

Learning place-dependant features for long-term vision-based localisation

Colin McManus¹ · Ben Upcroft² · Paul Newman¹

Received: 7 November 2014 / Accepted: 3 July 2015 / Published online: 23 July 2015
© Springer Science+Business Media New York 2015

Abstract This paper presents an alternative approach to the problem of outdoor, persistent visual localisation against a known map. Instead of blindly applying a feature detector/descriptor combination over all images of all places, we leverage prior experiences of a place to learn place-dependent feature detectors (i.e., features that are unique to each place in our map and used for localisation). Furthermore, as these features do not represent low-level structure, like edges or corners, but are in fact mid-level patches representing distinctive visual elements (e.g., windows, buildings, or silhouettes), we are able to localise across extreme appearance changes. Note that there is no requirement that the features possess semantic meaning, only that they are optimal for the task of localisation. This work is an extension on previous work (McManus et al. in Proceedings of robotics science and systems, 2014b) in the following ways: (i) we have included a landmark refinement and outlier rejection step during the learning phase, (ii) we have implemented an asynchronous pipeline design, (iii) we have tested on data collected in an urban environment, and (iv) we have implemented a purely monocular system. Using over 100 km worth of data for train-

ing, we present localisation results from Begbroke Science Park and central Oxford.

Keywords Feature learning · Appearance changes · Cross seasonal · Visual localisation · Long-term autonomy · Outdoor localisation

1 Introduction

Visual localisation across different lighting conditions, different weather conditions, and different seasons is a hard problem due to extreme appearance changes. Most vision systems use a feature-based front end, whereby salient image regions are both detected and described in a compact manner to enable the matching of these features across different images. The typical approach is to apply these feature detectors/descriptors [e.g., scale invariant feature transform (Lowe 2004) or speeded up robust features (Bay et al. 2008)] over the entire image at various scales. However, we believe there are some serious issues with just blindly applying these detectors across the entire image for the problem of localisation.

Firstly, matching point features typically fails under extreme appearance changes, such as different lighting conditions and weather conditions (Furgale and Barfoot 2010; McManus 2010; Churchill and Newman 2012). Recent attempts have been made to address the issue of lighting changes by using an illumination-invariant colour space (McManus et al. 2014a; Maddern et al. 2014) or learning a lighting-invariant descriptor space (Ranaganathan et al. 2013). However, dealing with gross appearance changes from night-to-day or summer-to-winter requires something beyond point features. Matching sequences using whole-image information has shown impressive results to this end

This is one of several papers published in *Autonomous Robots* comprising the “Special Issue on Robotics Science and Systems”.

✉ Colin McManus
colin@robots.ox.ac.uk

Ben Upcroft
ben.upcroft@qut.edu.au

Paul Newman
pnewman@robots.ox.ac.uk

¹ Mobile Robotics Group, University of Oxford, Oxford, UK

² CyPhy Lab, Queensland University of Technology, Brisbane, Australia

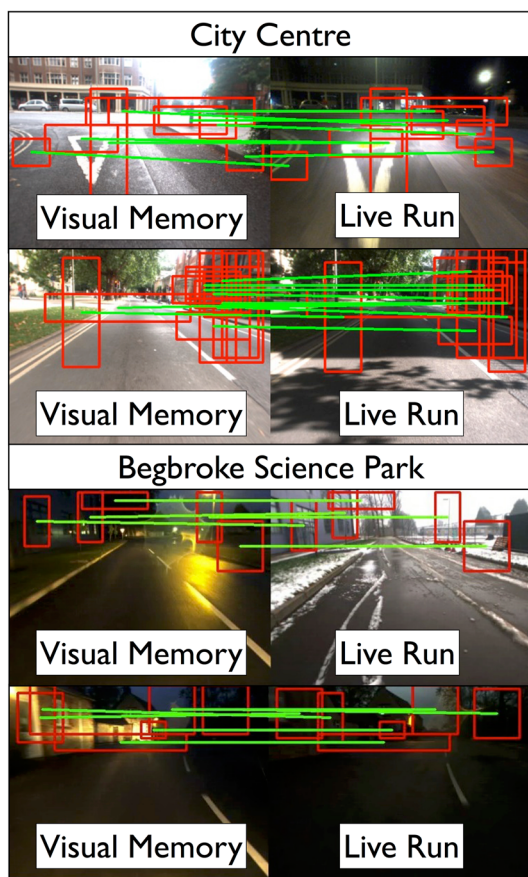


Fig. 1 Examples of our system running on challenging datasets with significant appearance changes; Oxford on the top and the Begbroke Science Park on the bottom. Our system learns unique visual elements that are place specific (i.e., distinctive to a particular place in the environment). These unique visual elements, called *scene signatures*, are used for realtime localisation to provide a rough metric estimate of the vehicle's position and orientation

(Milford 2013; Milford and Wyeth 2012), but is limited to topological localisation.

The second and perhaps most significant issue with the standard feature-based approach is that it seems somewhat naive with regards to the problem of localisation against a known map. If we have been to an environment several times under different appearance conditions, it must be possible to learn what is unique and important at a given place and to use this information for coarse metric estimation. From this *weak* localisation, other systems, such as lane/curb detectors, can be run in concert to refine the estimate for smooth local control.

We put forth an alternative approach to the problem of localisation against a known map and present an offline, unsupervised method to learn a bank of place-dependent feature detectors that fire on unique visual elements in the environment. These unique elements, called *scene signatures*, can be used online for rough metric localisation across a variety of challenging appearance changes (see Fig. 1). It is worth men-

tioning that there is no requirement that these features hold any semantic meaning; we are not trying to solve the problem of scene understanding. Instead, we are simply searching for mid-level patches that can be reliably observed across a variety of appearance changes, making them suitable for the task of localisation.

This work is an extension of our previous work (McManus et al. 2014b) in the following ways: (i) we have included an offline landmark-refinement step that optimises landmark positions in the map for improved localisation, (ii) we have implemented a realtime localisation pipeline, (iii) we have tested it on more data and in dynamic environments, and (iv) we have developed a monocular version of the system as opposed to stereo. We believe that our approach is a step in the right direction and moves away from the naivety of using one feature detector for all time. Instead, we believe that leveraging prior knowledge of appearance and/or structure to learn what is useful to us for navigation tasks is the way forward.

This paper is organised as follows. Section 2 reviews a variety of techniques that leverage knowledge of prior structure to aid in motion estimation and/or localisation. Section 3 provides a high-level system overview. First, we introduce the offline learning algorithm that is used to produce a set of place-dependent feature detectors, as well as the offline landmark refinement stage. We then describe how these scene signatures are used for localisation. Section 4 presents our feature stability and localisation experiments/results from the Begbroke Science Park and central Oxford. In total, we used over 100 km of data for the training. Finally, Sect. 6 presents a conclusion of our work.

2 Related work

The central idea of this paper is to challenge the traditional view of blindly applying one technique (e.g., out-of-the-box feature detector) for navigation tasks. Point-feature matching across extreme lighting and weather conditions often fails as finding associations of low-level structure is extremely challenging and sometimes not possible due to gross appearance changes. Although Valgren and Lilienthal (2010) showed that topological localisation across seasons could be feasible with point features, metric localisation was never examined and the experiments were conducted on a very limited set of images.¹ We wish to leverage knowledge of prior structure and/or appearance to improve our systems in terms of robustness and reliability. There have been several works that share

¹ In their earlier work, Valgren and Lilienthal (2007) originally concluded that it was not possible to perform localisation across seasons with point features. Their later work incorporated epipolar geometry constraints to make this possible over a limited set of images.

this belief and have been investigating alternative approaches to traditional problems like egomotion estimation and localisation.

Richardson and Olson (2013) presented an approach to learn an optimal feature detector, based on a family of convolutional filters, for visual odometry (VO) tasks. Lategahn et al. (2013) learn a whole-image descriptor to optimise place recognition outdoors. Milford and Wyeth (2012) presented SeqSLAM, which use sequences of whole-image information for topological localisation across challenging appearance conditions. Neubert et al. (2013) showed how to improve SeqSLAM to localise across seasons by learning a dictionary to translate between a winter/summer visual vocabulary. However, these approaches are very sensitive to viewpoint changes and are purely topological. Naseer et al. (2014) presented another sequence-based localisation system, but instead of performing direct image-to-image matching, they used HOG-based description of the images and formulated the problem as a minimum cost network flow to find the best matching sequence. Their method outperforms SeqSLAM, but is again purely topological. Johns and Yang (2013) learned visual and spatial feature co-occurrence maps to perform localisation across different lighting conditions throughout the day. Their method relies on visual words quantised from local features, but was able to outperform SeqSLAM in their datasets. However, their method is again purely topological and did not test across different weather conditions or seasons.

Several researchers have investigated the idea of semantic localisation, which shares a similar viewpoint that higher-level information is useful for localisation. Atanasov et al. (2014) present a system that uses random finite sets (RFS) to represent semantic information from their object detector, which allows them to account for missed detections, false positives, and perform data association. They show how the RFS observation model is equivalent to a matrix permanent computation, which makes the filtering problem tractable. Renato et al. (2013) present SLAM++; an *object oriented* approach to simultaneous localisation and mapping (SLAM) that uses 3D models of common indoor objects, such as chairs and tables, to perform realtime, full 6 degree-of-freedom (DOF) SLAM. Ko et al. (2013) and Yi et al. (2009) present an approach for semantic mapping, active localisation, and local navigation and planning. Their system abstracts spatial relationships and actions to higher level concepts (e.g., object is near or distant). Anati et al. (2012) side stepped the problem of data association by using the dense heat maps produced by the object detectors and incorporate a per-pixel likelihood score for observing a particular class. They incorporated these soft detections using a particle filter and demonstrated the system working in a large indoor environment. Bao and Savarese (2011) introduced the

concept of semantic structure from motion, which attempts to find the optimal maximum-likelihood estimate of camera poses, objects, and points. They show that in addition to outperforming a point-feature-based structure-from-motion system, they can also improve object detection due to extra geometric information when compared to detecting objects in images alone. Castle et al. (2007) developed a hybrid monocular SLAM system that combined traditional sparse features with known planar objects.

Although all of these aforementioned approaches shared the view that matching low-level structure isn't always the best approach for localisation, they introduce the challenging problem of scene understanding. As we will show, it is not necessarily the case that a visual element must have semantic meaning to be valuable for localisation tasks. For instance, our algorithm is able to find unique rectangular strips that encompass various structures, such as a building, road, and vegetation. This on its own is not a singular class, but simply a unique visual strip associated with that place. We are therefore not limited to a predefined set of classes, but instead let the algorithm find what is unique in a given place.

Note that our approach is very different from the localisation and mapping systems of Davison et al. (2007) and Davison and Murray (2002), which use image patches as their landmarks. These methods still rely on interest-point detection to find the patches, which are relatively small (e.g., 11×11 pixels in size). By construction, scene signatures are large distinctive elements in the scene that can be matched across different appearance conditions.

We presented a proof-of-concept system in our earlier work (McManus et al. 2014a), which showed it was possible to produce metric localisation estimates across extreme appearance changes. Our training algorithm is inspired by the work of Doersch et al. (2012), which uses an iterative, discriminative clustering scheme similar to Singh et al. (2012). In this paper, we extend the training phase to include landmark refinement for each candidate scene signature for improved localisation performance.

We wish to emphasize that we are not taking the stance that point features are bad. There is a rich heritage of using point features in many successful visual navigation systems (Konolige et al. 2010; Sibley et al. 2010; Piniés et al. 2010; Kaess et al. 2012) and they certainly have their place for a number of applications. For relative-motion estimation, like Visual Odometry, point features work extremely well as viewpoint and lighting conditions typically do not change much from frame to frame. Or if one is localising in an environment without much visual change, then point features could be a good solution. However, for outdoor localisation over long periods of time, we believe that point features may not be the best solution for the problem at hand.

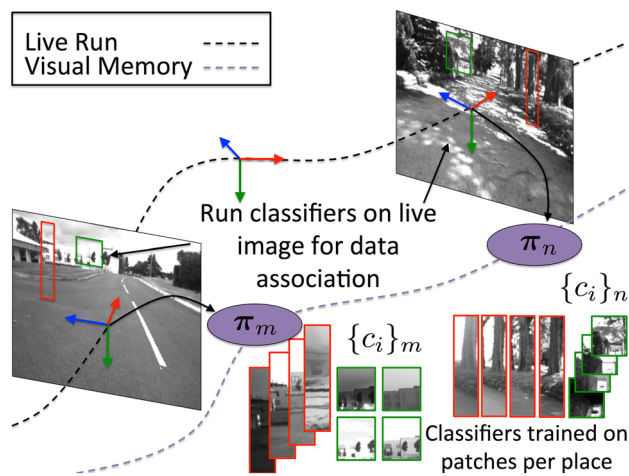


Fig. 2 Offline, we learn scene signatures in the form of SVM classifiers, where each classifier is associated with a particular place, π_p . At run-time, we use the bank of pre-trained classifiers associated with the nearest place, π_p , to perform data association and then localisation. By using larger, distinctive visual elements, we are able to localise in regions with extreme appearance change, where the point-feature-based counterpart fails

3 System overview

This section presents an overview of the two primary components to our system: (i) unsupervised learning of place-dependent features and (ii) our online localisation system. At a high level, our system works as follows. Offline, we learn a bank of *place-dependent* classifiers, $\{c_i\}_p$, that can detect unique visual elements specific to a particular place, π_p , such as trees, buildings, or unique strips in the image. At runtime, the vehicle can load the set of classifiers, $\{c_i\}_p$, associated with the closest location, π_p , and use them for rough, metric pose estimation (see Fig. 2). We say “rough”, metric pose estimation because these features typically represent distant objects, meaning that the translational component of the estimate is not well constrained. However, with a motion prior to guide the system, we show that the accuracy is comparable if not better than our inertial navigation system (INS) system.

Having said this, it’s important to note that like all localisation systems, our system does become lost at times (i.e., fail to localise). In these situations, we require a seed to relocalise the vehicle, such as GPS or a place recognition system. However, as we will show, the likelihood of traveling large distances without localising is lower with our approach than with a standard point-feature-based approach, meaning that the number of resets is minimised.

Nonetheless, it is true that we assume that this oracle (i.e., purely topological localiser) is available when the system becomes lost. This is a reasonable assumption, since in practice, if GPS is available, we would certainly make use of it to keep track of the approximate location of the system. Addi-

tionally, this allows us to better assess the performance of the system over the entire range of datasets.

3.1 Offline learning

There are two steps in the learning phase. The first step involves training SVM classifiers to find a set of candidate scene signatures through unsupervised, iterative, discriminative-clustering training. The second step performs bundle adjustment to find optimal landmark locations for each scene signature (i.e., optimal in that the landmark location minimises reprojection error over a sequence of frames). The output of these two processes yield a bank of motion-consistent classifiers, $\{c_i\}_p$, for each place, π_p .

For clarity, we reiterate that a scene signature is the underlying visual element in the scene that can be identified across various appearances. Thus, we seek to train a classifier that will detect a given scene signature.

3.2 Training algorithm

Our training data consist of a collection of images at approximately the same location and viewpoint under a variety of appearances (see Fig. 3). We collected these data using a survey vehicle equipped with an INS system, and defined *places* as physical locations spaced 10m apart along the driven route according to INS. The important aspect of the training data is that the viewpoint is as similar as possible. Ideally, we would like the viewpoints to be identical, but this is not possible due to inaccuracies in the INS and because the driven routes vary from one dataset to the next.

