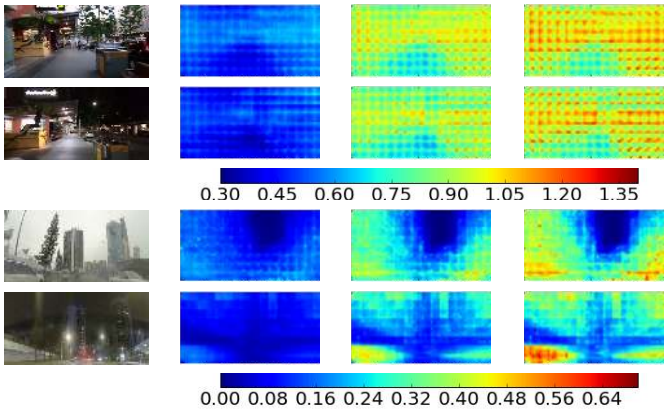
Straightening Sequence-Search for Appearance-Invariant Place Recognition using Robust Motion Estimation

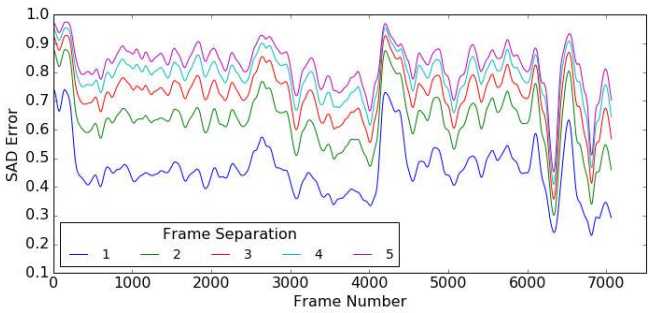
Sourav Garg and Michael Milford*

Abstract

Camera-based technologies have become a growing component of robotic and autonomous systems such as self-driving cars. Much recent research has investigated how key autonomous system capabilities such as visual place recognition can be made robust to real-world challenges such as extreme appearance change. The original sequence-based methods were incapable of handling platform velocity variability; to address this some follow-up techniques incorporated sophisticated sequence-matching search algorithms that also increased the likelihood of false positive matches. Other new methods solve this problem by explicitly incorporating motion information but either require a suitable motion encoding sensor, or use state-of-the-art visual odometry techniques that degrade in low light, motion blur, and rapid camera movement. In this work, we develop a solution for the middle ground by developing a robust translational motion estimator which is then used to normalize camera trajectories and improve place recognition performance. We also present a new odometry fusion method that seamlessly switches between state-of-the-art visual odometry and a more robust but cruder motion estimation pipeline, enabling continuity of motion estimation and hence high place recognition performance. We demonstrate the effectiveness of these contributions on several challenging datasets that explicitly test the appearance- and speed-invariance of place recognition techniques.



(a) Absolute Image Difference averaged over all image pairs separated by 1, 5 and 10 frames (left to right) for day and night traverses of two datasets.



(b) The image alignment error computed with a varying frame separation between image pairs for all the images in the database.

Figure 1: The spatial patterns in (a) show the regions within the images that are either persistent or variable locally and (b) shows that for any given image the pattern in change of image alignment error with increasing frame separation is generally consistent, irrespective of the environment or camera speed within a local temporal window.

*The authors are with the Australian Centre for Robotic Vision at the Queensland University of Technology. This work was supported by an ARC Centre of Excellence for Robotic Vision Grant CE140100016. SG is supported by QUT Postgraduate Research Award. MM is supported by an Australian Research Council Future Fellowship FT140101229. Email: sourav.garg@hdr.qut.edu.au

1 Introduction

Visual Place Recognition research has advanced rapidly in recent years, especially with the advent of deep-learning based convolutional networks. The use of hand-crafted features for large-scale place recognition methods like FAB-MAP [Cummins and Newman, 2011], and environment-invariant methods like SeqSLAM [Milford and Wyeth, 2012], has increasingly been replaced by deep-learned features [Vysotska and Stachniss, 2016; Naseer *et al.*, 2017]. These deep-learned place representations are generally more robust to variations in appearance of the environment due to changing season, time of day, and weather conditions. At the same time, the role of sequence-based approaches [Milford and Wyeth, 2012; Vysotska and Stachniss, 2016], leveraging the inherent temporal information, is equally vital for appearance-robust place recognition.

However, the robustness of these methods is often acquired at the cost of *viewpoint-dependence* and *velocity sensitivity* of camera trajectories. While the former has mostly been dealt using improved place representations [Hou *et al.*, 2017; Torii *et al.*, 2015], the latter has received significantly more attention, particularly for developing sophisticated velocity search techniques [Hansen and Browning, 2014; Pepperell *et al.*, 2016]. An alternative solution for velocity sensitivity is to leverage motion information using encoders [Pepperell *et al.*, 2014] or visual odometry [Glover *et al.*, 2015]. Although, state-of-the-art visual odometry solutions [Wang *et al.*, 2017; Mur-Artal *et al.*, 2015; Forster *et al.*, 2014] are capable of 6-DoF pose estimation, they often degrade in low light, motion blur and rapid camera movement [Glover *et al.*, 2015; Alismail *et al.*, 2016b].

In this work, we propose to use a learning-based motion estimator based on image alignment error computed within a local temporal window as shown in Fig. 1, contrary to the conventional use of alignment error for egomotion estimation using iterative optimization techniques often based on Lukas-Kanade framework [Lucas *et al.*, 1981]. The learned motion estimator is then used to speed-normalize the camera trajectories to tackle the velocity sensitivity problem of appearance-robust visual place recognition methods. We further demonstrate a hybrid visual odometry system for forward translational motion estimation that learns for as long as the state-of-the-art visual odometry system works and then estimates the motion until visual odometry system reinitializes correctly.

