Performance metric for vision based robot localization

Emanuele Frontoni, Andrea Ascani, Adriano Mancini, and Primo Zingaretti

***Abstract*—This paper aims to propose a roadmap for benchmarking vision based robot localization approaches. We discuss the need for a new series of benchmarks in the robot vision field to provide a direct quantitative measure of progress understandable to sponsors and evaluators of research as well as a guide to practitioners in the field. A first set of benchmarks in two categories is proposed: vision based topological and metric localization. In particular we present a novel way to measure performances with respect to the particular data ser used and we show the application of this method on 2 big dataset of omnidirectional indoor and outdoor images. We also present the web site where all data are collected and where every research group can announce their result using the proposed performance evaluator.**

***Index Terms*—Localization, Vision, Feature Extraction, Group Matching.**

I. Introduction

VISION based localization is one of the fundamental problems of mobile robots and until today, while tens of papers proposed different approaches to this problem, there are not real benchmarks in this field able to compare and evaluate advances in vision based robotics.

This paper discusses the strategy for devising and using a new series of benchmarks in the robot vision field and in particular in the field of vision based robot localization. We discuss the need for a new series of benchmarks in the robot vision field to provide a direct quantitative measure of progress as well as a guide to researchers in the field. In particular a first set of benchmarks in two categories is proposed: vision based topological and metric localization.

The knowledge about its position allows a mobile robot to efficiently fulfill different useful tasks like, for example, office delivery. In the past, a variety of approaches for mobile robot localization has been developed. They mainly differ in the techniques used to represent the belief of the robot about its current position and according to the type of sensor information that is used for localization. In this paper we consider the problem of vision-based mobile robot topological and metric localization. Compared to proximity sensors, which are used by a variety of successful robot systems, cameras have several desirable properties. They are low-cost sensors that provide a huge amount of information and they are passive so that vision-based navigation systems do not suffer from the interferences often observed when using active sound- or light-based proximity sensors (i.e. soft surfaces or glasses). Moreover, if robots are deployed in populated environments, it makes sense to base the perceptional skills used for localization on vision like humans do. Over the past years, several vision-based localization systems have been developed. They mainly differ in the features they use to match images or for the appearance based matching (i.e. color histograms, Fourier signatures, etc.) in opposition to metric based approaches (i.e. binocular vision, calibrated omnidirectional vision, etc.). 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. Our data sets, each consisting of a large number of omnidirectional images, have been acquired over different day time both in indoor and outdoor. Two different types of image feature algorithms, SIFT and the more recent SURF, have been used to compare the images, together with a variants of SURF, called U-SURF and a variant of SIFT called Adaptive SIFT2.

The paper is organized as follow. After discussing issues and methodologies in creating vision benchmark problems, this paper presents an evaluation criteria in the field of robot localization and a few example problems in the domain of vision based robot localization using omnidirectional vision.

Authors are with the Department of Ingegneria Infromatica, Gestionale e dell'Automazione (DIIGA), Universit Politecnica delle Marche, Ancona, ITALY. Contact author: frontoni@diiga.univpm.it. Authors thank Giorgio Ascani for his valuable support to this research.

II. Challenges in benchmarking robot vision approach

A pertinent benchmark approach is the field of robotics is a set of evaluations for all subsystems, including stereo vision, road-following, path planning, SLAM, grasping, etc. All this kind of test and modules are necessary to evaluate and compare the behavior of a robot system in different environments and different contours conditions.

The principal goals of the proposed vision benchmarks are to evaluate scientific progress in the specific mentioned problem area and to make the extent of such progress apparent to the evaluators of the research as well as to the scientists working in this field. Evaluation of a set of alternative solutions to a problem naturally involves comparing the resultant scores and to thus rank the techniques. We can't avoid competition. Nonetheless, it is very important to make sure that we are competing on the right problems, that the competition is fair, and that we don't poison the currently excellent cooperative atmosphere that exists in the robot vision research community. Among its positive benefits, benchmarking will promote collaboration. Many researchers will not be able to afford to develop all the system components themselves in order

to enter the evaluation. A module or component that has been proven to have high performance will be transferred from the hands of the developer to other sites whose main research focus is not the module, but rather access to its functionality. An example should be the probabilistic modules on the localization or SLAM problems in comparison with the specific vision problem. Further, the problem of automatically finding settings for adjustable parameters (present in almost every vision algorithm) is a key vision problem in which progress should be encouraged and the benchmark could be a positive influence in this regard.

Other positive effects of benchmark are: the promotion of research because there exists accepted criteria of progress; the establishment of some priorities on problems to be solved; and an increased awareness of the availability and usefulness of a broader range of techniques for performance improvement.

III. Performance evaluator and benchmarking procedure

All the experiments should be performed in different datasets provided that they are collected using the following criteria and published on line to permit comparison. Every dataset must provide a reference image database collected in different environment conditions (in particular light or weather conditions for indoor or outdoor images). Every dataset must provide at least five different test sets collected in the same environment, but in different positions with respect to the data set of reference images. If possible, test sets should be dynamic and cover the case of presence of occlusion.