Referring to Fig. 3, the reader will see three shapes drawn on every image (i.e., the red, blue, and green rectangles). The goal is to find out which set of patches represent “stable” visual elements across different appearances. By “stable,” we mean that if we trained a classifier to detect these types of patches, we would expect the classifier to find the same visual elements in a validation set, regardless of appearance conditions. In other words, we want a classifier that will always fire on the same set of trees, for example, in the same physical place, regardless of time of day, time of year, or weather. It is important to note that not all patches represent stable visual elements. For example, one can imagine that a classifier trained to detect the green rectangle may fire anywhere along the curb (i.e., this is not a locally distinctive feature as the surrounding region looks similar in appearance). However, the other two shapes (i.e., the red and the blue shape) would likely serve as good patches from which to train, because the underlying visual element appears very distinctive. In the red patch, we see a building with a window. This appears nowhere else locally and can be associated across all the images. The blue patch is a unique strip that transitions vertically from the ground to a wedge of brick

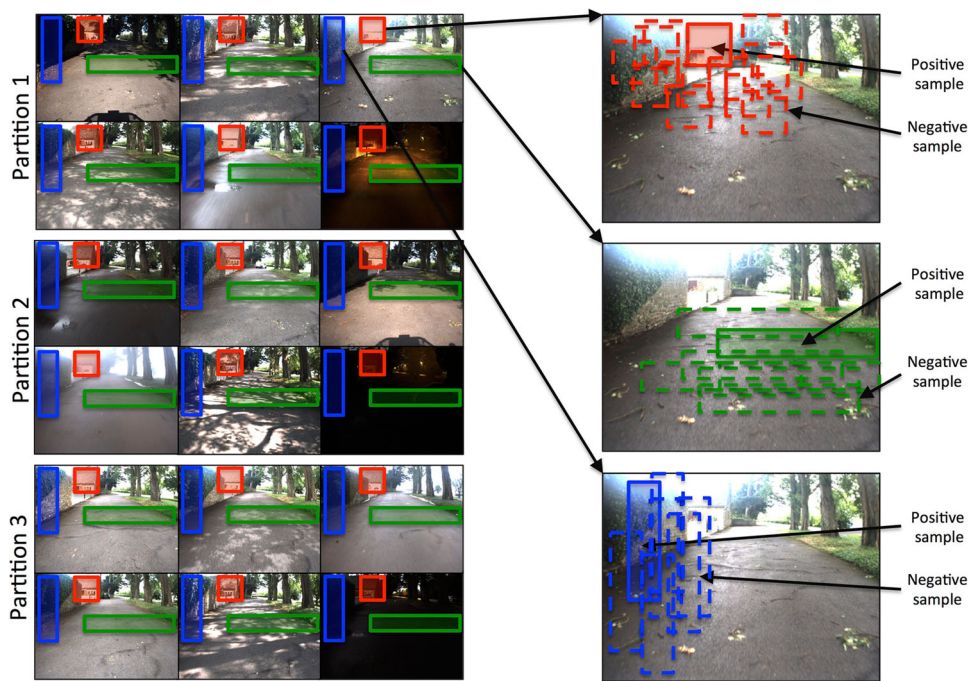


Fig. 3 Illustration of the first stage of our training algorithm. The training data consist of images of the same place under varying appearance conditions. These images are partitioned into three groups, such that each group has as much visual variability as possible. Then, we sample a set of shapes from each image, represented in this figure by *coloured rectangles*. The steps proceed as follows. Take a shape (e.g., *red*), compute HOG descriptors (Dalal and Triggs 2005) for each *red shape* in all images in partition 1. These will be the positive set. Sample the image around that shape and compute HOG descriptors (shown on the right column). These will be the negative set. Train a linear SVM classifier

and use this classifier on partition 2, which acts as a validation set. Take the top K firings in partition 2 and use these as positives to retrain the SVM. The new SVM, which was trained on just the subset of positives in partition 2 and its respective negative samples, is then applied to partition 3 in the same fashion. The validation set becomes the training set and the process repeats, wrapping around to the first partition again. If the top K firings in each respective partition remain unchanged, then we have converged to a discriminative feature. If we have not converged within three iterations, we terminate the process and reject the classifier as a candidate scene signatures (Color figure online)

wall to a bush. We might expect this to be very distinctive as well.

The question is, how do we determine which shapes serve as a basis for a stable classifier? We could train a classifier for each candidate set of shapes (e.g., the red, blue, and green) and then use the classifier on a hold-out set to see where the detections fire relative to the groundtruth location. However, this is unappealing as we would have to define a closeness threshold in image space to label a positive detection. Additionally, it fails to take into account that not all of the positive patches are informative. For instance, in Fig. 3, we can see that some of the blue patches encompass textureless, black regions in the image and would not be helpful to use as positive examples. Thus, we seek an unsupervised training technique that can accomplish the following two tasks: (i) it must be able to identify what types of patches represent “stable” elements (e.g., red, blue, or green?), and (ii) it must be able to select a discriminative subset of the positives for training (e.g., ignore the textureless blue patches or occluded patches). Fortunately, we can use an iterative training scheme similar to Doersch et al. (2012) and Singh et al. (2012) to accomplish this task.

The basic idea is as follows. Consider separating the images in Fig. 3 into three partitions as shown,² and training an SVM classifier on the red patches, which, initially, are all labeled as positives, with the negatives being sampled around the local region. After training this classifier, we apply it to the red rectangles in the second partition (i.e., a validation set). We then rank each red rectangle in the validation set according to its score and train a new SVM using the top K detections.³ This new classifier is then applied to the third partition in a similar manner. The top K firings from the validation set become the new positive examples from which to retrain. This new SVM would then be applied back to the first partition and the cycle would continue until convergence criteria are met. The convergence criteria require that the top K firings in each respective partition do not change, as this implies that we have found a subset of discriminative patches. Note that this is a very conservative approach, as it would select only the most representative examples of the visual element we seek to classify. However, we gladly trade

² As was done in Doersch et al. (2012).

³ We set $K = 5$ as done in Doersch et al. (2012).

recall for higher precision in this context, as we are concerned with limiting the number of mis-associations for pose estimation. If the convergence criteria are not met within three iterations,⁴ we reject the candidate classifier as representing a “stable” visual element.

In summary, the output of this procedure is a set of classifiers, each of which detects a region in the image that is constant under appearance changes and distinct in the image (i.e., a scene signature).

For our implementation, we used fixed set of 296 predefined shapes for every image. From this fixed template of shapes, the training algorithm selects the subset using the process described above. Increasing the sample size increases the training times required since our algorithm works by exhaustively training/testing each candidate shape. In the future, we wish to examine how to more intelligently pre-select these candidate shapes as a function of place. For example, having found a set of scene signatures for a given place, if we were to retrain with more data, we could focus on sampling in regions that produced the most scene signatures, as these are likely feature-rich areas in the image.

The next stage in this process looks at the stability of the classifiers over a local window of images and optimises a motion-consistent landmark location for each scene signature.

3.2.1 Landmark refinement

The final step involves landmark refinement for each scene signature, which represents one of the extensions to our previous work. The landmark refinement serves two purposes: (i) to eliminate bad candidate classifiers and (ii) compute a landmark position for each scene signature to enable metric localisation.

The contribution of this section comes purely in the application to our weak localisation system that uses scene signatures. As each scene signature represents a large patch in the image, assigning a singular point value is not a true representation of the depth since it can encompass many structures within the patch. We therefore attempt to minimise the error in this point-value assumption through a nonlinear, least-squares optimisation.

Consider one of the images in Fig. 3 and imagine taking a window of images forward and backward in time from the initial training image to test the classifier on each image in the window.⁵ As the vehicle moves smoothly over this image sequence, one would expect the feature detections to move

⁴ We chose three as was done in Doersch et al. (2012). Note that Singh et al. (2012) came to a similar conclusion that only 4–5 iterations are necessary.

⁵ In our experiments, the window was taken to be the distance between places, which is 10 m.

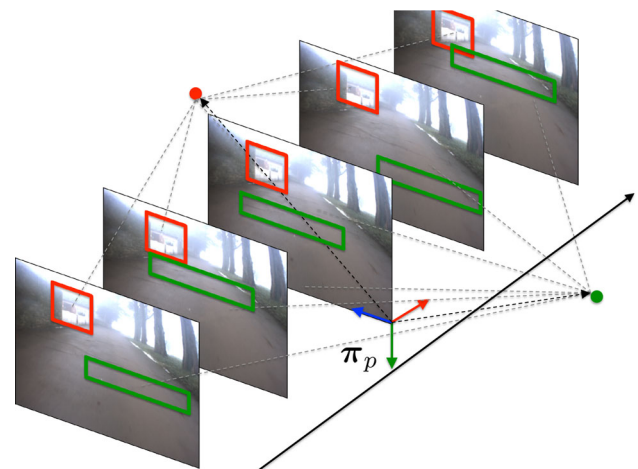


Fig. 4 Each candidate classifier is subjected to a round of temporal checks during the landmark refinement stage. A window of frames is taken around each place and the classifier is fired on all frames in order to compute statistics on the stability of the classifier. Inlier detections are then used for landmark refinement

smoothly over time as well. Figure 4 illustrates this process. We use 2-point RANSAC to compute a line of best fit for the temporal $x - y$ locations and reject any candidate classifiers if the ratio of inliers is less than half of the samples. If the inlier set is over half, we take this set of feature detections to perform landmark adjustment.

In our original work (McManus et al. 2014a), we estimated the landmark position by performing left-to-right stereo matching. However, as these features represent large patches in the image, template matching to obtain a single point estimate is not very sensible due to the large parallax. Instead, we first discretely sample a number of possible depths from 0–100 m with a resolution of 1 m and compute the total reprojection error over the window for each depth. We then take the best depth and perform a non-linear refinement around this initial guess, which is similar to a common method used in GPU depth computation (McKinnon et al. 2012).

Assuming that we have odometry along with the training data, such as wheel odometry or VO, we can use the known incremental transformations, $\{\mathbf{T}_{0,1}, \dots, \mathbf{T}_{k-1,k}\}$, between each image to reproject a landmark, \mathbf{p}_p^i , defined in π_p , into each frame in the window according to our monocular camera model:

$$\mathbf{z}_j^i := \mathbf{h}(\mathbf{T}_{j,p}, \mathbf{p}_p^i) + \mathbf{v}_j^i, \quad \mathbf{v}_j^i \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_j^i) \quad (1)$$

where $\mathbf{z}_j^i = [u, v]^T$ in the j th image and \mathbf{v}_j^i is zero-mean Gaussian noise. Our objective function is then just an uncertainty-weighted squared difference between the observed location, \mathbf{z}_j^i , (given by the classifier) and the predicted location, $\hat{\mathbf{z}}_j^i = \mathbf{h}(\mathbf{p}_p^i, \mathbf{T}_{j,p})$, given by our reprojection function:

$$J(\mathbf{p}_p^i) = \frac{1}{2} (\mathbf{z} - \mathbf{h}(\mathbf{p}_p^i))^T \mathbf{R}^{-1} (\mathbf{z} - \mathbf{h}(\mathbf{p}_p^i)) \quad (2)$$

where

$$\mathbf{z} := \begin{bmatrix} \mathbf{z}_0^i \\ \mathbf{z}_1^i \\ \vdots \\ \mathbf{z}_M^i \end{bmatrix}, \quad \mathbf{h}(\mathbf{p}_p^i) := \begin{bmatrix} \mathbf{h}(\mathbf{T}_{0,p}, \mathbf{p}_p^i) \\ \mathbf{h}(\mathbf{T}_{1,p}, \mathbf{p}_p^i) \\ \vdots \\ \mathbf{h}(\mathbf{T}_{M,p}, \mathbf{p}_p^i) \end{bmatrix},$$

$$\mathbf{R} := \text{diag}(\mathbf{R}_0^i, \mathbf{R}_1^i, \dots, \mathbf{R}_M^i).$$

The \mathbf{R}_j^i represent the measurement noise covariance for each detected scene signature. These were determined according to the approach described in McManus et al. (2014a). The uncertainty of a visual element in image space, \mathbf{R}_j^i , will be a function of the scale, s , at which it was detected, the area of the patch, a , the search resolution used when detecting the feature, r , and the SVM detection probability, λ :

$$\mathbf{R}_j^i = \mathbf{f}(a, r, s, \lambda). \quad (3)$$

The relationship between the scale and search resolution is given by:

$$\mathbf{R}_j^i \propto \frac{1}{s} \mathbf{Q}_j^i, \quad (4)$$

where \mathbf{Q}_j^i is the noise covariance on the search resolution, which is scaled according to the pyramid level at which the detector fires. The relationship with the other parameters is less clear. Intuitively, the lower the likelihood of being a scene signature and the larger the area of the patch, the less certain the keypoint position should be. We therefore use the following heuristic:

$$\mathbf{R}_j^i := \frac{a}{\lambda s} \mathbf{Q}_j^i. \quad (5)$$

As this is a nonlinear least-squares system, we take the standard approach and perform a first-order Taylor series expansion:

$$J(\bar{\mathbf{p}}_p^i + \delta \mathbf{p}^i) \approx (\bar{\mathbf{e}}_p - \mathbf{H}_p \delta \mathbf{p}^i)^T \mathbf{R}^{-1} (\bar{\mathbf{e}}_p - \mathbf{H}_p \delta \mathbf{p}^i), \quad (6)$$

where $\bar{\mathbf{e}}_p := \mathbf{z} - \mathbf{h}(\bar{\mathbf{p}}_p^i)$ and $\mathbf{H}_p := \partial \mathbf{h} / \partial \delta \mathbf{p}^i$. Taking the derivative of $J(\cdot)$ with respect to the perturbation, setting it to zero and solving yields the following system:

$$(\mathbf{H}_p^T \mathbf{R}^{-1} \mathbf{H}_p) \delta \mathbf{p}^i = -\mathbf{H}_p^T \mathbf{R}^{-1} (\mathbf{z} - \mathbf{h}(\bar{\mathbf{p}}_p^i)), \quad (7)$$

allowing us to iteratively update the landmark until convergence. For the non-linear optimisation we use Levenberg Marquardt (LM) (Levenberg 1944), which augments the

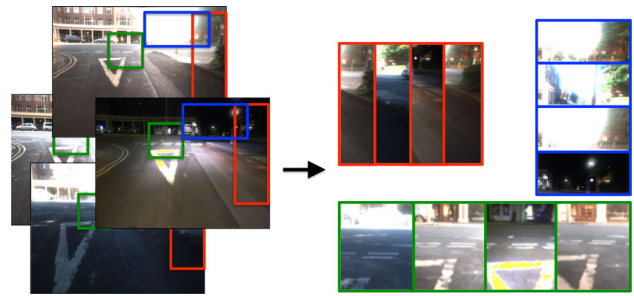


Fig. 5 Example scene signatures learned by our algorithm. Collections of images from the same place but different appearance conditions are used for training data (left) and the output of our training algorithm is a collection of SVM classifiers trained on distinctive image patches (right)

coefficient matrix by a block-diagonal matrix, $\lambda \mathbf{1}$, where $\mathbf{1}$ is the identity matrix and λ is used during the optimisation to control the convergence properties. We also use a Huber cost function for robustness. The Huber cost is given by the following (Hartley and Zisserman 2004):

$$\rho(c) = \begin{cases} c^2 & |c| < \alpha \\ 2\alpha|c| - \alpha^2 & \text{otherwise} \end{cases}$$

which results in just reweighting the inverse covariance terms in (8). The resulting system is given by:

$$(\mathbf{H}_p^T \tilde{\mathbf{R}}^{-1} \mathbf{H}_p + \lambda \mathbf{1}) \delta \mathbf{p}^i = -\mathbf{H}_p^T \mathbf{R}^{-1} (\mathbf{z} - \mathbf{h}(\bar{\mathbf{p}}_p^i)), \quad (8)$$

where $\tilde{\mathbf{R}} := \text{diag}(w_0 \mathbf{R}_0, \dots, w_M \mathbf{R}_M)$, with the weights being given by:

$$w_i = \begin{cases} 1 & |c| < \alpha \\ \frac{\alpha}{c_i} & \text{otherwise.} \end{cases} \quad (9)$$

To summarise, after finding a coarse initial guess from our discrete sampling we solve the above nonlinear least-squares problem using LM. This landmark position is then stored with the scene signature and used for pose estimation online. The end result of this procedure is a set of motion consistent scene signatures, $\{c_i\}_p$, defined for each place, π_p . Example scene signatures can be seen in Fig. 5.