Fig. 1 shows the key observations that set the premise of our work; that is, the image alignment error computed with a varying frame separation between the neighboring image pairs exhibits a consistent uniform pattern as shown in Fig. 1(b). Furthermore, the absolute image difference averaged over all the images exhibits spatial patterns that highlight the *persistent* regions, for example the sky and footpath and the *variable* regions. The *variable* regions are marked by uniform increase in pixel difference with increasing frame separation as shown in Fig. 1(a).

While conventional visual odometry systems continue to remain an obvious choice for egomotion estimation under ideal environmental conditions, our proposed odometry fusion method provides robust but cruder motion estimation pipeline enabling continuity of motion estimates. Instead of calculating the actual motion information, we learn a relationship between image alignment error and frame separation for estimating motion information which can be used in conjunction with any other source of egomotion. Apart from its utility for challenging environmental conditions, the appeal of the proposed system is more pronounced in such application domains where low-cost and limited-compute devices are often necessary. According to a recent survey [Bhowmick and Hazarika, 2017; Ye *et al.*, 2014], a compact wearable assistive device is often the first choice for visually impaired people. Similarly, assistive navigation systems [Oh *et al.*, 2017] that do not necessarily require a sophisticated, highly accurate, 6-DoF pose estimation, can sufficiently rely on robust forward translational motion information along with turn detection. These requirements are also akin to an appearance-robust route-based place recognition system as we demonstrate in this work.

Our key contributions are as below:

- Establishing the use of a learning-based motion estimator using image alignment error.
- Improving state-of-the-art visual place recognition performance by speed-normalizing camera trajectories using motion estimator.
- A new hybrid visual odometry system that supplements conventional state-of-the-art visual odometry with motion estimates to prevent system failures.

The paper proceeds as follows: Section 2 reviews the research work related to appearance-robust visual place recognition and attempts for improving it using visual odometry; Section 3 highlights the key components of our proposed approach; Section 4 describes the datasets, system parameters and performance measures used for experiments; Section 5 shows the quantitative and qualitative results obtained using the proposed methodology; Section 6 discusses the sensitivity of the proposed approach; finally, Section 7 concludes the paper with scope of future work.

2 Related Work

The focus of research in visual place recognition has recently moved from appearance-based techniques like

FAB-MAP [Cummins and Newman, 2008], BoWP [Kejriwal *et al.*, 2016], to appearance-robust methods like SeqSLAM [Milford and Wyeth, 2012], HMM Sequence Matching [Hansen and Browning, 2014], AMOSnet [Chen *et al.*, 2017], that enable place recognition even under challenging environmental conditions induced by change in season, time of day and weather. A number of these methods use sequence-based approach, leveraging the temporal information inherent in the problem, where weak matches are accrued over the time to find a strong matching sequence. However, the advances in appearance-robustness for place recognition are generally achieved at the cost of viewpoint dependence and velocity sensitivity.

The subsequent body of work has therefore focused on both the problems. The viewpoint-invariant techniques e.g. ConvNet Landmarks [Sunderhauf *et al.*, 2015; Hou *et al.*, 2017], NetVLAD [Arandjelovic *et al.*, 2016], View Synthesis [Torii *et al.*, 2015], have focused more on developing better representation of visual places that are less sensitive to viewpoint variations. However, for a significant subset of practical applications, for example, place recognition across route traversals, viewpoint dependence does not pose a great challenge. The alternative ways of dealing with viewpoint variations include the use of panoramic [Arroyo *et al.*, 2017; Hansen and Browning, 2015] or side-view [Pepperell *et al.*, 2015] imagery.

On the other hand, the velocity sensitivity problem has received much more attention in recent years. The original sequence-based methods assumed a constant velocity difference [Milford and Wyeth, 2012; Chen *et al.*, 2014] between the corresponding segments of reference and query traversals, where sequence searching was performed using line-fitting. A plethora of sophisticated search techniques have since been proposed for improved sequence matching like Dynamic Programming [Ho and Newman, 2007], Hidden Markov Model [Hansen and Browning, 2014], Network Flow [Naseer *et al.*, 2014], Directed Acyclic Graph [Vysotska and Stachniss, 2016], Particle Filter [Pepperell *et al.*, 2016], but at the cost of increased likelihood of false positive matches. Nevertheless, the local velocity difference constraints are still required to confine the velocity search range within practical limits.

The alternative way to approach the velocity sensitivity problem is to obtain motion estimates either through motion sensors or visual odometry. A variety of sensors like GPS [Saito and Kuroda, 2013], IMU [Lynen *et al.*, 2017] etc. as well as visual odometry [Floros *et al.*, 2013] have been proposed to improve place recognition and localization accuracy under ideal environmental settings. Specifically, the use of motion sensors to improve place recognition under challenging environmental conditions is demonstrated in SMART [Pepperell *et al.*, 2014; 2013]. However, the use of sensors is often limited as they are prone to errors, lack ubiquity unlike cameras, and necessitate sensor fusion techniques. Furthermore, the use of off-the-shelf visual odometry techniques based on binary feature matching like ORB-SLAM [Mur-Artal *et al.*, 2015], semi-direct matching [Forster *et al.*, 2014], or recent deep-learning based methods like Bayesian CNN [Peretroukhin *et al.*, 2017], DeepVO [Wang *et al.*, 2017], is not straightforward because the majority of these methods, though capable of estimating 6-DoF pose, have been demonstrated to work only within limited range of environmental conditions. The low-light VO solution recently proposed in [Alismail *et al.*, 2016b], though promising, requires depth information.

