Condition-Invariant, Top-Down Visual Place Recognition

Michael Milford, *Member, IEEE,* Walter Scheirer, Eleonora Vig, Arren Glover, Oliver Baumann,
Jason Mattingley and David Cox

Abstract— **In this paper we present a novel, condition-invariant 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, alleviating the need for either a velocity signal or image sequence, differentiating it from current state of the art methods. We conduct an exhaustive set of experiments evaluating the relationship between place recognition performance and computational resources using part of the challenging Alderley sunny day – rainy night dataset, which has only been previously solved by integrating over 320 frame long image sequences. We achieve recall rates of up to 51% at 100% precision, matching places that have undergone drastic perceptual change while rejecting match hypotheses between highly aliased images of different places. Human trials demonstrate the performance is approaching human capability. The results provide a new benchmark for single image, condition-invariant place recognition.**

I. Introduction

Visual sensors offer many advantages over traditional robotic mapping sensors, including low cost, small size, passive sensing and low power consumption. A large number of vision-based mapping systems have been developed over the past ten years, including FAB-MAP [1], MonoSLAM [2], FrameSLAM [3], V-GPS [4], Mini-SLAM [5], SeqSLAM [6, 7] amongst many others [8-13]. Yet as robots are tested over longer and longer time periods in real world environments, it is becoming clear that perceptual change, caused by factors such as day-night cycles, varying weather conditions and seasonal change, remains a significant challenge for vision-based methods. Current vision-based approaches to the problem are limited by one or more significant restrictions such as requiring hand-picked training data [14, 15], camera motion information, or long image sequences [7].

In this paper, we present a novel multi-step vision-based place recognition system inspired by the human visual processing pathway, and specifically the simultaneous increase in both matching *selectivity* and *tolerance* or invariance along the pathway [16]. 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 (Fig. 1c). We demonstrate the method achieving recall rates of up to 51% at 100% precision on the sunny day-rainy night Alderley dataset [7], creating a new benchmark in condition-invariant place recognition (Fig. 1). The approach is able to match very perceptually different images of the same place (Fig. 1a) while rejecting proposed matches between highly aliased images of different places (Fig. 1b). Finally we present a pilot human study that reveals algorithm performance is comparable to human performance.

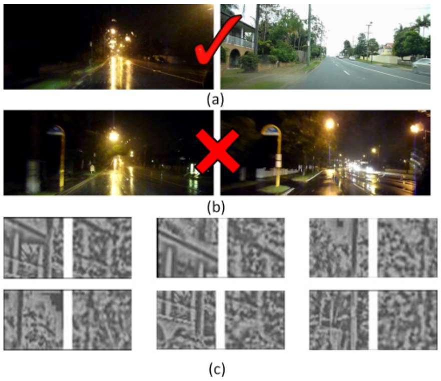


Figure 1: With no prior training, motion information or temporal filtering, the top-down, multi-stage place recognition algorithm presented here is able to perform instantaneous place recognition between (a) very perceptually different images while also rejecting (b) incorrect matches between aliased image pairs. The method uses a global whole image comparison stage followed by a (c) highly tolerant, patch-based comparison method.

Our primary focus in this paper is the condition invariance problem and not the pose invariance problem, which has been extensively addressed in less challenging environmental conditions [1-3, 17], typically using feature-based techniques such as SIFT [18] or SURF [19]. In the Discussion we detail possible methods for combining the pose invariance of current feature-based approaches with the methods presented here. The work presented here builds on an offline only pilot study implemented in [20]. We present the following new contributions: a new, parallelized and real-time capable implementation that scales to available compute and enables a 250% improvement in place recognition performance, human versus machine place recognition trials, and an extensive characterization of the relationship between place recognition performance and compute that provides an indication of the upper bounds on performance were unlimited compute available.

MM and AG are with the School of Electrical Engineering and Computer Science at the Queensland University of Technology, Brisbane, Australia, michael.milford@qut.edu.au. EV, WS and DC are with the School of Engineering and Applied Sciences and Department of Molecular and Cellular Biology at Harvard University. OB and JM are with the Queensland Brain Institute at the University of Queensland. This work was supported by an Australian Research Council Fellowship DE120100995 to MM.

