Scene Signatures: Localised and Point-less Features for Localisation

Colin McManus[1], Ben Upcroft[2], and Paul Newman[1]

***Abstract*—This paper is about localising across extreme lighting and weather conditions. We depart from the traditional point-feature-based approach since matching under dramatic appearance changes is a brittle and hard. Point-feature detectors are rigid procedures which pass over an image examining small, low-level structure such as corners or blobs. They apply the same criteria to all images of all places. This paper takes a contrary view and asks what is possible if instead we learn a bespoke detector for every place. Our localisation task then turns into curating a large bank of spatially indexed detectors and we show that this yields vastly superior performance in terms of robustness in exchange for a reduced but tolerable metric precision. We present an unsupervised system that produces broad-region detectors for distinctive visual elements, called *scene signatures*, which can be associated across almost all appearance changes. We show, using 21 km of data collected over a period of 3 months, that our system is capable of producing metric estimates from night-to-day or summer-to-winter conditions.**

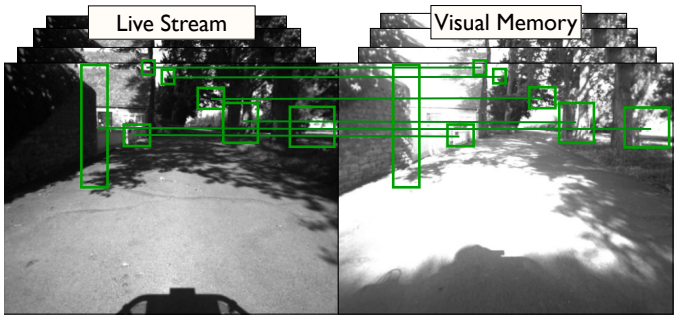
I. Introduction

Matching point features between different images is the standard approach in visual motion estimation and has led to a number of impressive systems for pose estimation and/or mapping over large scales (e.g., Visual Teach & Repeat (VT&R) [1], and Visual Simultaneous Localisation And Mapping (VSLAM) [2]). Matching low-level features such as edges, blobs, or corners works well when observing the same scene under similar conditions (e.g., for online ego-motion estimation). However, when trying to match images taken at different times of day or in different seasons, these low-level features often look utterly different. Larger image structures on the other hand, such as windows, signs, or doors, offer more hope as they capture shape and texture on a broader scale. We will show that if we are careful about how we identify suitable structures – which we shall refer to as *scene signatures* – then they can be reliably matched under large variations in appearance, thus opening the door to robust localisation.

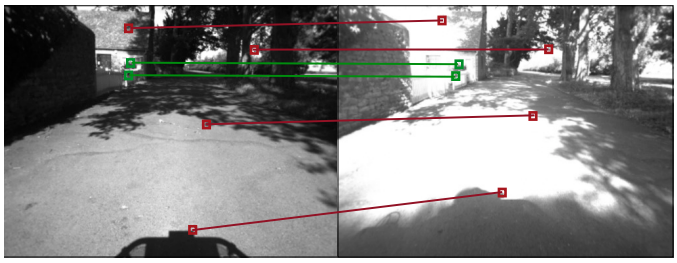
Low-level point features can be thought of as being on one extreme of a localisation paradigm, while using whole-image information, such as SeqSLAM [3], can be thought of as being on the opposite end of that spectrum. However, it should be noted that SeqSLAM just provides an estimate on the topological location of the vehicle and not a metric pose estimate. If we wish to provide a vehicle controller with a metric pose estimate, we need something in between these two approaches.

[1]Mobile Robotics Group, University of Oxford, Oxford, England; {colin,pnewman}@robots.ox.ac.uk

[2]Robotics and Aerospace Systems, Queensland University of Technology, Brisbane, Australia; ben.upcroft@qut.edu.au



(a) By matching *scene signatures* from a live stream (left) to a memory (right), we are able to *successfully* localise our vehicle.



(b) By matching point features from a live stream (left) to a memory (right), we are unable to successfully localise our vehicle.

Fig. 1. An illustration of the benefits of matching scene signatures, which are distinctive visual elements such as fences, windows, tree lines, etc., versus the traditional point-feature approach. Using point features for data association under extreme appearance changes often fails because point features only consider low-level structure, like edges, corners, or blobs. Scene signatures are more robust since they are large, distinctive elements.

In this paper, we present an unsupervised approach to find distinctive visual elements, such as windows, signs, doors, or tree silhouettes (see Figure 1), in a given place, $\boldsymbol{\pi}_p$, which is a node in a hybrid topological/metric map. These are distinctive signatures specific to the scene, and so we refer to them as *scene signatures*. We wish to stress that the benefit of using scene signatures over point features is that we can associate these scene signatures across extreme appearance changes, such as night to day or sunny to winter.

We shall constrain ourselves to the task of teaching a vehicle, for example an autonomous car, to localise using vision. We assume that the vehicle has or will be driven through the environment on multiple occasions and so we have many examples of the appearance of the places the vehicle drives through. Rather than building a map of point features against which to match point features detected at run time, we will construct, in an unsupervised way, a large

set of spatially indexed classifiers, which are associated with topological locations in the world. Each of these classifiers is carefully constructed to fire on a particular and distinctive aspect of the environment at that particular place, $\boldsymbol{\pi}_p$. As the vehicle progresses through its environment, we will retrieve the classifiers, $\{c_i\}_p$, relevant to its location, $\boldsymbol{\pi}_p$, and use them to identify known structure in the live image feed. These broad level features are used to create a "weak localiser" of sufficient accuracy to provide coarse local, metric information about the vehicle's pose.