Some preliminary investigations into using visual odometry under challenging environmental conditions have been demonstrated in [Milford and George, 2014] and [Glover *et al.*, 2015]. Our proposed approach uses patch-normalization of images similar to [Milford and George, 2014], assuming non-holonomic motion system, but differs in terms of the whole-image based dissimilarity measure that we learn through past traversal, with an aim to speed-normalize the camera trajectory, for improving place recognition. Glover et. al. in [Glover *et al.*, 2015] established that 6-DoF traditional visual odometry solutions [Geiger *et al.*, 2011] fail under extreme appearance variations, for example, night-time imagery; and that SIFT [Lowe, 2004] or patch-based motion estimation is more useful, given that only a forward translational motion estimation is generally needed for speed-normalization of the vehicle trajectory across route traversals. In this work, instead of calculating the actual motion information, we learn a relationship between image alignment error and frame separation for estimating motion information which can also be used in conjunction with [Milford and George, 2014; Glover *et al.*, 2015] or any other source of egomotion as demonstrated in this work using ORB-SLAM visual odometry [Mur-Artal *et al.*, 2015].

3 Proposed Approach

The image alignment error, or image registration, is often used to estimate geometric transformation between images using iterative optimization techniques [Lucas *et al.*, 1981; Alismail *et al.*, 2016b; Engel *et al.*, 2013]. Instead, we use image alignment error and frame separation for learning a motion estimator within a temporal window of past or current camera trajectory. The learned estimator is then used to predict the camera displacement in correspondence with changes in the alignment error for improving appearance-robust place recognition and supplementing state-of-the-art visual odometry.

3.1 Image Alignment Error

We use patch-normalized images [Milford and Wyeth, 2012], for calculating image alignment error using Sum of Absolute Difference between a given image pair. The error is computed for all the images in the reference database with varying frame separation and then averaged over all the images as given in Eq. 1.

$$\mu_e^s = \frac{\sum_{i}^{N-s} \sum_{k}^{K} |I'_k - I'^{i+s}_k|}{N} \quad (1)$$

where s is the frame separation between an image pair, varying from 0...10 (assuming camera frame rate to be 25-30 fps), N is the number of images in the database and I'_k is the patch normalized pixel intensity for pixel k in the image, calculated as below:

$$I'_k = \frac{I_k - \mu_p}{\sigma_p} \quad (2)$$

where I_k is raw pixel intensity and p refers to the patch centered at k for which mean μ_p and standard deviation σ_p is computed over the pixel intensities contained in the patch.

Fig. 2 shows average SAD error μ_e as a function of frame separation between pairs of image frames for both reference and query database. The curve saturates beyond certain frame separation, depending on the camera frame rate and egomotion, which is an expected outcome as the visual overlap between an image pair decreases with further camera movement. The pattern in Fig. 2 generated using patch-normalized images is similar to the quasi-convex cost surface generated in [Alismail *et al.*, 2016a] using bit-planes descriptor, demonstrated to be suitable for Lukas-Kanade algorithm [Lucas *et al.*, 1981], despite the discontinuous nature of both descriptor types.

3.2 Visual Motion Estimator

The relationship between mean SAD error and the frame separation values is learned using spline interpolation of second order, as shown in Fig. 3. The blue points are the actual mean error values at varying frame separations and the green curve shows the continuous output of the learned function. Now, for a given pair of consecutive images, having frame separation value of 1, SAD error is computed and fed to the learned relationship function to estimate a continuous set of values, termed as *visual motion* from here as below:

$$s'_i = F(e_i^{s=1}) \quad \forall i \in M, N \quad (3)$$

where s'_i is estimated visual motion, F is the second order spline interpolation function, $e_i^{s=1}$ is the SAD error for a consecutive image pair and M and N are the number of images in query and reference database.

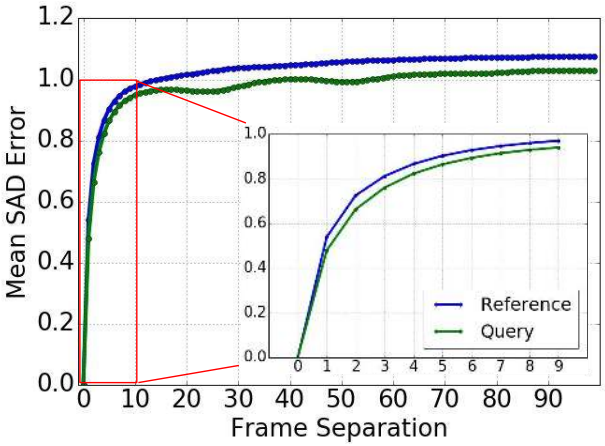


Figure 2: SAD Error as a function of Frame Separation between image pairs, averaged over all the images within a database.

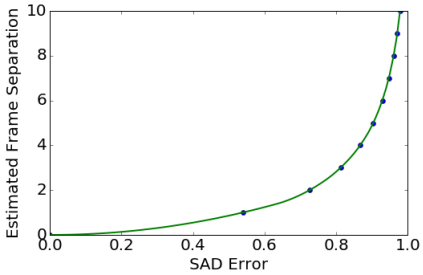


Figure 3: Visual Motion is estimated from the Mean SAD Error using second order spline interpolation. The blue points are the actual error values used for interpolation and green curve shows the continuous output of the relationship function.

