2018 IEEE International Conference on Robotics and Automation (ICRA)
May 21-25, 2018, Brisbane, Australia

Learning Place-and-Time-Dependent Binary Descriptors for Long-Term Visual Localization

Nan Zhang, Michael Warren, and Timothy D. Barfoot[1]

***Abstract*— Vision-based navigation is extremely susceptible to natural scene changes. This can result in localization failures in less than a few hours after map creation. To combat short-term illumination changes as well as long-term seasonal variations, we propose using a place-and-time-dependent binary descriptor that adapts to different scenarios in an online fashion. This is achieved by extending the GRIEF [6] evolution algorithm in two ways: correspondence generation using a known pose change and the inclusion of LATCH triplets in addition to BRIEF comparisons for descriptor generation. We show the adaptive descriptor outperforms a single descriptor scheme for localization within a single-experience Visual Teach and Repeat (VT&R) system while maintaining the efficiency of binary descriptors. By adapting the description function to different environmental conditions, it allows the system to operate for a longer period before a new experience is required. In the presence of extreme illumination changes from day to night, we obtain 40% more inlier matches compared to SURF. In the case of seasonal variations, a 70% increase is demonstrated. The increased correspondences result in more localizable sections along the paths, amounting to a 25% and 150% increase in the lighting and seasonal cases, respectively.**

I. INTRODUCTION

In the area of long-term visual navigation, a fundamental problem is localizing over time in the presence of natural scene changes as a result of illumination, seasonal, and weather variations. Light Detection and Ranging (LiDAR) systems overcome this limitation, but they are still relatively expensive and require large payload capacities not practical on mass-restricted systems.

Vision-based metric localization can be achieved by matching point features in images taken at different times and computing the relative pose change of the camera. To obtain these point features, a detection scheme is used to find the most salient points in the environment. These points should ideally correspond to the same triangulated landmarks in the environment irrespective of illumination or viewpoint changes. The information around these points can be summarized with a description function and then matched using a distance function. The inlier matches can then be used to estimate the six-degree-of-freedom (6DoF) transformation between the two camera positions.

Typically, visual descriptors are developed as one-size-fits-all methods of matching, with the goal of making a descriptor as generally applicable as possible. We seek to take a tangential approach: tuning descriptors at increasing levels of specificity to a particular location and time (see Figure 1). This is similar to the place-dependent features presented by McManus *et al.* [12] and Linegar *et al.* [8]. However, instead of training support vector machines (SVM) for each landmark, we use traditional binary descriptors. An evolutionary algorithm based on Generated BRIEF (GRIEF) [6] is used to learn an environment-dependent function for generating the descriptor. This allows the description function to be adapted to the appearance of the environment, tailoring it specific scenes.

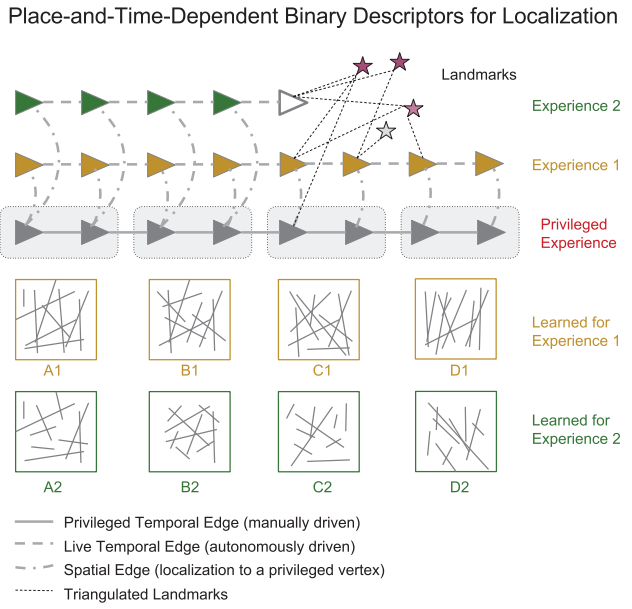

Fig. 1. Cartoon illustration of a place-and-time-dependent feature description scheme that adapts the matching function (A, B, C, D) to a segment of the path at a certain time using binary descriptors. The vertices (triangles) represent keyframes recorded during a traversal. They are connected to each other by spatial or temporal edges containing the estimated pose. The privileged experience is the manually driven path determined to be safe by the operator. The live experience is collected during autonomous repeats. The descriptors can be trained using either only the privileged experience or multiple experiences. Correspondences generated using the adaptive descriptor results in longer and improved localization performance in the presence of scene changes.

Binary descriptors such as BRIEF [1], ORB [18], and BRISK [7] are computed by comparing the intensity values at various positions within a patch around the image feature. The number of possible positions for such a computation can be substantial, especially for larger image patches. The authors of BRIEF drew positions from common distributions and chose the best ones. The authors of ORB chose com-

[1]All authors are with the University of Toronto Institute for Aerospace Studies (UTIAS), University of Toronto, 4925 Dufferin St, Ontario, Canada {nan.zhang, michaelwarren}@robotics.utias.utoronto.ca, tim.barfoot@utoronto.ca

parisons with high variance. BRISK uses a pattern that is composed of concentric rings. The sampling strategy has a significant effect on the result of matching, and so GRIEF [6] was devised to find the best positions within a patch given pre-labeled data.