Immediately we should ask, "for what tasks is such precision adequate?" We envision a hierarchical system in which at the top level we have very crude topological localiser which outputs the gross location of the vehicle. This output drives the localiser described in this work which takes a topological hint and returns a metric position accurate in orientation but with perhaps tens of centimeters in translational error. We assert that for autonomous road-vehicle navigation and control, we only need a coarse metric estimate of the vehicle's pose, after which, lower level lane following and/or curb detection algorithms can be applied to refine the estimate for a vehicle controller. This is a shift from the traditional methods that try and obtain centimeter-level accuracy. For a road vehicle with on-board obstacle avoidance and lane following software, global localisation accuracy to the half metre is sufficient.

The novel contributions of this paper are the following: (i) the introduction of "weak localisers," which use scene signatures to perform metric estimation, (ii) an unsupervised method which finds distinctive scene signatures, and (iii) the validation on challenging datasets displaying extreme appearance changes, from full light to deep darkness.

II. BACKGROUND

Decades of work have been focused on designing interest-point detectors and descriptors that can identify repeatable features and describe them using unique, compact representations. The output of these systems has enabled efficient feature correspondence across images taken at different viewpoints. Popular corner detectors include Harris Corners [4] and FAST [5], while blob detection can be performed with the Laplacian of Gaussian or MSER [6]. A range of image-point descriptors also exist, for example SIFT [7], SURF [8], BRIEF [9], and ORB [10] to name a few. However, all of these interest-point detectors/descriptors operate on small image patches, which can look entirely different under different lighting and/or weather conditions. We will show that scene signatures enable matching across extreme changes in appearance because they contain large, distinctive elements in the image. Note that our approach is very different from the localisation and mapping systems of Davison et al. [11], [12], which use image patches as their landmarks. These methods still rely on interest-point detection to find the patches and they use small patches (e.g., 11×11 pixels in size). By construction, scene signatures are large distinctive elements in the scene that can be matched across extreme appearance changes.

Recently, there has been a number of attempts to shift away from the traditional, straight forward approach of blindly applying an out-of-the-box point-feature detector/descriptor for egomotion estimation and/or localisation. Richardson et al. [13] present a method for learning an optimal feature detector for Visual Odometry (VO) tasks. Their method searches the space of convolution filters to find the detector that minimises reprojection error. Although this method is aimed at improving standard detection methods for an application specific task, it still focused on using point features, which works well for VO, but not for localisation (e.g., matching a sunny day against a rainy day).

Lategahn et al. [14] present a method for learning an optimal whole-image descriptor for place recognition. They use a genetic optimisation approach to find the optimal combination of fundamental feature blocks to construct their optimal descriptor. However, as with other methods, such as SeqSLAM [3], this can only inform the system of the topological position of the vehicle; it does not provide a metric estimate, which is important for us as we are interested in controlling a vehicle.

Rublee et al. [10] developed a new feature called ORB, which builds upon the FAST [15] detector and the BRIEF [9] descriptor. They use a greedy learning algorithm for de-correlating BRIEF features under rotational invariance. However, as this is still based on low-level structure, data association remains hard under extreme appearance change. Hundelshausen et al. [16] present a noteworthy descriptor that goes beyond point features and instead constructs a network of nodes and directed edges, where each edge is a descriptor in the network, referred to as a "d-token". However, because these descriptors directly sample pixel intensities, this would not be suitable for the types of extreme appearance changes we are considering.

Ultimately, we are concerned with the problem of long-term, robust localisation in outdoor environments, which experience a great deal of appearance changes (e.g., time of day and/or time of year). One approach to this problem would be a system like experience-based navigation [17], which records distinct visual experiences of the environment as the vehicle traverses. If the live video stream cannot be matched to a prior experience, it means the appearance of the world has changed enough to warrant the creation of a new experience. Although this is a feasible approach, we offer an alternative that tries to learn what elements in the environment are stable across all appearances. In this way, localisation is not done against numerous experiences, but rather just a collection of distinctive scene elements.

Recently, Doersch et al. [18] presented a method for extracting geo-distinctive image patches from a collection of images of London and Paris. Their method was able to find image patches, or *visual elements*, of windows, balconies, and street signs which clearly distinguished the Parisian streets from the London streets. The method is, in principle, very simple and relies on a large amount of data and a cross-validation training scheme. This will be discussed further in the next section. We have applied this idea to the localisation problem to find

distinctive visual elements that are stable across a wide range of appearance changes, such as lighting differences and/or seasonal changes. We call these scene signatures. The benefit of using scene signatures instead of low-level point features, which look for corners, edges, or blobs, is that the data association problem becomes less challenging, since scene signatures are very distinctive (e.g., doors, signs, windows, etc). As we use a stereo camera as the primary sensor, we can perform left-to-right matching and obtain 3D position information for each scene signature. This allows us to swap out point features for scene signatures in a VO framework, in order to produce metric pose estimates.

III. SYSTEM OVERVIEW

Here we describe the two main components of our system: (i) the notion of a "weak localiser" that uses scene signatures for pose estimation, and (ii) the offline training algorithm which produces the scene signatures. As we will show, using scene signatures instead of point features offers vast improvements in robustness to extreme appearance changes, resulting in a more robust localisation system.

At a high level, the steps involved in our localisation system work as follows:

1) Initialisation in the map (e.g., place recognition system),
2) Use dead reckoning (e.g., wheel odometry) to predict what place, $\boldsymbol{\pi}_p$, the vehicle is close to and load the bank of SVM classifiers, $\{c_i\}_p$, associated with that place,
3) Provided that the vehicle is sufficiently close to that place (e.g., within several meters), we use each SVM classifier at multiple scales to search for associations in the live image,
4) For each association, we compute the 3D stereo landmarks and solve for the optimal transformation estimate against the map.

