Visual place recognition techniques for pose estimation in changing environments

Stephanie Lowry and Henrik Andreasson
Centre for Applied Autonomous Sensor Systems, Örebro University
Örebro, SE
Email: stephanie.lowry@oru.se

***Abstract*—This paper investigates whether visual place recognition techniques can be used to provide pose estimation information for a visual SLAM system operating long-term in an environment where the appearance may change a great deal. It demonstrates that a combination of a conventional SURF feature detector and a condition-invariant feature descriptor such as HOG or conv3 can provide a method of determining the relative transformation between two images, even when there is both appearance change and rotation or viewpoint change.**

I. Introduction

Visual place recognition techniques for performing topological localization have been widely investigated and many successful systems have been demonstrated in environments that experience strong appearance change [18]. Topological localization allows a robot to perform loop closure – recognizing where it is when it returns to a place is has previously visited – which is a vital ingredient of any Simultaneous Localization and Mapping (SLAM) system, and it has been shown that topological loop closure is more efficient and scalable than using metric techniques [35]. However, once a loop closure candidate has been identified, a similarity transformation between the frames needs to be calculated to provide relative pose information [23].

This paper asks whether pose estimation can be performed even when the appearance of the environment has changed (see Figure 1). It demonstrates that using a conventional point feature detector such as SURF [3] and a condition-invariant feature descriptor such as Histogram of Oriented Gradients (HOG) [6] or conv3 [31] allows feature correspondences to be calculated. These feature matches can then be used as input to a SLAM system such as [7] or [23] to calculate the relative pose offset.

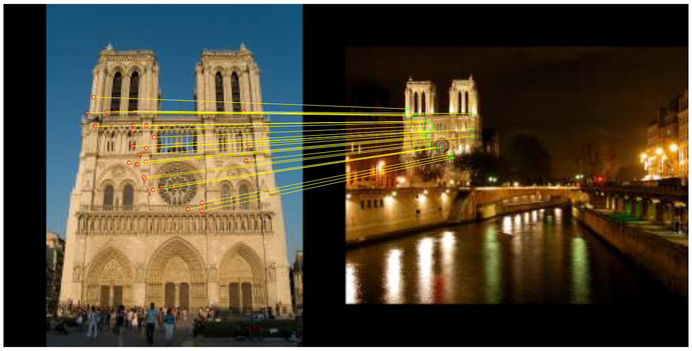


Fig. 1. Visual features can be matched across different viewpoints and appearance conditions, and feature correspondences used as input to loop closure algorithms. In this example, SURF keypoints are extracted and described using conv3 feature descriptors.

II. Prior work

Visual place recognition in changing environments can use a range of different feature types, but the most robust against condition change are features that describe the whole image in a single feature. Example of whole-image descriptors include WI-SURF [2], low-resolution images [22] or conv3 features [31]. However, describing the scene in a single feature means the system is vulnerable to viewpoint change [32].

The alternative viewpoint-invariant approach is to extract multiple features from each image that describe a portion of the scene. Such an approach requires a two-stage process – feature detection and feature description. Commonly used feature detection techniques include SIFT [16], SURF, FAST [30], and BRISK [15]. Such detectors are known as *point feature* detectors as they detect keypoints in the image and can be used to perform visual odometry as well as place recognition. However, it has been identified that point features often do not provide condition invariance [34, 29], and Furgale and Barfoot [11] noted that the non-repeatability of SURF features due to changing appearance, particularly lighting change, was a major cause of failure during visual-teach-and-repeat experiments.

A compromise solution between using a whole-image approach and using point feature detectors is to use parts of an image [33, 26, 21, 25]. A number of detection techniques can be used: Sünderhauf et al. [33] use EdgeBoxes [36], Neubert and Protzel [25] use SIFT keypoints, Neubert and Protzel [26] use SLIC superpixels [1] and McManus et al. [21] select distinctive image patches using an unsupervised learning technique. These image patch detectors provide both condition-invariance and enhanced viewpoint invariance.

Once the feature detector of choice has selected salient image regions from the image, whether these regions are defined using point features or patch features, these regions are described using feature descriptors, of which many different kinds exist. Point feature detectors including SIFT, SURF and BRISK each have an associated point feature descriptor, but it is also possible to mix and match between feature detectors and feature descriptors. SIFT and SURF descriptors have been used in a large number of visual localization tasks [18], but for appearance change other descriptors are preferred. HOG descriptors have been successfully used for place recognition in changing environments [21, 24] and features derived from Convolutional Neural Networks (CNNs) have also been shown to achieve impressive results in challenging environments [33, 26].

Visual place recognition solves the topological localization problem, where the system determines the most likely global location of the robot. There is a corresponding metric localization problem, where the system must compute the metric location of the robot relative to either a global coordinate frame or to a relative coordinate frame based on prior robot positions. There are a large number of visual localization systems that perform such pose estimation, using either sparse visual information such as point features [7, 23, 13] or edges [8], or dense pixel-level information [9].

There has been research into performing pose estimation in changing environments, in particular using geometric features such as edges [27] which are known to be relatively stable to appearance change such as lighting variance, and can be matched against a previously generated laser map [27, 4]. However, data association using edge features can be challenging [8]. This paper aims to contribute to this field by investigating the feasibility of using image features to perform pose estimation in changing environments.

III. Approach

The motivation behind this paper is to investigate the use of existing visual place recognition techniques to perform metric pose estimation for visual SLAM systems navigating long-term in a large outdoor environment. There are several aspects to this problem that need to be addressed. Firstly, the system must first perform loop closure. Loop closure can be performed using one of a number of condition-invariant and pose-invariant methods [33, 26]. The system must then perform a metric correction, by calculating the relative pose estimate based on extracted features from the loop closure candidate images. If the system is using a monocular rather than a stereo camera, 3D information for each feature can be inferred using tracking techniques [7] for pose estimation.

