SIFT, SURF and Seasons: Long-term Outdoor Localization Using Local Features

Christoffer Valgren   Achim Lilienthal
Applied Autonomous Sensor Systems
Örebro University
SE-70182 Örebro, Sweden
christoffer.wahlgren@tech.oru.se, achim@lilienthals.de

***Abstract*— Local feature matching has become a commonly used method to compare images. For mobile robots, a reliable method for comparing images can constitute a key component for localization and loop closing tasks. In this paper, we address the issues of outdoor appearance-based topological localization for a mobile robot *over time*. Our data sets, each consisting of a large number of panoramic images, have been acquired over a period of nine months with large seasonal changes (snow-covered ground, bare trees, autumn leaves, dense foliage, etc.). Two different types of image feature algorithms, SIFT and the more recent SURF, have been used to compare the images. We show that two variants of SURF, called U-SURF and SURF-128, outperform the other algorithms in terms of accuracy and speed.**

***Index Terms*— Outdoor Environments, Topological Localization, SIFT, SURF.**

I. Introduction

Local feature matching has become an increasingly used method for comparing images. Various methods have been proposed. The Scale-Invariant Feature Transform (SIFT) by Lowe [9] has, with its high accuracy and relatively low computation time, become the de facto standard. Some attempts of further improvements to the algorithm have been made (for example PCA-SIFT by Ke and Sukthankar [8]). Perhaps the most recent, promising approach is the Speeded Up Robust Features (SURF) by Bay et al. [3], which has been shown to yield comparable or better results to SIFT while having a fraction of the computational cost [3, 2].

For mobile robots, reliable image matching can form the basis for localization and loop closing detection. Local feature algorithms have been shown to be a good choice for image matching tasks on a mobile platform, as occlusions and missing objects can be handled. In particular, SIFT applied to panoramic images has been shown to give good results in indoor environments [1, 5] and also to some extent in outdoor environments [12]. However, outdoor environments are very different from indoor environments. There are a number of things that can alter the appearance of an outdoor scene: lighting conditions, shadows, seasonal changes, etc. All of these aspects makes image matching very difficult.[1] Some attempts have been made to match outdoor images from different seasons. Zhang and Kosecka [14] focus on recognizing buildings in images, using a hierarchical matching scheme where a “localized color histogram” is used to limit the search in an image database, with a final localization step based on SIFT feature matching. He et al. [7] also use SIFT features, but employ learning over time to find “feature prototypes” that can be used for localization.

In this paper, only local features extracted from panoramic images will be used to perform topological localization. Several other works rely on similar techniques to do topological mapping and localization, for example Booij et al. [5], Sagues et al. [11] and Valgren et al. [12, 13]. The most recent work related to this paper is a comparative study for the localization task in indoor environments, published by Murillo et al. [10], where it is found that SURF outperforms SIFT because of its high accuracy and lower computation time.

The rest of the paper is structured as follows. In Section II, the SIFT and SURF algorithms are discussed briefly. In Section III, the data sets used in the paper are described. In Section IV, the experiments are outlined and in Section V the results of the experiments are presented.

II. Feature Detectors and Descriptors

Both SIFT and SURF contain detectors that find interest points in an image. The interest point detectors for SIFT and SURF work differently. However, the output is in both cases a representation of the neighbourhood around an interest point as a descriptor vector. The descriptors can then be compared, or matched, to descriptors extracted from other images.

SIFT uses a descriptor of length 128. Depending on the application, there are different matching strategies. A common method, proposed by Lowe [9], is to compute the nearest neighbour of a feature, and then check if the second closest neighbour is further away than some threshold value. Other strategies consider only the nearest neighbour if the distance is smaller than a threshold, as in Zhang and Kosecka [15], or compute only the approximate nearest neighbour by using a kd-tree, as in Beis and Lowe [4].

SURF has several descriptor types of varying length. In this paper, we use regular SURF (descriptor length 64), SURF-128 (where the descriptor length has been doubled), and U-SURF (where the rotation invariance of the interest points have been left out, descriptor length is 64). U-SURF is useful for matching images where the viewpoints are differing by

[1] In some cases even impossible, since a snow-covered field might not *have* any features.