An illustration of our system is shown in Figure 2.

A. WEAK LOCALISERS

For a given place, $\boldsymbol{\pi}_p$, scene signatures represent distinctive visual elements, such as buildings, trees, or distinctive structure boundaries in the scene. Examples of scene signatures can be seen in Figure 3. Note that each place is associated with a set of SVM classifiers trained on distinctive scene signatures. We will show in the next section how these scene signatures can be learned offline in an unsupervised manner. However, let us assume that for now we have access to a bank of spatially indexed (for example by distance along a road) SVM classifiers of scene signatures, $\{c_i\}_p$.

Although these scene signatures represent large areas in the image, they can still provide a good metric idea of where the vehicle is locally. Additionally, because we can perform left-to-right matching between the stereo pair, each scene signature has an associated 3D point, allowing us to produce metric estimates local to each place.

In order to obtain sensible solutions, careful handling of the measurement uncertainties is required. The positional uncertainty of a visual element in image space, $\mathbf{P}_{z_i}$, will be

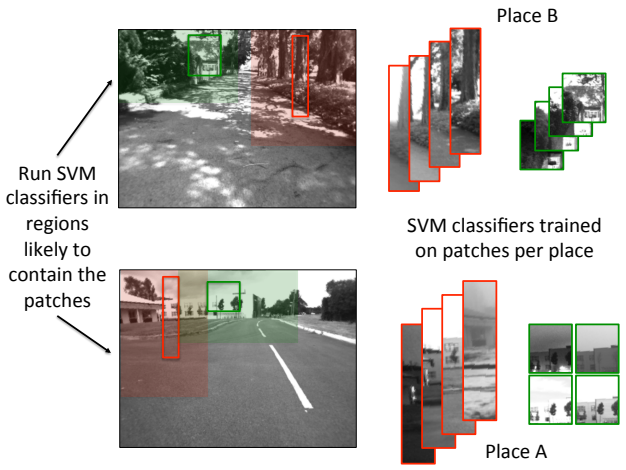


Fig. 2. Offline, we learn scene signatures in the form of SVM classifiers. Each classifier is associated with a particular place, $\boldsymbol{\pi}_p$, and spatial region in the image (i.e., the region that it's most likely to fire in). At run-time, we use the bank of pretrained classifiers associated with $\boldsymbol{\pi}_p$ 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.

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, $\mathbf{r}$, and the SVM detection probability, λ:

$$\mathbf{P}_{z_i} = \mathbf{f}(a, r, s, \lambda). \tag{1}$$

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

$$\mathbf{P}_{z_i} \propto \frac{1}{s}\,\mathbf{P}_r, \tag{2}$$

where $\mathbf{P}_r$ 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, however, is less clear. Intuitively, we expect that the lower the probability of being a scene signature and the larger the area of the patch, the less certain the keypoint position should be. Thus, as a heuristic, we assume that the covariance takes the following form,

$$\mathbf{P}_{z_i} := \frac{a}{\lambda s}\,\mathbf{P}_r. \tag{3}$$

Although not considered here, another factor that may be useful would be a level of confidence in the SVM score [19].

Since each patch feature, $\mathbf{z}^j$, has an associated 3D landmark, $\mathbf{p}^j$, we can use the standard stereo model, $\mathbf{h}(\cdot)$, to predict the location of a landmark in frame b relative to some other frame a, according to the transformation matrix, $\mathbf{T}_{a,b}$:

$$\mathbf{z}_a^j = \mathbf{h}(\mathbf{T}_{a,b}, \mathbf{p}_b^j) + \mathbf{n}_a^j, \qquad \mathbf{n}_a^j \sim \mathcal{N}(\mathbf{0}, \mathbf{P}_{z_a^j}). \tag{4}$$

Additionally, we use a strong prior, $\hat{\mathbf{T}}_{a,b}$, with a small uncertainty in the vertical offset between the live frame and the map, as well as roll and pitch, since we know that these positional differences would be small for a road vehicle. The prior is important because the translational component of the