This paper is concerned with extracting and matching features from the loop closure candidate images for input into the SLAM tracking and pose estimation modules. It compares the use of EdgeBoxes and SURF as feature detectors. It also tests a number of feature descriptors that have been widely investigated in the context of place recognition, particularly in changing environments.

To match the features, the approach from [33] was used. The cosine distance was calculated between the descriptors x_1, x_2 from the two images I_1 and I_2 respectively:

$$d(x_1, x_2) = 1 - \frac{(x_1 - \bar{x_1}) \cdot (x_2 - \bar{x_2})^T}{\sqrt{(x_1 \cdot x_1^T)(x_2 \cdot x_2^T)}}. \tag{1}$$

A cross-checked nearest-neighbor test was applied, so that two features x_1 and x_2 were only matched if they were each the nearest neighbor of the other: that is, if x_1 had the smallest cosine distance to x_2 of any descriptor in I_1 and x_2 also had the smallest cosine distance to x_1 of any descriptor in I_2. To determine a resulting correct inlier set, outliers were removed using RANSAC [10].

IV. Experiments

This section present results that investigate the use of various feature detectors and feature descriptors for visual localization in changing environments. We begin by investigating whether the notion of pose estimation across changing environments is feasible using existing feature detectors and descriptors. This simple test case uses aligned images so provides no viewpoint variation, but means that the relative pose between the two images is known.

We then investigate the effect of combining rotation variance and condition variance by artificially rotating the images relative to each other and testing the pose estimation results. Finally, we test the effect of a combination of viewpoint change and condition change on feature matching and visual localization.

A. Feature detectors – points and patches

The first experiment investigated the use of different local feature detectors to perform pose estimation across changing conditions. We tested two feature detectors – EdgeBoxes and SURF detectors. SURF feature detection has been used in many visual localization algorithms in the past [18], for example for topological place recognition [5] and for stereo visual odometry in [11]. However, concerns have been raised about SURF's inability to perform repeatably in changing environments [11, 34]. We note however that SURF is generally considered to encompass both the SURF feature detector and the SURF feature descriptor, while in this paper we decouple the two phases and investigate each separately: in this experiment only the feature detector was used. In contrast, EdgeBoxes have been demonstrated as a successful method of local feature extraction for topological place recognition in changing environments [33] but do not provide the point-level accuracy possible with SURF.

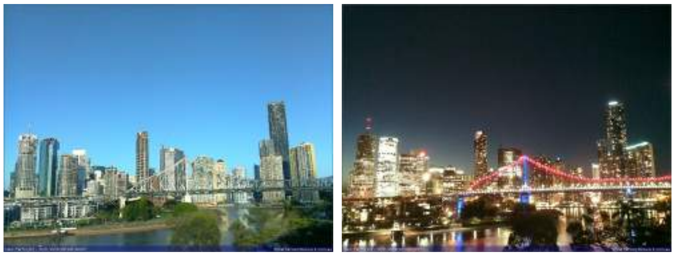


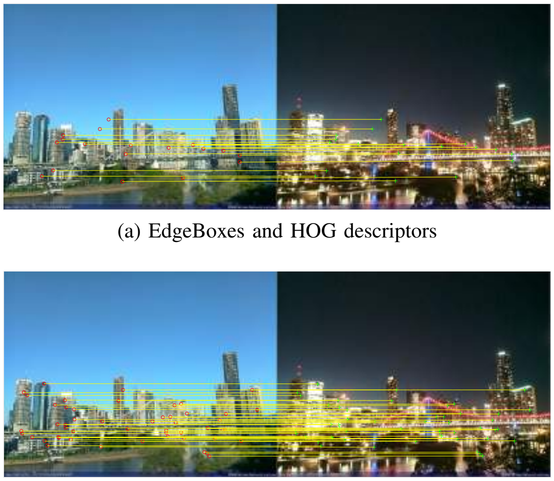
Fig. 2. Brisbane skyline images used to test feature matching across changing environments

Two images from a static city skyline webcam from the Australian city of Brisbane were used. One image was captured in the early morning at 6am and one during the night at 7pm (see Figure 2). As the images were from a static webcam, the viewpoint was known to be well-aligned, and an accurate measure of relative pose could be calculated. However, the near-perfect alignment means that the additional challenges that can be caused by viewpoint change, camera rotation, or scale change were not considered. These issues were investigated further in Sections IV-B and IV-D.

1) Setup: EdgeBoxes were extracted using the MATLAB code provided by the authors[1] using the pre-trained object models and default parameters, while the SURF keypoints were extracted using the inbuilt MATLAB functions using default parameters. The exception to the use of the default parameters for each feature descriptor was the choice of threshold; this meant that the number of features extracted per image could be controlled, and for both detectors, 3000 features were extracted from each image. While this is a large number of features, it is on a similar order of magnitude to the number of features extracted in for example [23], where 2000 features where extracted from the 1241×376 KITTI dataset images. For comparison, these images have size 982×737, or approximately 1.5 times more pixels.

Each detected feature was described using a HOG feature. While HOG features were originally designed for human detection they have also been successfully used for place recognition in changing environments [21, 24]. The HOG feature was extracted from an image patch defined by the size of the bounding box (for EdgeBoxes) and by the scale parameter (for SURF features). The patches were resized so that features of different scales could be compared, and in each case a HOG feature of 7056 dimensions was extracted.

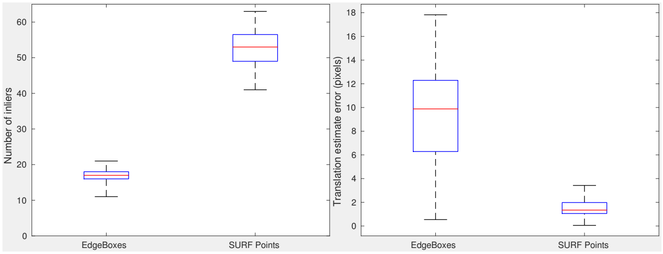
2) Results: Figure 3 displays example inlier matches from each feature detector. It can be seen that both versions find a considerable number of correct inlier matches. However, the SURF feature detector finds a larger number of matches.