Data set	Number of images	Main characteristics
A	131	No foliage. Winter, snow-covered ground. Overcast.
B	80	Bright green foliage. Bright sun and distinct shadows.
C	597	Deep green foliage. Varying cloud cover.
D1	250	Early fall, some yellow foliage. Partly bright sun.
D2	195	Less leaves on trees, some on the ground. Bright sun.
D3	291	Mostly yellow foliage, many leaves on the ground. Overcast.
D4	264	Many trees without foliage. Bright setting sun with some high clouds.

TABLE I

Reference table for the data sets. The number of images in the data sets only includes the outdoor images used in this paper.

a translation and rotation in the plane (i.e. planar motion). It should be noted that U-SURF is more sensitive to image acquisition issues, such as the omnidirectional camera axis not being perpendicular to the ground plane.

SURF can use the same matching scheme as SIFT, but has one additional improvement. It includes the sign of the Laplacian, i.e. it allows for a quick distinction between bright features on a dark background and dark features on a bright background. This allows for quicker feature matching than SIFT, even in the case of SURF-128.

III. The Data Sets

Seven data sets were acquired over a period of nine months. The data sets span a part of the campus at Örebro University, in both indoor and outdoor locations. For the purpose of this paper, all indoor images have been removed. The remaining images form data sets, ranging from 80 images up to 597 images, see Table I.

A. Details about the data sets

Data set A was acquired on a cloudy day in February, with bare trees and snow covered ground, see Figure 2 and 3.

Data set B was acquired on a warm May day, around noon, with a clear blue sky, see Figure 2.

Data set C, which is also the largest of the data sets and functions as our reference data set (it covers all places visited in the other data sets, see Figure 1), was acquired during two days in July, with a bright sky and varying cloud cover, see Figure 2 and Figure 3.

Data sets D1, D2, D3 and D4 were all acquired during October, with the purpose of capturing how the environment changes during Autumn. They have varying lighting conditions, and a different amount of leaves on the ground, see Figure 2 and Figure 3.

The images were acquired every few meters; the distance between images varies between the data sets. The data sets do not all cover the same areas. For example, data set D1 does not include Positions 1 and 2 in Figure 1.

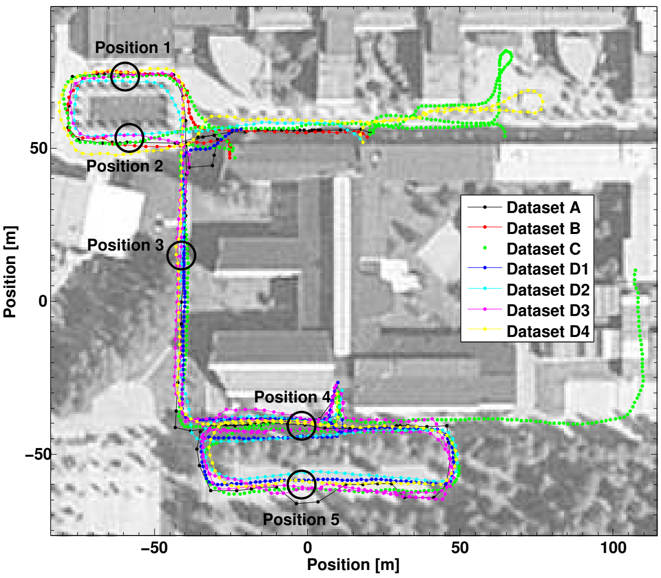

Fig. 1. Aerial image, illustrating the coverage of the data sets. The positions have been acquired by adjusting odometry measurements to fit the map. Approximate total path length for data set C is 1.1 km. Circles indicate the approximate positions 1 to 5 used in Experiment 2.

B. Data set acquisition

The data sets were acquired by an ActivMedia P3-AT robot (Figure 4) equipped with a standard consumer-grade SLR digital camera (Canon EOS350D, 8 megapixels) with a curved mirror from 0-360.com. This camera-mirror combination produces omnidirectional images that can be unwrapped into high-resolution panoramic images by a simple polar-to-Cartesian conversion. To reduce the number of detected features, the images were resized to one third of the original size, to about 800×240 pixels. The number of features for each image in data set C is shown in Figure 5.