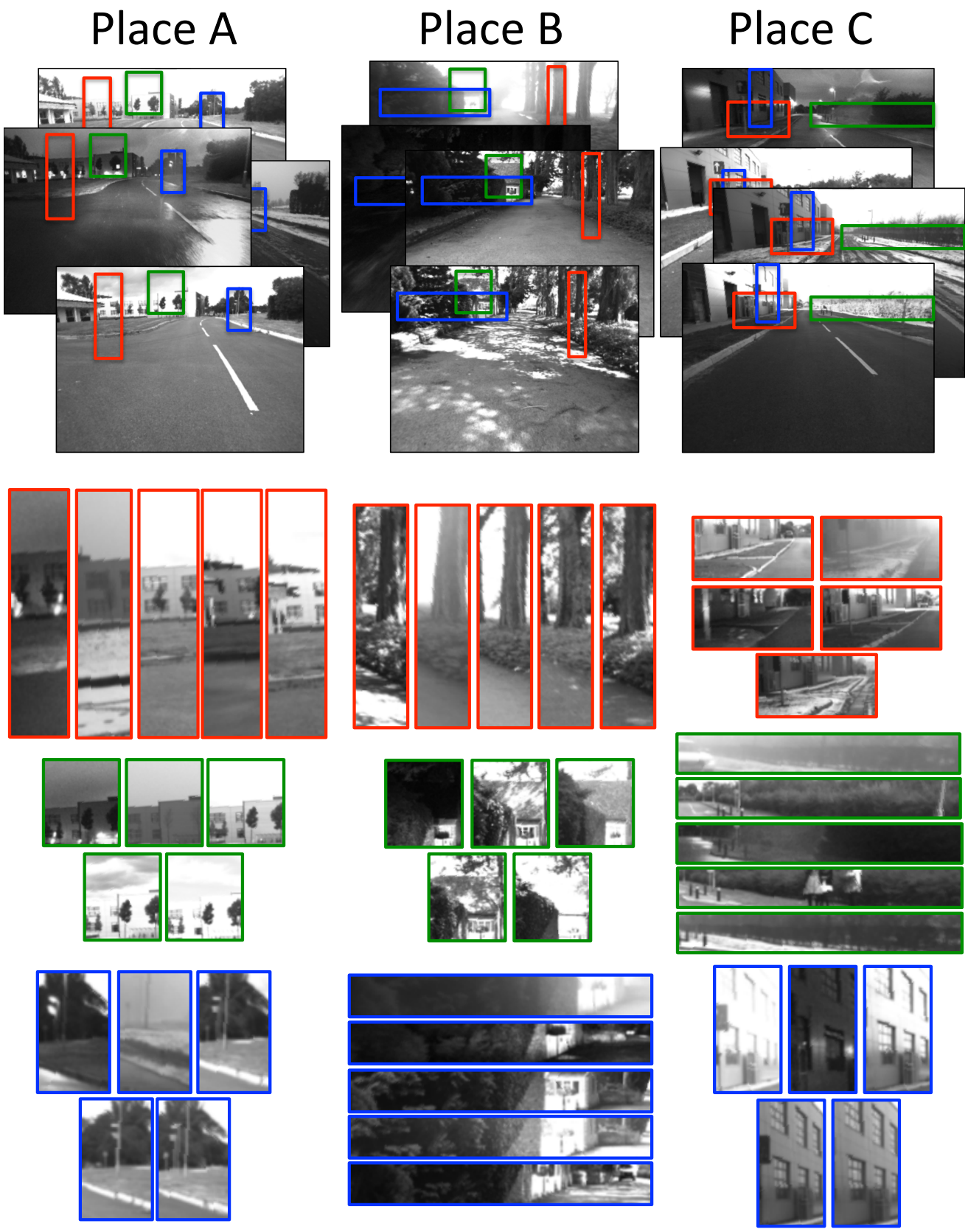


Fig. 3. Example scene signatures learned by our algorithm. Image sets from the same place with varying appearances (represented by run-times in this figure) are used offline to learn these distinctive scene signatures. SVM classifiers are trained for each cluster of scene signatures and can be used at run-time on the live image stream to perform data association, followed by metric pose estimation. Note that the shapes vary in size and dimension and tend to pick up things like changes in structure boundaries, as these are very distinctive (e.g., from road, to grass, to building, to sky).

localisation estimate will not be very accurate owing to the low number of patches in the foreground. However, we note that we obtain very good orientation estimates, so combined with a reasonable egomotion estimate from, say, wheel odometry or VO, the weak localisers are sufficient in providing pose estimates with similar accuracy as our INS system (i.e., sub meters) in large outdoor environments. Including the prior estimate, $\hat{\mathbf{T}}_{a,b}$, the final least-squares system we seek to optimize is given by the following:

$$\mathcal{O}(\mathbf{T}_{a,b}) = \frac{1}{2}\begin{bmatrix}\mathbf{q}(\mathbf{T}_{a,b},\hat{\mathbf{T}}_{a,b})\\ \mathbf{z}_a - \mathbf{g}(\mathbf{T}_{a,b},\mathbf{p}_b)\end{bmatrix}^T \begin{bmatrix}\mathbf{P}_x^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{P}_z^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{q}(\mathbf{T}_{a,b},\hat{\mathbf{T}}_{a,b})\\ \mathbf{z}_a - \mathbf{g}(\mathbf{T}_{a,b},\mathbf{p}_b)\end{bmatrix}, \tag{5}$$

where

$$\mathbf{z}_a := \begin{bmatrix}\mathbf{z}_a^0\\ \vdots\\ \mathbf{z}_a^M\end{bmatrix}, \quad \mathbf{p}_a := \begin{bmatrix}\mathbf{p}_a^0\\ \vdots\\ \mathbf{z}_p^M\end{bmatrix}, \quad \mathbf{P}_z := \text{diag}(\mathbf{P}_{z_a^0},\ldots,\mathbf{P}_{z_a^M}), \tag{6}$$

and $\mathbf{q}(\cdot)$ is a function that takes two SE3 transformation matrices and computes a 6×1 error vector, which depends on the choice of the orientation parameterisation. Note that we also use the Geman-McClure [20] robust cost function, which leaves us with the following objective function:

$$\mathcal{O}(\mathbf{T}_{a,b}) = \frac{1}{2}\sum_i \frac{\mathbf{e}_i^T\mathbf{P}_i^{-1}\mathbf{e}_i}{\sigma_i^2 + \mathbf{e}_i^T\mathbf{P}_i^{-1}\mathbf{e}_i}, \tag{7}$$

where σ_i are the M-estimator parameters and each $\mathbf{e}_i$ represents an error term (e.g., the prior or measurement). This has the effect of scaling the covariance to down weight the contribution of potential outliers during the iterative solve, which is done using Levenberg Marquardt [21].

B. UNSUPERVISED LEARNING OF SCENE SIGNATURES

This section will describe our unsupervised approach to learning scene signatures, which are locally distinctive and stable visual elements. *Locally distinctive* means that the visual element is distinct in a local region in image space. Given that we have a reasonable prior on the motion of the vehicle, it does not matter if the visual element occurs elsewhere in the image, it need only be locally distinctive for data association. *Stable* means that the visual element can be identified across multiple images of the same area, under a variety of appearances.