The paper proceeds as follows. In Section II we briefly review vision-based place recognition and mapping algorithms and recent attempts to improve their robustness to environmental change. Section III describes the approach taken in this paper. In Section IV we describe the experimental setup, with results presented in Section V. The paper concludes in Section VI with discussion including future research areas.

II. Background

Current vision-only mapping and place recognition techniques address the problem of perceptual change by either learning how the environment's appearance changes using training data, forming place recognition hypotheses over long sequences of images, or relying on "condition-invariant" features.

Learning how the appearance of the environment changes generally requires training data with known frame correspondences. [14] builds a database of observed features over the course of a day and night. [15] presents an approach that learns systematic scene changes in order to improve performance on a seasonal change dataset. 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.

Many probabilistic SLAM methods such as particle filter-based approaches build up place recognition hypotheses over time. Similarly, by explicitly matching sequences rather than individual frames, an image matcher must only report matches that are generally better than chance performance, rather than globally correct [7]. The SeqSLAM [7, 14, 15, 21] system established the current benchmark on the dataset used in this paper, but required very long sequences (320 frames) and constant camera velocities (or in the case of [14] a motion model), limiting its general applicability.

Lastly, SIFT [18], SURF [19] and a number of subsequent feature detectors have been demonstrated to display a significant degree of pose invariance but only a limited degree of condition-invariance (illumination, atmospheric conditions, shadows, seasons). Perceptual change as drastic as that shown in Fig. 1 has been shown to be challenging for conventional feature detectors [7, 22]. In this paper, we attempt to fill the capability gap between these methods by providing a training-free method that matches single images and does not require velocity information.

III. Approach

This section describes the place recognition components, overviewed in Fig. 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. The computationally intensive patch verification stage is parallelized and can be distributed amongst any number of processing units.

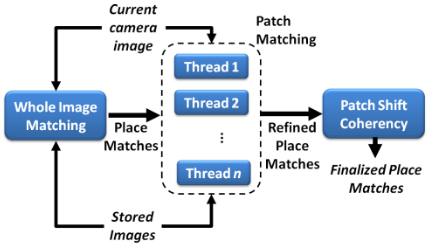


Figure 2: System architecture. A camera image is compared to stored images firstly at a whole image level, then at a patch-based level in a parallelized manner and finally at a patch-shift coherence level.

A. Whole Image Place Recognition

The whole image comparison method is similar to previously used methods [23], so we provide only a brief overview here. Camera images are resolution reduced (64×32 pixels) then patch normalized. Patch normalized pixel intensities, I', are given by:

$$I'_{xy} = \frac{I_{xy} - \mu_{xy}}{\sigma_{xy}} \tag{1}$$

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 absolute intensity differences, performed over a range of horizontal and vertical offsets:

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

where σ is the range of shift offsets, and $g(\,)$ is given by:

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

where s is the area in pixels of the template sub frame and p^i and p^j are the two images. The range of horizontal and vertical offsets provides some invariance to camera pose [6].

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 the presented dataset.

B. 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:

$$\hat{D}_i = \frac{D_i - \overline{\mathbf{D}}}{\sigma} \tag{4}$$

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 [24-26]. 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:

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

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 close very small loops.

C. *Sub-Image Patch Verification*

Patch verification is performed on images from the Z top ranked place match hypotheses proposed by the whole image matcher described in the previous section. Computation is split evenly over the number of available processing units (Fig. 2), leading to (along with code optimization) more than two orders of magnitude improvement in compute speed over the pilot study [27].