3.3 Online localisation

3.3.1 Weak localisers

This section introduces the notion of a *weak localiser*. As the scene signatures represent large image patches that are mostly located in the far field, the translational estimates from the nonlinear solver are not very accurate. As a result, we use a strong motion prior from an odometry source to bound the solution. We wish to stress however, that for our

application, we believe that a weak localiser which is accurate on the order of meters (i.e., similar to a commercial GPS system) is good enough to seed other onboard navigation systems for realtime control. In other words, we envision a hierarchical approach, whereby a crude topological localiser seeds our system, which produces an estimate accurate to within meters. We believe such an approach is adequate for autonomous navigation, since local observations onboard the vehicle can be used for planning and obstacle avoidance.

At runtime, we load the bank of classifiers, $\{c_i\}_p$, associated with the closest place π_p and use them on the live image at time t_k to produce a set of measurements, \mathbf{z}_k^i . As each classifier in the map has an associated landmark position, \mathbf{p}_p^i , we can use our camera model (1) to predict the location of a landmark in the live frame, according to the transformation matrix, $\mathbf{T}_{k,p}$.

As stated earlier, we use a strong motion prior to predict the transformation, $\hat{\mathbf{T}}_{k,p}$. Including the prior estimate, $\hat{\mathbf{T}}_{k,p}$, the final least-squares system we seek to optimise is given by the following:

$$J(\mathbf{T}_{k,p}) = \frac{1}{2} \mathbf{q}(\mathbf{T}_{k,p}, \hat{\mathbf{T}}_{k,p})^T \mathbf{P}_x^{-1} \mathbf{q}(\mathbf{T}_{k,p}, \hat{\mathbf{T}}_{k,p}) + \frac{1}{2} (\mathbf{z}_k - \mathbf{h}(\mathbf{T}_{k,p}, \mathbf{p}_p))^T \mathbf{R}^{-1} (\mathbf{z}_k - \mathbf{h}(\mathbf{T}_{k,p}, \mathbf{p}_p)) \tag{10}$$

where

$$\mathbf{z}_k := \begin{bmatrix} \mathbf{z}_k^0 \\ \vdots \\ \mathbf{z}_k^M \end{bmatrix}, \quad \mathbf{p}_p := \begin{bmatrix} \mathbf{p}_p^0 \\ \vdots \\ \mathbf{p}_p^M \end{bmatrix}, \quad \mathbf{R} := \text{diag}(\mathbf{R}_k^0, \dots, \mathbf{R}_k^M),$$

and $\mathbf{q}(\cdot)$ is a function that takes two SE3 transformation matrices and computes a 6×1 error vector. We also incorporate a Huber cost function and perform LM for the nonlinear optimisation.

3.3.2 Realtime system

To obtain realtime operation, an asynchronous pipeline design was used in order to fuse lower frequency localisation updates (~5 Hz) with high frequency odometry measurements (~20 Hz). The pipeline is illustrated in Fig. 6. The major processing blocks are described below.

- *Detector* This block performs the detection using a bank of classifiers, $\{c\}_p$, for the current place, which are trained on Histogram of Oriented Gradient (HOG) descriptors of each scene signature. Currently, OpenCV’s OpenCL GPU HOG is being used for the detection. Depending on the graphics card and the number of models being

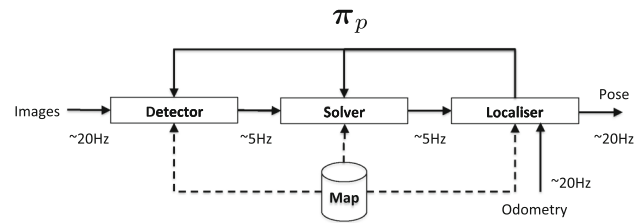


Fig. 6 Block-flow diagram of our localisation pipeline. In separate threads, the detection block and solver perform the scene-signature localisation as described in Sect. 3.3.1. The high-level localiser integrates the odometry measurements in-between localisation updates and also performs posegraph relaxation to smooth the output

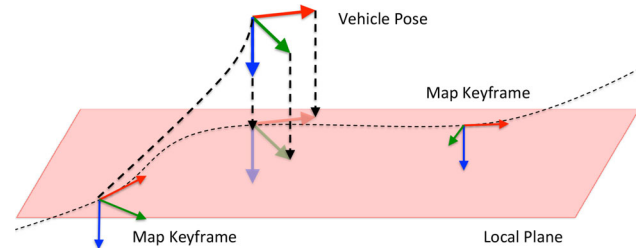


Fig. 7 Using the two nearest keyframes in the map, we define a local plane and project the estimate vehicle pose onto this plane. This augmented pose solution is then used in the next localisation cycle. This trick proved to work well in preventing drift in the z-direction. We also adjusted the roll and pitch to lay within the plane

used, this block runs anywhere from 2–5 Hz. Note that the detector block receives knowledge of π_p from the localiser block.

- *Solver* This block performs the estimation detailed in Sect. 3.3.1 and also requires knowledge of the current place, π_p , to load the associated landmarks for the pose solve. This block runs very fast but is limited by the rate at which the detections are given, and thus, runs at roughly the same rate as the detection block. Again, knowledge of the closest place is provided by the localiser.
- *Localiser* This is the high-level localiser that outputs vehicle pose relative to current place, π_p . The localiser block listens for high-rate odometry measurements that are used to predict the vehicle’s pose in between the low-rate localisation updates. As the localisation updates occur slower, we run a sliding-window posegraph relaxation technique.

The posegraph relaxation technique takes into account the following constraints: (i) relative constraints from odometry, (ii) sparse localisation constraints from the solver, and (iii) a prior constraint on the initial pose in the window. When localisation updates are not available, the system simply integrates the odometry for an estimate.

3.3.3 Projecting to SE(2)

The map, VO poses, localisation solver, and posegraph relaxation all operate in SE(3). However, we found the following “trick” proved useful in improving the localisation performance by preventing significant z-drift.

Since the vehicle is driving on a road, we know that at any point time it is reasonable to approximate the local vicinity as a plane. If we take the pose estimate in the map, we can project the pose down onto a local plane defined by the two closest keyframes. This is illustrated in Fig. 7. Essentially, we allow for the full 6 DOF pose, but then snap the solution down to SE(2). Again, this turned out to be a helpful trick to prevent slow drift in the z-direction.

4 Experiments and results

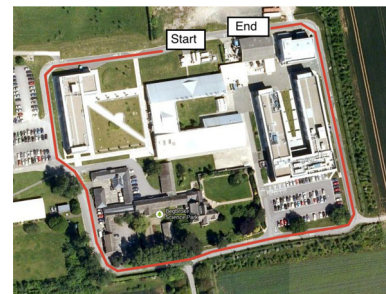
In this section we present two experiments. The first experiment analyses the matching performance of the scene signatures across extreme lighting and weather conditions (i.e., we focus purely on the front-end of the system). The second experiment looks at the performance of the complete end-to-end localisation system on kilometres of data collected from the Begbroke Science Park and central Oxford (see Fig. 8). Before proceeding with the experimental results, we first discuss the training and setup of the experiments.

4.1 Training and setup

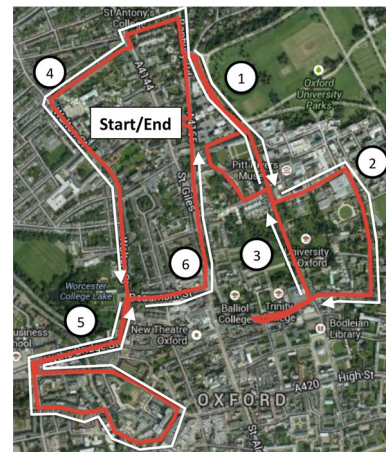
For Begbroke, we used 36 runs of a 650 m loop (see Fig. 8a), with places defined every 10 m along the respective routes according to our INS system. We note however, that places can be defined by other means, either manually or by using place recognition techniques. The only important factor is that the training images for a particular place have roughly the same viewpoint.

We trained our system with 31 datasets, which included images taken under different lighting and weather conditions. For our test data, we used five separate datasets with a wide range of visual variability, which included a sunny day run, an evening run, a rainy evening run, a snowy run, and a snowy and foggy run (example images are shown in Fig. 9). After training the scene signatures, we picked one reference run from the training data, which will be denoted as “the map,” and is the reference we localise against.

Similar to the Begbroke datasets, for Oxford, we defined places every 10 m along the 8 km route shown in Fig. 8b. Unfortunately, as our INS system is not reliable in urban environments, we were not able to automatically generate training data for certain sections of the route. We therefore only trained places on the 6 segments illustrated in the figure, which amounts to approximately 5.5 km. We used 15 datasets



(a) Begbroke route (approx. 650 m).



(b) Oxford route (approx. 8 km).

Fig. 8 Dataset routes used in our localisation experiments. Note that for the Oxford route (*right image*), we only report errors relative to each segment indicated with white arrows. This was a consequence of not having enough training data due to poor GPS measurements (recall that the training images are gathered from GPS-tagged surveys)

for training and 3 for testing. Example images of the map and live runs is shown in Fig. 10. Again, we wish to stress the selection of the live runs was hand-chosen to be the most challenging datasets for localisation, owing to their extreme differences in appearance from the map.

The learning phase took approximately 120 min per place, but were run as 5 separate processes, reducing the effective training time to 24 min. As each place is represented by a collection of SVM classifiers, the memory footprint is quite low at approximately 5 MB per place. This file size depends on the appearance of the scene (i.e., how many scene signatures are detected) and is therefore very environment dependent.

Since each place is separated by 10 m in our datasets, this is approximately 0.5 MB/m. If we were to train classifiers at a higher sampling density along the path, then it would be approximately $5/X$ MB/m, where X is the new separation distance between places. However, reducing the spacing between places is not necessarily advantageous. As mentioned earlier, the translational estimates are only accurate to within meters, so the system could incorrectly reference the

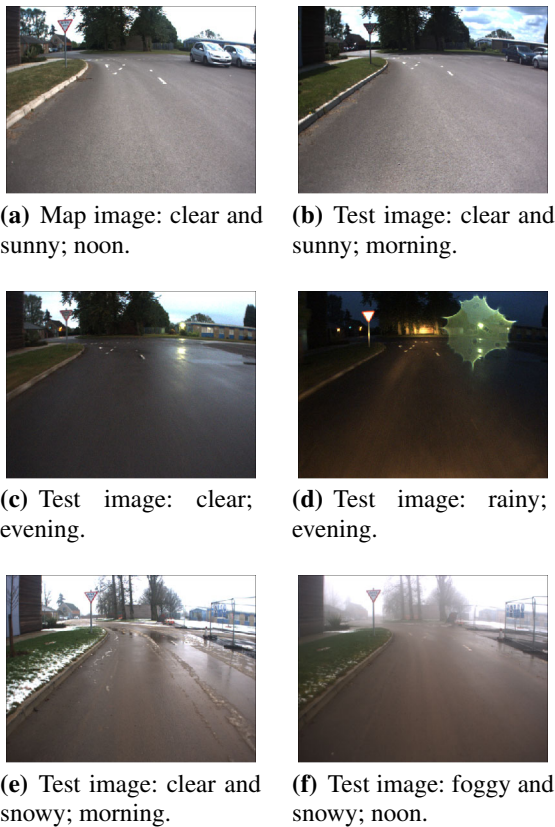


Fig. 9 Example test images used in our Begbroke localisation experiments. These were chosen due to their large visual variability

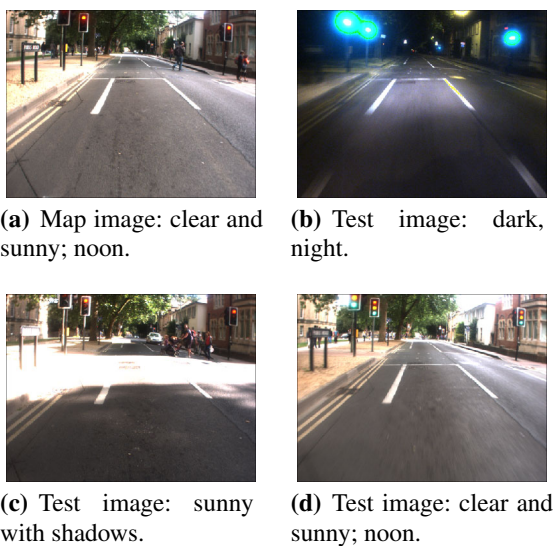


Fig. 10 Example test images used in our Oxford localisation experiments. These were chosen due to their large visual variability

wrong bank of classifiers if the places are too close together. Conversely, if they are separated too far apart, the appearance changes for each patch may be too significant to correctly

match. Thus, it would seem that a happy medium needs to be struck. We found a separation of 10m to work well in our experiments.

4.2 Feature matching experiments

The purpose of this section is to contrast the matching performance of the scene signatures against our point-feature system, in the absence of any geometric checks or motion-consistency checks that take place in the localisation pipeline. In other words, we wish to isolate the front-end of the system to see what matching potential is possible across extreme lighting and weather conditions.

For each INS-defined place in our training data, we took the corresponding groundtruth location of each test image so that the viewpoints of all images for every place are as similar as possible. By ensuring that the viewpoints of the test images and map image are well aligned, we can define a successful match as one in which the feature locations in both the live image and map image are in approximately the same location in image space. In other words, if we have a feature defined in the map image, \mathbf{y}_m^j , we would expect that the corresponding measurement in the live frame, \mathbf{y}_l^j , would be close in image space: $\|\mathbf{y}_m^j - \mathbf{y}_l^j\|_2 < \delta$, since we know that the transformation between the live and map frame is close to identity, by construction. In the following experiments, we defined $\delta = 15$ pixels. The same criteria applied to our point-feature system.

4.2.1 Begbroke

Figure 11 shows the number of feature matches per place around Begbroke for both our scene signatures and the point-feature system. As expected, we see that we are able to achieve correspondences across all appearance conditions with the scene signatures, but not with point features. In particular, point features fail to find enough matches for foggy and evening runs, due to motion blur, lack of texture, and environmental changes (e.g., snow on the road and buildings).