The odometry was stored for each run. The odometry was adjusted by hand to produce the paths shown in Figure 1.

IV. Experiments

Two experiments were designed to test the strength of SIFT and SURF.

- In Experiment 1, 40 images were chosen at random from each data set. These images were then matched against the reference data set C. The number of feature matches between the images was simply taken as a measure of similarity; the image with the highest similarity was considered to be the winner.[2] Note that this corresponds to *global topological localization*. The adjusted odometry shown in Figure 1 was used to determine if the localization was correctly performed or not; if the difference in position between test image and the image of the reference set with highest similarity was less than 10 m,

[2] For our images, the number of matches is small and there is little reason to involve a more complicated measure of similarity. Other, more sophisticated ways of determining similarity exist, for example by using the relation between the number of matches and the total number of features. See also the recent work by Cummins and Newman [6].

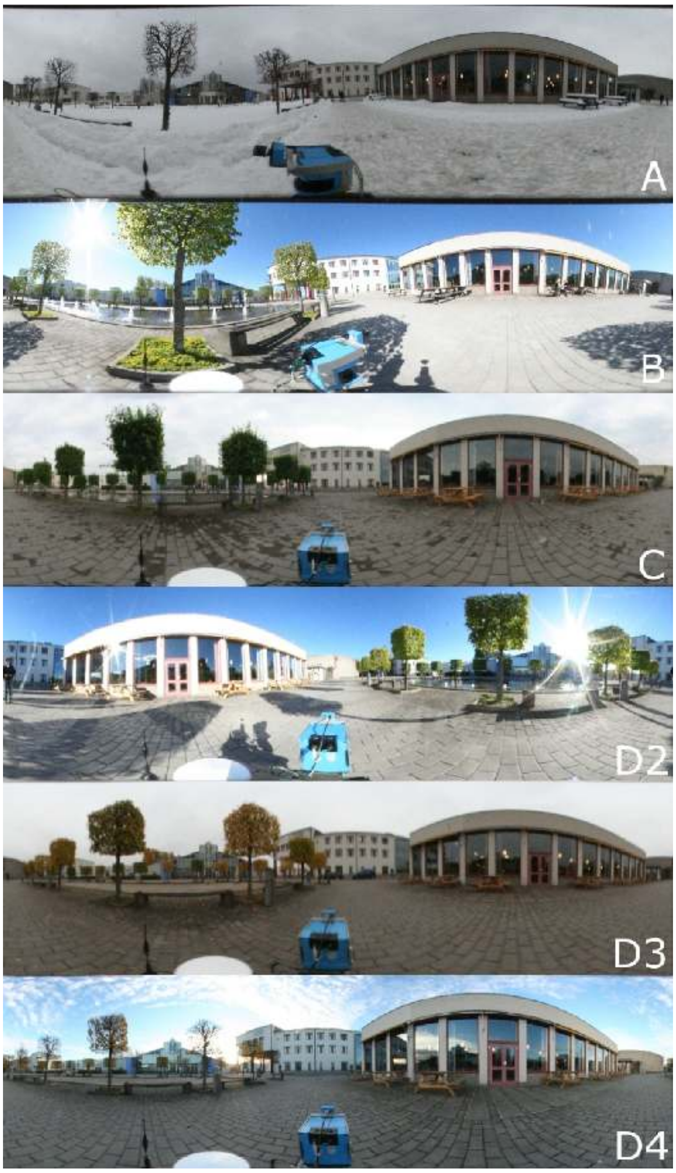

Fig. 2. Data sets A, B, C, D2, D3 and D4. Position 1.