3.3 Dynamic Frame Sampling

The visual motion s'_i estimated from the interpolation function F can be considered equivalent to some arbitrary distance in physical world and is used to re-sample the image frames in the database. The re-sampling is performed by using sampling factor u as a reference unit distance to select frames which are u units apart as defined below:

$$u = n(\mu_{s'} + 2\sigma_{s'}) \quad (4)$$

where n is a scalar to change the sampling rate, determined from the camera frame rate. The $\mu_{s'}$ and $\sigma_{s'}$ are the mean and standard deviation of the estimated visual motion values, s', computed over all the images in the reference database. The sampling factor, u, is used to re-sample both reference and query databases. u is chosen to be greater than $\mu_{s'}$ in order to avoid oversampling due to a possible underestimation of s'.

3.4 Visual Place Recognition

We use a state-of-the-art visual place recognition method SeqSLAM [Milford and Wyeth, 2012], demonstrated to work well under varying environmental conditions. The method, like other similar sequence-based techniques [Hansen and Browning, 2014; Naseer *et al.*, 2014], first calculates a cost matrix between the reference and query images and then searches for a matching sequence within an arbitrarily defined velocity search range. The length of sequence being searched and the velocity search range are two important parameters that determine the quality of matches. We conducted experiments for two scenarios using SeqSLAM: a vanilla approach, where the frames were processed with a fixed frame sampling rate, and the proposed approach, where frames were dynamically sampled using the visual motion estimator.

3.5 Hybrid Motion Estimator

The state-of-the-art visual odometry solutions are usually the preferred choice for egomotion estimation, but they need to be robust to challenging environmental conditions. We investigated the use of a hybrid motion estimation solution that combines the state-of-the-art visual odometry technique with the proposed motion estimator. For this purpose, we used the monocular ORB-SLAM Visual Odometry [Mur-Artal *et al.*, 2015] to initialize and estimate the camera trajectory and supplement it with the estimated motion values whenever it fails, until it reinitializes itself to start the process again. In this scenario, the displacement estimates are available from the visual odometry, therefore we can learn a relationship between SAD score and displacement values instead of frame separation values.

4 Experimental Setup

4.1 Datasets

We used small-scale datasets with high velocity-variability to evaluate the efficacy of our approach for relatively precise localization (within 15 meters for road-based Surfers Paradise Traffic and 4 meters for hand-held camera-based Kelvin Grove Footpath dataset). The traditional benchmark datasets generally overshadow the performance sensitivity due to variable velocity between the reference and query traversals. This mainly happens because the processed frames are traditionally sampled at larger distances in proportion to the size of the dataset. The ground truth aerial trajectories for KG Footpath and SP Traffic are shown in Fig. 4 along with some sample images in Fig. 5.

Surfers Paradise Traffic

The Surfers Paradise (SP) Traffic Dataset, used by authors in [Pepperell *et al.*, 2013; 2014], comprises two traverses of approximately 3.5 km each, from different

Figure 4: Ground Truth Trajectory marked over map for Kelvin Grove Footpath (left) and Surfers Paradise Traffic (right) datasets. Source: Google Map

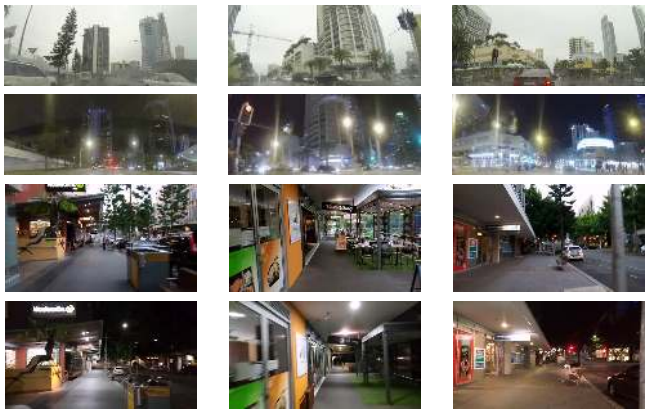


Figure 5: Sample images from day and night traverses of Surfers Paradise Traffic and Kelvin Grove Footpath datasets (from top to bottom respectively).

times of day (day and night), of an urban environment with varying traffic conditions throughout the traverse. We used a part of the traverse approximately 1 km long where the vehicle velocity varied the most between the two traverses due to heavy traffic conditions. Apart from the velocity-variability, the variations in environmental conditions between the two traverses are extreme as the naturally uniform day-time imagery transits into artificially bright and colorful night-time imagery. The authors in [Pepperell *et al.*, 2013; 2014] originally used OBD odometry to speed-normalize the vehicle trajectories for this dataset.

Kelvin Grove Footpath

The Kelvin Grove (KG) Footpath dataset comprises two traverses of a suburb of approximately 600 meters. The data was collected using a hand-held mobile phone camera while walking and running intermittently across the region, once during the day and then at night. A part of original data was removed which had similar velocity profiles for both reference and query databases and only the first and last part of the traverse, having opposite velocity profiles (slow vs fast and vice versa) were retained, hence forming an overall 300 meters long tra-

verse. The opposite velocity profiles can be visualized in Fig. 6. The full length traverses are used for hybrid visual odometry experiments to supplement ORB-SLAM as shown in Section 5.

Residence Indoor-Outdoor and Parking Lot

We use these two datasets, described in details in [Garg *et al.*, 2017], for the hybrid visual odometry experiments. Both the datasets exhibit changes in environmental conditions within the traverse. We use the day traverse of Residence Indoor-Outdoor which is captured using handheld camera and traverses the well-lit outdoor and unlit indoor areas of a residence. For the Parking Lot dataset, we use the night traverse captured from a vehicle traveling through a well-lit underground parking area to dark outdoor parking areas. These datasets exhibit frequent changes in lighting, bland regions, and frequent rotations and are therefore challenging for any monocular visual odometry method.