Figures 12 and 13 show heat maps of the locations in the image where matches are most likely to occur. To generate these heat maps, we added a count of +1 to each pixel contained within one of the matched shapes and averaged over the five test images for each place. Thus, we can compute an average distribution in image space for each place, as well as an average over the entire dataset. As expected, most of the matches occur in the far field, where we typically see distinctive structure, such as buildings and trees.

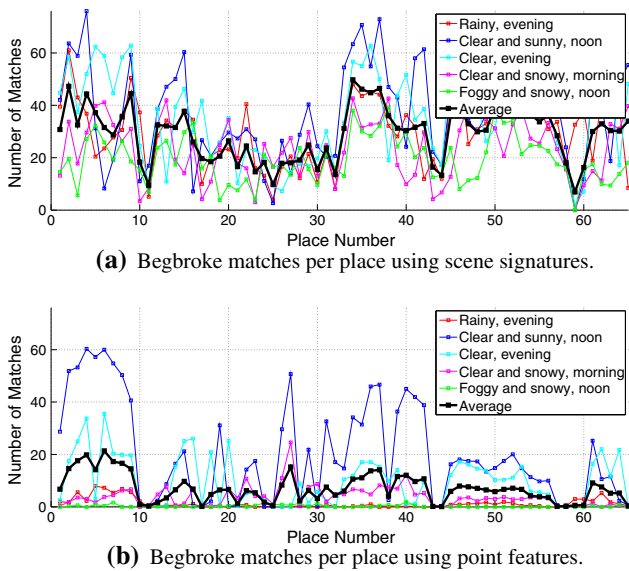


Fig. 11 Feature matches for each place in our Begbroke dataset using scene signatures (*top*) and point features (*bottom*). Places were defined as 10m segments along the reference trajectory shown in Fig. 8a. In this experiment, we used groundtruth aligned images at each place and performed feature matching against the map image for each respective live run. The results show that using scene signatures, we are able to match under all appearance conditions; not the case for the point-feature counterpart, which fails for the evening and foggy runs

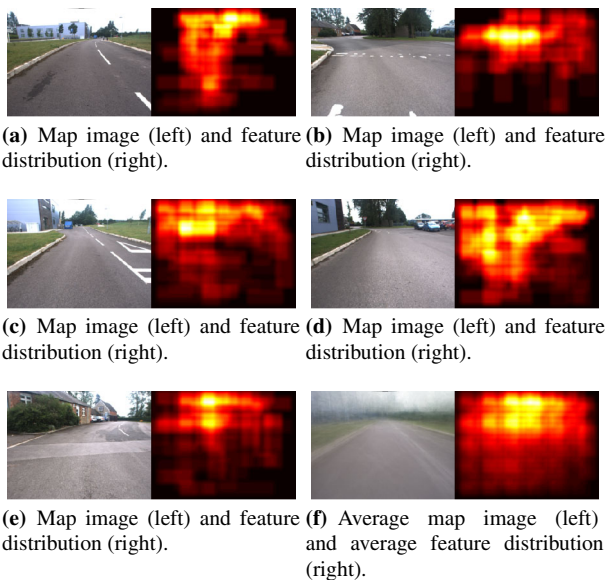


Fig. 12 Sample feature distributions per place, represented by heat maps. Each pair of images has the map image at a particular place (*left*) and the average heat map at that place (*right*), which was computed by adding a count of +1 for each pixel within a detection box and averaging over all detections. The *bottom right* figure shows the average place and average heat map over the entire dataset. Note that most of the detections are made on the upper half of the image, which is where we typically see distant buildings, trees, and distinctive structure

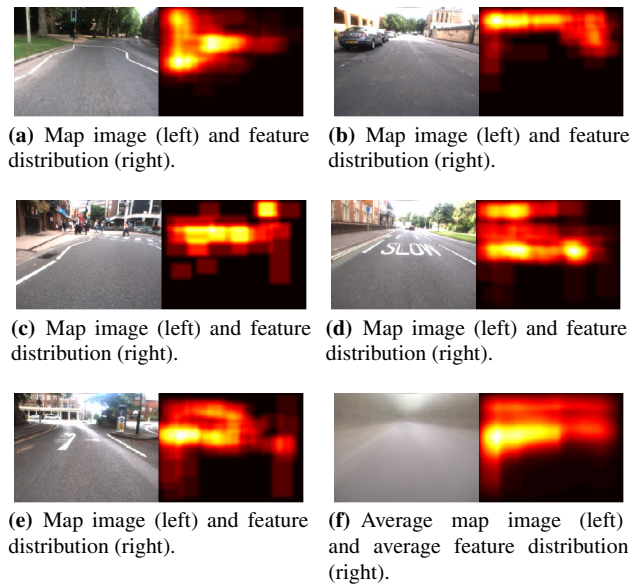


Fig. 13 Sample feature distributions per place, represented by heat maps. Each pair of images has the map image at a particular place (*left*) and the average heat map at that place (*right*), which was computed by adding a count of +1 for each pixel within a detection box and averaging over all detections. The *bottom right* figure shows the average place and average heat map over the entire dataset. These heat maps show more interesting structure than the Begbroke datasets, as we see density around traffic lights and road markings. This is the reason the average heat map has a bias towards the left side of the image

4.2.2 Oxford

Figure 14 shows the feature matches for the Oxford datasets using scene signatures and point features. It should be reiterated that not all the live runs followed the same trajectory as the map, so matches are only shown against the segments in common with the map (see Fig. 8b). Once again, the results show that using scene signatures is much more robust and we are able to match against all datasets for all places, while the point-feature approach fails over a number of places, especially for the nighttime dataset. This was the most challenging dataset as there is extreme motion blur and very little detail in the images.

4.3 Localisation experiments

In this section, we compare our weak-localiser approach to a point-feature-based system for the task of localisation. The goal of these experiments is to show that we can use scene signatures and a weak localiser to produce metric estimates even with some of the most challenging appearance changes. In order to ensure repeatable results, the processing was done offline to control the rate of localisation updates. We gave the system the opportunity to localise on every 4th image, which is equivalent to saying the localiser ran at 5 Hz for a 20 Hz camera feed.

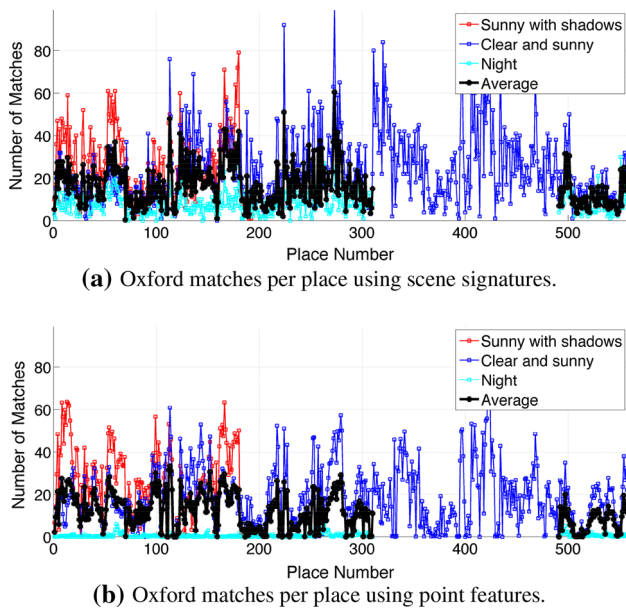


Fig. 14 Feature matches for each place in our Oxford dataset using scene signatures (*top*) and our point features (*bottom*). Note that gaps in the data represent regions where a live run did not intersect with the map. Failures are when the feature counts are at zero. Places were defined as 10m segments along the reference trajectory shown in Fig. 8b. In this experiment, we used groundtruth aligned images at each place and performed feature matching against the map image for each respective live run. The results show that using scene signatures, we are able to match under all appearance conditions, which is not the case for the point-feature counterpart, which fails for the nighttime run (the flat cyan line for the point feature approach) (Color figure online)

4.3.1 Begbroke

As our vehicle controller is only concerned with lateral and heading errors, we focus on these two metrics for assessing localisation performance. Figures 15, 16, 17, 18, 19 each show the following four plots for the 5 live runs: (i) lateral estimates, (ii) heading estimates, and (iii) speed estimates, and (iv) number of feature matches. The plots also show areas where our system was able to localise, represented by vertical red strips. Recall that our system localises at a rate of about 5 Hz and integrates odometry in between the updates, which results in small gaps between each localisation update. Some sample images of the feature matches have also been provided to give a visual interpretation of the system's performance.

A localisation failure was defined as travelling blind on odometry for more than 20 m (i.e., 2 segments in our case). If a failure occurred, we would reset the system using the INS. This criteria applied to both systems. Also note that simply seeding each respective system with the correct location in the map does not guarantee a localisation, due to significant appearance changes, which was the case with the point-feature system in most runs.

There are a number of interesting results from these plots. Firstly, one can see the scene-signature system works as well, if not better than our INS system and somewhat comparable to the point-feature system when it is working. As expected, the point-feature system was unable to localise in most of the cases where the appearance changes were drastic. Secondly, we see that there are two runs where the scene-signature system struggled: Figs. 16 and 17. This is most likely due to a lower number of feature matches during those runs and, more significantly, poor dead reckoning. Referring to the VO outputs for each run (Figs. 16a, 17c), we see that the two runs where the system struggled correspond to the two runs where the VO output was very noisy. Since the localisation system depends heavily on a strong motion prior, the runs where the motion priors were noisy most likely corresponded to suboptimal localisation estimates. To test this hypothesis, we swapped the VO output with the INS incremental transformations. These plots are shown in Figs. 20 and 21 and confirm the hypothesis. Although the average number of feature matches during those runs were lower due to the extreme appearance changes between the map image and the live run, we see that it was the poor dead reckoning estimates that contributed to the significant error.

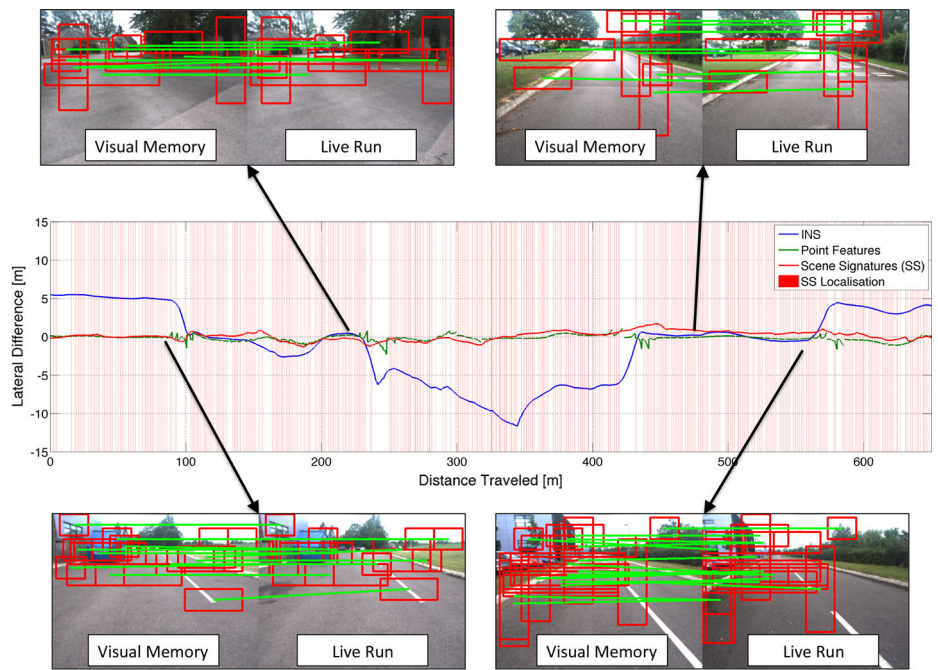
It is interesting to note that due to GPS drift, localisation estimates from one day to another with the INS system are only accurate to within meters; a common problem with groundtruthing localisation systems. The velocity estimates, however, are relatively smooth for a given run, as these are integrated from the IMU. The only time jumps are observed in the velocity estimates are when we lose GPS signal (e.g., under trees) and reacquire it at a later time. In these situations the INS attempts to reach a balance between the smooth estimates from the IMU and the apparent teleportation from the GPS.

4.3.2 Oxford

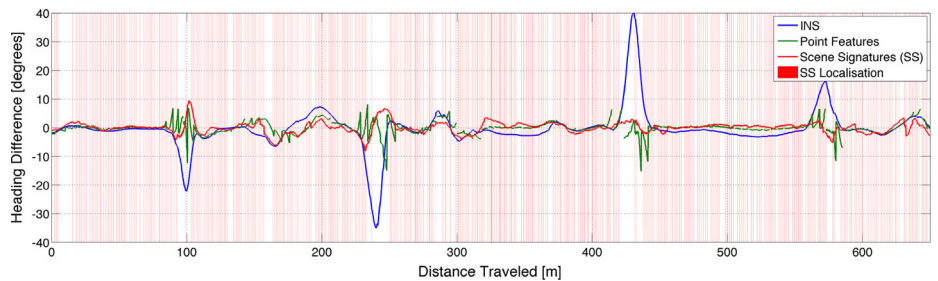
Figures 22, 23, and 24 show the lateral errors, heading errors, velocity profile, and number of feature matches for the three Oxford runs, which included a nighttime run, a sunny and shadowy run, and a clear sunny run. To reiterate, as each live run took a different path from the reference route, or because there were areas where we did not have training data, errors are only reported on the segments indicated in Fig. 8b.

The scene-signature system struggled during the nighttime run (Fig. 22) because of the extreme lack of any texture or detail in the images. This resulted in poor VO estimates and noisy feature matches. As a result, the system drifts in a number of locations, as indicated in the plots. We wish to stress that the point-feature system was unable to work under these conditions, further demonstrating the robustness of the scene-signature approach.

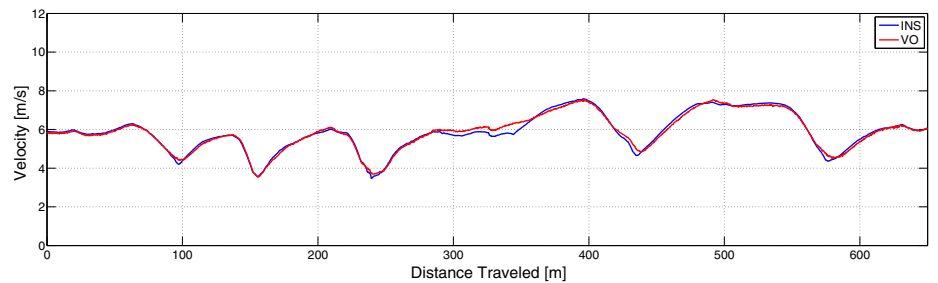
Fig. 15 Localisation results for the sunny afternoon run (Begbroke). The scene-signature system (*red*) performed comparably with the point-feature system (*green*) and outperformed the INS (*blue*), which drifted quite significantly from one dataset to the next. The *top* two plots represent the key error terms fed into our vehicle controller. The *vertical red lines* represent scene signature localisations (Color figure online)



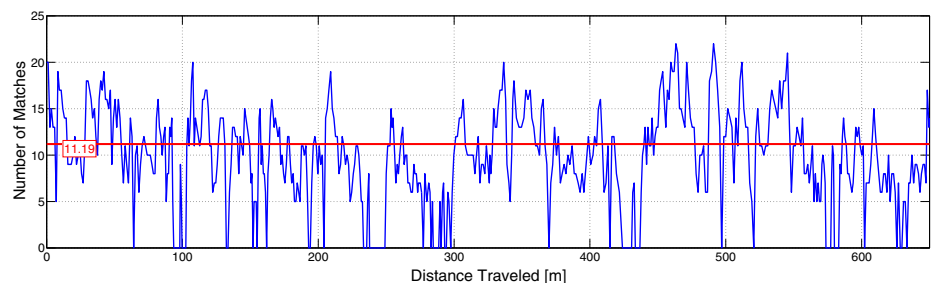
(a) Lateral estimates for a sunny visual memory vs. a sunny run.