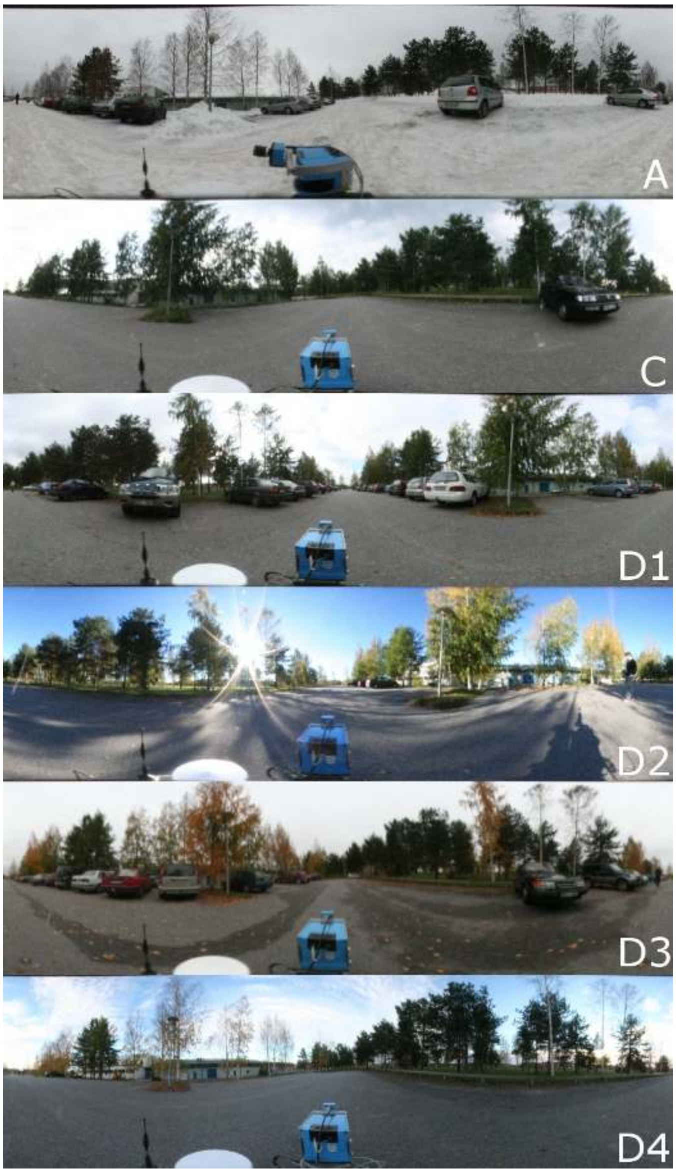

Fig. 3. Data sets A, C, D1, D2, D3 and D4. Position 5.

the localization was considered successful. A localization was also deemed correct if a set of images shared the highest score, and a correct image match was found in the set.

- In Experiment 2, five viewpoints (Position 1 through 5 in Figure 1) that occurred in several of the data sets were compared using SIFT and SURF. The number of correct correspondences and the total number of correspondences were recorded for each viewpoint. In this experiment, we rely on a human judge to determine the correctness of individual feature matches.

The binaries published on the web sites for SIFT[3] and SURF[4] were used to compute the feature descriptors. Both SIFT and SURF utilize feature vectors, so the same code was used to perform the image matching, with the exception that we introduced a check for the sign of the Laplacian for the SURF features. A simple brute force, nearest-neighbour search (using the Euclidean distance) was performed. Lowe found a value of 0.8 for the relation between the nearest and second nearest neighbour [9] to be suitable. In the paper by Bay et al. [3], a value of 0.7 is used for the SURF descriptor.

Since it is not the purpose of this paper to tune a particular matching scheme to our data sets, we have used both 0.7 and 0.8 as threshold in the experiments. However, it is likely that the threshold for the nearest-neighbour matching might influence the result; this is something that we leave for future work.

While an epipolar constraint (as in Booij et al. [5]) could be applied to improve the matching rate, this might give an unfair advantage to one of the algorithms. In particular SURF might suffer from this, since the SIFT keypoint detector returns many more interest points in the images, see Figure 5.

[3] *http://www.cs.ubc.ca/∼lowe/keypoints/*

[4] *http://www.vision.ee.ethz.ch/∼surf/*

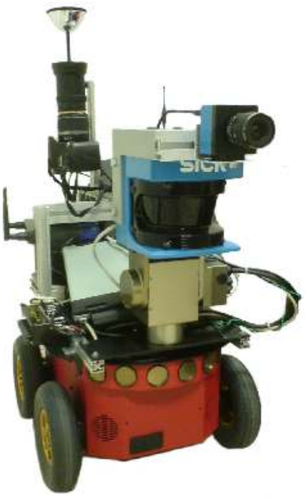

Fig. 4. The mobile robot platform, used to acquire the images used in this paper. The omnidirectional camera can be seen in the top left part of the image.