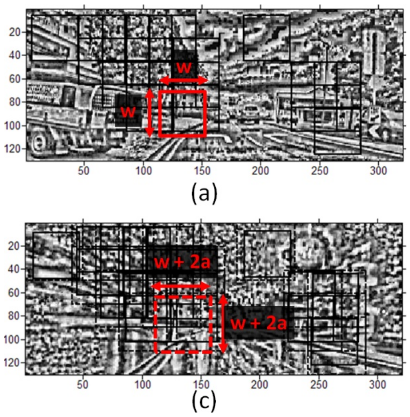
Small image patches at corresponding locations in the two images (see Fig. 3) are compared using a sum of absolute differences calculation similar to that described in Equations 2 and 3. Comparisons are performed over a sliding window centred on the patch location, but extending in both vertical and horizontal directions. We calculate a difference score ratio g_{rat}:

$$g_{rat} = \frac{g_2}{g_1} \tag{6}$$

where g_1 is the difference score for the best matching patch offset and g_2 is the (larger) 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:

$$q = \sum g_{rat} > g_m \tag{7}$$

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



(a)

(c)

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

D. *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 the previous section. The horizontal and vertical shifts are binned in a two-dimensional histogram $\mathbf{H}$ which is then smoothed using a moving summation window of radius s_{rng} (see Fig. 4). The peak shift count c:

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

provides an absolute measure of the number of spatially coherent patch matches. Precision recall curves are generated by sweeping over a range of threshold values of c over which a match is considered confirmed.

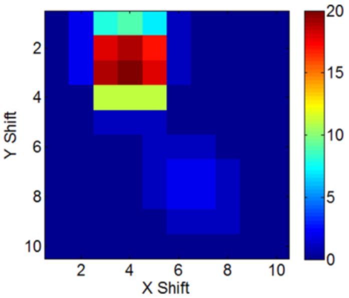


Figure 4: (a) Patch shift coherency verification involves creating a 2D histogram of the spatial shifts for patch matches, from which a coherency metric is calculated.

IV. Experimental Setup

This section describes the experimental environment, dataset acquisition and pre-processing, ground truth creation and key parameter values. All processing was performed on a Core i7-2620M 2.7 GHz laptop, running four parallel processing streams for the patch verification stage.

A. *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, which was cropped to remove the dashboard. The resulting video stream had significant and constant visual artifacts due to water streaming down the windscreen, windscreen wipers, compression artefacts and poor night-time illumination.

B. *Alderley Dataset*

The Alderley dataset comprises two 8 km journeys over the same route through the suburb of Alderley in Brisbane, Australia (Fig. 5). 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.

C. *Ground Truth*

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. All precision-recall 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.

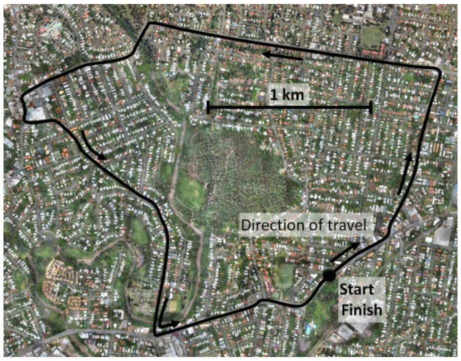


Figure 5: Aerial photo and camera path for the Alderley dataset. An 8 km long route was traversed twice, once during sunny day-time conditions and once during heavy rain at night. Copyright Nearmaps.

D. Image Pre-Processing

Image brightening and contrast enhancement was performed by exploiting the 12 bits of intensity information stored in YV12 video output by many consumer cameras including the one used here. Enhanced images were then down sampled to the resolutions required by each place recognition module and then patch normalized. 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.

E. Parameter Values

Parameter values are given in Table I. These parameters have been heuristically determined over a range of development datasets.

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
P_{size}	8	Patch-normalization radius
f_{jump}	2 frames (1.1 metres max)	Frame learning rate
Z	1-1000 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 radius
r_{peak}	2 pixels	Patch quality score peak search exclusion zone
g_m	1.04325	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

F. Human versus Machine Trial

From the full dataset, a selection of 200 pairs of frames were chosen such that they were evenly distributed between correct and incorrect matches and day-night/night-night pairs, making a total of four categories with 50 frame pairs in each. Half of the frames were selected to be particularly difficult due to perceptual aliasing and low image quality, and the other half were randomly selected evenly over the dataset.

