paul, United States, 2012. IEEE.

[Corke *et al.*, 2013] Peter Corke, Rohan Paul, Winston Churchill, and Paul Newman. Dealing with shadows: Capturing intrinsic scene appearance for image-based outdoor localisation. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 2085–2092. IEEE, 2013.

[Cummins and Newman, 2008] Mark Cummins and Paul Newman. FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance. *International Journal of Robotics Research*, 27(6):647–665, 2008.

[Cummins and Newman, 2009] Mark Cummins and Paul Newman. Highly scalable appearance-only SLAM - FAB-MAP 2.0. In *Robotics: Science and Systems*, Seattle, United States, 2009.

[Davison *et al.*, 2007] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. MonoSLAM: Real-Time Single Camera SLAM. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1052–1067, 2007.

[Johns and Yang, 2013a] Edward Johns and Guang-Zhong Yang. Dynamic Scene Models for Incremental, Long-Term, Appearance-Based Localisation. In *Proc. ICRA*, 2013.

[Johns and Yang, 2013b] Edward Johns and Guang-Zhong Yang. Feature Co-occurrence Maps: Appearance-based Localisation Throughout the Day. In *IEEE International Conference on Robotics and Automation*, Karlsruhe, Germany, 2013. IEEE.

[Johns and Yang, 2014] Edward Johns and Guang-Zhong Yang. Generative methods for long-term place recognition in dynamic scenes. *International Journal of Computer Vision*, 106(3):297–314, 2014.

[Klein and Murray, 2007] George Klein and David Murray. Parallel tracking and mapping for small AR workspaces. In *International Symposium on Mixed and Augmented Reality*, Nara, Japan, 2007. IEEE ACM.

[Konolige and Agrawal, 2008] K Konolige and M Agrawal. FrameSLAM: From Bundle Adjustment to Real-Time Visual Mapping. *IEEE Transactions on Robotics*, 24(5):1066–1077, 2008.

[Liu and Zhang, 2013] Yang Liu and Hong Zhang. Towards improving the efficiency of sequence-based SLAM. In *IEEE International Conference on Mechatronics and Automation (ICMA)*. IEEE, 2013.

[Lowry *et al.*, 2014] Stephanie Lowry, Michael J Milford, and Gordon F Wyeth. Transforming Morning to Afternoon using Linear Regression Techniques. In *IEEE International Conference on Robotics and Automation*, Hong Kong, 2014.

[Maddern *et al.*, 2014] Will Maddern, Alexander D Stewart, Colin McManus, Ben Upcroft, Winston Churchill, and Paul Newman. Illumination

Invariant Imaging: Applications in Robust Vision-based Localisation, Mapping and Classification for Autonomous Vehicles. *Proc. of Workshop on Visual Place Recognition in Changing Environments, IEEE International Conference on Robotics and Automation (ICRA)*, 2014.

[McManus *et al.*, 2014a] Colin McManus, Winston Churchill, Will Maddern, Alexander D Stewart, and Paul Newman. Shady Dealings: Robust, Long-Term Visual Localisation using Illumination Invariance. *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, 2014.

[McManus *et al.*, 2014b] Colin McManus, Ben Upcroft, and Paul Newman. Scene Signatures: Localised and Point-less Features for Localisation. In *Proceedings of Robotics Science and Systems (RSS)*, Berkley, CA, USA, 2014.

[Milford and Wyeth, 2012] Michael Milford and Gordon Wyeth. SeqSLAM: Visual Route-Based Navigation for Sunny Summer Days and Stormy Winter Nights. In *IEEE International Conference on Robotics and Automation*, St Paul, United States, 2012. IEEE.

[Milford *et al.*, 2013] Michael J Milford, Ian Turner, and Peter I Corke. Long Exposure Localization in Darkness Using Consumer Cameras. In *IEEE International Conference on Robotics and Automation*, Karlsruhe, Germany, 2013.

[Milford, 2013] Michael Milford. Vision-based place recognition: how low can you go? *The International Journal of Robotics Research*, 32(7):766–789, 2013.

[Naseer *et al.*, 2014] Tayyab Naseer, Luciano Spinello, Wolfram Burgard, and Cyrill Stachniss. Robust Visual Robot Localization Across Seasons using Network Flows. 2014.

[Pepperell *et al.*, 2013] Edward Pepperell, Peter I Corke, and Michael J Milford. Towards Persistent Visual Navigation using SMART. In *Australasian Conference on Robotics and Automation (ACRA)*, Sydney, Australia, 2013.

[Pepperell *et al.*, 2014] Edward Pepperell, Peter I Corke, and Michael J Milford. All-Environment Visual Place Recognition with SMART. In *IEEE International Conference on Robotics and Automation*, Hong Kong, 2014.

[Sünderhauf *et al.*, 2013a] Niko Sünderhauf, Peer Neubert, and Peter Protzel. Are We There Yet? Challenging SeqSLAM on a 3000 km Journey Across All Four Seasons. In *Proc. of Workshop on Long-Term Autonomy, IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2013.

[Sünderhauf *et al.*, 2013b] Niko Sünderhauf, Peer Neubert, and Peter Protzel. Predicting the Change A Step Towards Life-Long Operation in Everyday Environments. In *Robotics: Science and Systems (RSS) Robotics Challenges and Vision Workshop*, 2013.

[Valgren and Lilienthal, 2007] C Valgren and A Lilienthal. SIFT, SURF and Seasons: Long-term Outdoor Localization Using Local Features, 2007.