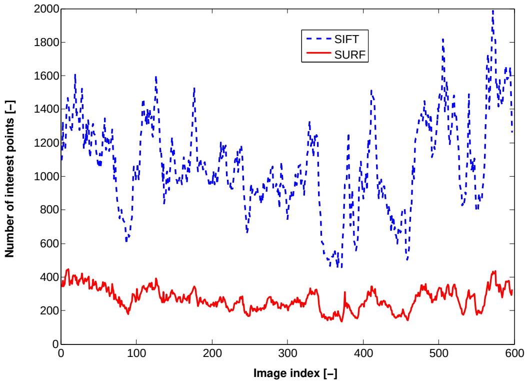

Fig. 5. The number of interest points for reference data set C. Note that the different versions of SURF use the same feature detector and therefore have the same number of interest points.

V. Results

A. Experiment 1

The charts in Figure 6 and 7 show the results from Experiment 1, using thresholds of 0.7 and 0.8, respectively. A striking result is that feature matching alone *cannot*, in general, be used to perform correct single image localization when the data sets are acquired over a longer period of time. The localization rate is too low – even the best algorithm does not reach higher than about 40% in one of the cases. On the other hand, this is perhaps not surprising, since the dynamics of outdoor environments sometimes can leave even humans lost.

Another interesting result is that SIFT performs worst for two of the data sets, and never gives the highest localization rate (exclusively).

SURF-128 outperforms the other algorithms, giving the highest localization rate for all data sets except data set B.

It is notable that the low localization rates of data set B and D2 coincide with a qualitative change in the appearance of the environment; both data set B and D2 were acquired in direct sun light that casts long shadows and causes the white buildings to be very bright, see Figure 2.

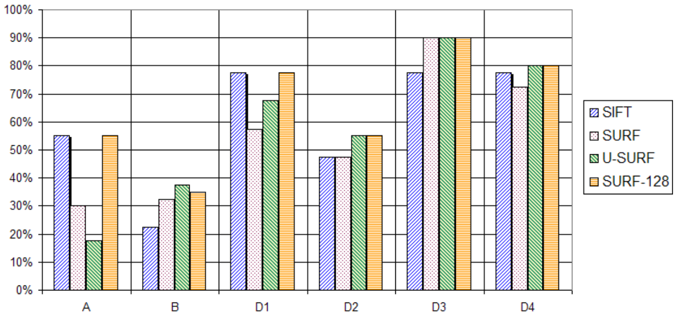

Fig. 6. Result for Experiment 1. 40 random images from each data set were localized with respect to the reference data set C. The bar chart shows the localization rate, using a threshold of 0.7.

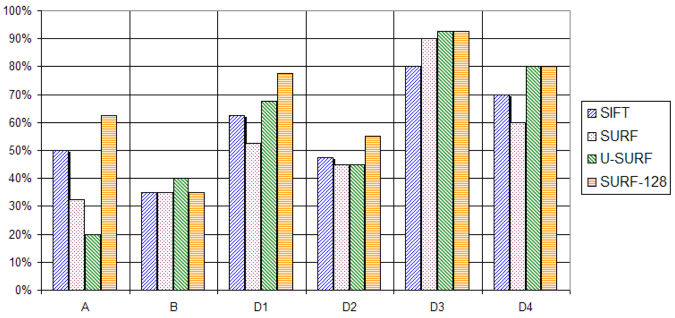

Fig. 7. Result for Experiment 1. 40 random images from each data set were localized with respect to the reference data set C. The bar chart shows the localization rate, using a threshold of 0.8.

B. Experiment 2

The chart matrices shown in Figures 8 through 12 show the results from Experiment 2 using a threshold of 0.7 (results by using a threshold of 0.8 are omitted for space reasons, but are qualitatively the same). Again, data sets B and D2 have a low number of matches. It is also, as one might expect, fairly hard to match the snow-covered data set A to the other data sets.

