Towards Condition-Invariant, Top-Down Visual Place Recognition

Michael Milford[1], Eleonora Vig[2], Walter Scheirer[2] and David Cox[2]
[1]School of Electrical Engineering and Computer Science
Queensland University of Technology
[2]School of Engineering and Applied Sciences and Department of Molecular and Cellular Biology
Harvard University
[1]michael.milford@qut.edu.au

Abstract

In this paper we present a novel place recognition algorithm inspired by recent discoveries in human visual neuroscience. The algorithm combines intolerant but fast low resolution whole image matching with highly tolerant, sub-image patch matching processes. The approach does not require prior training and works on single images (although we use a cohort normalization score to exploit temporal frame information), alleviating the need for either a velocity signal or image sequence, differentiating it from current state of the art methods. We demonstrate the algorithm on the challenging Alderley sunny day – rainy night dataset, which has only been previously solved by integrating over 320 frame long image sequences. The system is able to achieve 21.24% recall at 100% precision, matching drastically different day and night-time images of places while successfully rejecting match hypotheses between highly aliased images of different places. The results provide a new benchmark for single image, condition-invariant place recognition.

1 Introduction

As camera technology has matured and dropped rapidly in price over the past decade, there has been a proliferation of vision-based robotic mapping and navigation algorithms including FAB-MAP [Cummins and Newman, 2009], MonoSLAM [Davison, et al., 2007], FrameSLAM [Konolige and Agrawal, 2008], V-GPS [Burschka and Hager, 2004], Mini-SLAM [Andreasson, et al., 2007], SeqSLAM [Milford, 2013, Milford and Wyeth, 2012] and others [Andreasson, et al., 2008, Paz, et al., 2008, Royer, et al., 2005, Zhang and Kleeman, 2009, Konolige, et al., 2008, Milford and Wyeth, 2008]. Many of these systems are capable of mapping performance that rivals or exceeds range-based systems, including mapping of routes as long as 1000 km [Cummins and Newman, 2009]. However, it is becoming increasingly apparent that vision-based approaches have at least one very significant disadvantage – their susceptibility to changing environmental conditions. If the many advantages of visual sensors - low cost, small size, passive sensing and low power consumption – are ever to be exploited on mobile robots and in personal navigation systems operating over long periods of time in real-world, unstructured environments, this challenge must be solved. Current vision-based approaches to the problem are limited by one or more significant restrictions such as requiring hand-picked training data [Johns and Yang, 2013, Sunderhauf, et al., 2013], camera motion information, or long image sequences [Milford and Wyeth, 2012].

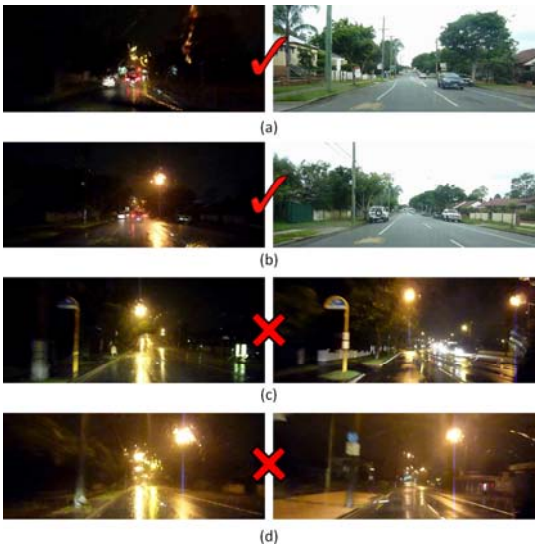

Figure 1: Using single frames only, with no prior training, motion information or temporal filtering, the top-down, multi-stage place recognition algorithm presented here is able to perform instantaneous place recognition between (a-b) very perceptually different images while also rejecting (c-d) incorrect matches between aliased image pairs.

In this paper, we present a novel multi-step vision-based place recognition system inspired by the recent discovery in human neuroscience [Rust and DiCarlo,

This work was supported by an Australian Research Council DECRA grant DE120100995, awarded to Michael Milford.

2010] that as visual information travels along the human visual cortical area, the brain simultaneously increases both its selectivity and matching *tolerance* or invariance. We extend this concept to the domain of place recognition, by implementing an initial low resolution, low tolerance whole image matcher followed by a higher resolution, highly tolerant patch matching stage. We test the system on the highly challenging Alderley dataset [Milford and Wyeth, 2012], which comprises both sunny day-time and rainy night-time footage. The system is able to perform error-free single-image place recognition at a 21% recall rate, matching places that have undergone huge perceptual change while correctly rejecting perceptually similar but different places, especially similar night-night scenes (Figure 1).

Because of the ambitious nature of the challenge, we make some significant assumptions about scope. The primary scope limitation is that we are only attempting to address the condition invariance problem and not the pose invariance problem. There is a large body of existing research on pose invariant recognition [Cummins and Newman, 2009, Davison, et al., 2007, Konolige and Agrawal, 2008, Klein and Murray, 2007] generally based on feature-based techniques like SIFT [Lowe, 2004] and SURF [Bay, et al., 2006], albeit in less challenging environmental conditions than shown in Figure 1, which has been shown to be difficult for conventional feature detectors [Milford and Wyeth, 2012, Valgren and Lilienthal, 2007]. In the Discussion section, we describe potential ways in which we can expand the approach to provide varying degrees of pose invariance. Furthermore, the current system is purely a place learning and recognition system, and the rate of learning is fixed. Integrating it into an existing mapping framework such as RatSLAM [Milford and Wyeth, 2010] would provide mechanisms for bounding learning and producing a spatial map.

The paper proceeds as follows. In Section 2 we review vision-based place recognition and mapping algorithms and detail recent attempts to improve their robustness to environmental change. Section 3 describes the approach taken in this paper. In Section 4 we describe the experimental setup, with results presented in Section 5. The paper concludes in Section 6 with discussion including future research areas.

2 Background

Vision-based place recognition is an integral component of many robotic mapping systems. After the initial drive towards mapping larger environments ever more accurately, attention has now turned towards dealing with the problem of dealing with environmental change. Current vision-only approaches generally fall into one or more of three different categories; approaches which attempt to learn how the appearance of an environment changes, approaches which use temporal filtering over long sequences of images, and approaches which attempt to develop condition-invariant features or image descriptors.