(a) EdgeBoxes and HOG descriptors

(b) SURF points and HOG descriptors

Fig. 3. Sample inlier sets for (a) the EdgeBoxes feature detector and (b) the SURF feature detector when combined with HOG features across a day-night image comparison task. The SURF feature detector finds more matched features.



(a) Number of inliers

(b) Translation estimate error

Fig. 4. Matching results for day-night comparison for EdgeBoxes and SURF feature detectors, summarized using (a) number of inliers and (b) translation estimate error. The SURF feature detector finds more inliers than the EdgeBoxes feature detector, and has a narrower distribution of translation errors.

As RANSAC is non-deterministic, 100 trials of each experiment were run. As there could be wide variability in the RANSAC output, particularly in the most challenging cases, a distribution of results gives a fairer picture of the reliability of each approach. Figure 4 presents the distribution of inliers found by each method (see Figure 4a), and the error in the translation estimate for each (see Figure 4b). These results show that SURF points find more inlier matches than EdgeBoxes (mean number of inliers: 53 vs. 17).

The minimum number of inliers found with SURF was 41, which was substantially greater than the maximum number of 21 found with EdgeBoxes. Similarly, the mean translation error found was 1.5 pixels for SURF features with a maximum translation error of 3.4 pixels. In contrast, the mean translation error for EdgeBoxes was 9.4 pixels with a maximum error of 17.8 pixels.

As can be seen, the EdgeBoxes provide an approximate estimate of the pose difference. However, the SURF points, when combined with HOG features, find more inlier matches, and provide a more accurate result. Based on these results, SURF points were used as the chosen feature extraction technique for the remainder of the experiments.

B. Appearance change and rotation

Place matching in the presence of camera rotation is an area that has been identified as challenging for visual place recognition in changing environments. Valgren and Lilienthal [34] noted that upright SURF (that is, SURF without the orientation component used), performed more reliably than other SURF variants. The improvement of condition-invariance by the removal of rotation-invariance is not exclusive to SURF, with upright SIFT being found to provide better lighting invariance than the rotation-invariant SIFT feature [29]. In general, ground-based robotic vehicles are not required to handle large camera rotations; however, some rotation may be caused by traversing uneven ground and thus slight rotation invariance may be required. This experiment investigated whether features that are condition-invariant can simultaneously handle some rotation invariance, and vice versa. It tests a known rotation-invariant feature descriptor (SURF) with poor condition invariance with a feature descriptor known to

[1] https://github.com/pdollar/edges

be robust to changing environments (HOG) but not typically defined in a rotation-invariant way.

1) Setup: The experiment initially performed a baseline test between two images with little appearance change between them, using the Brisbane skyline image set described in Section IV-A with images captured at 6am and 7am. This comparison demonstrated the capability of the feature descriptors to manage rotation variance in the absence of appearance variation. Then the 6am and 7pm images were compared as in Section IV-A, to investigate the effect of combined rotation and appearance change. The 7am and 7pm images were rotated at 5° intervals from 0° to 50°.

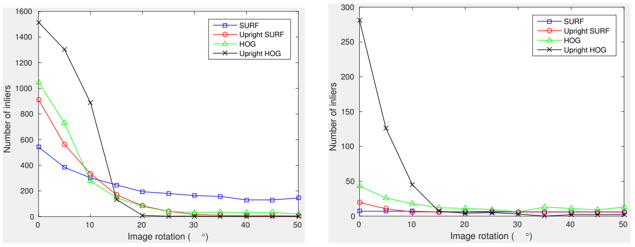
Keypoints were selected using the SURF feature detector with default parameters as in Section IV-A. The number of allowed features per image was 3000. The associated 128-dimension SURF descriptors were extracted, both the upright version which ignores the SURF point orientation and which will be referred to as U-SURF, and the full rotation-invariant version which will be referred to as R-SURF. U-SURF features are claimed to have some robustness to minor rotations in the order of ±15° [3] and so this approach tested whether upright descriptors could be adequate for low-rotation scenarios.

Similarly, a 7056-dimension HOG descriptor was extracted for each keypoint, as in Section IV-A. As for the SURF features, an upright version (U-HOG) was extracted, while a rotation-aware version (R-HOG) was calculated by rotating the local image region according to the calculated SURF point orientation.

2) Results: This experiment was evaluated using a similar approach to Section IV-A. For each test case, the distribution of inliers over 100 RANSAC trials was calculated. The estimated rotation was also computed and used for evaluation, in a similar manner as the translation estimate was used in Section IV-A.

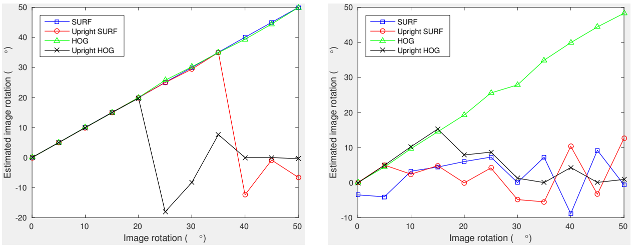
Figure 5 displays the median number of inliers for each image matching test. In both cases (with and without appearance change) the number of matched features is highest for U-HOG features at low rotations, followed by R-HOG, U-SURF and R-SURF. However, at higher image rotations (15° and above), R-SURF features perform best when there is no appearance change (Figure 5a) and R-HOG features perform best when there is appearance change (Figure 5b). These results agree with the expected strengths of each of the tested features – HOG features match upright image regions well, even in changing environments, but are not highly rotation invariant, while SURF features demonstrate the opposite strengths and weaknesses.