(b) Heading estimates for a sunny visual memory vs. a sunny run

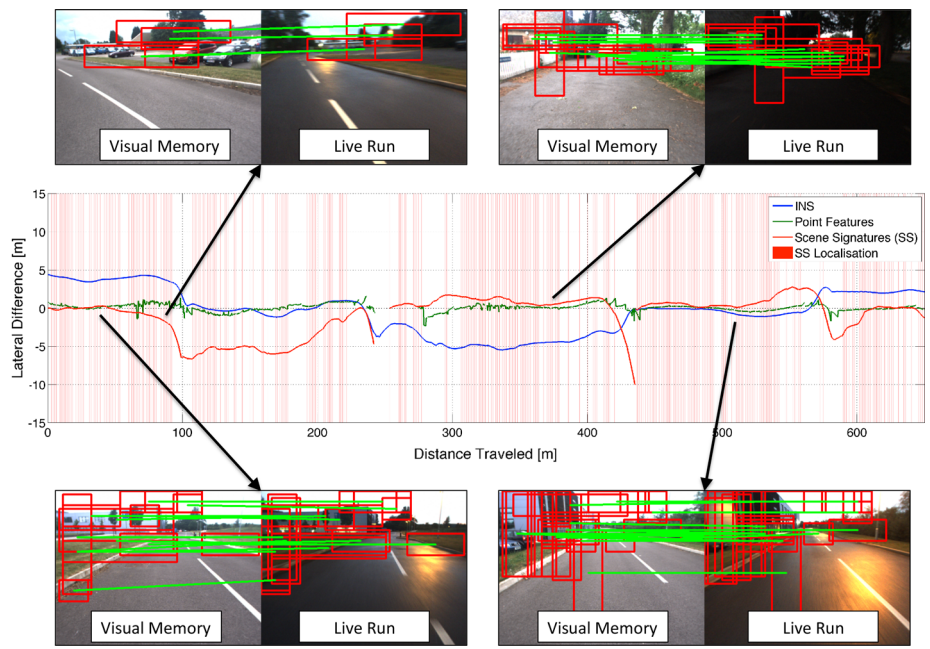


(c) Live VO profile against groundtruth.

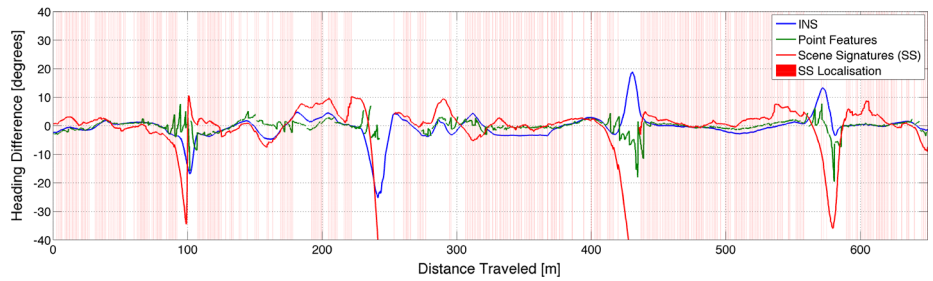


(d) Number of feature matches.

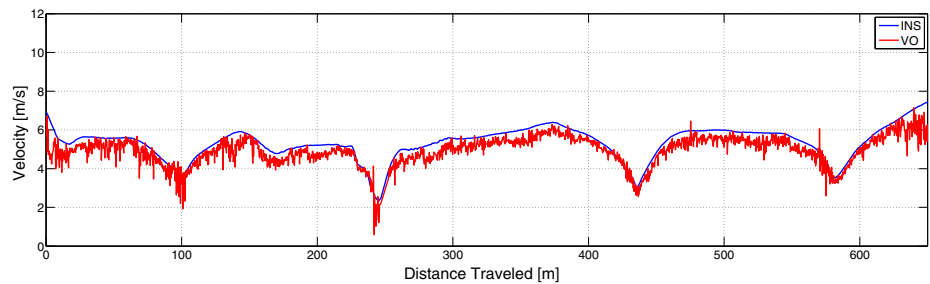
Fig. 16 Localisation results for the clear, evening run (Begbroke). The scene-signature system experienced greater lateral deviations in this run due to the noisy VO output (third subfigure from the top). Note that when the VO was substituted for the INS relative poses, the accuracy significantly improved (shown later in Fig. 20)



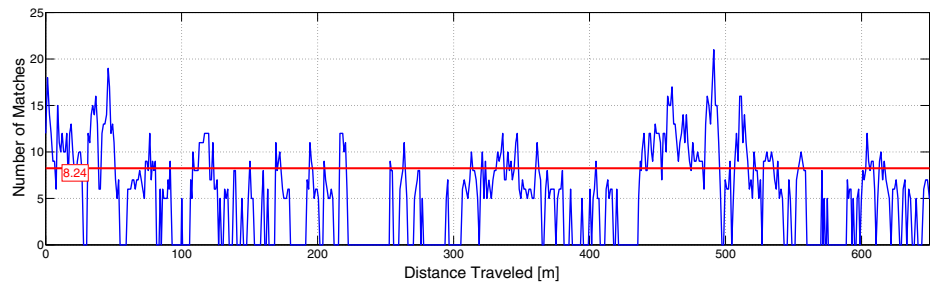
(a) Lateral estimates for a sunny visual memory vs. a clear, evening run.



(b) Heading estimates for a sunny visual memory vs. a clear, evening run

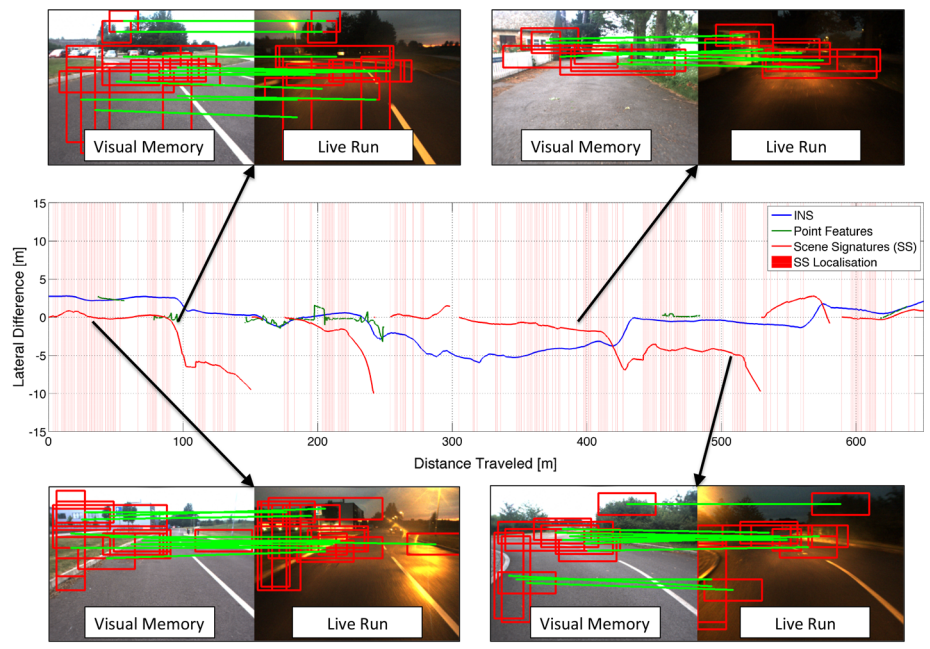


(c) Live VO profile against groundtruth.

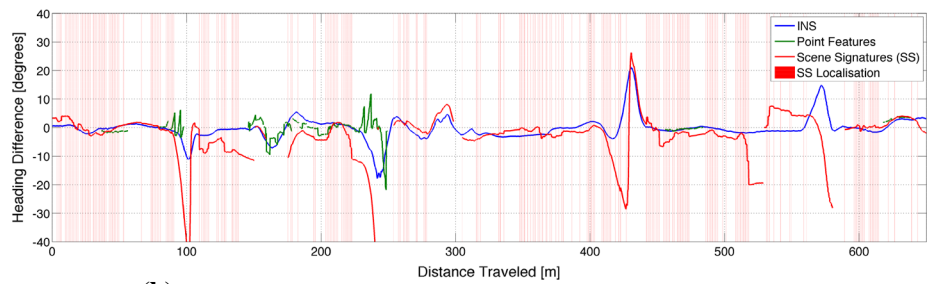


(d) Number of feature matches.

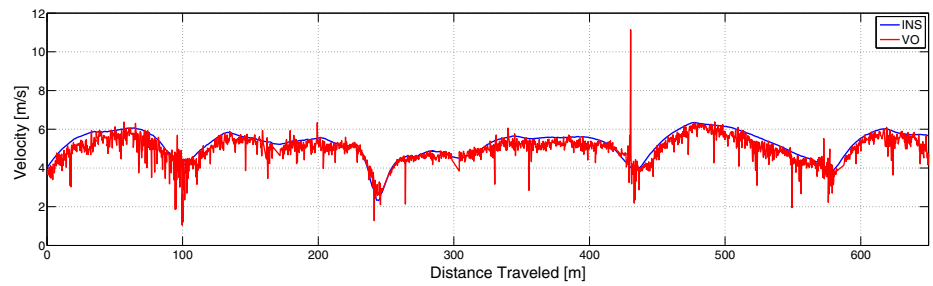
Fig. 17 Localisation results for the rainy, evening run (Begbroke). Similar to the other evening run, the VO output was very noisy and as a result, the localisation performance suffered. However, when the VO was substituted for the INS relative poses, we observed a significant improvement in accuracy (shown later in Fig. 21). We did, however, outperform the point-feature system (*green*), which was unable to cope with such extreme appearance changes (Color figure online)



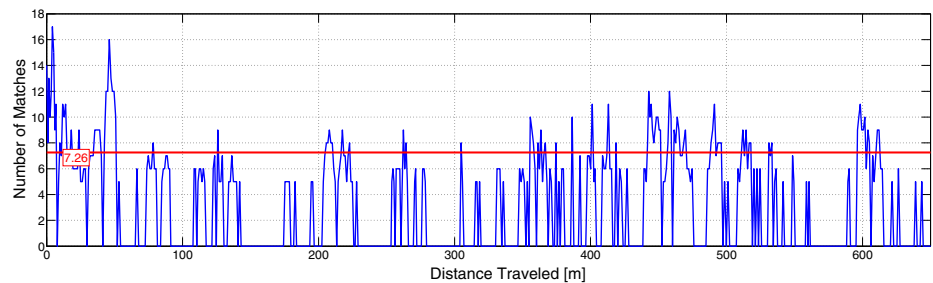
(a) Lateral estimates for a sunny visual memory vs. a rainy, evening run.



(b) Heading estimates for a sunny visual memory vs. a rainy, evening run

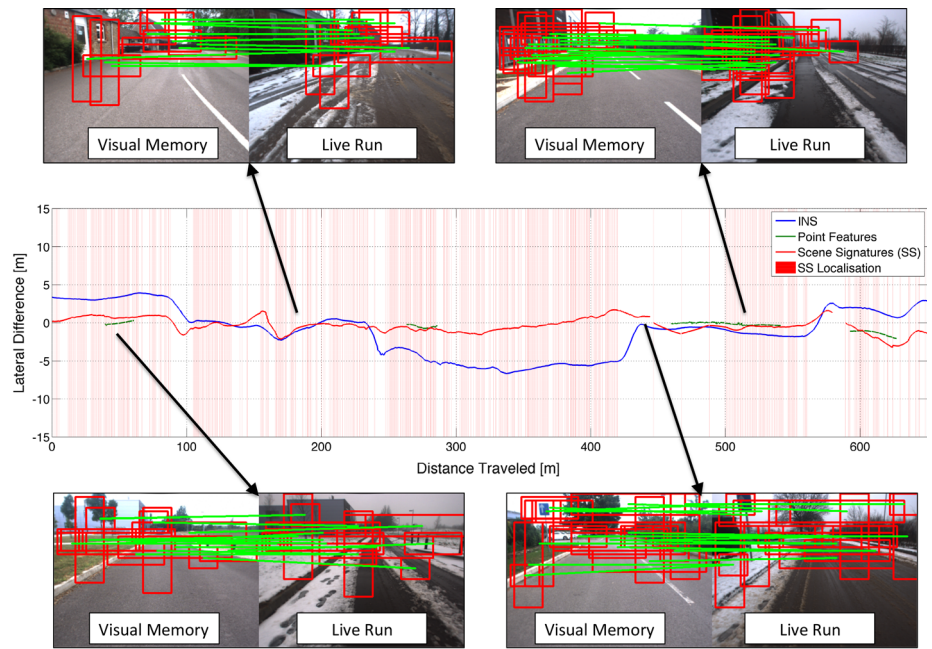


(c) Live VO profile against groundtruth.

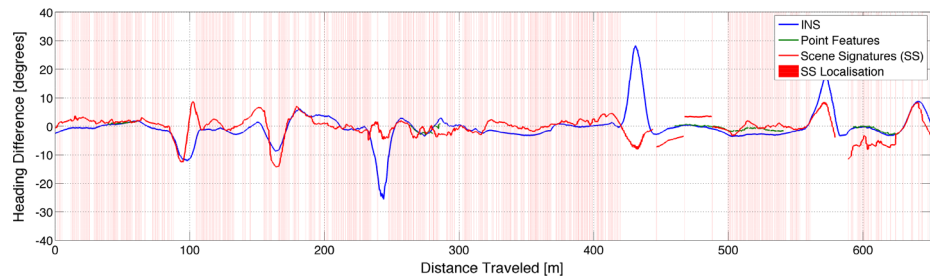


(d) Number of feature matches.