To learn how the appearance of the environment changes, training data with established frame correspondences is required. [Sunderhauf, et al., 2013] presents an approach that learns systematic scene changes in order to improve performance on a seasonal change dataset. [Johns and Yang, 2013] builds a database of observed features over the course of a day and night. These current approaches have at least two significant limitations; they require appropriate training data for a particular environment, and learnt change information has not yet been show to generalize to different environments or different, unwitnessed types of change.

Using more than a single image to form place recognition hypotheses reduces the requirements of the place matching algorithm; instead of reporting globally correct matches, the algorithm must only generally report matches that are significantly better than chance. This approach, used in SeqSLAM [Milford and Wyeth, 2012] and follow up work [Johns and Yang, 2013, Sunderhauf, et al., 2013, Sunderhauf, et al., 2013] enables place recognition in challenging conditions including the dataset presented in this paper. The disadvantage is that quite long sequences (320 frames) were required to generate good (35% recall at 100% precision) performance. For non-constant robot speed applications, this long sequence requirement in turn requires either the use of velocity information, or a larger sequence match search space [Johns and Yang, 2013] that allows for possible velocity changes, but leads to both greater computation time and the increased risk of finding false positives.

Finally, attempts to generate truly invariant feature detectors have met with limited success. SIFT [Lowe, 2004], SURF [Bay, et al., 2006] 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) far less than that shown in Figure 1.

In this paper, we attempt to fill a capability gap by providing a training-free method that can match single images and does not require velocity information.

3 Approach

This section describes the place recognition components, overviewed in Figure 2. A camera image is compared to all stored images, first at a whole image matching stage, then at a patch matching stage, with the output evaluated using a patch shift coherency calculation.

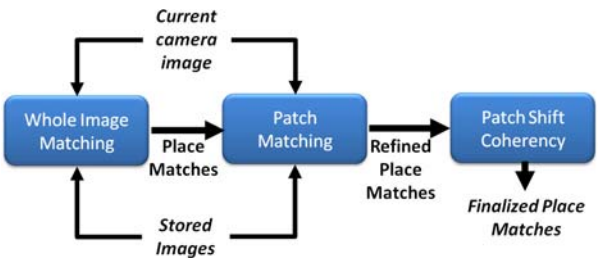

Figure 2: System architecture. A camera image is compared to stored images firstly at a whole image level, then at a patch-based level and finally at a patch-shift coherence level.

3.1 Whole Image Place Recognition

Camera images are resolution reduced (64×32 pixels) then patch normalized (all pixels). Patch normalized pixel intensities, I', are given by:

$$I'_{xy} = \frac{I_{xy} - \mu_{xy}}{\sigma_{xy}} \quad (1)$$

where μ_{xy} and σ_{xy} are the mean and standard deviation of pixel values in the patch of size P_{size} that (x, y) is located within. Mean image differences between the current image and all stored images are calculated using a normalized sum of intensity differences, performed over a range of horizontal and vertical offsets:

$$D_j = \min_{\Delta x, \Delta y \in [-\sigma, \sigma]} g(\Delta x, \Delta y, i, j) \quad (2)$$

where σ is the template offset range, and $g()$ is given by:

$$g(\Delta x, \Delta y, i, j) = \frac{1}{s} \sum_{x=0} \sum_{y=0} \left(p^i_{x+\Delta x, y+\Delta y} - p^j_{x,y} \right) \quad (3)$$

where s is the area in pixels of the template sub frame. The range of horizontal and vertical offsets provides some invariance to camera pose.

In this implementation, we simply add new images to the library of stored images at a fixed rate (1 every 2 frames, corresponding to a maximum inter-frame separation of 1.1 metres for presented dataset.

3.2 Cohort-based Normalization

The vector of difference scores output by Equation 2 is normalized twice. Firstly, the difference score matches between the current camera frame and all stored frames are normalized as follows:

$$\hat{D}_i = \frac{D_i - \overline{\mathbf{D}}}{\sigma} \quad (4)$$

where D_i is the original difference score for the match between the current frame and the i^{th} frame.

The second normalization is based on the standard speaker recognition and computer vision technique of normalizing scores by cohort [Furui, 1997, Aggarwal, et al., 2008, Tulyakov, et al., 2008]. We use a modified version that uses video frame time-stamps to normalize different scores by time. Datasets are "chunked" into r temporally contiguous frame groups. Each difference score D is then normalized as follows:

$$\hat{D}_{ij} = \frac{D_{ij} - \overline{\mathbf{D}_j}}{\sigma_j} \quad (5)$$

where D_{ij} is the i^{th} difference score within the j^{th} dataset chunk. As a point of clarification, cohort normalization only uses *past* camera frames so the method is real-time capable – "future" frame information is not used.

Finally, to stop the system matching the current frame to the immediately preceding frame, we truncate cohort normalization and place matching l frames from the current frame. In a full SLAM system, this same outcome would be achieved using odometry and a particle cloud; in our place recognition-only system, the implication is that the system is unable to match the current place to itself.

3.3 Sub-Image Patch Matching

Whole image matching performance on low resolution images degrades significantly when perceptual change becomes large enough (such as over day-night cycles), which is why previous work has focused on matching using long sequences of images [Milford and Wyeth, 2012]. The novel patch verification process presented here is performed on images from the Z top ranked place match hypotheses proposed by the whole image matcher described in the previous section. Small image patches at corresponding locations in the two images (see Figure 3) are compared using a sum of absolute differences calculation similar to that described in Equations 2 and 3. Comparisons are performed over a sliding window centred on the patch location, but extending in both vertical and horizontal directions. However, rather than just finding the maximal patch match and its associated offsets, the entire set of difference scores for each patch comparison are used to create a difference score ratio g_{rat}:

$$g_{rat} = \frac{g_1}{g_2} \quad (6)$$

where g_1 is the difference score for the best matching patch offset and g_2 is the score for the next best matching offset located outside a range of r_{peak} from the first score. A count of patch matches with difference score ratios exceeding a minimum score requirement g_m (value given in Table 1) produces an overall patch match count q:

$$q = \sum g_{rat} > g_m \quad (7)$$

Examples of patch matches meeting the quality requirements are shown in the Section 5.