The rotation estimate is shown in Figure 6. When there is no appearance change (Figure 6a), both R-SURF and R-HOG accurately estimate the true rotation for all the tested angles. U-SURF also accurately estimates the true rotation for all offsets up to 35°, while U-HOG is accurate to 20° before failing. This result suggests that the HOG descriptor is more sensitive to rotation variance than SURF. However, when there is no appearance change either upright descriptor performs well over a relatively broad range of rotation offsets.



(a) Rotation variance only (b) Rotation and appearance variance

Fig. 5. Median number of inliers over 100 RANSAC trials for image matching with (a) rotation change only and (b) both appearance and rotation change. The x axis displays the relative rotation between the tested images. Upright HOG features perform strongly when the relative rotations are small (less than 15°) but drop off rapidly at higher rotations. SURF features perform well when there is only rotation variance but perform worse than HOG features when there is also appearance change.



(a) Rotation variance only (b) Rotation and appearance variance

Fig. 6. Median rotation estimate over 100 RANSAC trials for image matching with (a) rotation change only and (b) both appearance and rotation change. The x axis displays the relative rotation between the tested images, so a line $y = x$ demonstrates a correct rotation estimate. SURF features match or outperform HOG features when there is only rotation variance, but when there is also condition change HOG features perform best.

When there is both rotation and appearance change (see Figure 6b), the performance of each descriptor is quite different. R-SURF does not correctly match the images, even when there is no rotation offset, while U-SURF does correctly estimate the rotation offset for 0° and 5° but fails for larger rotation offsets. This result agrees with the earlier results that demonstrated that U-SURF is more repeatable in changing environments than other SURF varieties [34]. However, R-HOG features accurately estimate the rotation offset for all the tested rotation values. U-HOG also accurately estimates the offset up to 15° before failing. This is slightly poorer performance than for the similar appearance case, but once again demonstrates that an upright descriptor may be robust in cases where the rotation offset is small, even when the environmental appearance has changed considerably.

In summary, these results demonstrate that a condition-robust descriptor such as HOG can be combined with a SURF orientation calculation to perform rotation-invariant feature matching in a visually changing environment. An important aspect to note, however, is the scale on the y-axis of each plot of Figure 5, particularly Figure 5b. At best, fewer than 300 inlier matches are found (using U-HOG) when the appearance of the environment has changed, even when the relative rotation between the two images is 0°. As the rotation offset increases, this number rapidly decreases to fewer than 30 for higher rotation offsets, for all the tested descriptors. As the number of features per image is 3000, in the simplest case only 10% of features are being matched, and in the more challenging cases it can drop below 1% of features, although R-HOG still manages to calculate the relative offset.

C. Dimensionality

The previous experiment demonstrated that HOG features outperformed SURF features in changing environments. However, the two features have substantially different dimensionality – SURF features possess 128 dimensions while the HOG features used above contain 7056 dimensions. This section investigates whether this significant disparity in dimensionality might have an effect on the resulting feature matching capability.

1) Setup: This experiment uses U-HOG and U-SURF and performs feature matching between the unrotated 6am and 7pm images, varying the dimensionality of the U-HOG feature by resizing the image box around each keypoint to different sizes. Note that the size of the region extracted from each image remains unchanged, and is determined by the scale of the SURF keypoint. However, by resizing the image box after extraction it is possible to artificially vary the HOG feature generated, without changing any of the default feature parameters. Performance is evaluated by calculating the median number of inliers over 100 RANSAC trials.

2) Results: Figure 7 shows the median number of inliers for HOG features with different numbers of dimensions. The performance of the HOG features can be seen to decrease rapidly as the number of dimensions decreases below 3000, although it can also be observed that the relationship between inlier number and dimensionality is not monotonic, but in fact decreases after approximately 3600 dimensions. The 128-dimension SURF features, marked in red on this plot, performs much more competitively when compared to HOG features of similar dimensionality.

This experiment does not imply that there are no differences between features of the same dimension. For example, the next section (Section IV-D) demonstrates that SURF features of dimension 128 and image patches of dimension 121 behave quite differently, with each performing better in a different context. However, it does suggest that higher-dimension features can in general can out-perform those with fewer dimensions, suggesting the likelihood of a trade-off between condition robustness and storage efficiency.

D. Appearance change and viewpoint change

The previous experiments used images that were captured from the same viewpoint, so that evaluation of the inlier matches could be easily performed. However, the real concern is the system performance when the viewpoint of the camera has changed. The final experiment demonstrates performance on a dataset that demonstrates both viewpoint change and appearance change. The image set was a set of images of Notre Dame Cathedral from the well-known Paris Buildings dataset [28] as shown in Figure 8. The images used represent a range of different viewpoint offsets and day-night appearances.

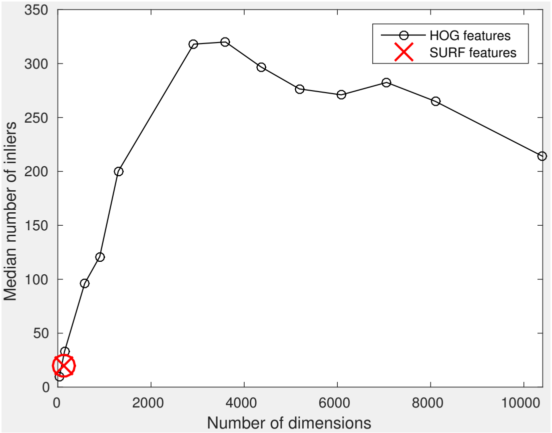


Fig. 7. Median number of inliers for HOG features (black circles) matched across appearance change on the Brisbane skyline images. The performance of the HOG features decreases rapidly as the number of dimensions decreases below 3000. The performance disparity between the SURF feature (red marker) and a HOG feature of a similar dimensionality is less pronounced than when compared to a HOG feature of larger dimensionality.