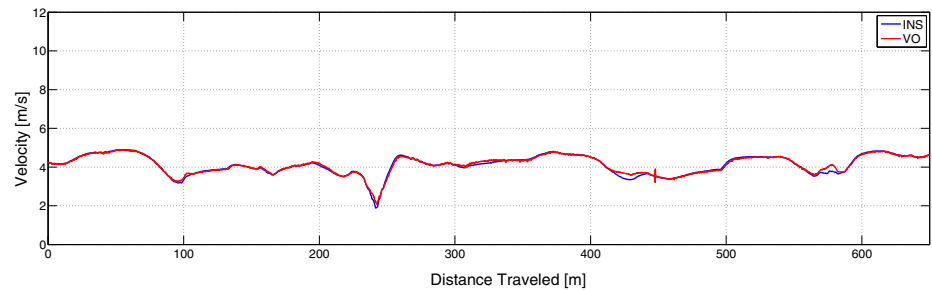
Fig. 18 Localisation results for the a clear, snowy run (Begbroke). As the VO output was quite good for this run, even with significant appearance differences to the map, the scene-signature system (*red*) was able to successfully localise, whereas the point-feature system failed over most of the run (*green*) (Color figure online)



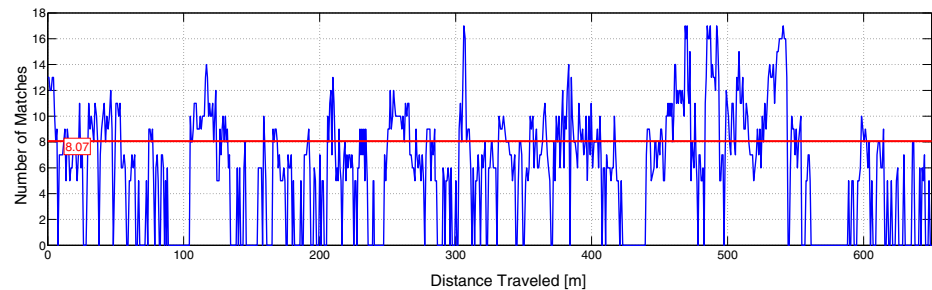
(a) Lateral estimates for a sunny visual memory vs. a clear, snowy run.



(b) Heading estimates for a sunny visual memory vs. a clear, snowy run

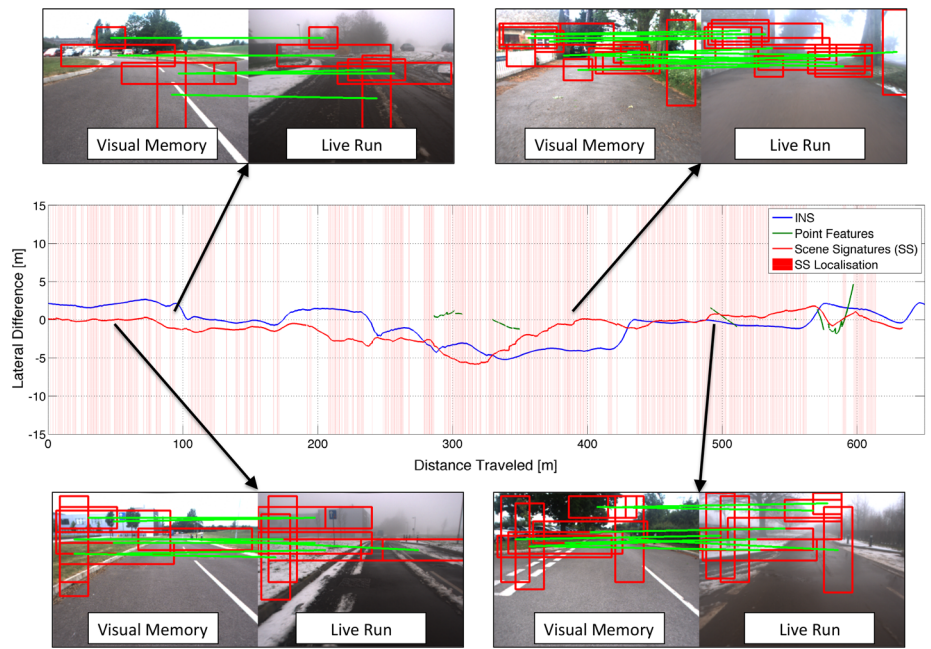


(c) Live VO profile against groundtruth.

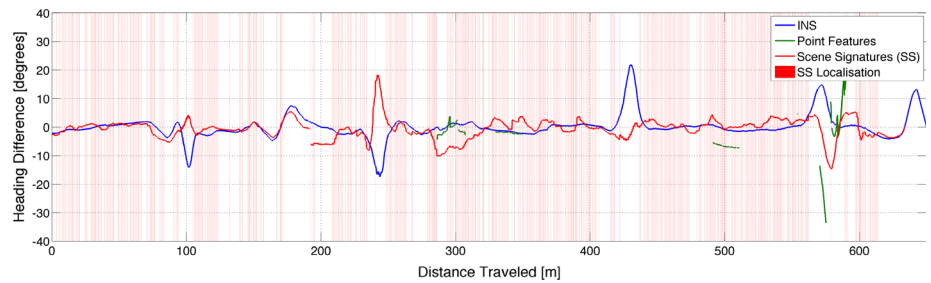


(d) Number of feature matches.

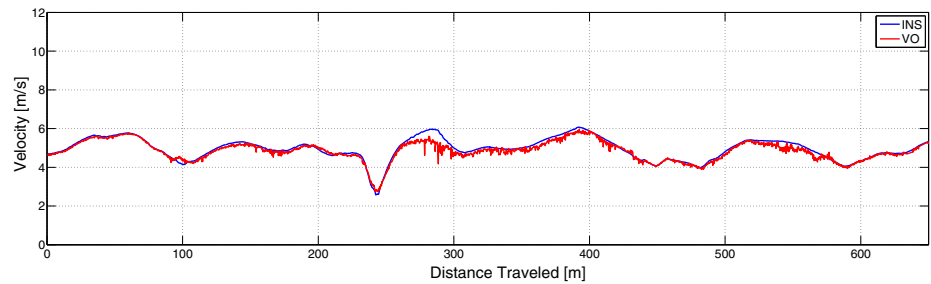
Fig. 19 Localisation results for the misty, snow run (Begbroke). Another example where the scene-signature system (*red*) was able to localise over the entire run despite significant appearance differences, while the point-feature system (*green*) (Color figure online)



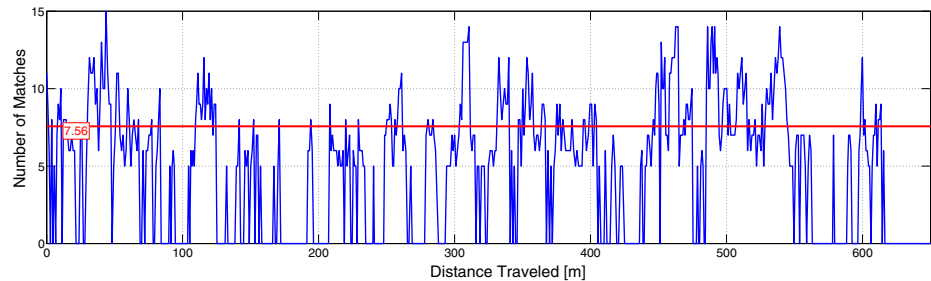
(a) Lateral estimates for a sunny visual memory vs. a misty, snow run.



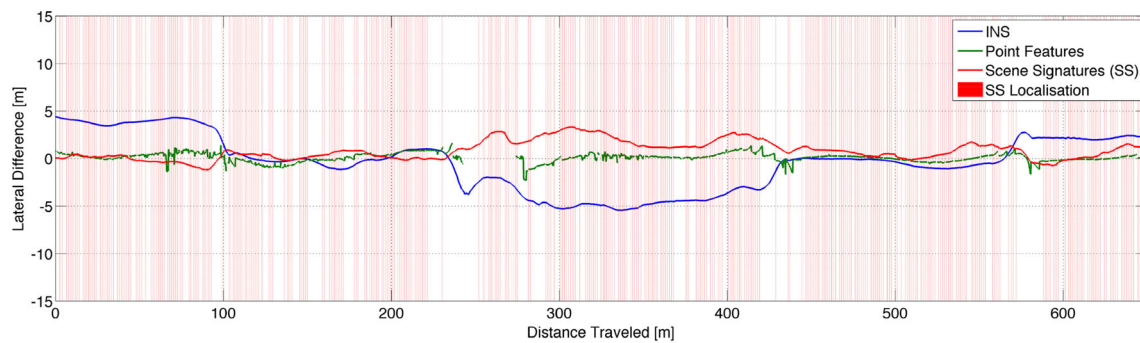
(b) Heading estimates for a sunny visual memory vs. a misty, snow run



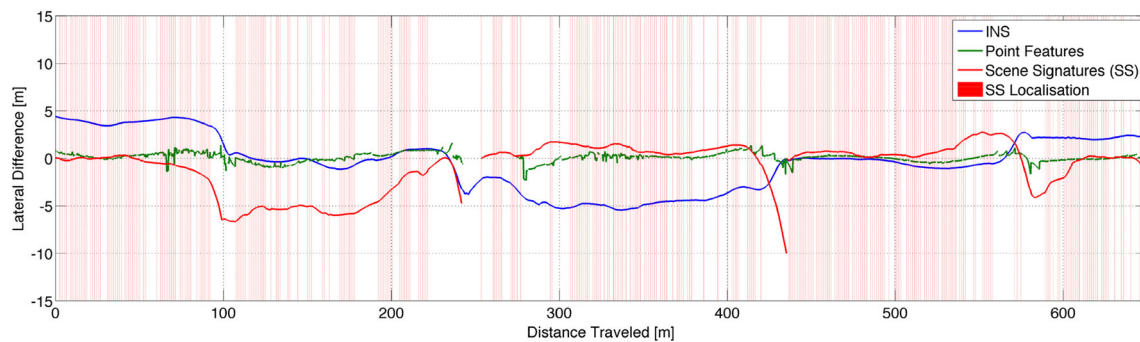
(c) Live VO profile against groundtruth.



(d) Number of feature matches.



(a) Lateral estimates for a sunny visual memory vs. a clear, evening run using the INS for dead reckoning.



(b) Lateral estimates for a sunny visual memory vs. a clear, evening run using VO for dead reckoning.

Fig. 20 Begbroke localisation results for the clear, evening run using the INS for dead reckoning (*top plot*). By replacing the noisy VO relative poses (*bottom plot*) for this run with the smoother INS measurements, we see that the scene-signature system is able to localise with a com-

parable accuracy to both systems. Again, this is due to the fact that the system relies on a strong motion prior for localisation, so if this motion prior is noisy, the estimates will suffer

The shadowy daytime run (Fig. 23) went well, producing estimates commensurate to the point-feature system and better than the INS. As mentioned earlier, the point-feature system works well when the appearance conditions are similar, meaning that it serves as better groundtruth for the like-against-like runs, owing to the unreliability of the INS. The clear daytime run (Fig. 24) shows similar performance again to the point-feature system.

4.4 Distance traveled on dead reckoning

The previous subsections presented a number results analysing the performance of the system over a variety of lighting, weather, and seasonal conditions. In order to provide the reader with a concise summary of the results for both Begbroke and Oxford, we turn our attention to Fig. 25, which shows the likelihood of traveling blind on odometry in between localisation updates.

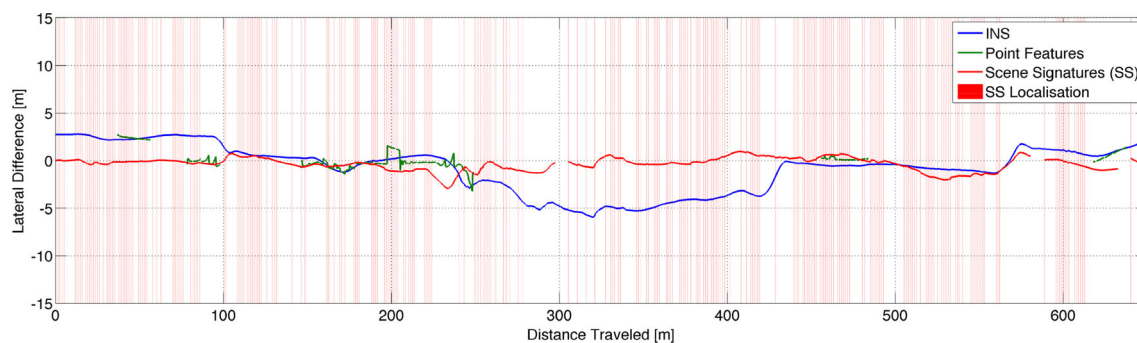
When the system fails to localise, we wish to know how far the vehicle is likely to travel on dead reckoning alone. This is a very interesting way of framing the performance of the system, as it implies that a localiser that fails frequently but

with short intervals between failures is preferable over one that fails infrequently but over large distances. The results show that the likelihood of traveling on dead reckoning is significantly less with our approach, which is an ideal characteristic of a localisation system and the key result of our experiments.

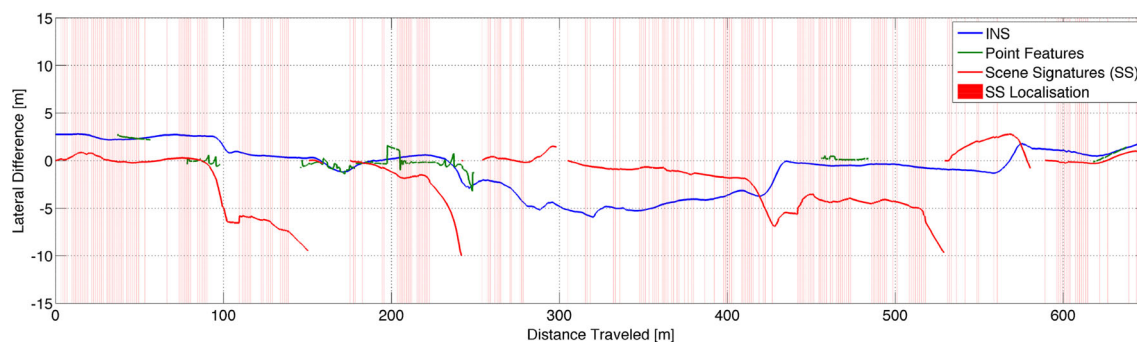
5 Discussion

Referring to the results in the previous section, we observed a significant increase in the robustness of the system for a slightly reduced metric accuracy when compared to our point-feature system. Having said this, we also saw that for sunny-against-sunny experiments, the scene signature approach did comparably to point features and better than the INS on average.

We also demonstrated the important role accurate dead reckoning has with regards to localisation performance. During the nighttime runs, the VO output suffered due to motion blur and low feature matching counts. However, when we swapped out the incremental transformations with the INS,



(a) Lateral estimates for a sunny visual memory vs. a rainy, evening run using the INS for dead reckoning.



(b) Lateral estimates for a sunny visual memory vs. a rainy, evening run using VO for dead reckoning.

Fig. 21 Begbroke localisation results for the rainy, evening run using the INS for dead reckoning (*top plot*). Results using VO poses are shown in the bottom plot. Another example where accurate dead reckoning improved the system's localisation performance despite drastic differences in appearance. The point-feature system was unable to cope and failed over a majority of the run

we saw that the localisation performance improved significantly. As our system is agnostic to the relative motion source, one could use wheel odometry and/or an IMU as opposed to VO. In our experiments however, we did not have access to this, which is why we used the VO output.