The environmental-dependence aspect of the proposed descriptor makes it the ideal candidate for localization within a Visual Teach and Repeat (VT&R) framework [5]. The goal is to follow the original commanded path as closely as possible. This means the viewpoint change during successive repeats is expected to be relatively small. We can also leverage the various repeats over the same path as training data to improve the description function. After each repeat, the evolutionary algorithm can be applied to evolve the description function for localization in future repeats.

This work extends GRIEF [6] and is inspired by place-dependent features [11], [8]. Unlike in [6], we estimate the full 6DoF pose of the vehicle rather than only the heading error. Ultimately, this work demonstrates:

- An unsupervised learning scheme for data labeling that can be extended to other classes of description functions
- Adapting the description function to the environment leads to improved matching performance compared to a fixed matching function
- Increased operation time and localization success rate of single experience VT&R

In section III we describe the method for generating labeled data under the VT&R framework. We also present the evolutionary algorithm used for adapting the descriptor pattern. Using this process, we evaluate our method on two different datasets in section IV: *In The Dark* and *UTIAS Snow*, both collected at University of Toronto Institute of Aerospace Studies (UTIAS). The first deals with illumination change over a 24 hour period. The second deals with seasonal variations from fall to spring.

II. RELATED WORK

Localization problems can be grouped based on how precise the desired result is: metric localization and topological localization. The approach of interest is the former, which produces an estimated state and uncertainty relative to some internal representation of the environment for the purpose of visual route following. Through the use of triangulated point features in the environment, an estimated pose can be obtained by solving the Perspective-n-Point (PnP) problem. Other approaches use higher dimensional constructs such as lines [17], and objects (SLAM++) [19] for localization in the same manner.

Valgren *et al.* [22] examined the use of SIFT and SURF descriptors for long-term navigation. They conclude that U-SURF resulted in the best performance, but ultimately using local feature matching alone is not sufficient for cross-seasonal metric localization. It is difficult to deal with seasonal changes, but different techniques have been explored to deal with illumination and shadows. Techniques such as illumination-invariant images [10] and colour-constant images [9] result in more stable but often a smaller number of feature matches. Other image processing techniques such as contrast limited adaptive histogram equalization (CLAHE) [23] creates an order of magnitude more matches by bringing out more details in the image.

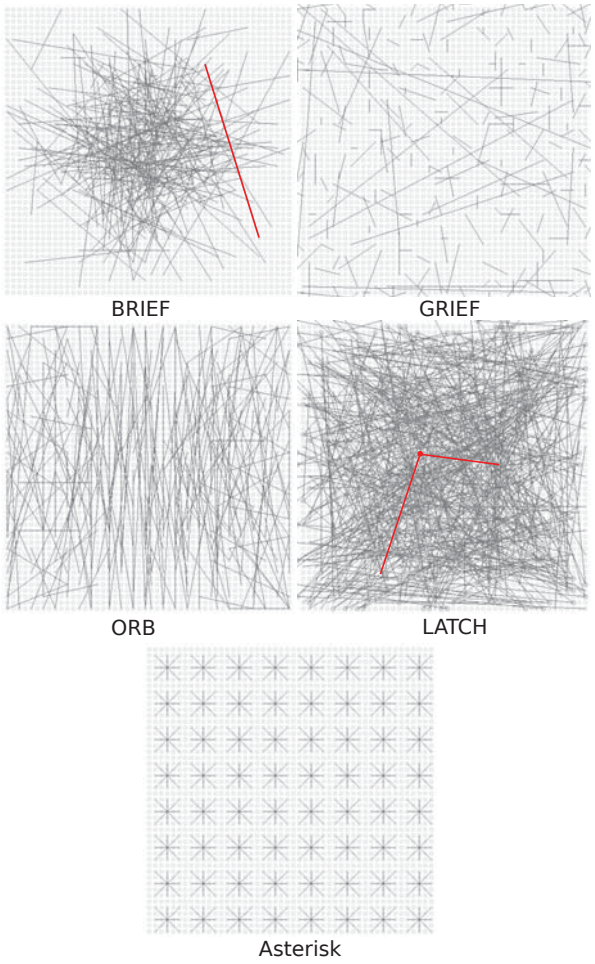

Fig. 2. The patterns used for BRIEF, ORB, GRIEF, and LATCH. For the BRIEF inspired descriptors, the pixel comparisons are displayed as a line connecting the pixel being compared. For LATCH, the positions of the three sub-patches are connected using two lines with the center position denoted using a bold circle. An example of one comparison in each case is highlighted in red. A hand-crafted descriptor (Asterisk) inspired by SURF and the results from GRIEF is also shown.

Different techniques have been applied for learning better visual descriptors to improve their performance. Two examples are: convex optimization [21] and convolution neural networks (CNN) [2], [20]. These learned descriptors are quite robust to lighting, viewpoint, deformations, and small seasonal changes. The only drawback is that they are much more computationally intensive than binary descriptors.