1) Setup: Once again the SURF feature detector was used, with 1000 features extracted per image. Four feature description techniques were tested – 128-dimension SURF features, image patches resized to 121 dimensions to provide a comparably sized feature to the 128-dimension SURF features, 7056-dimension HOG features as used in the previous experiments, and 64896-dimension `conv3` features were extracted as in [33] using the Caffe framework [12] and the default pre-trained ImageNet model [14]. In each case, the upright feature was used. Following [17], we learned a PCA decomposition for each feature type on a training image and post-processed the extracted features using Zero Components Analysis (ZCA) to

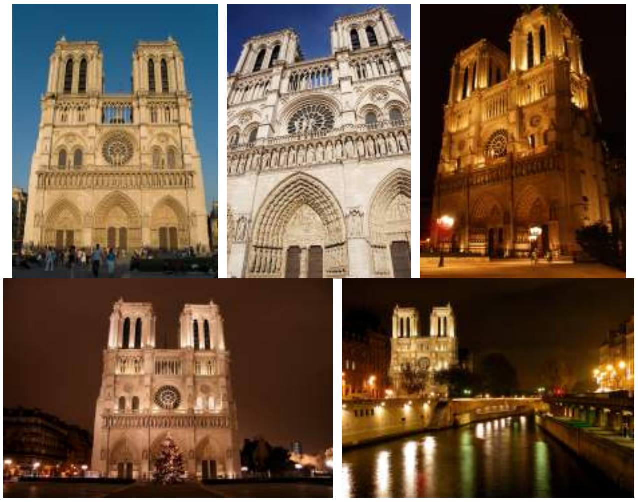


Fig. 8. Sample images of Notre Dame from the Paris image dataset, demonstrating a number of different viewpoints and day-night appearances.

improve condition robustness. If U are the PCA eigenvectors learned during training, and λ_i are the corresponding eigenvalues, then ZCA is calculated for the set of descriptors x from an image I via:

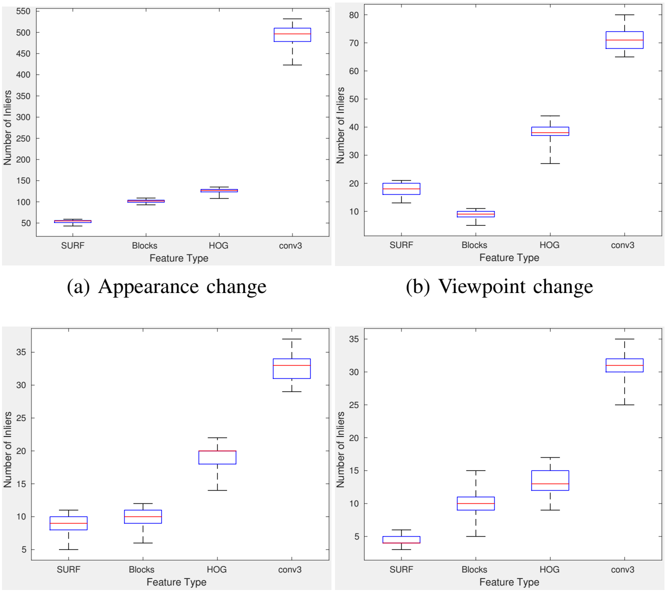
$$x_{\text{ZCA}} = \frac{U^T x}{\sqrt{\lambda_i + \epsilon}}. \tag{2}$$

The regularization parameter ϵ avoids division by zero, and a value of 10^{-3} was used. Results were evaluated using a combination of metrics. As in the previous experiments, the distribution of inlier matches over 100 RANSAC trials were calculated. However, in these cases there was no relative pose ground truth as the images were all taken from disparate viewpoints and the accuracy of the pose estimate could not be calculated in a simplified way. However, we hypothesized that a "good" set of feature matches would provide a more repeatable performance over the RANSAC trials, and thus while the ground truth accuracy of the transformation would be unknown, the variance in the distribution of pose estimate would give a qualitative estimate of the quality of the matching set.

2) Results: Figure 9 shows examples of calculated inlier sets for a number of images that demonstrate both viewpoint and appearance change, using SURF features, image patches, HOG features and `conv3` features as descriptors. All the features, including the SURF features, perform adequately in the case when there is only appearance change and minor scale change (first row of Figure 9). There are fewer SURF matches and image patch matches in the other test cases, while both HOG features and `conv3` features perform reasonably in all tested scenarios, particularly the `conv3` feature.

Figure 10 shows the distribution of inliers calculated for all features types. We note that the `conv3` feature outperforms the other three features in all cases, as suggested by Figure 9. The SURF feature performs worst in all three cases involving appearance change. When there is no appearance change but only viewpoint change (Figure 10b), SURF features outperform the block image patches, which are not suited for such transformations. Importantly, these results suggest that the condition-invariant HOG and `conv3` features also display some viewpoint invariance, which is important for feature matching for pose estimation.

As always, it is important to consider the scale on the y axis of each plot, as this shows that the different images are not all equally challenging. When there is only appearance change (Figure 10a), an average of 491 `conv3` inlier pairs are found out of 1000 features per image; that is, a correct match is found for nearly half of all features. When there is only viewpoint change (Figure 10b), the average is 71 `conv3` inlier pairs, while for the more challenging combinations of viewpoint and appearance change in Figures 10c and 10d the mean decreases to around 30 `conv3` inlier pairs; that is, only 3% of all features are matched.