Data sets C, D3 and D4 were in general easy to match, while data set D2 was the hardest to match. Again, the weather conditions (visible sun, overcast, etc.) seem to be a very important factor. It is of interest that Position 5 (see Figure 3), which should be a very difficult case since it is a parking lot with very few stable, salient features, is not much harder to match than the other cases. For Position 5, U-SURF performs best with the highest number of correct matches.

It is hard to decide which algorithm performs best in Experiment 2. SIFT gives a larger number of matches (perhaps because the number of detected interest points is higher), but more matches are also wrong. With a percentage of only 67% correct matches, it is clearly the worst of the algorithms. Again, SURF-128 comes out on shared first position with 85% correct matches. However, SURF-128 finds fewer matches than the other algorithms. In this regard, U-SURF that has

Algorithm	Total matches	Total correct matches	Percentage correct
SIFT	1222	824	67%
SURF	598	473	79%
U-SURF	760	648	85%
SURF-128	531	452	85%

TABLE II

TOTAL NUMBER OF MATCHES FOR EXPERIMENT 2, WITH THRESHOLD 0.7.

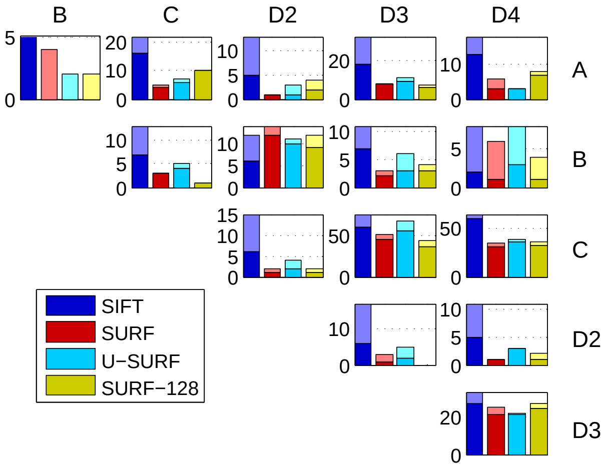

Fig. 8. Result for Experiment 2, position 1. The same approximate viewpoint was selected from several data sets, and the matches were evaluated by a human judge. Each bar chart shows the matches between two data sets (indicated by labels on rows/columns). In the charts, darker color indicates correct matches, brighter color indicates total number of matches.

a high number of both found *and* correct matches will be the winner. See Table II.

C. *Time consumption*

The computational cost for SIFT is much greater than SURF.

Figure 13 shows the computation time required for the detection of features for the different algorithms. U-SURF is fastest, followed by SURF, SURF-128 and finally SIFT, which is nearly three times as slow.

Figure 14 shows the computation time required to do feature matching for the different algorithms. The times required for the SURF variants are all at around 0.25 seconds, while SIFT varies from 0.5 to 2 seconds (with a mean of about 1.1 seconds).

VI. CONCLUSIONS

In this paper, we have investigated how local feature algorithms can cope with large outdoor environments that change over the year. The results from our experiments can be summarized as follows:

- It is not, with the current algorithms, possible to do single panoramic image localization based only on appearance in large outdoor environments with seasonal changes.
- SURF-128 outperforms the other algorithms for these data sets, at least in terms of localization. SURF-128 also has the highest percentage of correct feature matches

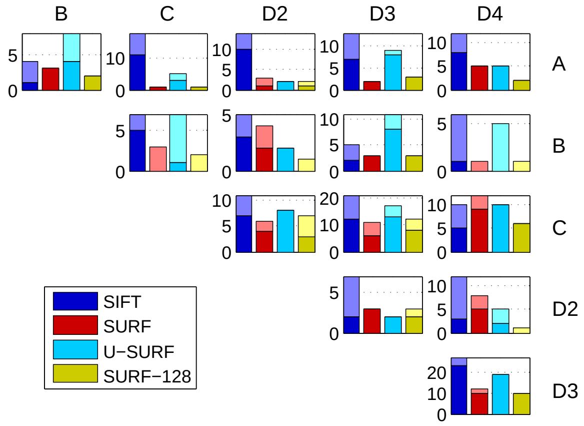

Fig. 9. Result for Experiment 2, position 2.