4.2 Ground Truth & Performance Measure

The ground truth matches for both the datasets were generated manually for visual place recognition by considering a few select matching landmark locations within the traverse and then interpolating the ground truth for the entire traverse. A place recognition match is considered a true positive if it lies within a range of its ground truth: 15 meters and 4 meters respectively for SP Traffic and KG Footpath dataset. The chosen criteria is stricter than 40 meters used in [Cummins and Newman, 2011; Milford and Wyeth, 2012] for road-based scenarios. The precision-recall values along with max-F1 score were generated by varying the sequence uniqueness parameter as a threshold as described in [Milford and Wyeth, 2012].

The ground truth displacement, required for hybrid motion estimation experiments, was obtained using Google Maps with the help of metric distances between different landmark locations within the traverse which was then interpolated for the entire traverse.

5 Results

5.1 Visual Place Recognition

We used two parameters of the SeqSLAM place recognition system for performance comparison: Sequence Length and Search Range [Milford and Wyeth, 2012]. The sequence length determines the extent of the sequence to be searched within the cost matrix and the search range defines the range of slopes at which the sequence is searched. The former is varied from 1 frame, which means single-frame matching, to 100 frames, which is almost half the size of the cost matrix. The latter is varied from 0 radians, which means to search strictly across the diagonal at 45°, up to 0.8 radians,

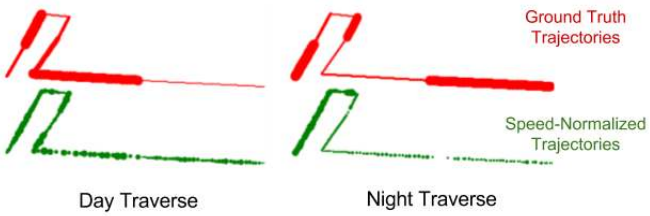
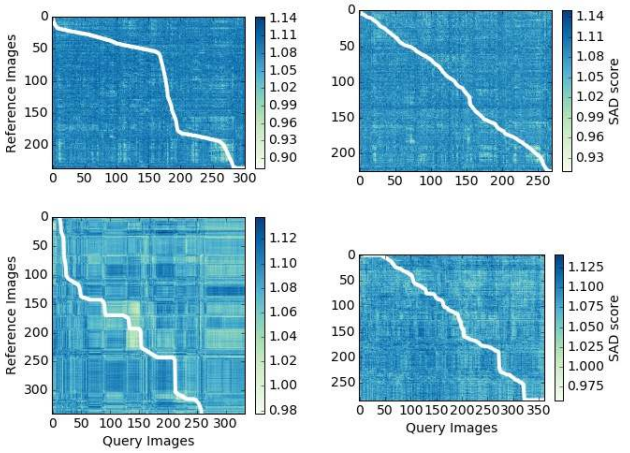


Figure 6: The ground truth (at top in red) and speed-normalized (at bottom in green) camera trajectories, using proposed approach, for KG Footpath dataset. The thickness of a sample point is determined using instantaneous speed (thicker is faster). The uniformity across the speed-normalized trajectories for day and night traverse creates a relatively straight matching path for place recognition cost matrix as shown in Fig. 7.

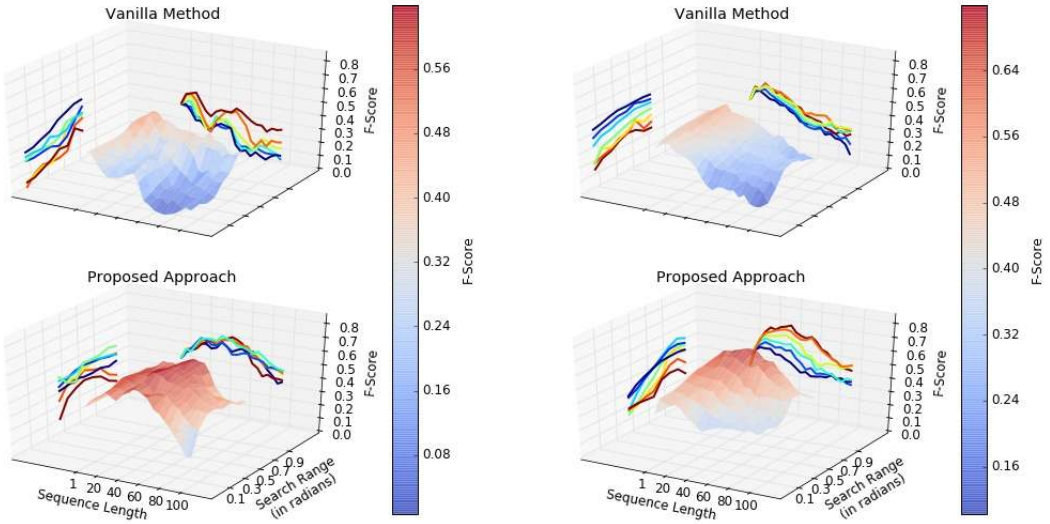


(a) Vanilla (b) Proposed

Figure 7: Ground Truth Matches (white) plotted over the SAD error matrix between query and reference images for (a) vanilla and (b) proposed method for KG Footpath (top) and SP Traffic (bottom) datasets.

which implies searching within a range of ±45° with respect to the diagonal.