All of the above approaches rely upon little-to-no physical changes occurring between autonomous navigation and map creation. Milford *et al.* [13] proposed the use a sequence of images to localize rather a single one. Neubert *et al.* [14] proposed a method of predicting the appearance change and matching the predicted image against the live images. While effective, these systems are not able to provide the precision required for vision-in-the-loop navigation.

Dayoub *et al.* [4] proposed a system that employs the idea of short and long-term memory to forget old features and add new ones. Churchill *et al.* [3] introduces the notion of saving multiple experiences in problematic areas and localizing to them all in parallel. Multi-experience VT&R [15] similarly uses bridging experiences to overcome the presence of natural scene changes. Every time a vehicle drives through the environment, a completely new map is generated with respect to the original or privileged experience. This allows the system to match to any of the stored experiences. It has been demonstrated to work across seasons from fall to winter and into the spring time [16].

This work differs from the above methods by tailoring the description function to the environment. We hypothesize that adapting the description function to the environment leads to improved localization performance. This is similar in principle to using individual SVMs for each landmark, which proves to be quite robust [8]. However, we demonstrate this within a traditional feature-based navigation system with binary descriptors.

III. METHODOLOGY

This work is presented within the VT&R system, specifically the descriptor matching portion of the localization subsystem. As the name suggests, there are two components to VT&R: teach and repeat. During the teach phase, a user commands a vehicle through the environment. A map is built using stereo visual odometry (VO) and stored in the form of a spatial-temporal pose graph [15] (see Figure 1). Windowed bundle adjustment (BA) is performed periodically to optimize the landmark positions and vehicle positions. Each vertex in the graph corresponds to a keyframe containing all the observed landmarks. The edges contain the estimated transformations and uncertainties between vertices. During the repeat phase, newly observed landmarks are matched against the map and passed through random sample consensus (RANSAC) to obtain a pose estimate.

To isolate the performance of using different description functions along the path, we only localize back to the privileged experience. This is reflective of situations such as GPS-denied emergency return of unmanned aerial vehicles (UAV) where the scene change can be dramatic less than an hour after the original pass. It could also be beneficial in scenarios where frequent traversal of the path is difficult to achieve. The proposed scheme should result in a more robust visual-based localization system that can wait longer periods of time before a new experience is required. Ultimately, this can be combined with multi-experience localization (MEL) to reduce the storage and computation cost of the system.

VT&R normally uses GPU-accelerated SURF descriptors and detectors for both visual odometry (VO) as well as localization. For consistency, we maintain the use of SURF for VO, but localization is performed using the proposed environment-dependent binary descriptor. We keep the same detections from SURF for localization but re-compute the descriptors. The low computational time of binary descriptors makes it possible to achieve real-time performance.

Taking the environment-dependence idea to the extreme, a unique pattern can be used at every keyframe. We stop at the keyframe level, but one can extend this method for generating a unique pattern for parts of an image or even every landmark. This would require a change in the matching framework and could be explored in future work. Practically, learning a different descriptor for every keyframe leads to poor performance due to the small amount of training data that is available.

A. Descriptor Computation

Given an image $\mathbf{I}$ and a keypoint of interest at $\mathbf{x}_k$, the ith bit of the descriptor can be computed from either a BRIEF comparison (1) or a LATCH comparison (2). Like Calonder *et al.* [1], we maintain a 256-bit descriptor using a fixed 48×48 pixel patch computed after applying a 9×9 box filter on the image. Each bit of the BRIEF descriptor results from an intensity comparison of two points $(\mathbf{x}_a, \mathbf{x}_b)$ with the center of the patch as the origin. Similarly, each LATCH comparison results from a comparison of the Frobenius norm between three sub-patches of size $S \times S$ pixels centered around the points $(\mathbf{x}_a, \mathbf{x}_b, \mathbf{x}_c)$. For simplicity, we take the value of S to be unity as it improves the run time efficiency of the descriptor without sacrificing much performance. The comparisons are of the following form:

$$b^i_{\text{brief}}(\mathbf{I}, \mathbf{x}_k) = \mathbf{I}(\mathbf{x}_k + \mathbf{x}_a) > \mathbf{I}(\mathbf{x}_k + \mathbf{x}_b) \tag{1}$$

$$\begin{aligned} b^i_{\text{latch}}(\mathbf{I}, \mathbf{x}_k) &= \|\mathbf{I}(\mathbf{x}_k + \mathbf{x}_a) - \mathbf{I}(\mathbf{x}_k + \mathbf{x}_b)\| \\ &> \|\mathbf{I}(\mathbf{x}_k + \mathbf{x}_c) - \mathbf{I}(\mathbf{x}_k + \mathbf{x}_b)\| \end{aligned} \tag{2}$$

The intensity information varies considerably with natural scene changes. The 'gradient information' used by BRIEF and ORB is robust to some of these changes. It is reasonable to assume the 'Hessian information' used by LATCH should be more robust. In evolutionary terms, we introduce a new species into the gene pool and allow them to compete for survival. The locations of the comparisons can be thought of as the entire gene pool.