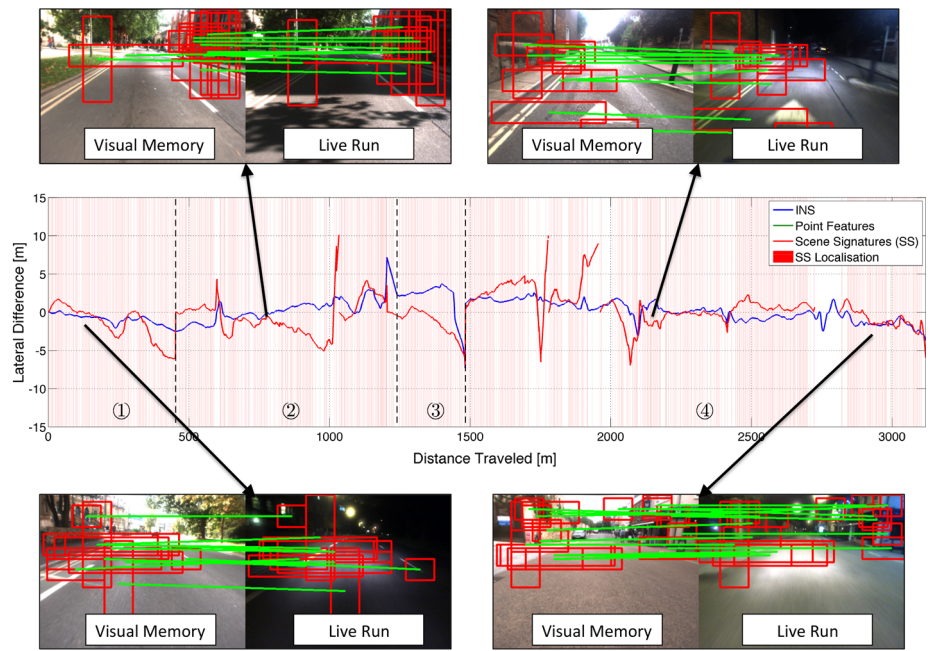
As we begin to move to more heavily cluttered urban environments, accounting for distractions (e.g., large busses) will likely be an important factor. In [McManus et al. \(2013\)](#), we presented a method for both stereo and monocular systems called distraction suppression, which leverages knowledge of prior 3D structure to generate distraction masks (see [Fig. 26](#)). We envision this type of system being critical in busy urban environments, as it will restrict the attention of our localiser to only search in areas of the image that belong to the static background.

In the future, we also wish to address the issue that most of the detected scene signatures represent distance visual elements, meaning that the translational estimate is not well constrained. The primary reason we tend to only identify far-field features is because the training data is not perfectly aligned due to variations in the driving for each data collection and the fact that we select images based on their GPS location, which can be off by several meters in some cases. Since our training algorithm works by assuming that

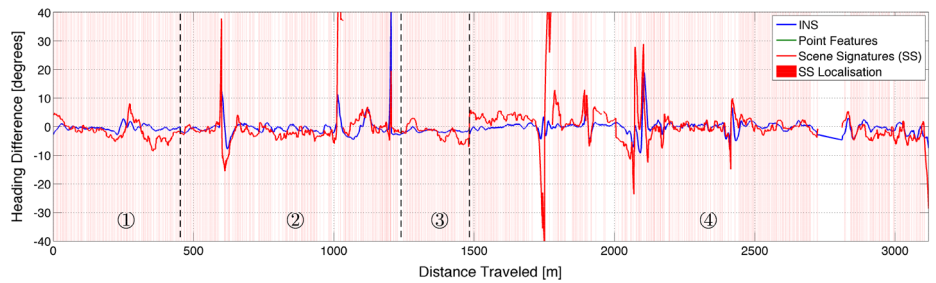
samples in the same location in image space represent the same visual element, the lateral/longitudinal offsets result in failures to converge on many near-field objects. The data association problem is difficult since the images are not perfectly aligned. One approach may be to use GPS as a seed for a sequence-based topological localiser to try and find better image matches for a given place. However, even if this is accomplished, it is unclear how we would be able to cope with the lateral offsets that result from variations in driving the same route. This is certainly an open research problem that we wish to pursue.

It is worth taking a moment to contrast the scene signature approach with other localisation system our group have developed, such as LAPS ([Stewart and Newman 2012](#)) and experience-based navigation (EBN) ([Churchill and Newman 2012](#)). LAPS is a monocular-based localisation method that leverages appearance of prior structure, in the form of coloured point clouds. The technique estimates the 6DOF pose of the vehicle by minimising the normalised information distance between the appearance of 3D lidar data reprojected into overlapping images. Results on 57 km of driving in central Oxford ([Stewart 2015](#)) demonstrate successful localisation over a variety of lighting conditions. However, as with the scene-signature localisation system, when LAPS became

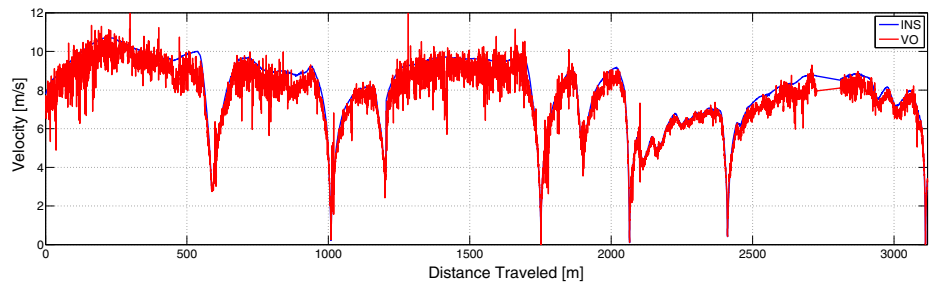
Fig. 22 Localisation results for the night run through segments 1–4 (Oxford). The localisation performance for this run was poor due to extremely low-light conditions and the lack of texture in the images. However, we note that the estimates seem commensurate with the INS and that the point-feature system failed on this run which is why there is no green line



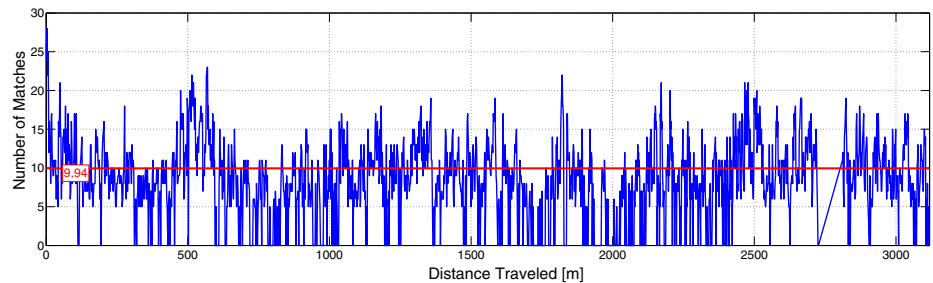
(a) Lateral estimates for a sunny visual memory vs. a dark, night run.



(b) Heading estimates for a sunny visual memory vs. a dark, night run.

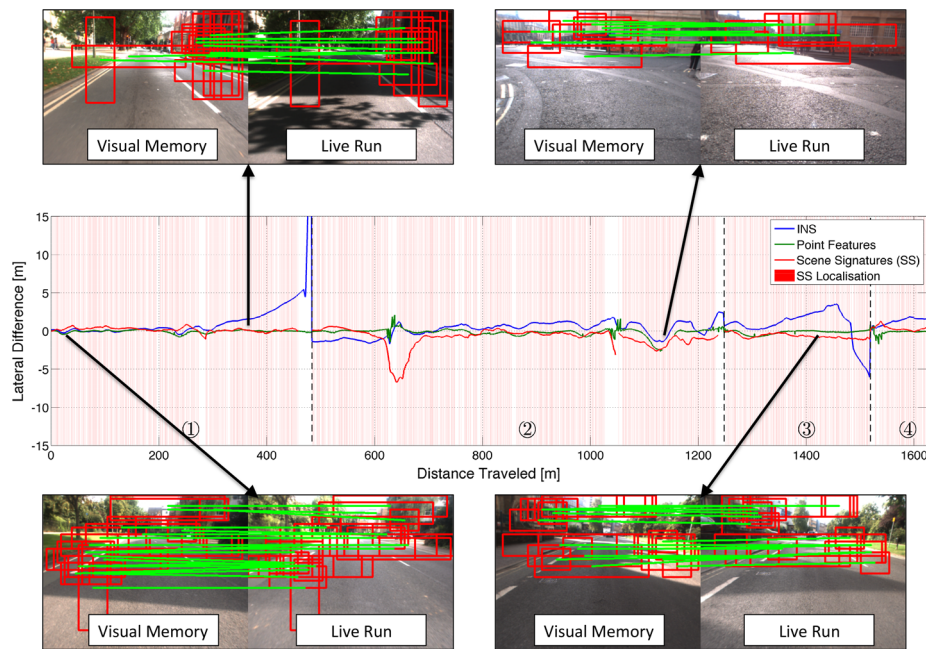


(c) Live VO profile against groundtruth.

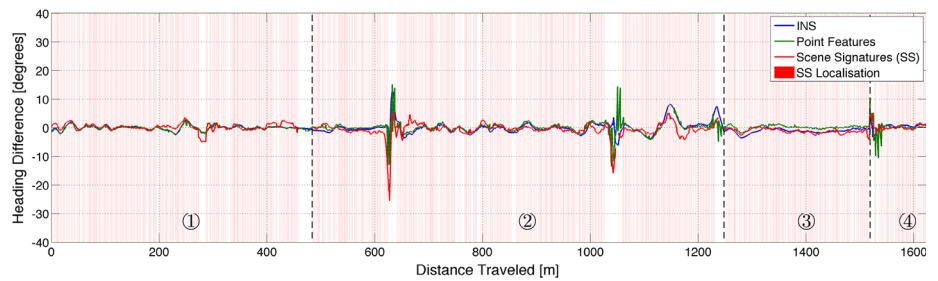


(d) Number of feature matches.

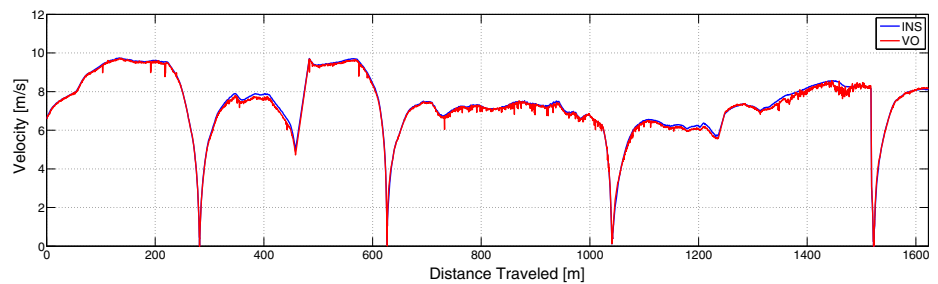
Fig. 23 Localisation results for a sunny, shadowy run through segments 1–4 (Oxford). The scene-signature system performed well on this run, with one dip around the 600 m mark, which corresponded to a localisation drift during a turn. The rest of the run followed the point-feature estimate closely, which is more trustworthy as groundtruth than the INS in these runs, since the conditions were similar



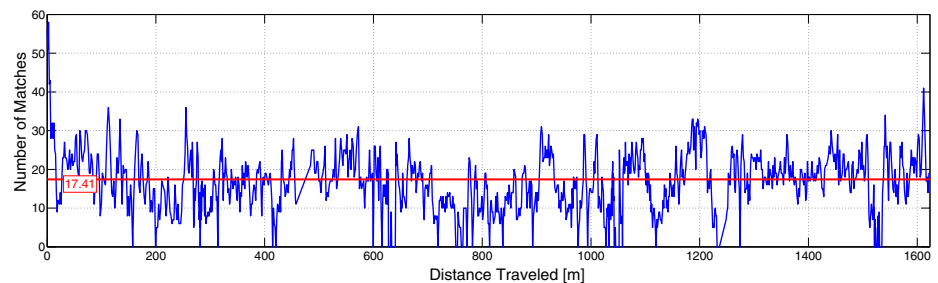
(a) Lateral estimates for a sunny visual memory vs. a sunny run.



(b) Heading estimates for a sunny visual memory vs. a sunny run.

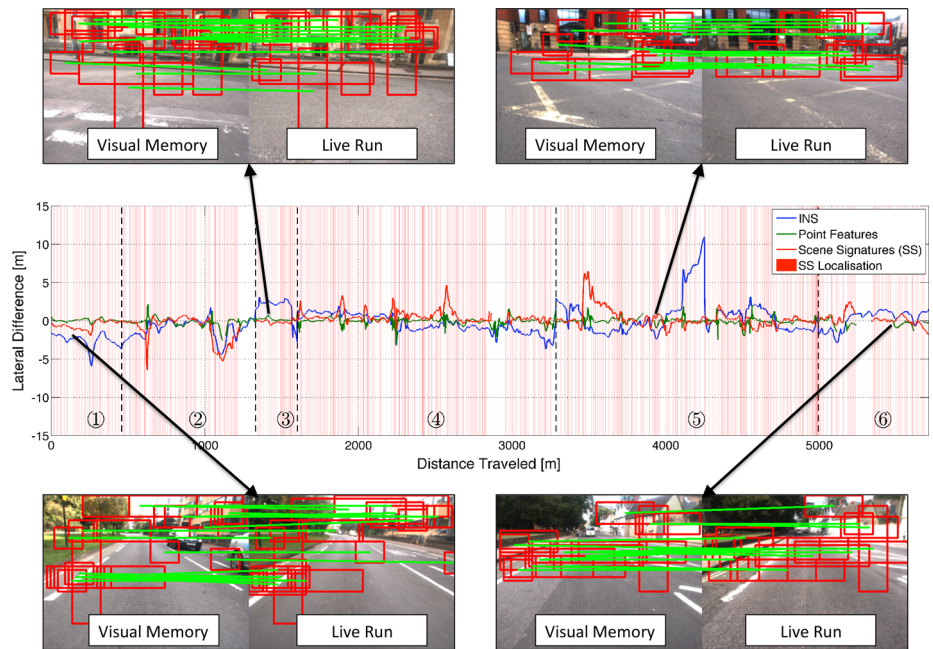


(c) Live VO profile against groundtruth.

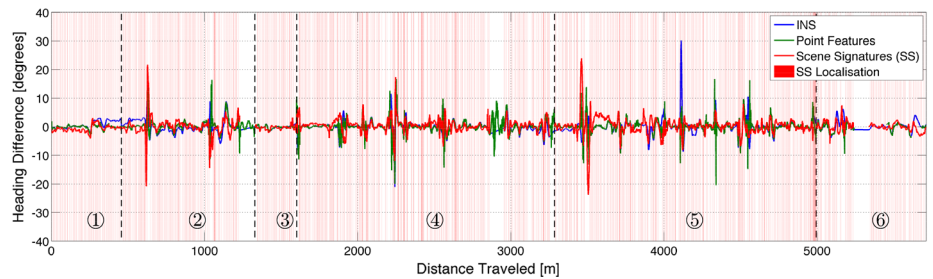


(d) Number of feature matches.