The training algorithm can be divided into the following steps, where the main adaption of the algorithm described in [18] occurs in steps 1-3d.

1) Collect a set of N images with a variety of appearances at a particular location in the world (we have N=31 in our experiments).
2) For each image, partition it into M tiles (see Figure 4; we have M=4 in our experiments).
3) For all M tiles, do the following (we refer to each iteration here as a "round"):
   a) Set the same tile in all N images as the "positive" tile, meaning that all patches sampled from this tile are in the positive set (light brown in Figure 4).
   b) Sample K patches of varying dimension from the positive tile (i.e., positive set for this round).
   c) Select a band around the positive tile; patches sampled in these regions are put in the negative set (the blue band in Figure 4).
   d) Sample K patches of varying dimension from the negative band (i.e., the negative set for this round).
   e) Split the positive and negative patches sampled from all the tiles into l positive/negative datasets (we use $l = 3$ in our method).
   f) For each patch, compute a HOG descriptor [22] and compute the top 20 nearest neighbours for each of the positive patches. Note that each grouping of 20 patches represents the seed for a candidate scene signature cluster.
   g) Prune by discarding candidates where more than half of the nearest neighbours belong to the negative set, or if there are more than 2 overlaps with any of the other clusters. This reduces the number of candidate clusters to approximately 200.
   h) Perform cross-validation training on l datasets to produce a set of SVM classifiers.

In each iteration, we train an SVM detector using the top k nearest neighbours for each candidate cluster and all of the negative-set patches as negative examples ($k = 5$ in our experiments). Then, the newly trained detectors are applied to one of the other datasets to select the top k detections for retraining the SVM. After retraining the SVM, the process continues: applying the detectors trained on the previous round to the other datasets. If the top k detections for each dataset stop changing, then the detector has converged and the image patches are deemed sufficiently distinctive and stable. As in [18], we use a maximum of three iterations for convergence. Note that the output of this process is a set of bespoke SVM classifiers tuned to a particular place, $\{c_i\}_p$ per region in our map, $\boldsymbol{\pi}_p$.

Figure 3 shows example clusters that were generated from this algorithm (note that this is a small subset; typically we find around 30-40 patches for each image). Each scene signature has been colour coded to show which cluster it belongs to. As can be seen, many of the clusters are things like distinctive tree silhouettes, corners of doors, fences, windows, etc. Another interesting observation to make is that the road is typically ignored, which is to be expected as it is mainly a homogeneous texture and not very distinctive between datasets.

IV. EXPERIMENTS AND RESULTS

In this section, we compare our weak-localiser approach to a typical point-feature-based approach (SURF [8] in our case) for the task of localisation. We trained our system with 31 datasets of a 750 m outdoor loop. Places were defined every 20 m along the route according to an INS system, which resulted in 31 locations. Thus, for each of the 31 locations there were 31 training images. Scene signatures were then generated according to Section III-B. We note however, that places can be defined by other means, either manually or by

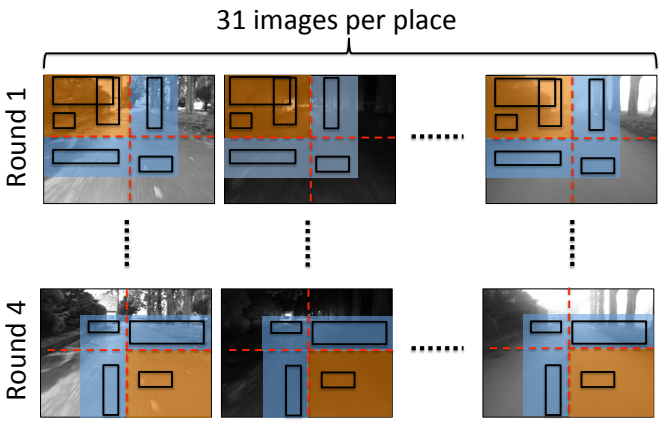


Fig. 4. Our strategy for partitioning the data to produce scene signatures. We take a collection of images at a particular location in the world and partition each image into a number of tiles. In this example, patches (black rectangles) drawn from the light brown regions are placed in the positive set and patches drawn from the blue regions are placed in the negative set. Since we have the same pattern for every image and each image has roughly the same viewpoint, we are able to seed the training algorithm with elements subject to varying appearance changes.

(a) Canonical map image: clear and sunny; noon.

(b) Test image: clear and sunny; morning.

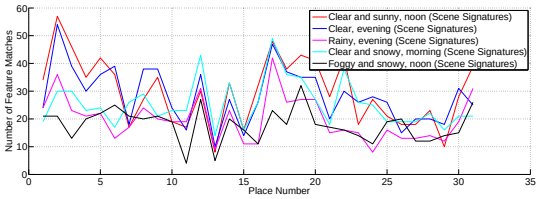
(c) Test image: clear; evening.

(d) Test image: rainy; evening.

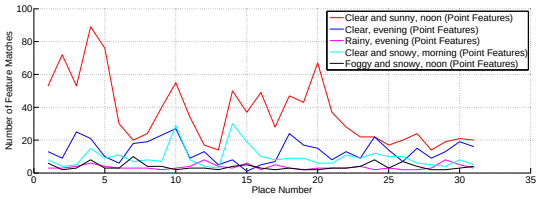
(e) Test image: clear and snowy; morning.

(f) Test image: foggy and snowy; noon.

Fig. 5. Example test images used in our localisation experiments. These were chosen due to their large visual variability.

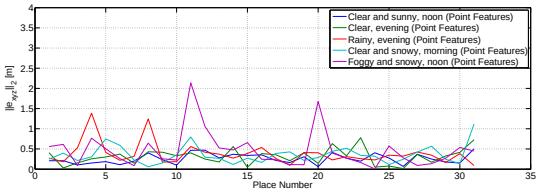


(a) Feature matches using scene signatures.