Fig. 6 shows the ground truth and speed-normalized trajectories for KG Footpath dataset for both reference and query traverses. The thickness of any point on the trajectory is proportional to the instantaneous speed of the camera. The speed-normalized trajectories have very low velocity-variability as compared to the ground truth. The effects of speed-normalization on place recognition cost matrix can also be noted in Fig. 7, where the wavy matching path gets relatively straightened due to dynamic frame sampling. Ideally, an increase in straightening of the matching path helps in narrowing the search space for searching a matching sequence locally. It further allows a longer sequence length to be used to improve matching performance, especially for linear search



(a) Kelvin Grove Footpath (b) Surfers Paradise Traffic

Figure 8: F-Score performance for vanilla and proposed approach for (a) KG-Footpath and (b) Surfers Paradise dataset. The performance consistently increases with increase in search range; however, proposed approach is further able to utilize the benefits of relatively longer sequence lengths due to straighter matching paths in cost matrix.

operations. Fig. 8 shows a performance comparison between the fixed frame sampling based vanilla approach and dynamic frame sampling based proposed approach for both Kelvin Grove Footpath and Surfers Paradise Traffic Datasets. The vanilla approach is unable to reach peak performance because of the wavy nature of the ground truth matching path as shown in Fig. 7 as compared to a relatively straight path obtained using the proposed approach. A longer sequence length for either of the methods decreases performance because the matching diagonal is not perfectly straight. The increase in sequence search range parameter consistently increases the performance as expected. The key performance gains are mentioned in Table 1. The values for sequence length and search range in the table were selected as found optimal from the performance curves in Fig. 8.

Table 1: Max F-Score comparison between vanilla fixed frame sampling and proposed dynamic frame sampling approach.

Dataset	Vanilla	Proposed	Sequence Length	Search Range
KG Footpath	0.24	0.59	25 meters	±0.5 radians
SP Traffic	0.48	0.67	90 meters	±0.5 radians

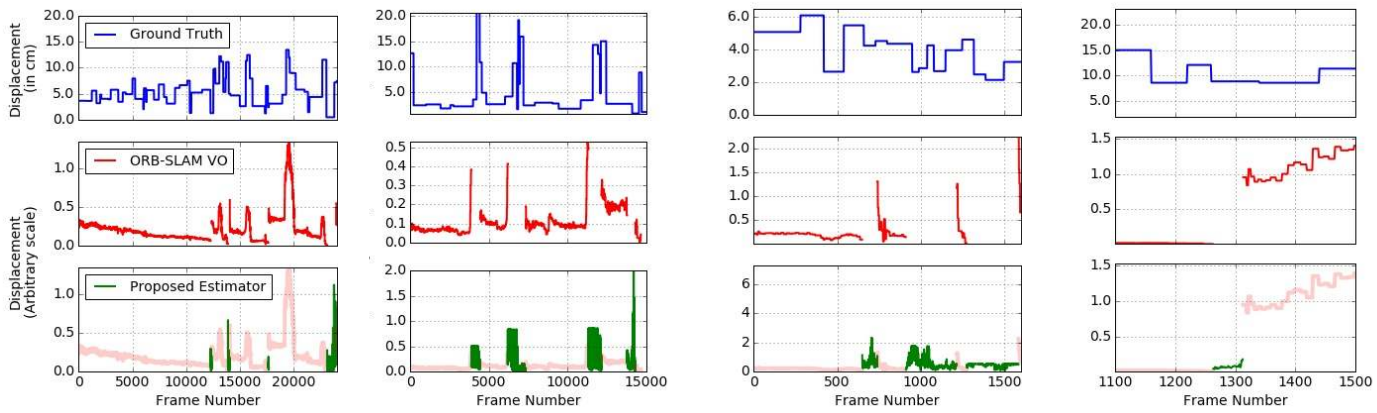
5.2 Hybrid Motion Estimation

Fig. 9 shows the comparison between the ground truth displacement, ORB-SLAM motion estimates and the proposed motion estimator learned using the current camera trajectory. It can be seen that the failures in visual odometry (discontinuities in red in the middle) can be supplemented with the proposed estimator (green trajectory overlaid on red estimates at the bottom). The reinitialization of visual odometry causes the displacement values to be scaled differently. The discontinuities in Fig. 9(a) and (b) are mainly because of rapid motion, whereas in (c) and (d), it is attributed to transitions in environmental conditions within the traverse, featureless regions, and pure rotations.

6 Discussion

6.1 SAD Error Behaviour Across Images

We used SAD error averaged over all the images in a local temporal window, but it is important to understand how the error changes with respect to the images. Fig. 10 shows that image alignment error, though occasionally variable across the images, is consistent across varying frame separations. The SAD error for any image increases uniformly with an increase in frame separation but it can be at times significantly different for different images. This can happen because of an abrupt change in environment, fairly large turns and speedy camera motion. However, such variations are usually contained within relatively small temporal windows corresponding to high SAD error. These temporal regions (peaks in Fig. 10 bottom graph) are captured by computing mean error vertically as shown in Fig. 10 and removing the corresponding images from the calculations in Eq. 1. Alternatively, these peaks could be used to explicitly detect turns with image offset matching [Milford and George,



(a) KG Footpath Night (b) KG Footpath Day (c) Residence Indoor-Outdoor (d) Parking Lot Night

Figure 9: The conventional state-of-the-art visual odometry is prone to failures due to high camera speed, sharp turns and low-light or featureless environments. It is shown here how the proposed motion estimator can be used to learn and supplement the translational motion estimation pipeline.

2014].