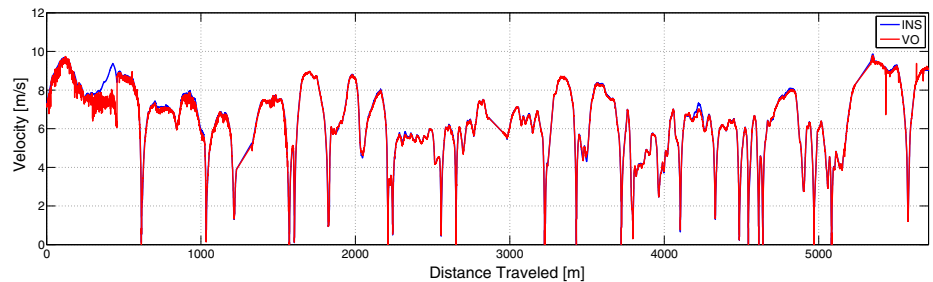
Fig. 24 Localisation results for a sunny run through segments 1–6 (Oxford). This was a longer run and again, we see that the scene-signature system performs as well, if not better than the INS, and commensurate in a number of locations with the point-feature system, which performs well because the appearance conditions are similar



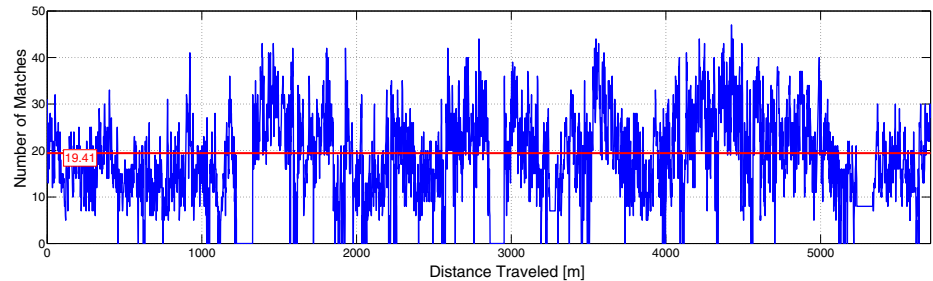
(a) Lateral estimates for a sunny visual memory vs. a sunny run.



(b) Heading estimates for a sunny visual memory vs. a sunny run.



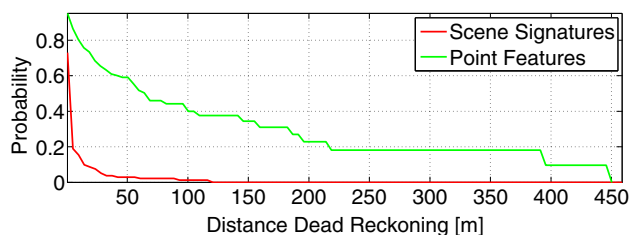
(c) Live VO profile against groundtruth.



(d) Number of feature matches.



(a) Begbroke results.



(b) Oxford results.

Fig. 25 When a localisation failure occurs, these plots show the likelihood of traveling blind on dead reckoning by distance. The ideal system would push the lines towards the bottom left-hand corner. These plots represent the key summary of our results, which is that scene signatures enable vastly superior robustness over point features



(a) Image taken in central London during the Olympics, where large parts of the scene are occupied by dynamic objects, which can distract and impede egomotion estimation.



(b) Using knowledge of prior 3D structure, we can generate probability masks that indicate which regions in the image are likely to belong to the static background (white).

Fig. 26 Example of a distraction mask (right image) that can be generated using the methods described in McManus et al. (2013). The *white regions* represent areas most likely to belong to the background, while the *black* represents areas that most likely belong to ephemeral objects of any class. Image credit: McManus et al. (2013)

lost, GPS resets were used to reinitialise the system. Again, this is not a unique problem to scene signatures or LAPS. Any localisation system that becomes lost will need a seed from an external system.

Although LAPS is fairly robust to appearance changes caused by different lighting conditions,⁶ it is not robust to appearance changes caused by different weather or seasons (i.e., when there is gross change in appearance). However, LAPS is able to produce more accurate localisation estimates,

⁶ Maddern et al. (2014) demonstrated improved robustness to LAPS by using an illumination-invariant colour space.

so as with the point-feature-based system, there is a tradeoff between accuracy and robustness with our system.

EBN presents a different approach to the long-term localisation problem. It is a more a framework that can be applied to any localisation system, such as LAPS, scene signatures, or a point-feature-based system, as was presented originally (Churchill and Newman 2012). The concept is to record distinct visual experiences of an environment, and continually grow this database over time to capture all the different visual modes. For example, consider a vehicle traversing its workspace for the first time and recording stereo image sequences of its environment. When the vehicle revisits the same environment, it can refer to the archived image sequences to attempt to localise. If the appearance of the scene has changed significantly and localisation is not possible, the system archives a new visual experience of that region until it is able to relocalise or it is finished its journey. In this way, the system only records experiences where the appearance of the world has changed enough as to prevent localisation. Over time, the system will hopefully saturate and will be able to represent all the different appearances of that environment. Recently, Linegar et al. (2015) examined the problem of knowing which subset of experiences to use online for localisation, taking into consideration finite computational resources.

The concept of EBN could also be applied to scene signatures. For example, imagine training scene signatures for nighttime experiences or daytime experiences, with the aim of discovering visual elements that are unique to night but perhaps not day and vice versa. However, issues surrounding the question of what to do when lost still remain (e.g., requiring an oracle topological localiser like GPS). EBN will record new visual experiences if lost, but during that time, it is using odometry and therefore will drift relative to the map.

To summarise, although EBN has been presented with a point-feature-based system, the concept of archiving only distinct visual experiences can be applied to any localisation system such as LAPS or scene signatures. Both LAPS and the point-feature-based approach aim to provide centimetre-level accuracy in pose estimates, but are unable to cope with gross appearance changes such as different weather or seasons. As we have shown in this paper, using scene signatures enables much more robust localisation performance over a traditional point-feature approach.

6 Conclusion

This paper presented a new approach to metric pose estimation in outdoor environments. We use training data in the form of image sequences of the same environment collected under varying appearance conditions and learn unique, place-

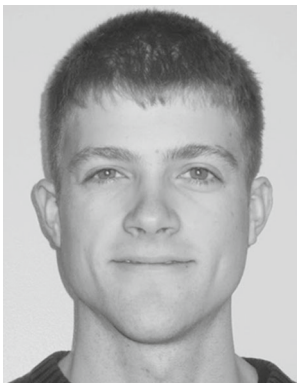
dependent feature detectors that fire on distinctive visual elements in the scene. These feature detectors, called scene signatures, enable feature matching across extreme appearance conditions. We presented a pipeline design that uses these scene signatures at runtime to perform coarse, metric localisation, with an accuracy commensurate with an expensive INS system. We believe that the idea of leveraging knowledge of prior data to learn what is important or distinctive for application-specific tasks is the way forward for reliable navigation systems.

Acknowledgments This work would not have been possible without the financial support from the Nissan Motor Company, the EPSRC Leadership Fellowship Grant (EP/J012017/1), and V-CHARGE (Grant Agreement Number 269916).

References

- Anati, R., Scaramuzza, D., Derpanis, K., & Daniilidis, K. (2012). Robot localization using soft object detection. In *Proceedings of the IEEE international conference on robotics and automation (ICRA)*, St. Paul.
- Atanasov, N., Zhu, M., Daniilidis, K., & Pappas, G. J. (2014). Semantic localisation via the matrix permanent. In *Proceedings of robotics science and systems (RSS)*, Berkeley.
- Bao, S. Y., & Savarese, S. (2011). Semantic structure from motion. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2025–2032.
- Bay, H., Ess, A., Tuytelaars, T., & Gool, L. (2008). Surf: Speeded up robust features. *Computer Vision and Image Understanding (CVIU)*, 110(3), 346–359.
- Castle, R. O., Gawley, D. J., Klein, G., & Murray, D. W. (2007). Towards simultaneous recognition, localization and mapping for hand-held and wearable cameras. In *Proceedings of the IEEE international conference on robotics and automation (ICRA)*, Rome.
- Churchill, W., & Newman, P. (2012). Practice makes perfect? Managing and leveraging visual experiences for lifelong navigation. In *Proceedings of the international conference on robotics and automation*, Saint Paul.
- Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the conference on computer vision and pattern recognition* (pp. 886–893), San Diego.
- Davison, A., & Murray, D. (2002). Simultaneous localization and map-building using active vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7), 865–880.
- Davison, A., Reid, I., Motlon, N., & Stasse, O. (2007). Monoslam: Real-time single camera slam. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6), 1052–1067.
- Doersch, C., Singh, S., Gupta, A., Sivic, J., & Efros, A. (2012). What makes paris look like Paris? *ACM Transactions on Graphics*, 31(4), 101.
- Furgale, P., & Barfoot, T. (2001). Visual teach and repeat for long-range rover autonomy. *Journal of Field Robotics, Special Issue on "Visual Mapping and Navigation Outdoors"*, 27(5), 534–560.
- Hartley, R., & Zisserman, A. (2004). *Multiple view geometry in computer vision* (2nd ed.). Cambridge: Cambridge University Press, ISBN: 0521540518.
- Johns, E., & Yang, G.-Z. (2013). Feature co-occurrence maps: Appearance-based localisation throughout the day. In *Proceedings of the international conference on robotics and automation*.
- Kaess, M., Johannson, H., Roberts, R., Ila, V., Leonard, J., & Dellaert, F. (2012). isam2: Incremental smoothing and mapping using the bayes tree. *International Journal of Robotics Research*, 31(2), 216–235.
- Ko, D. W., Yi, C., & Suh, I. H. (2013). Semantic mapping and navigation: A bayesian approach. In *Proceedings of the IEEE/RSJ international conference on intelligent robotics and systems (IROS)*, pp. 2630–2636.
- Konolige, K., Bowman, J., Chen, J., Mihelich, P., Calonder, M., Lepetit, V., et al. (2010). View-based maps. *The International Journal of Robotics Research*, 29(8), 941–957.
- Lategahn, H., Beck, J., Kitt, B., & Stiller, C. (2013). *How to learn an illumination robust image feature for place recognition*. IEEE intelligent vehicles symposium, Gold Coast.
- Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. *The Quarterly of Applied Mathematics*, 2, 164–168.
- Linegar, C., Churchill, W., & Newman, P. (2015). Work smart, not hard: Recalling relevant experiences for vast-scale but time-constrained localisation. In *IEEE international conference on robotics and automation (ICRA)*, Seattle.
- Lowe, D. (2004). Distinctive image features from scale-invariant key points. *International Journal of Computer Vision*, 60(2), 91–110.
- Maddern, W., Stewart, A., McManus, C., Upcroft, B., Churchill, W., & Newman, P. (2014). Illumination invariant imaging: Applications in robust vision-based localisation, mapping and classification for autonomous vehicles. In *Proceedings of the visual place recognition in changing environments workshop, IEEE international conference on robotics and automation*, Hong Kong.
- McKinnon, D., Smith, R., & Upcroft, B. (2012). A semi-local method for iterative depth-map refinement. In *Proceedings of the IEEE international conference on robotics and automation (ICRA)*.
- McManus, C. (2010). The unscented kalman filter for state estimation. *Presented at the simultaneous localization and mapping (SLAM) workshop, 7th Canadian conference on computer vision (CRV)*.
- McManus, C., Churchill, W., Maddern, W., Stewart, A., & Newman, P. (2014a). Shady dealings: Robust, long-term visual localisation using illumination invariance. In *Proceedings of the IEEE international conference on robotics and automation (ICRA)*, Hong Kong.
- McManus, C., Churchill, W., Napier, A., Davis, B., & Newman, P. (2013). Distraction suppression for vision-based pose estimation at city scales. In *Proceedings of the IEEE international conference on robotics and automation*, Karlsruhe.
- McManus, C., Upcroft, B., & Newman, P. (2014b). Scene signatures: Localised and point-less features for localisation. In *Proceedings of robotics science and systems (RSS)*, Berkeley.
- Milford, M. (2013). Vision-based place recognition: How low can you go? *The International Journal of Robotics Research*, 32(7), 766–789.
- Milford, M., & Wyeth, G. (2012). Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights. In *Proceedings of the IEEE international conference on robotics and automation (ICRA)*, Saint Paul.
- Naseer, T., Spinello, L., Burgard, W., & Stachniss, C. (2014). Robust visual robot localization across seasons using network flows. In *AAAI conference on artificial intelligence (AAAI)*, Quebec.
- Neubert, P., Sunderhauf, N., & Protzel, P. (2013). Appearance change prediction for long-term navigation across seasons. In *European Conference on mobile robotics (ECMR)*.
- Piniés, P., Paz, L. M., Gálvez-López, D., & Tardós, J. D. (2010). Ci-graph simultaneous localisation and mappin for three-dimensional reconstruction of large and complex environments using a multi-camera system. *Journal of Field Robotics*, 27(5), 561–586.

- Ranaganathan, A., Matsumoto, S., & Ilstrup, D. (2013). Towards illumination invariance for visual localization. *Proceedings of the IEEE international conference on robotics and automation (ICRA)* (pp. 3791–3798), Karlsruhe.
- Renato F. Salas-Moreno, Richard A. Newcombe, H. S. P. H. J. K. & Davison, A. J. (2013). Slam++: Simultaneous localisation and mapping at the level of object. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.
- Richardson, A. & Olson, E. (2013). Learning convolutional filters for interest point detection. In *Proceedings of the IEEE international conference on robotics and automation (ICRA)*.
- Sibley, G., Mei, C., Reid, I., & Newman, P. (2010). Vast-scale outdoor navigation using adaptive relative bundle adjustment. *The International Journal of Robotics Research*, 29(8), 958–980.
- Singh, S., Gupta, A., & Efros, A. A. (2012). Unsupervised discovery of mid-level discriminative patches. In *Proceedings of the European conference on computer vision (ECCV)*.
- Stewart, A. & Newman, P. (2012). Laps - localisation using appearance of prior structure: 6-dof monocular camera localisation using prior pointclouds. In *Proceedings of the international conference on robotics and automation*, Saint Paul.
- Stewart, A. D. (2015). *Localisation using the appearance of prior structure*. PhD thesis, University of Oxford.
- Valgren, C., & Lilienthal, A. (2007). Sift, surf & seasons: Long-term outdoor localization using local features. In *Proceedings of the 3rd European conference on mobile robotics (ECMR)*.
- Valgren, C., & Lilienthal, A. (2010). Sift, surf and seasons: Appearance-based long-term localization in outdoor environments. *Robotics and Autonomous Systems*, 58(2), 149–156.
- Yi, C., Suh, I. H., Lim, G. H., & Choi, B.-U. (2009). Active-semantic localization with a single consumer-grade camera. In *Proceedings of the IEEE international conference on systems, man and cybernetics (SMC)*, pp. 2161–2166.



Colin McManus holds a B.A.Sc. (Hons., 2009) in Engineering Science (Aerospace Option) from the University of Toronto, a M.A.Sc (2011) in Aerospace Engineering from the University of Toronto Institute for Aerospace Studies, and is currently a D.Phil. student working with the Mobile Robotics Group at the University of Oxford.



Ben Uproft holds a B.Sc. (1999) and Ph.D. (2004) in Physics from the University of Queensland. He was a post-doctoral fellow (2004–2007) researching Robotics at the University of Sydney. He was a senior lecturer at the University of Queensland from 2007–2011 and is currently a senior lecturer and member of the CyPhy Lab at the Queensland University of Technology.



Paul Newman holds a M.E. (1995) in Engineering Science from University of Oxford, and a Ph.D. (1999) in Autonomous Navigation at the Australian Center for Field Robotics, University of Sydney. He was a post-doctoral fellow and later research scientist with the Department of Ocean Engineering at M.I.T. from 2000–2003. Currently, he is the BP Professor of Information Engineering at the University of Oxford and an EPSRC Leadership Fellow. He heads the Mobile Robotics Group within the Department of Engineering Science.