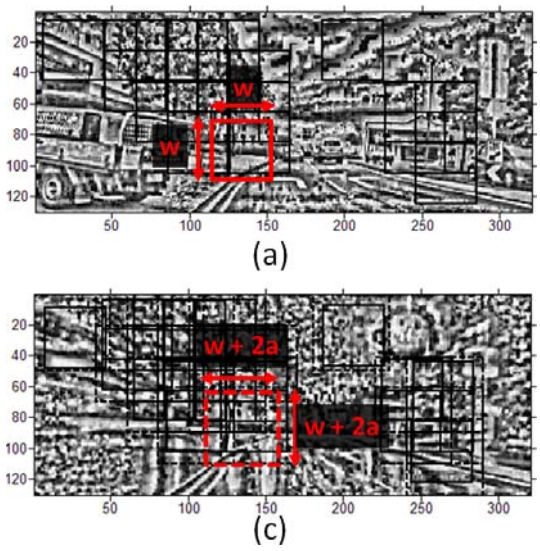

Figure 3: Patch verification involves comparing (a) small patches at (b) corresponding locations in a proposed matching image over a local sliding window [-*a, a*].

3.4 Patch Shift Coherency

To further evaluate the place match likelihood, a coherency check is performed on the reported shift offsets for the q matching patches meeting the quality requirement set in Section 3.3. The horizontal and vertical shifts are binned in a two-dimensional histogram **H** which is then smoothed using a moving summation window of radius s_{rng} (see Figure 4). We define two shift coherency metrics, the peak shift count c:

$$c = \max(\mathbf{H}) \tag{8}$$

and the peak-mean ratio r_{pm}:

$$r_{pm} = \frac{c}{\overline{\mathbf{H}}} \tag{9}$$

The peak shift count provides an absolute measure of the number of spatially coherent patch matches, while the peak-mean ratio provides a patch consensus measure independent of the total patch match count.

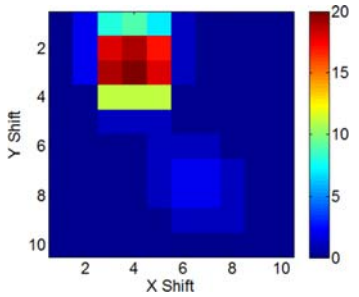

Figure 4: (a) Patch shift coherency verification involves creating a 2D histogram of the spatial shifts for patch matches, from which coherency metrics are calculated.

4 Experimental Setup

This section describes the experimental environment, dataset acquisition and pre-processing, ground truth creation and key parameter values.

4.1 Camera Equipment

A Panasonic Lumix DMC-TZ7 digital snapshot camera was mounted forward facing on the car dashboard, recording 720p video at a frame rate of 25 frames per second. The video was cropped as shown in Fig. 5. Due to heavy rain, the resulting video stream had significant and constant visual artefacts due to water streaming down the windscreen, windscreen wipers, compression artefacts and poor night-time illumination.

Figure 5: (a) Video acquisition for the Alderley dataset. A forward facing Panasonic Lumix DMC-TZ7 camera captured 720p video at 25 frames per second through the windshield. Images were cropped to remove the main dashboard areas.

4.2 Alderley Dataset

The Alderley dataset comprises two 8 km journeys over the same route through the suburb of Alderley in Brisbane, Australia (Fig. 6). The first run was gathered in the middle of the night during a severe storm with very heavy rain and low visibility. The second run was gathered during a bright clear morning. The car's velocity was typically between 45 and 60 km/hr throughout the dataset except when slowing down to stop due to traffic.

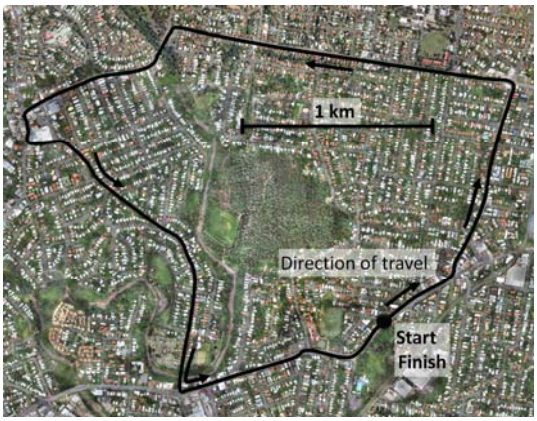

Figure 6: Aerial photo and camera path for the Alderley dataset. An 8 km long route was traversed twice, once during sunny day-time conditions and once during heavy rain at night.

4.3 Ground Truth

GPS was not gathered during acquisition of the Alderley dataset. Consequently, to obtain ground truth, the videos were manually parsed frame by frame to pick key frame correspondences. Points were selected based on video frames that showed prominent, unambiguous features and were more densely sampled around transition points (such as the car stopping and starting at traffic lights). 93 locations were tagged in the two Alderley datasets. The manually selected frame pairs can be considered correct to within 5 frames in the original 25 fps video, corresponding to a maximum ground truth error (at 60 km/hr) of approximately 3 metres.

4.4 Image Pre-Processing

Image contrast enhancement was performed on the day- and night-time road datasets (although the day dataset did not "need" image enhancement, the same enhancement was applied for the purposes of consistency). Many consumer cameras, including the ones used in this experiment, capture video in YV12 format (chroma sampling scheme 4:2:0), which provides a useful 12 bits of intensity information per pixel, while sacrificing color representation. Often this extra intensity information is lost in a standard processing chain, but we applied brightening and histogram equalization to the original YV12 format images before converting them into standard grayscale images for use by the place recognition and visual odometry algorithms. These images were then down sampled to the resolutions required by each place recognition module and then patch normalized (Fig. 7).

Images were processed at a rate of 15 Hz by the place recognition algorithms and a simple no-motion detector (based on image change) was used to pause processing at extended stoppages at traffic lights, as in the original SeqSLAM study. The day dataset was processed first, meaning the place recognition algorithm was required to match night-time images back to day-time images *and* avoid incorrectly matching night-time images to night-time images.