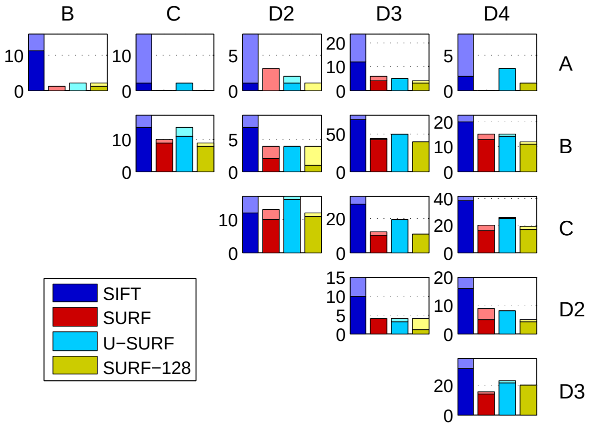

Fig. 10. Result for Experiment 2, position 3.

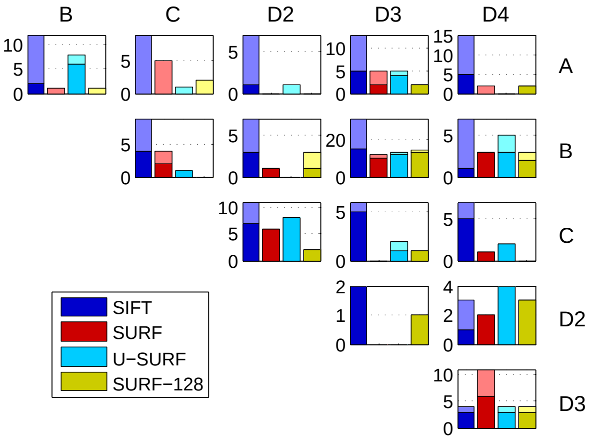

Fig. 11. Result for Experiment 2, position 4.

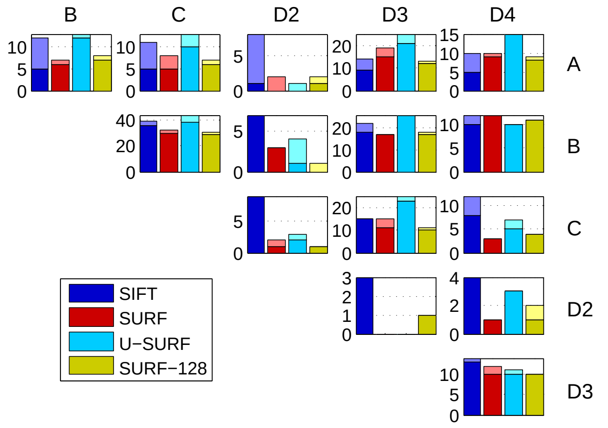

Fig. 12. Result for Experiment 2, position 5.

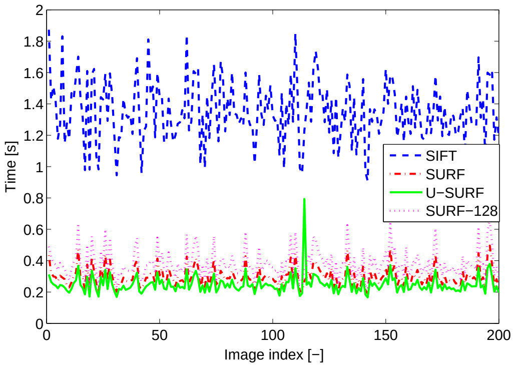

Fig. 13. The time required to compute the descriptors for 200 random images chosen from all data sets.

- when comparing images taken during different seasons, but also returns the lowest number of matches.
- U-SURF had severe problems with the data set containing snow (A) for the localization task, but otherwise performs nearly as well as SURF-128.
- For our data sets, standard SURF has approximately the same performance as SIFT.

None of the SURF variants give results considerable worse than SIFT, yet SURF has a considerably lower computation time. It should therefore be clear that SURF is better suited for the task of localization in outdoor environments – but this is only when one considers the matching technique used in this paper. It might be of interest to apply an epipolar constraint (using RANSAC, for example) which might give SIFT – with its higher number of features – back a slight advantage. In fact, we believe that while SURF might be useful for doing “coarse” topological localization, there are cases when SURF does not return a sufficiently high number of correspondences in order to allow precise pose estimation. In these cases, SIFT, which in general returns a higher number of correspondences, might be a better choice.