Eleven healthy participants gave informed written consent to the experimental procedures, as approved by The University of Queensland Human Research Ethics Committee, and in accordance with the Helsinki Declaration of 1975. Six of the participants were male and five female. The participants' ages ranged from 20 to 30 years (mean age=23.2, SD=3.0 years) and their education level varied from 15 to 22 years (mean=17.4, SD=2). Trials were performed by displaying an image pair on a screen for 5 seconds and having the participant decide whether the images were the same or a different place. Participants were given a training phase of 8 images (not included in the main test set) to prepare for the timing of the experiment before being tested on all 200 frame pairs. The patch-verification stage of the algorithm was tested on the same 200 frames with varying values of *c*.

V. RESULTS

In this section we present precision-recall curves and ground truth plots, with comparison to both a whole image-only approach and the SeqSLAM algorithm, and human trial results. 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.

A. Performance versus Number of Evaluated Hypotheses

We conducted an exhaustive study of the relationship between place recognition performance and the number of place match hypotheses (output by the whole image matcher) considered at each place, from 1 to 1000. The maximum hypothesis count of 1000 corresponds to testing approximately only 15% of all possible place matches. Due to the significant computational demands of the study, it was performed on two 1000 frame subsets of the overall dataset. We selected subsets for which performance was similar to that on the overall dataset. Obviously as hypothesis counts increased, the approach deviated somewhat from the initial concept of having a low tolerance initial matching stage.

Precision-recall curves for 1 to 1000 place hypotheses are shown in Fig. 6. Recall at all precision levels increases as the number of hypotheses considered for each place increases, but with diminishing returns. For example, increasing the number of evaluated hypotheses per place (and hence compute) by two orders of magnitude from 5 to 500 leads to an improvement in maximum recall of about 80%.

Figure 7 shows the maximum recall performance at 100% precision for each number of hypotheses. Increasing the number of considered hypotheses from 1 to 50 leads to a tripling in recall at 100% precision, but performance gains diminish beyond that number. The maximum recall rate at 100% precision is 51%, for the 500 hypothesis case. Performance at 100% precision is often sensitive to parameter variation, but it is clear that any number of hypotheses from 100 to 1000 results in 43-50% recall at 100% precision.

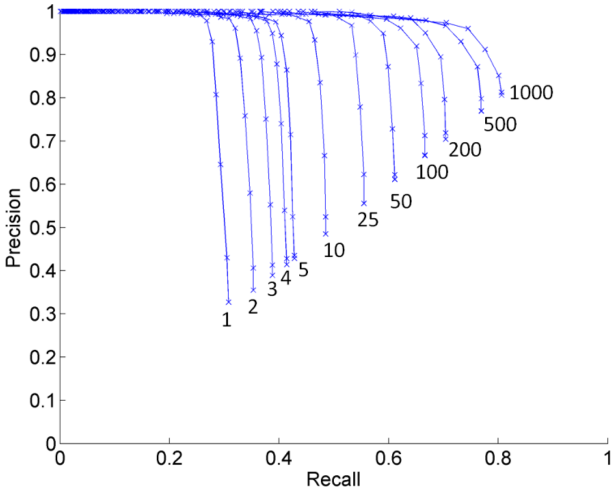


Figure 6: Precision-recall performance for varying hypothesis counts (labelled). Recall at all precision levels increases with diminishing returns as the number of place recognition hypotheses being evaluated increases.

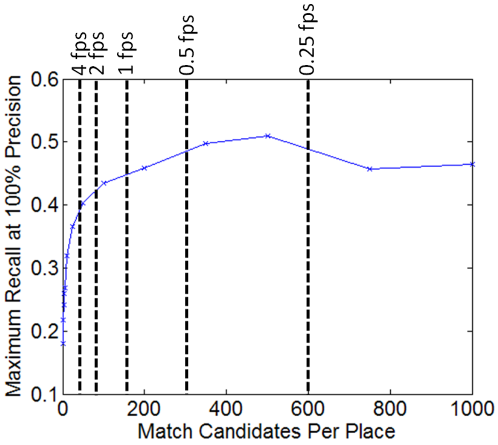


Figure 7: Maximum recall rate at 100% precision for varying numbers of evaluated place match hypotheses per place. The vertical dashed lines indicate the real-time achievable number of place match hypothesis evaluations that can be performed per second. For example if new places must be matched at 1 fps, each new place can be evaluated against 150 place match hypotheses in real-time.