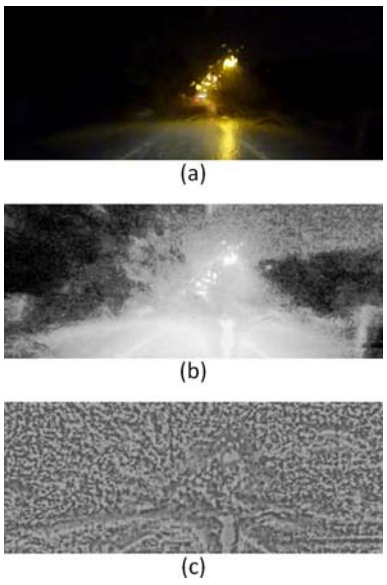

Figure 7: (a) Raw YV12 camera frames with 12 bits of intensity information per pixel were (b) histogram equalized then resolution reduced and (c) patch normalized to produce the input images for both the whole image and patch-based place recognition algorithms.

4.5 Parameter Values

Parameter values are given in Table I:

TABLE I
PARAMETER LIST

Parameter	Value	Description
R_x, R_y	64,32	Whole image matching resolution
R_x, R_y	320,130	Patch-normalized image resolution for patch verification
f_{jump}	2 frames (1.1 metres max)	Frame learning rate
Z	5 top matches	Number of place match hypotheses evaluated by the patch verification process
w	40×40 pixels	Patch size for patch verification
a	5 pixels	Patch verification local search range
r_{peak}	2 pixels	Patch quality score peak search exclusion zone
g_m	1.125	Minimum difference score ratio for an accepted patch match
l	75 frames	Recently visited place matching exclusion zone
S_{rng}	1 pixel	Sliding summation window radius for patch shift histogram

5 Results

In this section we present precision-recall curves and ground truth plots and compare performance to a whole image-only approach and the SeqSLAM algorithm. We also present patch matches and patch shift coherency histograms for both accepted and rejected place matches that illustrate how the system works. A video accompaniment to this paper further demonstrates the methodology and results.

5.1 Precision-Recall Curves

Figure 9 shows the precision-recall curves with (solid blue line) and without (dashed red line, whole image only matching) patch verification. These curves were generated with a false positive distance threshold of 13 metres, which is a third of the 40 metre distance used in the original SeqSLAM study.

Due to the perceptual difficulty of the dataset, the low resolution, whole image matching technique is never able to achieve 100% precision at any recall level,

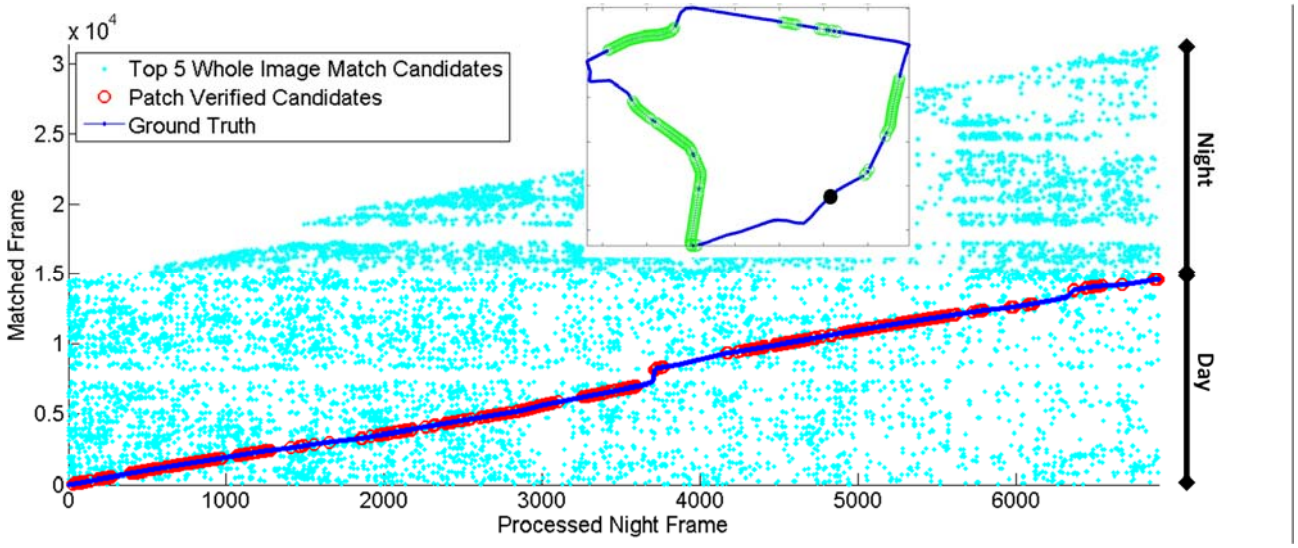

Figure 8: Ground truth plot at 100% precision and 21.24% recall, with inset from the 2012 SeqSLAM [Milford and Wyeth, 2012] results at 100% precision. The vertical axis represents stored frames from the initial day-time run (0 – 14500) and night-time run (14500 – ~31000). Qualitatively the environmental coverage is much more consistent, with the largest place recognition gap being approximately 280 metres, compared with approximately 1400 m in the original result. The solid black circle in the inset represents the starting location, with the route traversed in an anti-clockwise direction.

peaking at approximately 93.6% precision. In contrast, the top-down method achieves 100% precision up to a maximum recall rate of 21.2%, and then drops to 38.5% recall and precision. The top-down method is able to achieve higher absolute recall levels because it is able to discard incorrect but top-ranked place match candidates output by the whole image matcher and instead confirm lower ranked but correct whole image match candidates. As can be seen in the attached video and the results figures, the dataset is extremely challenging and we would argue that even a human exploiting semantic information would find it difficult to achieve a high level of recall at 100% precision.