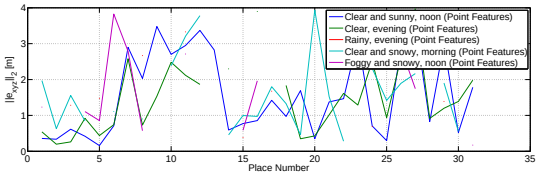


(b) Feature matches for use point features.

Fig. 7. Feature matches for each place over all 5 runs for both the point-feature system and our scene-signature system. Note how the number of matches using scene signatures stays relatively consistent across all 5 runs while point features struggle greatly for the foggy and rainy runs.



(a) Translational estimation error using scene signatures.



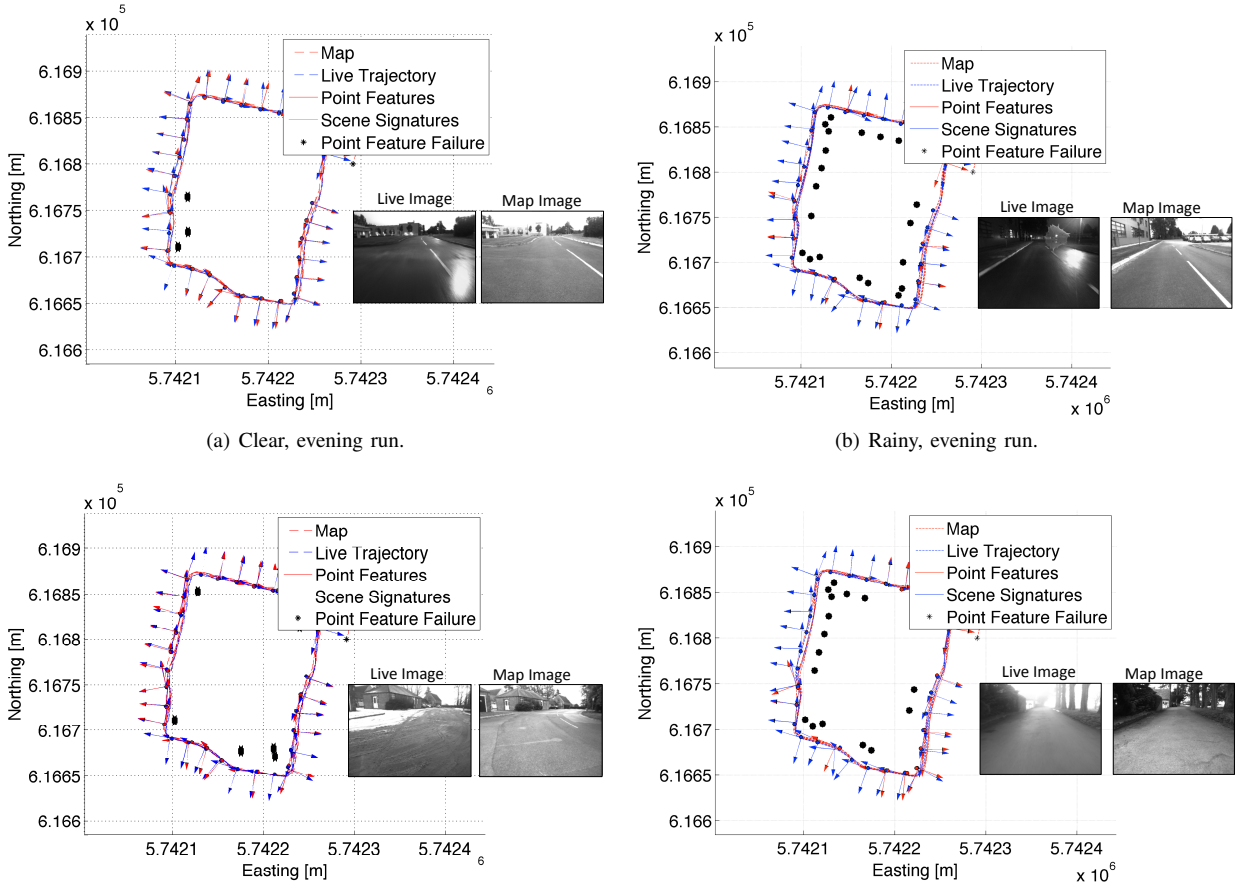
(b) Translational estimation error using point features.

Fig. 8. Translational estimates errors for the scene signatures and point features at each place. Note that gaps in the point feature plots represent localisation failures (i.e., either a failure to match or a divergent solution).

using place recognition techniques. The only important factor is that the training images for a particular place have roughly the same viewpoint in the area.

Our code was implemented in MATLAB and takes approximately one hour of learning per place. We note that we did not exploit any parallel processing, which would significantly reduce processing times as the training in each image tile can be done independently. As each place is represented by a collection of SVM classifiers, the total storage size per place is $10 \sim 15$ MB.

For our test data, we used 5 separate datasets that included a sunny day run, an evening run, a rainy evening run, a snowy run, and a snowy and foggy run. Examples of some test images are shown in Figure 5. The goal of these experiments is to show that we can use scene signatures and a weak localiser to produce metric estimates regardless of the appearance changes.



(a) Clear, evening run.

(b) Rainy, evening run.

(c) Clear and snowy run.

(d) Foggy and snowy run.

Fig. 6. Localisation results, where the arrows represent 2D projections of the vehicle coordinate frames for both methods. As both the point-feature and our scene-signature method were able to localise all frames for the first dataset, we omitted the plots here and turn to the more challenging cases. Our scene-signature approach was able to localise all frames, whereas the point-feature-based system failed on 33% of the places. These failures have been indicated on the plots with a large black circle. Note that almost all estimates agree to the INS ground truth within meters.