B. *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 51%. The small cyan dots indicate the 500 top ranking place match hypotheses after the initial whole image matching stage, with the solid dark blue line indicating ground truth. The power of the patch verification process is clearly shown – of the 500,000 place hypotheses considered, the system correctly chooses 510 (1000 being perfect) without making any mistakes. From a practical navigation perspective, the matches have good coverage over the route – the longest segment without place matches measures approximately 44 metres (40 frames).

C. *Sample Place Matches*

Figures 9 and 10 show examples of two place match hypotheses output by the whole image matcher that were correctly accepted (Fig. 9) and rejected (Fig. 10) by the patch verification process. One of the more challenging successful place matches is shown in Fig. 9. Despite vastly different perceptual conditions the algorithm is able to find a large number of (highly tolerant) patch matches (Fig. 9e). The smoothed histogram of patch match shifts (Fig. 9f) is highly coherent with a peak matching score of 25.

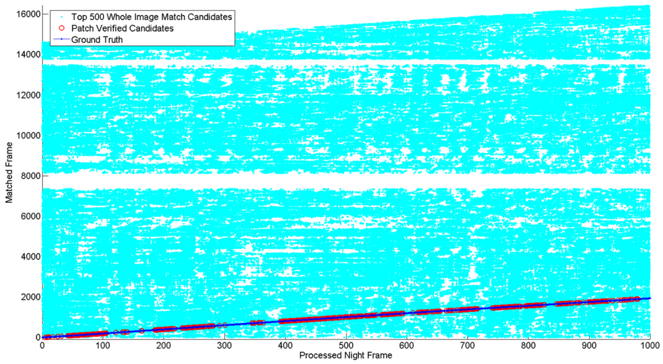


Figure 8: Ground truth plot at 100% precision and 51% recall. The patch verification stage confirms 510 correct matches and successfully rejects half a million incorrect match hypotheses.

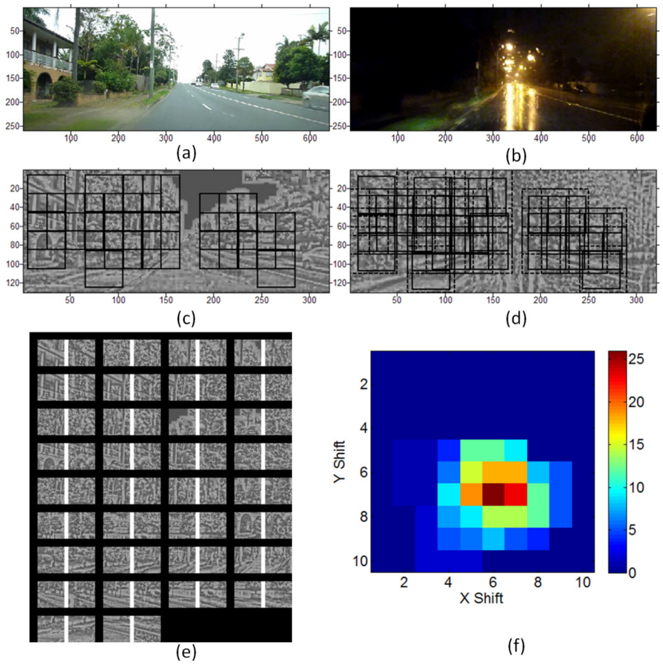


Figure 9: One of the more challenging place matches correctly identified by the system. (a-b) Original images. (c-d) Patch normalized images with black rectangles indicating the patch matches exceeding the quality threshold. (e) The smoothed 2D histogram of patch match shifts. The overall matching score for this image was 25.

Figure 10 shows two perceptually similar images of *different* places that were matched by the initial whole image matcher. The patch verification process finds a moderate number of patch matches exceeding the minimum difference score ratio threshold, but the shift histogram is less coherent than in Fig. 9f, with a maximum matching score of only 11. Most incorrect match hypotheses had much lower matching scores than this one.