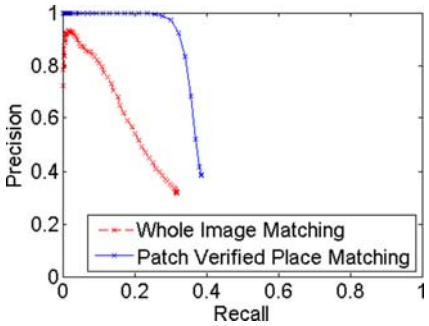

Figure 9: Precision-recall curves for the whole image matching and top-down matching method presented here. The patch verification step results in a significant improvement in precision-recall performance at precision levels. Perhaps unsurprisingly given the nature of the dataset, single frame-based whole image matching is incapable of reaching 100% precision at any recall level.

5.2 Place Recognition Distribution

Figure 8 shows the distribution of patch-verified place recognition hypotheses (red hollow circles) for a precision level of 100% and recall rate of 21.24%. The small cyan dots indicate the 5 top ranking place match hypotheses after the initial whole image matching stage, with the solid dark blue line indicating ground truth. The inset shows the distribution of place matches in the original SeqSLAM implementation at 100% precision. Although the overall recall rate of our current approach is lower (21% versus 35% recall), qualitatively the environmental coverage is much more even, a desirable characteristic for robot mapping systems [Cummins and Newman, 2009]. The longest segment of no reported matches is approximately 280 metres, versus approximately 1400 metres in the SeqSLAM result.

5.3 Sample Place Matches

Figures 10 to 13 show sample place matches and rejected place matches for the system operating at 100% precision and 21.24% recall. One of the more challenging successful place matches is shown in Figure 10. Despite vastly different perceptual conditions, and the movement of some vehicles, the algorithm is able to find a large number of high quality patch matches. The smoothed histogram of patch match shifts (Figure 10f) is highly coherent with a peak matching score of 15.

In contrast, Figure 11 shows two perceptually similar images of different places that were matched by the initial whole image matcher. The patch verification process finds a significant number of patch matches exceeding the minimum difference score ratio threshold, but the shift histogram is less coherent than in Figure 10f, with a maximum matching score of only 10. We have handpicked this example since it was one of the most challenging – at higher recall rates, this place match is one of the first to be incorrectly accepted.

Figures 12 and 13 show two of the five place matches output by the whole image matching process for the place shown in the night image (Figure 12b and 13b are the same place). The patch matcher finds a large number of matches in both cases, but after histogram binning the matching score for the match in Figure 12 ends up highest (20 versus 13), and hence the image shown in Figure 12a was chosen as the best match. However, even the second best patch verified match has a higher matching score than the highly aliased match shown in Figure 11, demonstrating the advantages of verifying multiple place match hypotheses from the whole image matching stage.

5.4 Compute

The algorithms are currently primarily implemented in Matlab and are not capable of processing images at real-time speed e.g. 15 frames per second. The primary computational load is due to the initial whole image matching process followed by the patch verification process. The calculation below gives the approximate computational requirements for these operations at the end of the dataset used in this paper:

Whole Image Comparisons

$$64 \text{ pixels} \times 32 \text{ pixels} \times 3 \text{ xshift} \times 3 \text{ yshift} \times 13576 \text{ frames} \times 15 \text{ fps} = \mathbf{3.8 \times 10^9 \text{ pixel comparisons/s}}$$

Patch Verification

$$40 \text{ pixels} \times 40 \text{ pixels} \times 10 \text{ xshift} \times 10 \text{ yshift} \times 14 \text{ patches across} \times 5 \text{ patches down} \times 15 \text{ fps} \times 5 \text{ candidates/comparison} = \mathbf{8.4 \times 10^8 \text{ comparisons/s}}$$

At least a one-to-one ratio between 8 bit pixel comparisons and nominal computer clock speed is usually achievable using just the CPU, suggesting that simply porting the algorithms to C code should render them close to real-time for datasets of this size on a modern PC, without resorting to GPU-based computation.

The compute growth of calculating low resolution image matches is already detailed in [Milford, 2013]. Compute scales linearly with the number of images stored and the square of the degree of pose invariance that is required by the matching process.

To obtain further computational speed-ups, we are currently examining several possibilities: the use of a hierarchical spatial pyramid, where only a fraction of promising image matches at each resolution are then verified at a higher resolution; the use of human-inspired saliency measures, in order to identify and compute only the most salient image regions; and finally use of specialist hardware such as GPU computation.