B. Data Labeling

Given only a single experience, VO matches can be used to evolve the descriptor. With multiple experiences, localization matches can also be incorporated. Both positive, S_p, and negative, S_n, correspondences are important in the evolutionary process. To obtain the set S_p, we used the estimated 6DoF pose of the vehicle, $\mathbf{T}_{ab}$, relative to an earlier vertex, and transform all the landmarks in homogeneous coordinates, $\mathbf{p}$, back into the map frame,

$$\mathbf{p}' = \mathbf{T}_{sv}\mathbf{T}_{ab}\mathbf{T}^{-1}_{sv}\mathbf{p} \tag{3}$$

and then reproject them into the image plane. The transform from the vehicle frame to sensor frame is given by $\mathbf{T}_{sv}$. Any reprojected landmarks, $\mathbf{p}'$, that fall within 3 pixels of a map feature are labelled as a correspondence. These geometric correspondences, S_p, are the set of all possible matches that should have occurred given an ideal description function.

Fig. 3. (Left) Sample images for the *In The Dark* dataset. Each row shows the same location at various times during a 24-hour cycle. We see the presence of large shadows, lens flares, and poorly illuminated scenes. 20 repeats are used to validate the environment-dependent descriptor with at most half of them being used for training and the other half for testing. (Right) An aerial view of the path traversed for the *In The Dark* dataset. Each repeat totals to about 250 meters of driving around the UTIAS Dome. The first half of the path is over paved roads and the second half over grass. This path was driven approximately every hour over a span of 24 hours using multi-experience VT&R.

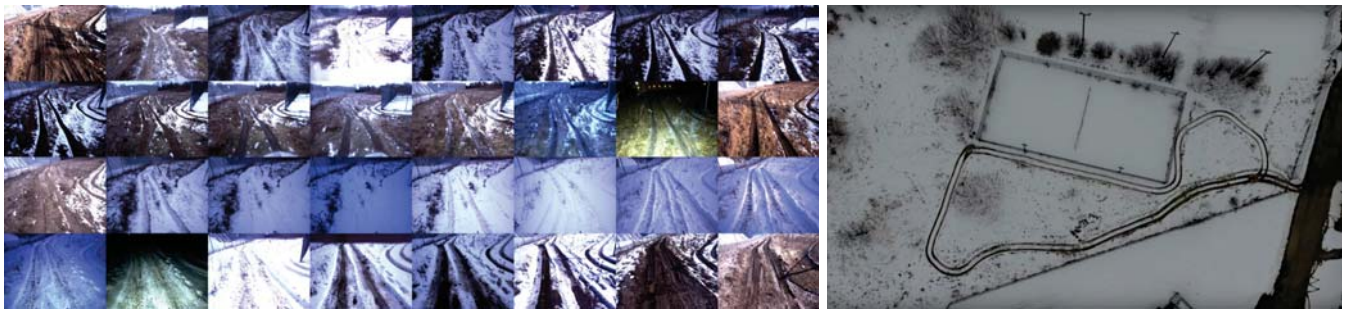

Fig. 4. (Left) Sample images for the *UTIAS Snow* dataset. All the images show the same location at various times during the data collection process. The proposed system fails to localize when the snow completely covers the ground. About 50 repeats are used to validate the environment-dependent descriptor with at most half of them being used for training and the other half for testing. (Right) An aerial view of the path traversed for the *UTIAS Snow* dataset. Each repeat totals to about 250 meters through tall grass and rough terrain beside the tennis court at UTIAS. This path was driven at regular intervals from late January into early May using multi-experience VT&R.

Next, we match the descriptors between the live and map images using Hamming distance. This set includes both true positive, D_{tp}, and false positive, D_{fp} matches. We can also obtain false negatives, D_{fn}, by finding elements in S_p, but not in D_{tp}. The set, S_n, is essentially equal to D_{fp}. The true negative, D_{tn}, should not matter as they do not affect the matching performance. Usually, there are far more elements in S_n compared to S_p. We find it is better to keep the two sets in roughly equal proportion, so the effect of negative correspondences does not overpower the correct correspondences.

C. Evolutionary Algorithm

The process of evolving the descriptor uses the genetic algorithm described in [6]. The one addition is that we filter the set, D_n, so that it is equal in size to D_p. This balances out the evolution so that it converges faster. The fitness of the ith comparison is calculated based on the sets, S_p and S_n, given in (4). The fitness score and inlier matches are shown in Figure 6 along with a visualization of the descriptor patterns during the evolution. The fitness score is important as it allows us to determine which comparisons positively contribute to the true positive matches and negatively to the false positives matches. The expectation is that as total fitness increases, the number of matches should also increase. This is true when the minimum matching threshold is set to a reasonable value. This is why we base the convergence criteria on the number of true positive matches instead of the fitness score:

$$f_i(S_p, S_n) = \sum_{S_p}(1 - 2d_i) + \sum_{S_n}(2d_i - 1) \tag{4}$$

$$d_i = \begin{cases} 0, & \text{if } b^i = b'^i \\ 1, & \text{otherwise} \end{cases} \tag{5}$$

The evolutionary algorithm is as follows:

1) Compute all the descriptor matches from map images to live images using the current pattern
2) Re-project all live landmarks into map images using estimated transforms
3) Generate $D_{tp}, D_{tn}, D_{fp}, D_{fp}$ using geometric matches and descriptor matches
4) Add D_{tp} and D_{fn} into the set S_p, and D_{fp} into S_n
5) Filter the set S_n using a minimum matching threshold, then randomly sample it so that it is equal in size to the set S_p
6) Compute the fitness of each comparison
7) Replace the worst 20% of comparisons drawn from an uniform distribution with equal probability of either a BRIEF or LATCH comparision