To reiterate, our localisation strategy is as follows. After initialising in the map, we use dead reckoning to predict when we are within a couple meters of a place, after which we load the SVM classifiers associated with that place and run them on the live image to detect the scene signatures. Once we have associated these scene signatures, we can perform local, metric, pose estimation. This approach of predicting where the nearest topological node is and then localising against the map is similar to teach-and-repeat systems such as McManus et al. [23] and Furgale and Barfoot [1], except that our map keyframes are separated by larger distances.

Figure 6 presents the localisation results for the 5 live runs against our map that contains a bank of trained classifiers per place. The results show the live and reference INS trajectories as well as the localisation estimates for both our system and the baseline. Unfortunately, as our INS system drifts on the order of meters from one dataset to the next, it proved to be ill-suited to asses the accuracy of the estimates. We instead exploit the fact that the training images are gathered at approximately the same position and use a generous tolerance on the translational/rotational estimates to define a localisation failure. Letting $\hat{\mathbf{x}} := [\hat{\mathbf{t}}^T, \hat{\boldsymbol{\theta}}^T]^T$ represent our estimate, we define a localisation failure if $||\hat{\mathbf{t}}||_2 > \alpha$ or $||\hat{\boldsymbol{\theta}}||_2 > \beta$, where we chose $\alpha = 4\,\text{m}$ and $\beta = 30°$.

TABLE I
NUMBER OF FRAMES LOCALISED AGAINST OUT OF 31 PLACES.

Live Run	Scene Signatures	Point Features
Clear and sunny morning	31	31
Clear evening	31	28
Rainy evening	31	9
Clear and snowy morning	31	24
Foggy and snowy afternoon	31	12

Table I shows the number of frames localised against for each run (according to INS ground truth), where we see that our system was able to localise all frames in all 5 datasets, despite extreme variations in appearance. The point-feature-based system was unable to localise a majority of the frames for the rainy evening run and the foggy snowy run. Figure 7 shows the number of feature matches for each place over all 5 runs. Figure 8 shows the estimation errors for each place. Figure 9 shows examples of our system succeeding where the point-feature system failed.

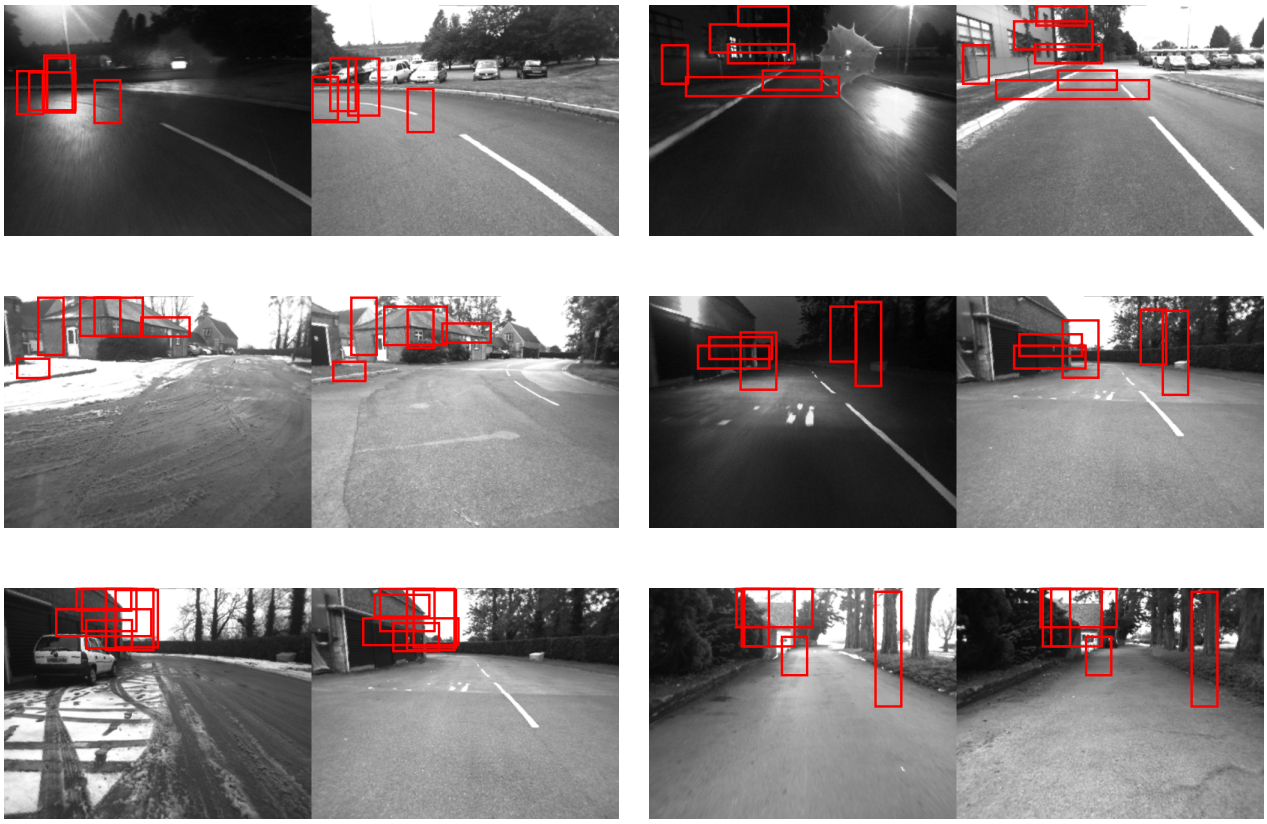


Fig. 9. Examples where our method was able to localise the live run (left image of each pair) with the map (right image of each pair), while point-features failed. Only a subset of the matches are being shown for clarity. On average, we obtained 24 matches per place across all five datasets using scene signatures.