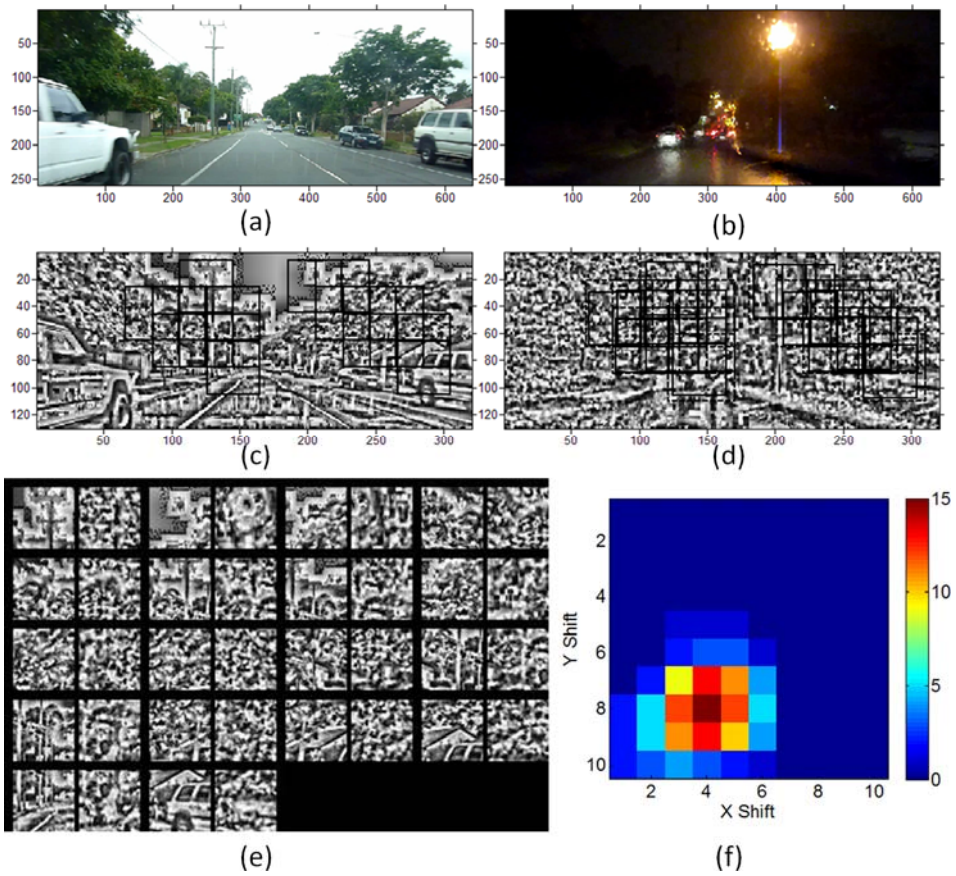

Figure 10: One of the more challenging place matches correctly identified by the system. (a-b) Original images. (c-d) Patch normalized images with black rectangles indicating the patch matches exceeding the quality threshold, with patches shown in (e). (f) The smoothed 2D histogram of patch match shifts. The overall matching score for this image was 15.

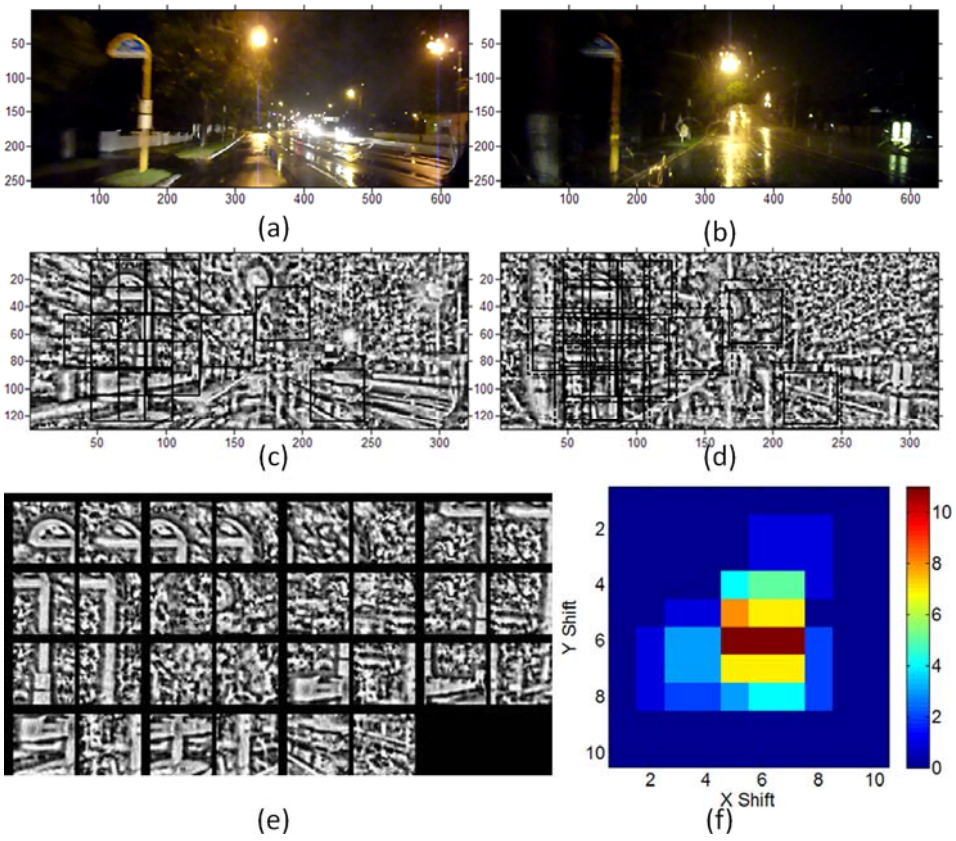

Figure 11: Two highly aliased but spatially separate places that were matched by the whole image matcher but then successfully rejected by the patch verification method. The matching score for this image pair was 11. As well as having a lower absolute value, the histogram peak is less sharply defined than for the other correct matches presented in here. We note that this image pair was one of the most challenging to reject – most incorrect image pair candidates output by the whole image matcher resulted in far lower matching scores and evenly distributed histograms.