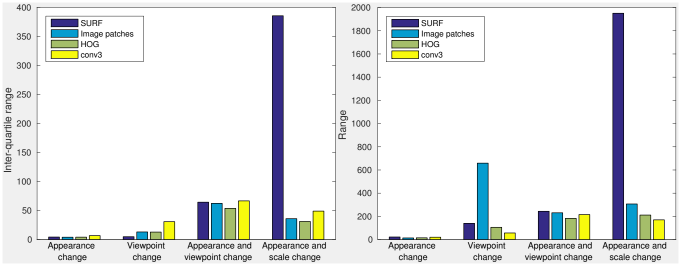
(a) Appearance change

(b) Viewpoint change

(c) Appearance and viewpoint change

(d) Appearance and scale change

Fig. 10. Distribution of inliers for all tested feature descriptors between test images displaying (a) appearance change only, (b) viewpoint change only, (c) appearance change and viewpoint change, and (d) appearance change and large scale change. SURF features perform worst when there is any appearance change, while image patches perform worst when there is viewpoint change only.



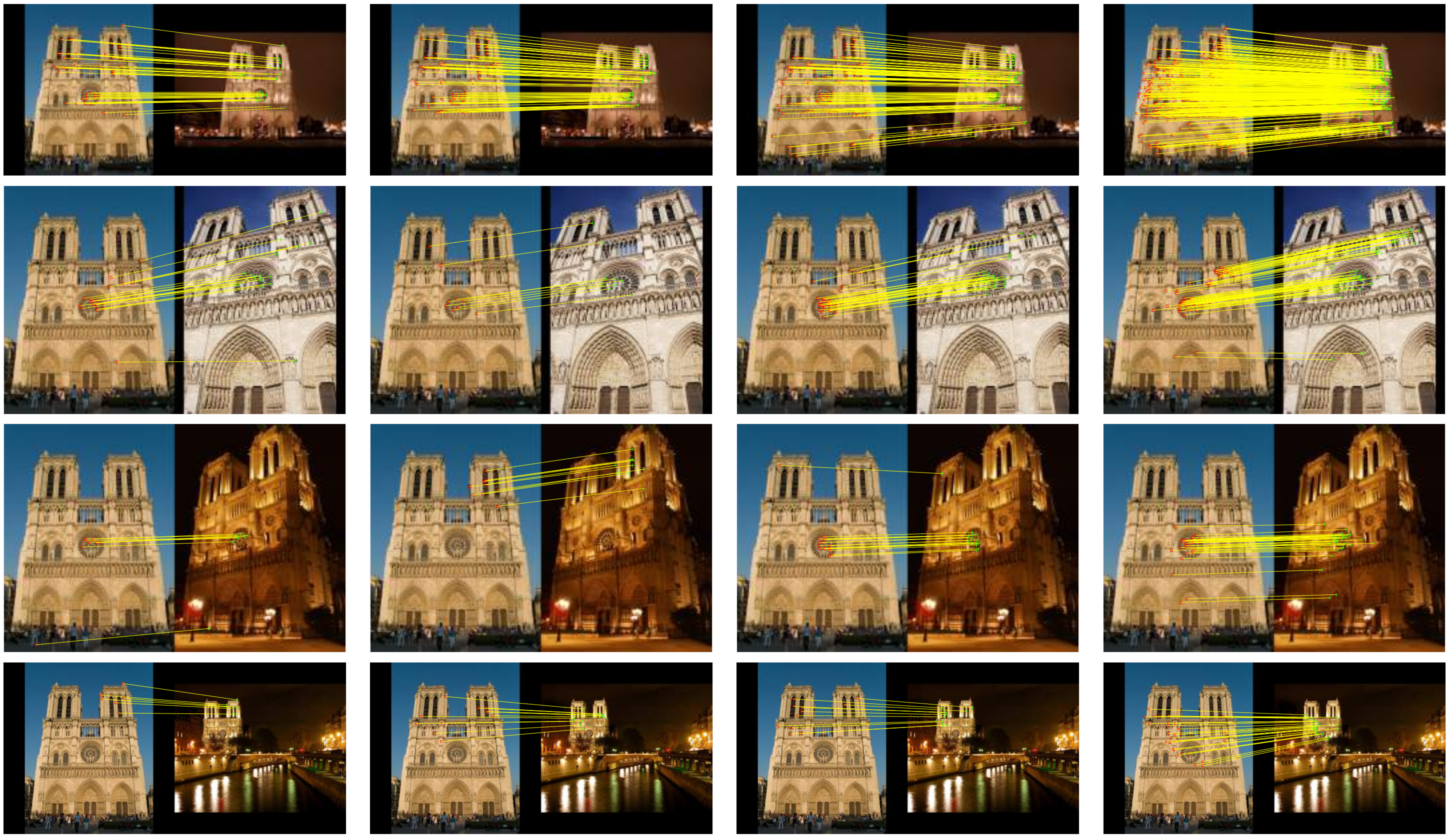
(a) Inter-quartile range

(b) Total range

Fig. 11. Distribution of transformation estimates for each tested feature descriptor and image tests across 100 RANSAC trials, summarized using the (a) inter-quartile range, and (b) total range. The widest distribution is for SURF features when there is appearance change and scale change (dark blue bars), while the image patches perform poorly when there is viewpoint change.

The number of inliers gives an impression of the quality of the match, as do the visual examples in Figure 9. To support this information, we consider the transformation estimate calculated for each image. Figure 11 displays the distribution of transformation estimate for each image type and each feature type across 100 RANSAC trials. Figure 11a displays the interquartile range (to eliminate outliers from the results) and Figure 11b displays the total range (incorporating outliers). These results are also summarized in Tables I and II. Although the ground truth estimate is unknown, we would expect that a successful feature set would compute the same transformation estimate more reliably than a poor quality set.