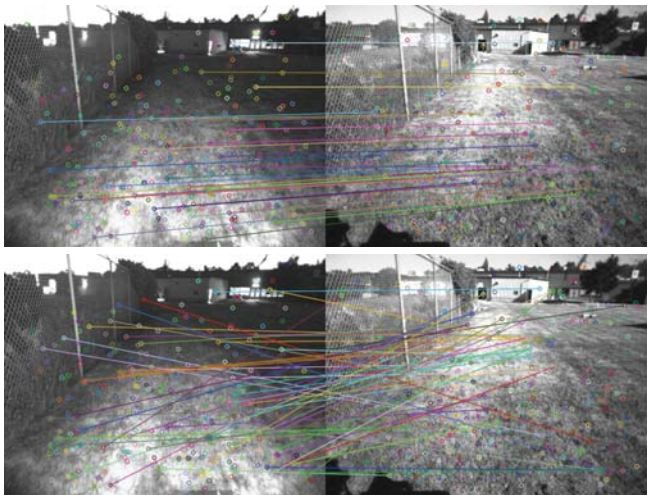

Fig. 5. The top image shows all the possible landmark correspondences, S_p. The bottom image shows the correspondences generated using the descriptor containing the sets D_{tp} and D_{fp}. These labels can be used in the evolutionary algorithm to maximize the total fitness and therefore the number of elements in D_{tp}.

8) Repeat until number of true positive matches converges or for a set number of iterations

We initialize with a random pattern drawn from a uniform distribution. Using a pre-trained pattern could lead to faster convergence. The comparison pattern for the descriptor is evolved offline using the above algorithm and written back into the corresponding vertex in the graph.

The training process takes a few minutes using an Intel i7-3720QM without any multi-threading or GPU acceleration. The maximum number of iterations is limited to 200 from experimentation, and we terminate if the number of correct matches stops increasing for 10 iterations. The authors of GRIEF trained their pattern for an hour. Presumably, we could have achieved slightly better results if we allow the algorithm to run for a longer period but this has diminishing returns. The fitness score and inlier matches are shown in Figure 6 along with a visualization of the descriptor pattern.

D. Datasets

In The Dark deals with illumination changes and *UTIAS Snow* deals with seasonal changes. Both datasets were collected using the Clearpath Grizzly rover shown in Figure 11 at UTIAS. For *In The Dark*, the Grizzly was driven over the path shown in Figure 3 20 times over a period of 24 hours at approximately equal intervals. This totals to about 5 km of driving over both paved roads as well as grass.

For *UTIAS Snow*, the Grizzly was driven over the path shown in Figure 4 over 100 times from late January into early May. Only the first 50 experiences are examined as single experience localization fails past that point. Without the intermediate bridging experiences, the scene change becomes too drastic for proper landmark correspondence. This dataset is entirely over grass, but some buildings are visible. In both cases, half of the experiences are used for training E_{tr}, and the other half for testing E_{te}. Both datasets were collected autonomously using the multi-experience VT&R system as presented in [15].

IV. RESULTS

Due to differences between binary descriptors and floating-point descriptors such as SURF, we must impose different minimum matching thresholds before RANSAC. We experimentally determine the optimal thresholds that produce the most matches for both classes of descriptors. For binary descriptors, we assign a max threshold of 0.3 and SURF a value of 0.12. For binary descriptors the value is computed as the fraction of bits that differ to the total number of bits. For SURF, it is calculated by subtracting the cosine distance from one.

We try three different schemes for training: using VO matches from the privileged experience (*pe*), using a temporally close experience from earlier in time (*se*), and using all experiences from the training set (*ae*). For each of these schemes, we also try learning a single pattern over the entire path (*s*) and learning a different one every 15 meters (*m*). This was chosen arbitrarily and splits the paths into 16 sections. Together, this creates six different scenarios: *pe-s*, *pe-m*, *se-s*, *se-m*, *ae-s*, *ae-m*. The descriptor patterns are evolved using localization results from multi-experience VT&R in each scenario.

As an example, for *In The Dark*, the privileged (teach) experience can be considered to be *exp0*. We refer to the 20 repeats as: *exp1, exp2, . . . , exp20* in chronological order. The testing set, E_{te}, and training set, E_{tr}, correspond to the odd numbered and even numbered experiences. This means both sets contain the full 24 hours of illumination changes. We give an example of how the adaptive description pattern is generated in each case:

- pe : train on *exp0*, test on E_{te}
- se : train on *exp1*, test on *exp2*
- se : . . .
- se : train on *exp19*, test on *exp20*
- ae : train on E_{tr}, test on E_{te}

To obtain a baseline for comparison, we use the SURF descriptor and try to localize all the repeats from the test set (E_{te}) back to the privileged experience. We also do the same for other common binary descriptors such as ORB, BRIEF, and LATCH. A random pattern (*rand*) generated using a uniform distribution and the pattern that was trained in [6] (*grief*) are also tested. Finally, a hand-crafted pattern inspired by SURF is also examined (*asterisk*) (see Figure 2). All these descriptors are compared using the six schemes noted above for both datasets.