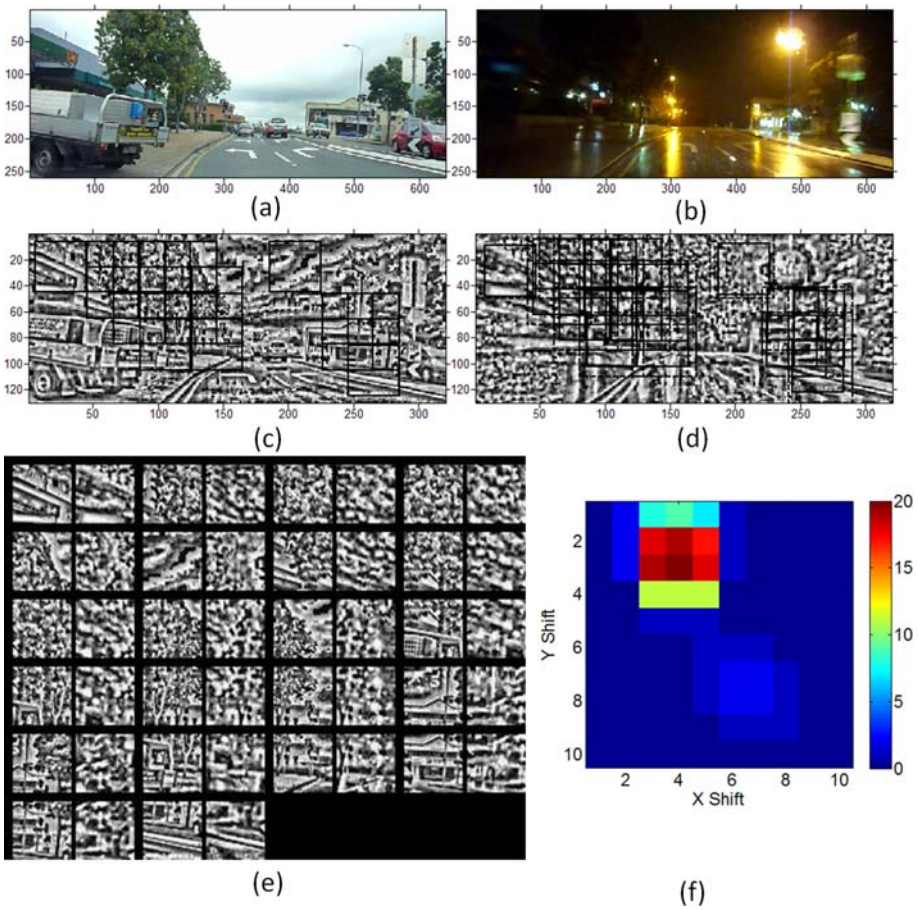

Figure 12: These two images were successfully matched with a matching score of 20. (a, c) This image was one of five candidate image matches for the (b, d) image on the right, and resulted in the highest matching score.

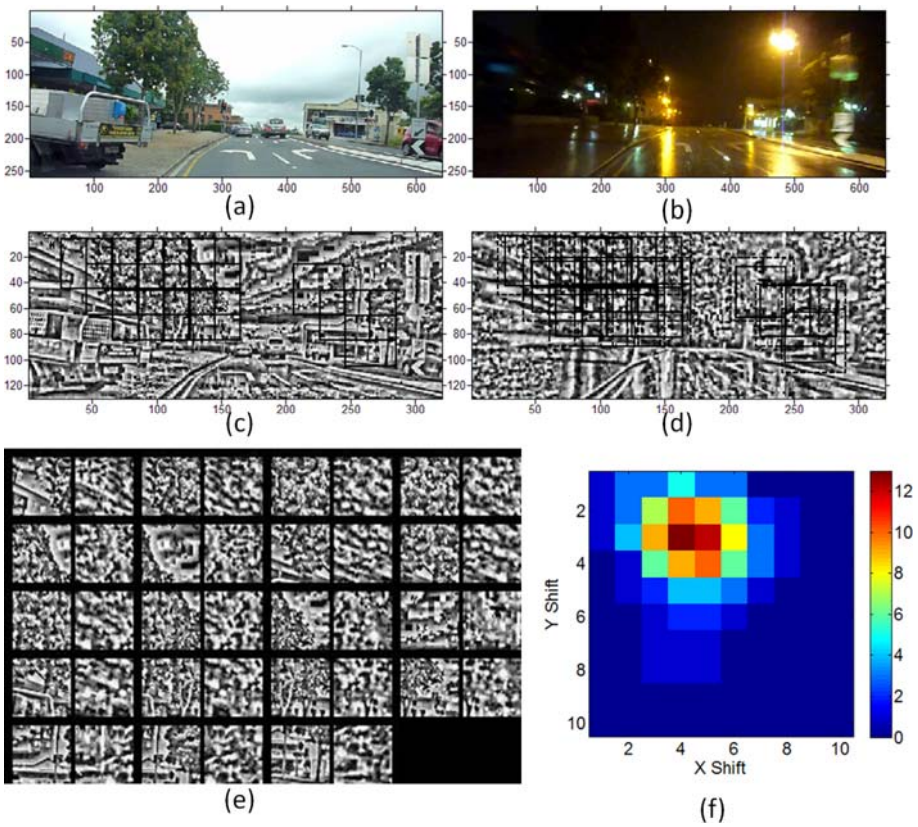

Figure 13: (a, c) This image was one of the other candidate matches for the (b, d) image on the right, which is the same as that shown in Figure 12b. Although the system had already found a successful match, it was able to find this secondary correct match as well, with a matching score of 13 which is higher than that of highly aliased places such as shown in Figure 11.

6 Discussion and Future Work

In this paper we have presented a novel top-down, multi-step visual place recognition system. The overall matching process is inspired by the increasingly selective and tolerant processing stream in the human brain; the low tolerance initial matching stage outputs a small number of candidate match hypotheses, which are then verified or rejected by a highly tolerant patch-based matching method. Results on a challenging dataset demonstrate that the method is capable of producing comparable performance to the current sequence-based state of the art algorithm, but without requiring sequences. Although we do not yet have comprehensive results, parameter sweeps over multiple, different datasets suggest that similar parameter values will provide optimal performance across multiple datasets, and that the system is not overly specializing on the dataset presented in this paper.

The patch verification approach improves matching performance so drastically because, somewhat like other verification techniques such as geometric verification [Cummins and Newman, 2009], it detects the small number of false positive matches reported by the low resolution whole image matcher *and simultaneously* finds a larger number of true positives. Although we did not investigate it here, it may be feasible in future to parallelize the patch verification process and perform it on all possible place matches output by the initial matching stage, rather than just the top few candidates.

We have focused almost entirely on the problem of condition invariance. Future work will investigate how to provide a higher degree of pose invariance, a task that traditional feature-based recognition methods excel at. Researchers have shown that whole image-based image comparison can degrade gracefully as camera pose changes, especially when using panoramic images [Sturzl and Zeil, 2007], suggesting that the problem could be partially addressed simply by expanding the number of candidate matches output by the whole image matcher, at the cost of increased computational load. At the patch verification level, drawing upon techniques used in related fields such as face recognition may also help. Introducing a deformable graph (rather than the current rigid grid) over which patch matching is performed, it may be possible to achieve significantly greater degrees of pose invariance.

We speculate that the place recognition performance achieved here is, at least within this very specific, constrained task, starting to enter the ballpark of human capability. We are examining how to compare, on an even playing field, human and algorithmic performance on this task.

The presented method can likely be improved by the addition of motion information, temporal filtering over image sequences and prior training, and many researchers are currently developing these techniques. However, we believe it is also important to keep pushing the boundaries of what can be achieved in a “pure” vision-based place recognition sense, and hope that the results presented here spur further interest in this challenge.