It is instructive to compare the results from Figure 11 to the inlier distributions from Figure 10. As expected, the most variable results are seen in Figure 11 when the median number of features is smaller than 10, namely SURF features



(a) SURF features (b) Image patches (c) HOG features (d) `conv3` features

Fig. 9. Examples of inlier matches for (a) SURF features, (b) image patches, (c) HOG features and (d) `conv3` features for a number of different appearance conditions and viewpoint positions. In general, the condition-invariant `conv3` features finds the most inlier matches, followed by the HOG features. All the features perform well on the first row comparison, despite the day-night appearance change. The image patches struggle when there is a viewpoint change, even when the conditions have not changed a great deal (second row). The SURF features struggle on the large scale change and appearance change on the final row.

TABLE I
INTER-QUARTILE RANGE OF TRANSFORMATION ESTIMATES IN PIXELS FOR EACH TESTED FEATURE DESCRIPTOR AND IMAGE TESTS ACROSS 100 RANSAC TRIALS.

	SURF features	Image patches	HOG features	`conv3` features
Appearance change	4.4	**4.1**	4.2	6.8
Viewpoint change	**5.0**	13.1	13.0	30.8
Appearance change and viewpoint change	64.4	62.3	**53.7**	66.7
Appearance change and scale change	385.5	36.0	**31.2**	49.0

TABLE II
RANGE OF TRANSFORMATION ESTIMATES IN PIXELS FOR EACH TESTED FEATURE DESCRIPTOR AND IMAGE TESTS ACROSS 100 RANSAC TRIALS.

	SURF features	Image patches	HOG features	`conv3` features
Appearance change	22.4	**14.3**	15.4	20.5
Viewpoint change	139.9	658.8	105.9	**57.1**
Appearance change and viewpoint change	243.9	231.1	**182.9**	215.7
Appearance change and scale change	1950.8	307.0	211.5	**170.2**

with appearance and scale change, and image patches with viewpoint change. For appearance change only, all the features perform well, with an inter-quartile range of between 4 and 7 pixels and a range of less than 23.

For viewpoint change only, SURF features achieve a top inter-quartile range of 5 pixels, compared to a range of 13 pixels for both image patches and HOG features, and 30 pixels for `conv3` features, although as mentioned above the image patches estimate has many extreme outliers and a range of 658 pixels. Curiously, the SURF features have a wider range than the `conv3` features. The ranges become increasingly wide as the images matching tasks becomes increasingly challenging. For the images that combine appearance and viewpoint or scale change, HOG features achieve the best inter-quartile range.

An interesting note is that the `conv3` feature does not have the narrowest inter-quartile range for any of the tested scenarios, but it does achieve a narrower range in two cases. This result suggests that `conv3` features are less accurate than the other features in terms of precise positioning, but may find a valid set of approximate inliers more reliably.

V. Conclusion

These results show that the combination of a local feature detector such as SURF with condition-invariant feature descriptors such as HOG or `conv3` provide a promising combination to estimate pose more precisely than a topological place recognition calculation, even when the appearance of the environment has changed considerably. This paper is by no means an exhaustive survey of feature types or appearance changes, but instead presents representative examples of some of the classes of possible tools that can be applied to the problem using example scenarios to motivate some of the challenges and potential solutions.

There are a number of issues to be resolved. In challenging conditions, even the most successful feature descriptors may only achieve inlier sets containing as few as 1% of all matches. As a result, many features need to be kept for each image, and there may be cases where even fewer good feature matches may be found, and so no accurate pose estimate can be calculated at all. There are also efficiency considerations – more features need greater storage capacity and comparison steps take longer. Furthermore, as Section IV-C suggests, features with more dimensions are likely to display greater condition robustness and so even greater efficiency sacrifices may be necessary to achieve condition-invariance. Once interesting avenue to pursue is to develop a more sophisticated way of selecting features, so fewer features are retained, but the ones that are kept are more likely to be useful for feature matching.

Future work includes developing a fully integrated visual localization system that performs both global place recognition and local pose estimation via a coarse-to-fine approach. For example, image feature detection using EdgeBoxes has been shown to provide good topological localization results in changing environments with viewpoint change [33], but as we show here it is outperformed in terms of local pose estimation by SURF features. These results display another trade-off in visual place recognition – in this case between condition-invariance and viewpoint-invariance in terms of feature size. Large features display condition-invariance to a greater degree than small features, but provide less flexibility for drastic viewpoint changes – whole-image descriptors being the most extreme example [2, 32]. Ideally, the system would use the same features for topological and metric localization, but such trade-offs mean a compromise approach may be necessary.

These results can be extended and combined with other techniques such as lidar-intensity images [20], illumination invariant images [19] and high dynamic range cameras, and future work is investigating these avenues. Other future work includes introducing stereo cameras to improve pose estimates, and extension of this research to yet more challenging environments and more extreme viewpoint change.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *Patt. Anal. Mach. Intell. IEEE Trans.*, 34(11):2274–2282, Nov 2012. ISSN 0162-8828.

[2] H. Badino, D. Huber, and T. Kanade. Real-time topometric localization. In *Robot. Autom. (ICRA 2012), IEEE Int. Conf.*, pages 1635–1642, May 2012.

[3] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110(3):346–359, June 2008. ISSN 1077-3142.

[4] T. Caselitz, B. Steder, M. Ruhnke, and W. Burgard. Matching geometry for long-term monocular camera localization. In *ICRA Workshop on AI for Long-term Autonomy*, Stockholm,Sweden, 2016.

[5] Mark Cummins and Paul Newman. FAB-MAP: Probabilistic localization and mapping in the space of appearance. *Int. J. Rob. Res.*, 27(6):647–665, 2008.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Comput. Vis. Patt. Recognit. (CVPR 2005), IEEE Comput. Soc. Conf.*, volume 1, pages 886–893 vol. 1, June 2005.