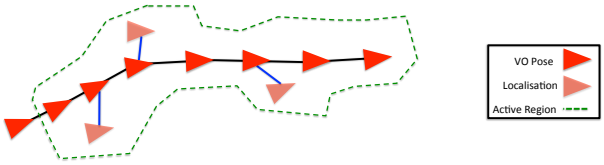


Fig. 10. Illustration of our runtime localisation approach. Localisation updates occur at 2-5 Hz, while Visual Odometry updates occur at 15-20 Hz. After we receive a localisation update, we perform posegraph relaxation over a sliding window, indicated by the active region in green.

V. SYSTEMS WORK

The results and method described in this paper represent our initial iteration of the scene-signature approach, which was coded in Matlab and ran offline. We have since ported the code to C++ to obtain realtime performance. Our linear-SVM detection class uses OpenCV's OpenCL HOG for feature extraction. Our scene-signature detection block runs in a separate thread at approximately 2-5 Hz. Our main thread runs Visual Odometry at approximately 15-20 Hz to predict poses in between localisations. As the localisation updates occur at a slower rate, we perform posegraph relaxation over a sliding window to obtain our final estimate (see Figure 10).

VI. DISCUSSION AND CONCLUSION

We have demonstrated a new approach to the localisation task, which departs from the traditional point-feature system by learning spatially indexed classifiers of distinctive visual elements called *scene signatures*. Although we are unable to obtain accuracy on the order of centimeters, we are more robust to extreme appearance change and obtain the same type of coverage as a topological localisation system, like SeqSLAM [3], but with the added benefit of metric pose.

Scene signatures enable robust, metric localisation where traditional systems simply fail. Each bank of place-dependant SVM classifiers is run on the live image stream to perform the data association, and a standard frame-to-frame localisation framework is used to obtain the metric pose estimate. We have shown that our approach can successfully localise the vehicle across very challenging lighting and/or weather conditions. We believe that point features alone are simply not enough for robust, long-term localisation systems and that our approach is a step in the right direction.

VII. ACKNOWLEDGEMENTS

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

[1] P. Furgale and T. Barfoot, "Visual teach and repeat for long-range rover autonomy," *Journal of Field Robotics, special issue on "Visual mapping and navigation outdoors"*, vol. 27, no. 5, pp. 534–560, 2010.

[2] K. Konolige, J. Bowman, J. Chen, P. Mihelich, M. Calonder, V. Lepetit, and P. Fua, "View-based maps," *The International Journal of Robotics Research*, vol. 29, no. 8, pp. 941–957, 2010.

[3] M. Milford and G. Wyeth, "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, Minnesota, USA, 14-18 May 2012.

[4] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proceedings of the 4th Alvey Vision Conference*, 1988, pp. 147–151.

[5] E. Rosten, G. Reitmayr, and T. Drummond, "Real-time video annotations for augmented reality," in *Advances in Visual Computing*, 2005.

[6] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust Wide Baseline Stereo from Maximally Stable Extremal Regions," in *Proceedings of the British Machine Vision Conference*. BMVA Press, 2002, pp. 36.1–36.10, doi:10.5244/C.16.36.

[7] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, 2004.

[8] H. Bay, A. Ess, T. Tuytelaars, and L. Gool, "Surf: Speeded up robust features," *Computer Vision and Image Understanding (CVIU)*, vol. 110, no. 3, pp. 346–359, 2008.

[9] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua, "Brief: Computing a local binary descriptor very fast," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1281–1298, 2012.

[10] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 2564–2571.

[11] A. Davison, I. Reid, N. Motlon, and O. Stasse, "Monoslam: Real-time single camera slam," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, 2007.

[12] A. Davison and D. Murray, "Simultaneous localization and map-building using active vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, 2002.

[13] A. Richardson and E. Olson, "Learning convolutional filters for interest point detection," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2013.

[14] H. Lategahn, J. Beck, B. Kitt, and C. Stiller, "How to learn an illumination robust image feature for place recognition," in *IEEE Intelligent Vehicles Symposium*, Gold Coast, Australia, 2013.

[15] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *European Conference on Computer Vision*, 2006.

[16] F. von Hundelshausen and R. Sukthankar, "D-Nets: Beyond Patch-Based Image Descriptors," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2012.

[17] W. Churchill and P. Newman, "Practice makes perfect? managing and leveraging visual experiences for lifelong navigation," in *Proceedings of the International Conference on Robotics and Automation*, Saint Paul, Minnesota, USA, 14-18 May 2012.

[18] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. Efros, "What makes paris look like paris?" *ACM Transactions on Graphics*, 2012.

[19] A. Pronobis and B. Caputo, "Confidence-based cue integration for visual place recognition," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robotics and Systems*, San Diego, California, USA, Oct 29 - Nov 2 2007.

[20] S. Geman and D. McClure, "Statistical method for tomographic image reconstruction," in *Proceedings of the 46th Session of the International Statistical Institute, Bulletin of the ISI*, vol. 52, 1987, pp. 5–21.

[21] K. Levenberg, "A method for the solution of certain non-linear problems in least squares," *The Quarterly of Applied Mathematics*, vol. 2, pp. 164–168, 1944.

[22] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, San Diego, California, USA, 2005, pp. 886–893.

[23] C. McManus, P. Furgale, B. Stenning, and T. D. Barfoot, "Visual Teach and Repeat Using Appearance-Based Lidar," in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, 2012.