Further work involves expanding the experiments, in part to further investigate the impact of the thresholds, but in particular to understand *why* a particular local feature algorithm can successfully handle a case for which another fails. It would also be of interest to see if there is a way to successfully handle the cases with bright sun and distinct shadows.

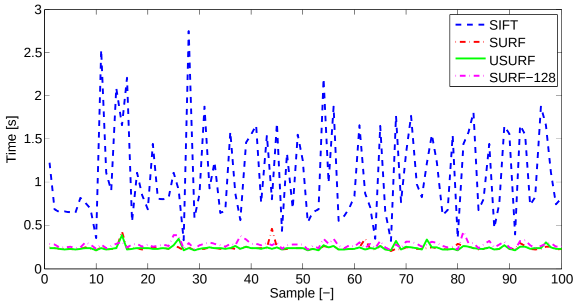

Fig. 14. The time required to match 100 random image pairs chosen from all data sets.

Acknowledgment

This work is partially supported by The Swedish Defence Material Administration.

Many thanks to Henrik Andreasson, Martin Magnusson and Martin Persson for their great help with the data acquisition.

References

[1] H. Andreasson and T. Duckett. Topological localization for mobile robots using omni-directional vision and local features. In *Proc. IAV 2004, the 5th IFAC Symposium on Intelligent Autonomous Vehicles*, Lisbon, Portugal, 2004.

[2] H. Bay, B. Fasel, and L. Van Gool. Interactive museum guide: Fast and robust recognition of museum objects. In *Proc. Int. Workshop on Mobile Vision*, 2006.

[3] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded Up Robust Features. In *Ninth European Conference on Computer Vision*, 2006.

[4] J. Beis and D. Lowe. Shape indexing using approximate nearest-neighbor search in highdimensional spaces. In *Proc. IEEE Conf. Comp. Vision Patt. Recog.*, pages 1000–1006, 1997.

[5] O. Booij, Z. Zivkovic, and B. Kröse. From sensors to rooms. In *Proc. IROS Workshop From Sensors to Human Spatial Concepts*, 2006.

[6] M. Cummins and P. Newman. Probabilistic appearance based navigation and loop closing. In *Proc. IEEE Int. Conf. on Robotics and Automation*, pages 2042–2048, 2007.

[7] X. He, R. S. Zemel, and Volodymyr Mnih. Topological map learning from outdoor image sequences. *Journal of Field Robotics*, 23:1091–1104, 2006.

[8] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 506–513, 2004.

[9] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2:91–110, 2004.

[10] A.C. Murillo, J.J. Guerrero, and C. Sagues. SURF features for efficient robot localization with omnidirectional images. In *Proc. IEEE Int. Conf. on Robotics and Automation*, pages 3901–3907, 2007.

[11] C. Sagues, A.C. Murillo, J.J. Guerrero, T. Goedeme, T. Tuytelaars, and L. Van Gool. Localization with omnidirectional images using the radial trifocal tensor. In *Proc. IEEE Int. Conf. on Robotics and Automation*, pages 551–556, 2006.

[12] C. Valgren, T. Duckett, and A. Lilienthal. Incremental spectral clustering and its application to topological mapping. In *Proc. IEEE Int. Conf. on Robotics and Automation*, pages 4283–4288, 2007.

[13] C. Valgren, A. Lilienthal, and T. Duckett. Incremental topological mapping using omnidirectional vision. In *Proc. IEEE Int. Conf. On Intelligent Robots and Systems*, pages 3441–3447, 2006.

[14] W. Zhang and J. Kosecka. Localization based on building recognition. In *Workshop on Applications for Visually Impaired, IEEE Int. Conf. on Computer Vision and Pattern Recognition*, volume 3, pages 21–21, 2005.

[15] W. Zhang and J. Kosecka. Image based localization in urban environments. In *International Symposium on 3D Data Processing, Visualization and Transmission*, pages 33–40, 2006.