The upper plot in Figure 7 shows the percentage of post-RANSAC inlier matches for each of the 10 test experiences (E_{te}) back to the map. These values are normalized based on the total number of landmarks saved during the privileged experience. The bottom plot in Figure 7 shows the fraction of vertices that are successfully localized. Success is defined as greater than 10 matches at a vertex.

The percentage of landmarks that can be matched drops to below 40% when repeating immediately after the teach

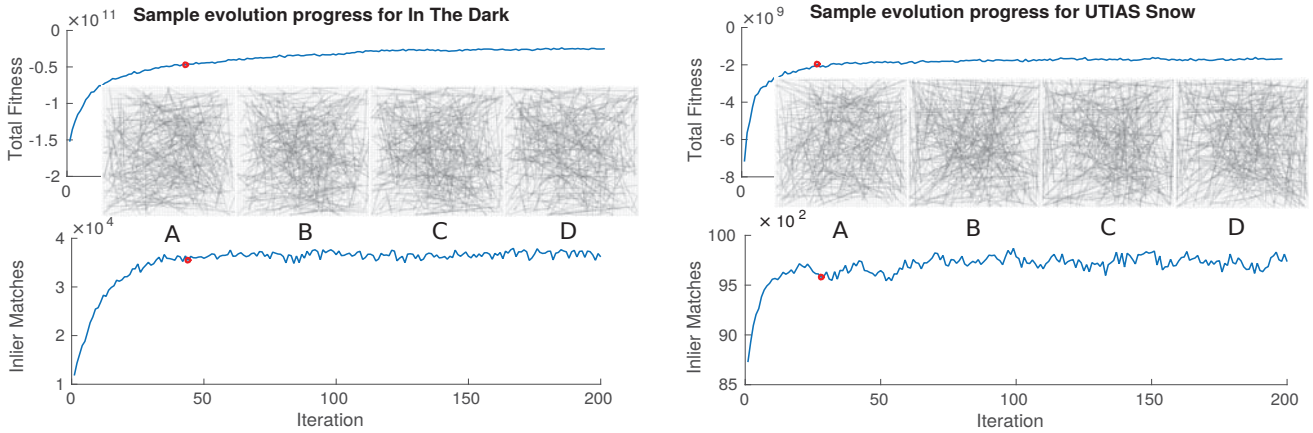

Fig. 6. The evolution of the descriptor pattern over time for *In The Dark* and *UTIAS Snow* over 200 iterations. The total fitness asymptotically converges in both cases. The red dots denote the point at which the evolution process would be normally terminated and saved to the graph.

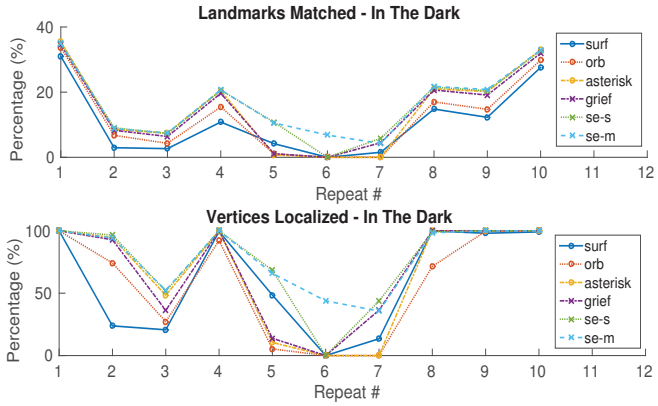

Fig. 7. Localization results of *In The Dark* in chronological order. The top plot shows the percentage of landmarks from the privileged experience successfully matched over 10 repeats using each descriptor. The time difference between the repeats is approximately 2 hours. The bottom plot shows the percentage of vertices that were successfully localized (more than 10 matches). Only a subset of the relevant descriptors is shown.

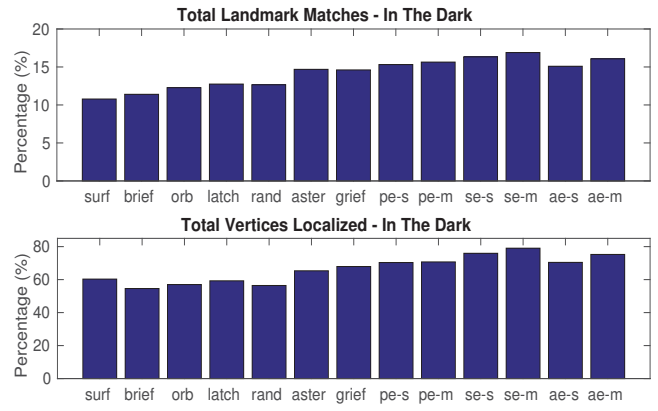

Fig. 8. Localization results of *In The Dark* over 10 repeats. The top plot shows the total percentage of landmarks matched to the privileged experience. The bottom plot shows the total percentage of vertices localized. The evolved descriptors out perform other methods with multiple descriptors learned from similar experiences with *se-m* resulting in the best performance.