[7] A.J. Davison, I.D. Reid, N.D. Molton, and O. Stasse. MonoSLAM: Real-time single camera SLAM. *Patt. Anal. Mach. Intell. IEEE Trans.*, 29(6):1052–1067, June 2007. ISSN 0162-8828. doi: 10.1109/TPAMI.2007.1049.

[8] Ethan Eade and Tom Drummond. Edge landmarks in monocular SLAM. In *Br. Mach. Vis. Conf. (BMVC)*. BMVA Press, 2006.

[9] J. Engel, T. Schöps, and D. Cremers. LSD-SLAM: Large-scale direct monocular SLAM. In *Computer Vision – ECCV 2014*, 2014.

[10] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981. ISSN 0001-0782.

[11] Paul Furgale and Timothy Barfoot. Visual teach and

repeat for long-range rover autonomy. *J. F. Robot.*, 27 (5):534–560, 2010. ISSN 1556-4967.

[12] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[13] K. Konolige and M. Agrawal. FrameSLAM: From bundle adjustment to real-time visual mapping. *Robot. IEEE Trans.*, 24(5):1066–1077, Oct 2008. ISSN 1552-3098.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. ImageNet classification with deep convolutional neural networks. In *Neural Inf. Proc. Syst. (NIPS)*, pages 1097–1105. 2012.

[15] S. Leutenegger, M. Chli, and R. Y. Siegwart. Brisk: Binary robust invariant scalable keypoints. In *2011 International Conference on Computer Vision*, pages 2548–2555, Nov 2011.

[16] D. Lowe. Object recognition from local scale-invariant features. In *Comput. Vis. (ICCV), 1999 IEEE Int. Conf.*, volume 2, pages 1150–1157 vol.2, 1999.

[17] S. Lowry and M. J. Milford. Supervised and unsupervised linear learning techniques for visual place recognition in changing environments. *IEEE Transactions on Robotics*, PP(99):1–14, 2016. ISSN 1552-3098.

[18] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford. Visual place recognition: A survey. *IEEE Transactions on Robotics*, 32(1):1–19, Feb 2016. ISSN 1552-3098.

[19] Will Maddern, Alex Stewart, Colin McManus, Ben Upcroft, Winston Churchill, and Paul Newman. 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 (ICRA)*, Hong Kong, China, May 2014.

[20] Colin McManus, Paul Furgale, and Timothy Barfoot. Towards lighting-invariant visual navigation: An appearance-based approach using scanning laser-rangefinders. *Robot. Auton. Syst.*, 61(8):836 – 852, 2013. ISSN 0921-8890.

[21] Colin McManus, Ben Upcroft, and Paul Newmann. Scene signatures: Localised and point-less features for localisation. In *Robot. Sci. Syst.*, Berkeley, USA, July 2014.

[22] M. Milford and G. Wyeth. SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. In *Robot. Autom. (ICRA 2012), IEEE Int. Conf.*, pages 1643–1649, May 2012.

[23] R. Mur-Artal, J. M. M. Montiel, and J. D. Tards. Orb-slam: A versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5):1147–1163, Oct 2015. ISSN 1552-3098.

[24] Tayyab Naseer, Luciano Spinello, Wolfram Burgard, and Cyrill Stachniss. Robust visual robot localization across seasons using network flows. In *Artif. Intell. (AAAI), Nat. Conf.*, 2014.

[25] P. Neubert and P. Protzel. Local region detector + cnn based landmarks for practical place recognition in changing environments. In *Mobile Robots (ECMR), 2015 European Conference on*, pages 1–6, Sept 2015.

[26] P. Neubert and P. Protzel. Beyond holistic descriptors, keypoints, and fixed patches: Multiscale superpixel grids for place recognition in changing environments. *IEEE Robotics and Automation Letters*, 1(1):484–491, Jan 2016. ISSN 2377-3766.

[27] Stephen Nuske, Jonathan Roberts, and Gordon Wyeth. Robust outdoor visual localization using a three-dimensional-edge map. *J. F. Robot.*, 26(9):728–756, 2009. ISSN 1556-4967. doi: 10.1002/rob.20306.

[28] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[29] P. Ross, A. English, D. Ball, and P. Corke. A method to quantify a descriptor's illumination variance. In *Robot. Autom. (ACRA 2014), Australas. Conf.*, 2014.

[30] Edward Rosten and Tom Drummond. Machine learning for high-speed corner detection. In *Computer Vision – ECCV 2006*, 2006. ISBN 978-3-540-33832-1. doi: 10.1007/11744023_34.

[31] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford. On the performance of ConvNet features for place recognition. In *Intell. Robot. Syst. (IROS 2015), IEEE/RSJ Int. Conf.*, pages 4297–4304, Sept 2015.

[32] Niko Sünderhauf, Peer Neubert, and Peter Protzel. Are we there yet? Challenging SeqSLAM on a 3000 km journey across all four seasons. In *ICRA Work. Long-Term Autonomy*, Karlsruhe,Germany, 2013.

[33] Niko Sünderhauf, Sareh Shirazi, Adam Jacobson, Feras Dayoub, Edward Pepperell, Ben Upcroft, and Michael Milford. Place recognition with ConvNet landmarks: Viewpoint-robust, condition-robust, training-free. In *Robot. Sci. Syst.*, Rome, Italy, July 2015.

[34] Christoffer Valgren and Achim Lilienthal. SIFT, SURF & seasons: Appearance-based long-term localization in outdoor environments. *Robot. Auton. Syst.*, 58(2):149–156, 2010.

[35] Brian Williams, Mark Cummins, José Neira, Paul Newman, Ian Reid, and Juan Tardós. A comparison of loop closing techniques in monocular slam. *Robot. Auton. Syst.*, 57(12):1188–1197, December 2009. ISSN 0921-8890.

[36] C. Zitnick and P. Dollár. Edge Boxes: Locating object proposals from edges. In *Computer Vision – ECCV 2014*, 2014.