References

[M. Cummins and P. Newman, 2009] M. Cummins and P. Newman, "Highly scalable appearance-only SLAM - FAB-MAP 2.0," in *Robotics: Science and Systems*, Seattle, United States, 2009.

[A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, 2007] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "MonoSLAM: Real-Time Single Camera SLAM," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 29, pp. 1052-1067, 2007.

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

[D. Burschka and G. D. Hager, 2004] D. Burschka and G. D. Hager, "V-GPS (SLAM): Vision-based inertial system for mobile robots," 2004, pp. 409-415 Vol. 1.

[H. Andreasson, T. Duckett, and A. Lilienthal, 2007] H. Andreasson, T. Duckett, and A. Lilienthal, "Mini-SLAM: Minimalistic Visual SLAM in Large-Scale Environments Based on a New Interpretation of Image Similarity," in *International Conference on Robotics and Automation*, Rome, Italy, 2007, pp. 4096-4101.

[M. Milford, 2013] M. Milford, "Vision-based place recognition: how low can you go?," *International Journal of Robotics Research,* vol. 32, pp. 766-789, 2013.

[M. Milford and G. Wyeth, 2012] M. Milford and G. 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.

[H. Andreasson, T. Duckett, and A. Lilienthal, 2008] H. Andreasson, T. Duckett, and A. Lilienthal, "A Minimalistic Approach to Appearance-Based Visual SLAM," *IEEE Transactions on Robotics,* vol. 24, pp. 1-11, 2008.

[L. M. Paz, P. Pinies, J. D. Tardos, and J. Neira, 2008] L. M. Paz, P. Pinies, J. D. Tardos, and J. Neira, "Large-Scale 6-DOF SLAM With Stereo-in-Hand," *IEEE Transactions on Robotics,* vol. 24, pp. 946-957, 2008.

[E. Royer, J. Bom, M. Dhome, B. Thuilot, M. Lhuillier, and F. Marmoiton, 2005] E. Royer, J. Bom, M. Dhome, B. Thuilot, M. Lhuillier, and F. Marmoiton, "Outdoor autonomous navigation using monocular vision," in *IEEE International Conference on Intelligent Robots and Systems*, 2005, pp. 1253-1258.

[A. M. Zhang and L. Kleeman, 2009] A. M. Zhang and L. Kleeman, "Robust Appearance Based Visual Route Following for Navigation in Large-scale Outdoor Environments," *The International Journal of Robotics Research,* vol. 28, pp. 331-356, 2009.

[K. Konolige, M. Agrawal, R. Bolles, C. Cowan, M. Fischler, and B. Gerkey, 2008] K. Konolige, M. Agrawal, R. Bolles, C. Cowan, M. Fischler, and B. Gerkey, "Outdoor mapping and navigation using stereo vision," 2008, pp. 179-190.

[M. Milford and G. Wyeth, 2008] M. Milford and G. Wyeth, "Mapping a Suburb with a Single Camera using a Biologically Inspired SLAM System," *IEEE Transactions on Robotics,* vol. 24, pp. 1038-1053, 2008.

[E. Johns and G. Z. Yang, 2013] E. Johns and G. Z. Yang, "Feature Co-occurrence Maps: Appearance-based Localisation Throughout the Day," in *International Conference on Robotics and Automation*, Karlsruhe, Germany, 2013.

[N. Sunderhauf, P. Neubert, and P. Protzel, 2013] N. Sunderhauf, P. Neubert, and P. Protzel, "Predicting the Change – A Step Towards Life-Long Operation in Everyday Environments," in *Robotics Challenges and Vision Workshop at Robotics Science and Systems*, Berlin, 2013.

[N. C. Rust and J. J. DiCarlo, 2010] N. C. Rust and J. J. DiCarlo, "Selectivity and Tolerance ("Invariance") Both Increase as Visual Information Propagates from Cortical Area V4 to IT," *The Journal of Neuroscience,* vol. 30, pp. 12978-12995, 2010.

[G. Klein and D. Murray, 2007] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *International Symposium on Mixed and Augmented Reality*, Nara, Japan, 2007.

[D. G. Lowe, 2004] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision,* vol. 60, pp. 91-110, 2004.

[H. Bay, T. Tuytelaars, and L. Van Gool, 2006] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded Up Robust Features," in *Computer Vision – ECCV 2006*, ed, 2006, pp. 404-417.

[C. Valgren and A. Lilienthal, 2007] C. Valgren and A. Lilienthal, "Sift, surf, and seasons: Long-term outdoor localization using local features," presented at the Proc. of 3rd European Conference on Mobile Robots, Freiburg, Germany, 2007.

[M. Milford and G. Wyeth, 2010] M. Milford and G. Wyeth, "Persistent Navigation and Mapping using a Biologically Inspired SLAM System," *International Journal of Robotics Research,* vol. 29, pp. 1131-1153, 2010.

[N. Sunderhauf, P. Neubert, and P. Protzel, 2013] N. Sunderhauf, P. Neubert, and P. Protzel, "Are We There Yet? Challenging SeqSLAM on a 3000 km Journey Across All Four Seasons," in *International Conference on Robotics and Automation*, Karlsruhe, Germany, 2013.

[S. Furui, 1997] S. Furui, "Recent Advances in Speaker Recognition," *Pat. Rec. Letters,* vol. 18, pp. 859-872, 1997.

[G. Aggarwal, N. Ratha, R. Bolle, and R. Chellappa, 2008] G. Aggarwal, N. Ratha, R. Bolle, and R. Chellappa, "Multi-biometric Cohort Analysis for Biometric Fusion," in *IEEE Conf. on Acoustics, Speech and Signal Processing*, 2008.

[S. Tulyakov, Z. Zhang, and V. Govindaraju, 2008] S. Tulyakov, Z. Zhang, and V. Govindaraju, "Comparison of Combination Methods Utilizing t-normalization and Second Best Score Models," in *IEEE Workshop on Biometrics*, 2008.

[W. Sturzl and J. Zeil, 2007] W. Sturzl and J. Zeil, "Depth, contrast and view-based homing in outdoor scenes," *Biological Cybernetics,* vol. 96, pp. 519-531, 2007.