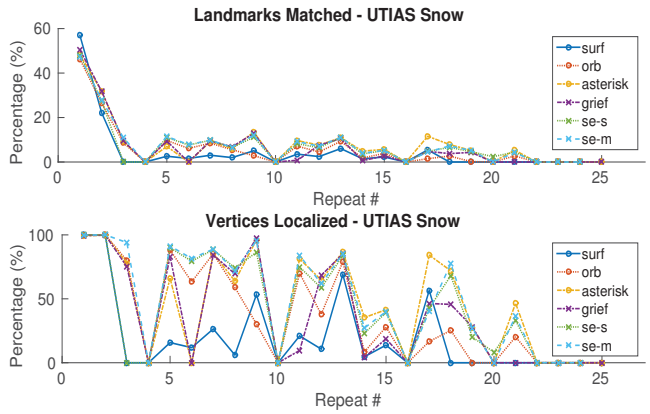

Fig. 9. Localization results of *UTIAS Snow* in chronological order. The top plot shows the percentage of landmarks from the privileged experience successfully matched over 25 repeats using each descriptor. The time difference between the repeats is approximately every 2-3 days. The bottom plot shows the percentage of vertices that were successfully localized.

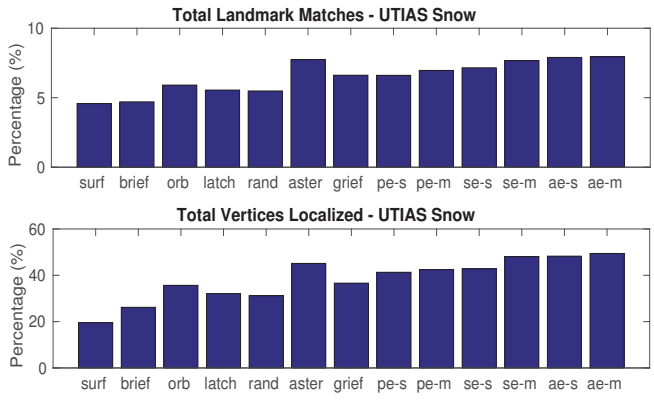

Fig. 10. Localization results of *UTIAS Snow* over 25 repeats. The top plot shows the total percentage of landmarks matched to the privileged experience. The bottom plot shows the total percentage of vertices localized. The evolved descriptors out perform other methods with multiple descriptors learned from all experiences *ae-m* resulting in the best performance.

experience. This means the majority of stored landmarks will never get matched, either because the feature detector is unable to detect them again or they are not distinctive enough. A feature detector that can consistently produce the same detections is crucial to localization performance.

During repeats 2 and 3, the number of matches drops due to the presence of long shadows and lens flares. Repeats 5, 6, and 7 correspond to nighttime repeats. Coming back to the same time the next day, the number of matches increases back to around 40%. It is important to note that the only scheme that produces matches during repeat 6 is *se-m*, corresponding to using different description functions along the path trained using visually similar experiences.

Examining the results at a higher level in Figure 8, we see the learned descriptors outperform the traditional descriptors, increasing the percentage of localizable vertices from around 60% to 75%. Using multiple descriptors along the path results in slightly improved localization results across the board, *pe-m, se-m, ae-m*. By changing the comparison patterns along the path, it restricts the range of the description function, making it more discriminative to the visual information at specific locations. As expected, training the descriptor using visually similar experiences results in the best performance (*se-m*).

Notably, training using only the privileged experience produces a similar matching performance to training using all experiences. In this case, it means the evolution is mainly increasing the robustness of the descriptor to viewpoint changes. Binary descriptors effectively handle large illumination change by design.

Similar improvements are seen for the *UTIAS Snow* dataset, shown in Figure 9 and 10. The effect of using similar experiences for training is less effective than the other schemes compared to the results obtained with *In The Dark*. This is likely due to the substantial physical changes in the environment during successive experiences. By training on specific experiences, the evolutionary algorithm allows the description function to over-fit to the location. This is not a problem for illumination changes due to the robustness of binary descriptors in that particular case. However as the scene physically changes, this over-fitting becomes problematic.

Compared to SURF, the percentage of localizable path increases from 20% to close to 50%. The number of matches increases by more than 70%. The performance fluctuates as snow falls and melts. In repeats 4, 10, and 16, localization fails over the entire path. A significant amount of snowfall was accumulated after repeat 22 and the system was no longer able to localize.

It is interesting to note the GRIEF pattern shown in Figure 2 does extremely well in our dataset. This may be attributed to the shorter comparison patterns that were observed by the authors of GRIEF. Motivated by this and taking inspiration from the sub-regions used in SURF we create a hand-crafted binary pattern called Asterisk (see Figure 2). It performs exceptionally well on the snow dataset coming very close to the performance of the evolved descriptors. It can be used as an initial pattern for evolution or simply as is.

Fig. 11. The Clearpath Grizzly rover fitted with a Bumblebee XB3 stereo camera. The stereo images are logged at 10 Hz for both the *In The Dark* and *UTIAS Snow* datasets. Multi-experience localization is used to establish data correspondence.

This demonstrates that certain patterns are better than others for localization and a single pattern does not necessarily generalize to all environments, hence the proposed system. It would be interesting to see if the matching performance of Asterisk holds up in other types of environments.