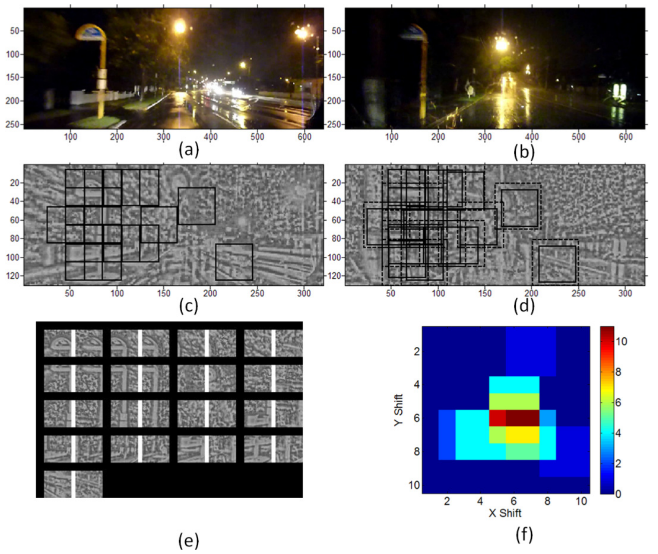


Figure 10: Two visually similar but spatially separate places that were matched by the whole image matcher but then successfully rejected by the patch verification method. The matching score for this image pair was 11. This image pair was one of the most challenging to reject – most incorrect image pair candidates output by the whole image matcher resulted in much lower matching scores.

D. *Full Dataset Study*

Figure 11 shows the precision-recall curves with (solid blue line) and without (dashed red line, whole image only matching) patch verification for a 5 hypothesis count system (which ran faster than real-time when processing frames at 15 fps) on the full (2 × 8 km) dataset. Although the maximum recall (21%) at 100% precision was lower than in the initial SeqSLAM study (35%), place match coverage over the environment was more even, with the longest segment without place matches measuring approximately 280 metres, versus 1400 metres in the original SeqSLAM result [7].

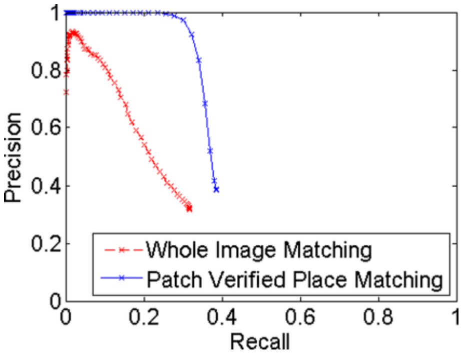


Figure 11: Precision-recall curves for the whole image matching-only and top-down matching method applied to the entire Alderley dataset using a 5 hypothesis count system. The patch verification step results in a significant improvement in precision-recall performance at all precision levels. Perhaps unsurprisingly given the nature of the dataset, single frame-based whole image matching is incapable of reaching 100% precision at any recall level.

E. *Computational Efficiency*

The current algorithms are implemented in a mixture of optimized C and unoptimized Matlab code. For datasets such as the one presented here, the primary computational load is due to the patch verification process, rather than the whole image matching process (see [6] for a discussion of low resolution image matching compute growth). Running on a Core i7-2620M 2.7 GHz laptop over 4 parallel threads, the current implementation is capable of performing patch verification on approximately 150 place match candidates per second. Further computational speed-ups could likely be obtained through several means. Firstly, a hierarchical spatial pyramid approach would ensure only a fraction of promising image matches at each resolution were verified at a higher resolution. Secondly, identifying salient image regions through a bio-inspired saliency model may reduce compute without adversely affecting performance [28]. Finally, widely available GPU hardware offers the potential for further parallelization of compute and consequent speed ups.

F. *Human versus Machine Trial*

Figure 12 compares the performance of the eleven human participants over the 200 frame test to the algorithm. At its best operating point, the algorithm's performance was comparable to human performance.