7 Conclusion and Future Work

We demonstrated the utility of image alignment error for learning a motion estimator using the past or current camera trajectory. The learned motion estimator was used to dynamically sample image frames to improve velocity sensitivity of the appearance-robust place recognition method. We also demonstrated a new hybrid visual odometry system that learns from current camera trajectory generated using state-of-the-art monocular visual odometry method and then estimates motion during system failure. While the conventional visual odometry systems can be accurately employed for egomotion estimation under ideal environmental conditions, our proposed odometry fusion method provides robust but cruder motion estimation pipeline enabling continuity of motion estimates. Furthermore, it can be potentially used for low-cost, low-compute assistive devices, where forward translational motion along with turn detection is often sufficient, for example, assistive technologies for visually-impaired people.

The straightening of sequence-search using robust motion estimation helps in finding a better matching sequence within the place recognition cost matrix by dynamically sampling the image frames. It can be seamlessly used with a variety of place representation options, including the use of deep convolutional features [Naseer *et al.*, 2017], as well as different methods of performing the sequence search, for example, the use of graph-based search [Vysotska and Stachniss, 2016].

The current work can be extended to include a learning method that can utilize the spatial information from the image differences across varying frame separations.

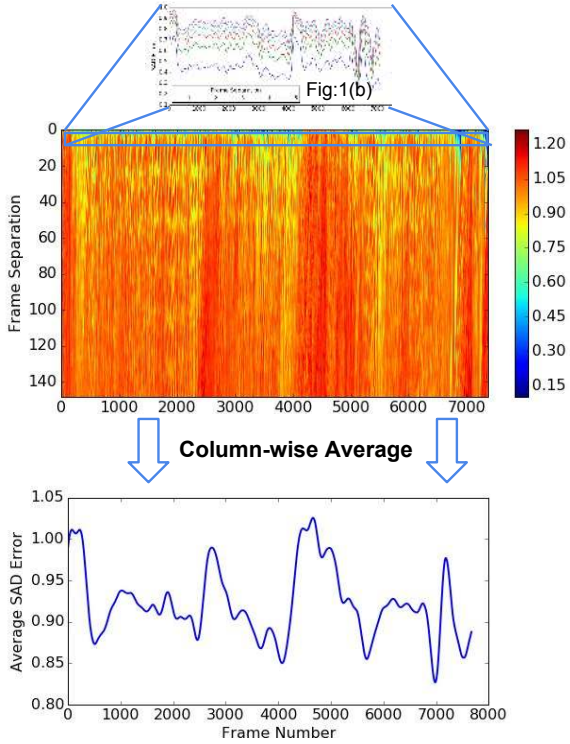


Figure 10: The heat map (top) shows that it is possible to capture the temporal regions that lead to a high SAD error by averaging it column-wise. The peaks (bottom) in the curve arise due to fairly large turns, abrupt environment change, and speedy motion and can be thresholded to remove variance in the calculation of Eq. 1. (The bottom curve has been smoothed using a sliding window of 5 seconds.)

Some of the regions within the image tend to be more stable in a local temporal window, which can be leveraged for more accurate motion estimation. Furthermore, the deep-learned visual semantics can be employed to either mask or locate objects and patches within the environment that are invariant to changes in environmental conditions, for example, roads and building facades. In its current form, the proposed approach can be used for improving appearance-robust place recognition and as a robust translational motion estimator which can be used in conjunction with state-of-the-art visual odometry techniques.

References

[Alismail *et al.*, 2016a] Hatem Alismail, Brett Browning, and Simon Lucey. Bit-planes: Dense subpixel alignment of binary descriptors. *arXiv preprint arXiv:1602.00307*, 2016.

[Alismail *et al.*, 2016b] Hatem Alismail, Brett Browning, and Simon Lucey. Direct visual odometry using bit-planes. *arXiv preprint arXiv:1604.00990*, 2016.

[Arandjelovic *et al.*, 2016] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5297–5307, 2016.

[Arroyo *et al.*, 2017] Roberto Arroyo, Pablo F Alcantarilla, Luis M Bergasa, and Eduardo Romera. Are you able to perform a life-long visual topological localization? *Autonomous Robots*, pages 1–21, 2017.

[Bhowmick and Hazarika, 2017] Alexy Bhowmick and Shyamanta M Hazarika. An insight into assistive technology for the visually impaired and blind people: state-of-the-art and future trends. *Journal on Multimodal User Interfaces*, 11(2):149–172, 2017.

[Chen *et al.*, 2014] Zetao Chen, Obadiah Lam, Adam Jacobson, and Michael Milford. Convolutional neural network-based place recognition. *arXiv preprint arXiv:1411.1509*, 2014.

[Chen *et al.*, 2017] Zetao Chen, Adam Jacobson, Niko Sunderhauf, Ben Upcroft, Lingqiao Liu, Chunhua Shen, Ian Reid, and Michael Milford. Deep learning features at scale for visual place recognition. *arXiv preprint arXiv:1701.05105*, 2017.

[Cummins and Newman, 2008] Mark Cummins and Paul Newman. Fab-map: Probabilistic localization and mapping in the space of appearance. *The International Journal of Robotics Research*, 27(6):647–665, 2008.

[Cummins and Newman, 2011] Mark Cummins and Paul Newman. Appearance-only slam at large scale with fab-map 2.0. *The International Journal of Robotics Research*, 30(9):1100–1123, 2011.

[Engel *et al.*, 2013] Jakob Engel, Jurgen Sturm, and Daniel Cremers. Semi-dense visual odometry for a monocular camera. In *Proceedings of the IEEE international conference on computer vision*, pages 1449–1456, 2013.

[Floros *et al.*, 2013] Georgios Floros, Benito van der Zander, and Bastian Leibe. Openstreetslam: Global vehicle localization using openstreetmaps. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pages 1054–1059. IEEE, 2013.

[Forster *et al.*, 2014] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. Svo: Fast semi-direct monocular visual odometry. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 15–22. IEEE, 2014.