V. CONCLUSION & FUTURE WORK

We presented an unsupervised method of feature matching using learned place-and-time-dependent descriptors. It is demonstrated that this increases the localization ability of single-experience VT&R while maintaining similar computational complexity. We demonstrate day-to-night localization without the use of expensive low-light cameras and pre-processing of the images, which will further improve localization performance. In the case of extreme environmental changes, the representational power given by binary descriptors is insufficient for long-term operation. However, we do see improved matching and localization performance compared to other descriptors.

The performance of the proposed method is affected by the training data used. For testing, we set a fixed interval for switching to a new descriptor and tried a variety of training strategies. It is best to use all the training data that is available, but an intelligent method of determining how often to learn a new descriptor along a path is crucial for optimal matching performance. It is a trade-off between the generality of the descriptor across scene changes and its specificity to a particular location and time.

An interesting extension might be to replace the description function with a neural network. This could offer much more representational power and could be trained using a Siamese network. The inference time for a shallow multilayer perceptron is computationally cheap especially with GPUs and allow it to handle not only illumination but seasonal changes as well. With the addition of convolutional layers, it could start to learn the appearance of dominant landmarks as more experience is gathered.

REFERENCES

[1] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. Brief: Binary robust independent elementary features. *Computer Vision–ECCV 2010*, pages 778–792, 2010.

[2] Nicholas Carlevaris-Bianco and Ryan M Eustice. Learning visual feature descriptors for dynamic lighting conditions. In *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*, pages 2769–2776. IEEE, 2014.

[3] Winston Churchill and Paul Newman. Practice makes perfect? managing and leveraging visual experiences for lifelong navigation. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 4525–4532. IEEE, 2012.

[4] Feras Dayoub and Tom Duckett. An adaptive appearance-based map for long-term topological localization of mobile robots. In *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, pages 3364–3369. IEEE, 2008.

[5] Paul Furgale and Timothy D Barfoot. Visual teach and repeat for long-range rover autonomy. *Journal of Field Robotics*, 27(5):534–560, 2010.

[6] Tomáš Krajník, Pablo Cristóforis, Keerthy Kusumam, Peer Neubert, and Tom Duckett. Image features for visual teach-and-repeat navigation in changing environments. *Robotics and Autonomous Systems*, 88:127–141, 2017.

[7] Stefan Leutenegger, Margarita Chli, and Roland Y Siegwart. Brisk: Binary robust invariant scalable keypoints. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2548–2555. IEEE, 2011.

[8] Chris Linegar, Winston Churchill, and Paul Newman. Made to measure: Bespoke landmarks for 24-hour, all-weather localisation with a camera. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pages 787–794. IEEE, 2016.

[9] Kirk MacTavish, Michael Paton, and Timothy D Barfoot. Beyond a shadow of a doubt: Place recognition with colour-constant images. In *Field and Service Robotics*, pages 187–199. Springer, 2016.

[10] Colin McManus, Winston Churchill, Will Maddern, Alexander D Stewart, and Paul Newman. Shady dealings: Robust, long-term visual localisation using illumination invariance. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 901–906. IEEE, 2014.

[11] Colin McManus, Ben Upcroft, and Paul Newman. Learning place-dependant features for long-term vision-based localisation. *Autonomous Robots*, 39(3):363–387, 2015.

[12] Colin McManus, Ben Upcroft, and Paul Newmann. Scene signatures: Localised and point-less features for localisation. 2014.

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

[14] Peer Neubert, Niko Sunderhauf, and Peter Protzel. Appearance change prediction for long-term navigation across seasons. In *Mobile Robots (ECMR), 2013 European Conference on*, pages 198–203. IEEE, 2013.

[15] Michael Paton, Kirk MacTavish, Michael Warren, and Timothy D Barfoot. Bridging the appearance gap: Multi-experience localization for long-term visual teach and repeat. In *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, pages 1918–1925. IEEE, 2016.

[16] Michael Paton, François Pomerleau, and Timothy D. Barfoot. *In the Dead of Winter: Challenging Vision-Based Path Following in Extreme Conditions*, pages 563–576. Springer International Publishing, Cham, 2016.

[17] Eduardo Perdices, Luis M López, and José M Canas. Lineslam: Visual real time localization using lines and ukf. In *ROBOT2013: First Iberian Robotics Conference*, pages 663–678. Springer, 2014.

[18] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *Computer Vision (ICCV), 2011 IEEE international conference on*, pages 2564–2571. IEEE, 2011.

[19] Renato F Salas-Moreno, Richard A Newcombe, Hauke Strasdat, Paul HJ Kelly, and Andrew J Davison. Slam++: Simultaneous localisation and mapping at the level of objects. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1352–1359, 2013.

[20] Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, Pascal Fua, and Francesc Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 118–126, 2015.

[21] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Learning local feature descriptors using convex optimisation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1573–1585, 2014.

[22] Christoffer Valgren and Achim J Lilienthal. Sift, surf and seasons: Long-term outdoor localization using local features. In *EMCR*, 2007.

[23] Stephen Williams and Ayanna M. Howard. Developing monocular visual pose estimation for arctic environments. *Journal of Field Robotics*, 27(2):145–157, 2010.