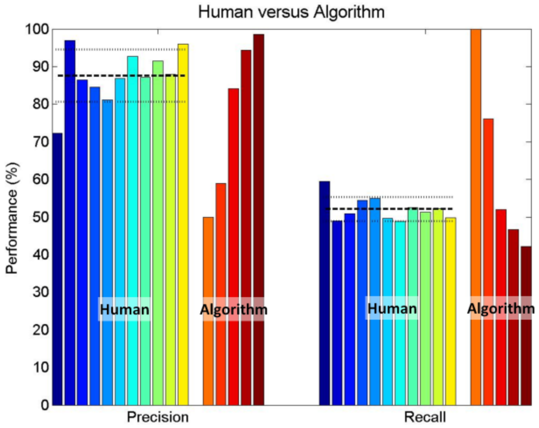


Figure 12: Peak algorithm performance is comparable to human performance. Dashed lines indicate mean and standard deviation.

VI. 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 relatively 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 or superior performance to the current sequence-based state of the art algorithm, but without requiring sequences or prior training. The patch verification step reliably confirms correct matches output by the whole image matcher while detecting and removing false positive matches. The benefits of the multi-step top down approach were perhaps most clear in the 500 hypothesis count study, in which the system found 510 correct place match hypotheses out of a possible 1000 (51% recall at 100% precision) while rejecting half a million false positive matches. Near maximal performance is reached when evaluating a relatively small number of match hypotheses, supporting the concept of an initial low tolerance matching stage. In addition, the place recognition performance achieved here on this specific task is starting to be comparable to human performance, as shown by our pilot human study.

We have focused almost entirely on the problem of condition invariance, which restricts the applicability of the current method to scenarios where camera poses tend to be repeatable such as car navigation systems and indoor robotics. To investigate the pose invariance problem, we will draw upon computer vision techniques such as deformable graphs to replace the current rigid grid over which patch verification is performed; this change will in turn likely require a more sophisticated patch shift coherency technique. At the whole image level, researchers have shown that place recognition degrades remarkably gracefully with pose change [29], especially with panoramic cameras. To leverage this property, the detected pose changes at the whole image level will need to be applied at the patch verification stage.

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 [30] would provide mechanisms for bounding learning and producing a spatial map. Finally, it may be possible to achieve significant further performance improvements through the use of exhaustive hyper parameter searches [28] over parameters such as patch matching quality thresholds. We will develop a representative range of place recognition datasets and investigate such an approach using cluster computing.

Using motion information, temporal filtering over image sequences and prior training are all well developed techniques that could be used to improve the performance of this method, at the cost of versatility. However, by attempting to push the boundaries of what can be achieved using just single images, it may be possible to reveal new insights into the problem that would otherwise be obscured by these additional processes. We hope that the strengths and weaknesses of the work presented here serve as a point of discussion for future research in this area.

ACKNOWLEDGEMENTS

We acknowledge Adam Bulley, Freya Young and Sarah Catchpoole for their help in acquiring the human performance data.

REFERENCES

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

[2] 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.

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

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

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] I. Biederman, "Aspects and extension of a theory of human image understanding.," *Computational processes in human vision: An interdisciplinary perspective,* 1988.

[16] 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.

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

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

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

[20] M. Milford, E. Vig, W. Scheirer, and D. Cox, "Towards Condition-Invariant, Top-Down Visual Place Recognition," in *Australasian Conference on Robotics and Automation*, Sydney, Australia, 2013.

[21] 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.

[22] 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.

[23] M. Milford, F. Schill, P. Corke, R. Mahony, and G. Wyeth, "Aerial SLAM with a Single Camera Using Visual Expectation," in *International Conference on Robotics and Automation*, Shanghai, China, 2011.

[24] M. C. Potter, "Meaning in visual search," *Science* vol. 187, pp. 965-966, 1975.

[25] V. Vapnik, "The support vector method of function estimation," *Nonlinear Modeling,* pp. 55-85, 1998.

[26] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision,* vol. 42, pp. 145-175, 2001.

[27] B. P and A. P, "topological mobile robot localization using fast vision techniques," presented at the IEEE ICRA, 2002.

[28] N. Pinto, D. Doukhan, J. DiCarlo, and D. Cox, "High-Throughput Screening Approach to Discovering Good Forms of Biologically Inspired Visual Representation," *PLoS Computational Biology,* vol. 5, 2009.

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

[30] 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.