[Garg *et al.*, 2017] Sourav Garg, Adam Jacobson, Swagat Kumar, and Michael Milford. Improving condition-and environment-invariant place recognition with semantic place categorization. *arXiv preprint arXiv:1706.07144*, 2017.

[Geiger *et al.*, 2011] Andreas Geiger, Julius Ziegler, and Christoph Stiller. Stereoscan: Dense 3d reconstruction in real-time. In *Intelligent Vehicles Symposium (IV), 2011 IEEE*, pages 963–968. Ieee, 2011.

[Glover *et al.*, 2015] Arren Glover, Edward Pepperell, Gordon Wyeth, Ben Upcroft, and Michael Milford. Repeatable condition-invariant visual odometry for sequence-based place recognition. In *Proceedings of the Australasian conference on robotics and automation (ACRA)*, 2015.

[Hansen and Browning, 2014] Peter Hansen and Brett Browning. Visual place recognition using hmm sequence matching. In *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*, pages 4549–4555. IEEE, 2014.

[Hansen and Browning, 2015] Peter Hansen and Brett Browning. Omnidirectional visual place recognition using rotation invariant sequence matching. 2015.

[Ho and Newman, 2007] Kin Leong Ho and Paul Newman. Detecting loop closure with scene sequences. *International Journal of Computer Vision*, 74(3):261–286, 2007.

[Hou *et al.*, 2017] Yi Hou, Hong Zhang, Shilin Zhou, and Huanxin Zou. Use of roadway scene semantic information and geometry-preserving landmark pairs to improve visual place recognition in changing environments. *IEEE Access*, 2017.

[Kejriwal *et al.*, 2016] Nishant Kejriwal, Swagat Kumar, and Tomohiro Shibata. High performance loop

closure detection using bag of word pairs. *Robotics and Autonomous Systems*, 77:55–65, 2016.

[Lowe, 2004] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[Lucas *et al.*, 1981] Bruce D Lucas, Takeo Kanade, et al. An iterative image registration technique with an application to stereo vision. 1981.

[Lynen *et al.*, 2017] Simon Lynen, Michael Bosse, and Roland Siegwart. Trajectory-based place-recognition for efficient large scale localization. *International Journal of Computer Vision*, 124(1):49–64, 2017.

[Milford and George, 2014] Michael Milford and Ashley George. Featureless visual processing for slam in changing outdoor environments. In *Field and Service Robotics*, pages 569–583. Springer, 2014.

[Milford and Wyeth, 2012] Michael J Milford and Gordon F Wyeth. Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 1643–1649. IEEE, 2012.

[Mur-Artal *et al.*, 2015] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015.

[Naseer *et al.*, 2014] Tayyab Naseer, Luciano Spinello, Wolfram Burgard, and Cyrill Stachniss. Robust visual robot localization across seasons using network flows. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[Naseer *et al.*, 2017] Tayyab Naseer, Gabriel L Oliveira, Thomas Brox, and Wolfram Burgard. Semantics-aware visual localization under challenging perceptual conditions. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[Oh *et al.*, 2017] Yeonju Oh, Wei-Liang Kao, and Byung-Cheol Min. Indoor navigation aid system using no positioning technique for visually impaired people. In *International Conference on Human-Computer Interaction*, pages 390–397. Springer, 2017.

[Pepperell *et al.*, 2013] Edward Pepperell, Peter Corke, and Michael Milford. Towards persistent visual navigation using smart. In *Proceedings of Australasian Conference on Robotics and Automation*. ARAA, 2013.

[Pepperell *et al.*, 2014] Edward Pepperell, Peter I Corke, and Michael J Milford. All-environment visual place recognition with smart. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 1612–1618. IEEE, 2014.

[Pepperell *et al.*, 2015] Edward Pepperell, Peter I Corke, and Michael J Milford. Automatic image scaling for place recognition in changing environments. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pages 1118–1124. IEEE, 2015.

[Pepperell *et al.*, 2016] Edward Pepperell, Peter Corke, and Michael Milford. Routed roads: Probabilistic vision-based place recognition for changing conditions, split streets and varied viewpoints. *The International Journal of Robotics Research*, 35(9):1057–1179, 2016.

[Peretroukhin *et al.*, 2017] Valentin Peretroukhin, Lee Clement, and Jonathan Kelly. Reducing drift in visual odometry by inferring sun direction using a bayesian convolutional neural network. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 2035–2042. IEEE, 2017.

[Saito and Kuroda, 2013] Takato Saito and Yoji Kuroda. Mobile robot localization using multiple observations based on place recognition and gps. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pages 1548–1553. IEEE, 2013.

[Sunderhauf *et al.*, 2015] Niko Sunderhauf, Sareh Shirazi, Adam Jacobson, Feras Dayoub, Edward Pepperell, Ben Upcroft, and Michael Milford. Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free. *Proceedings of Robotics: Science and Systems XII*, 2015.

[Torii *et al.*, 2015] Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1808–1817, 2015.

[Vysotska and Stachniss, 2016] Olga Vysotska and Cyrill Stachniss. Lazy data association for image sequences matching under substantial appearance changes. *IEEE Robotics and Automation Letters*, 1(1):213–220, 2016.

[Wang *et al.*, 2017] Sen Wang, Ronald Clark, Hongkai Wen, and Niki Trigoni. Deepvo: towards end-to-end visual odometry with deep recurrent convolutional neural networks. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 2043–2050. IEEE, 2017.

[Ye *et al.*, 2014] Hanlu Ye, Meethu Malu, Uran Oh, and Leah Findlater. Current and future mobile and wearable device use by people with visual impairments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3123–3132. ACM, 2014.