The performance evaluator takes into account different aspects of the vision based localization problem:

- the size of the dataset S (considering both the extension of the area covered by the robot $S1$ and the number of images collected in the reference image database NS);
- the aliasing of the dataset A (a detailed explication of the aliasing measure will be given in next session);
- the percentage of correct localization, in the case of topological localization, or the average localization error, both represented by L, in the case of metric localization (average of at least ten trials, measured in percentage or millimetres respectively);
- the resolution R used in the image processing module, measured using the total number of pixels of the image (lowest resolutions are considered better in robotics due to the purpose of needing low cost commercial robotics systems);

The evaluator E results in the formula:

$$E = \frac{A \cdot L}{S \cdot R}$$

Where $S = S1/NS$ represents the fact that good results on a big area covered by a few images represent a better performance than the result obtained on a same area covered by a huge number of reference images.

The aliasing of a dataset is a very important parameter when we want to compare results among different approaches using also different data. A place in the environment can be defined aliased when there are other places the look like it. The aliasing should be a measure of the self-similarity of the environment and of the refence dataset (in the case of visual robot navigation). To deal with this problem we define aliasing of a dataset the medium value of the matrix representing the similarity, evaluated with one of the used approach, between all reference images. The highest is the aliasing, the hardest is the localization using those set of reference images.

The web site *vrai.diiga.univpm.it* contains all datasets used in this paper and a couple of forms to collect other datasets and to collect results obtained by different researchers using different approaches to solve the problem of robot localization.

IV. Experiments and results

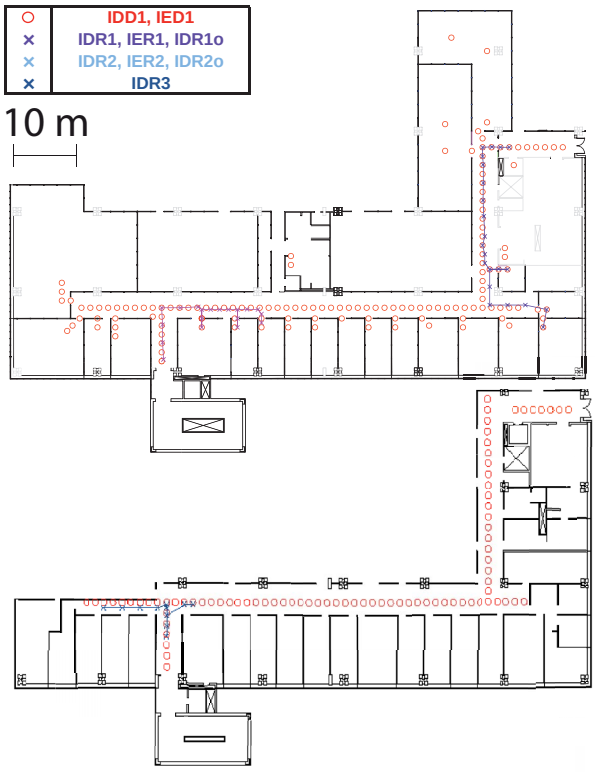

Figure 1: Details about Indoor datasets and routes

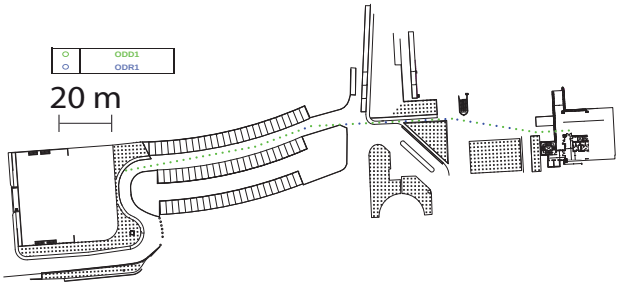

Figure 2: Details about Outdoor datasets

Here we present a comparison, using the proposed evaluation method, among different feature based localization

methods, using two different methods to derive localization from image similarity.

Localization performances are evaluated using two different approaches.

- Winner-Takes-All Localization (WTA): the estimated robot position is the position of the most similar image in the dataset with respect to the actual robot view. This position is compared with the ground truth verified during the test route. If the label is the same the robot is considered correctly localized.
- Monte Carlo Localization (MCL): the estimated robot pose is the result of the probabilistic particle filter applied using a vision similarity based sensor model and a classic motion model.

Figure 3: ActivMedia P3-AT Robot, equipped with omnidirectional camera

Figure 4: Kraun webcam (1280x1024 pixels) and omnidirectional mirror used to collect images

We are interested in an appearance-based place recognition system for topological localization. **Topological localization** algorithms are based on adjacency maps. Nodes represent locations, while arcs represent the adjacency relationships between locations. In our case we rely only on place labels of every reference image: if the label of the estimated (both with WTA and MCL) robot position is the same of the real robot one we consider the topological localization successful.

On another side we can consider a **metric localization** during which we are interested on a minimization of the localization error between the estimated (using MCL) position of the robot and the real robot pose. In this case the robot position is estimated using the error between the highest weighted particle and the real robot position or between the weighted mean particle position and the real robot pose.

A Monte Carlo particle filter could also be used to compare localization performances with classical localization approaches used in literature and to deal better with perceptual aliasing.

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 [1] has become, with its high accuracy and relatively low computation time, the de facto standard. Some attempts of further improvements to the algorithm have been made (for example PCA-SIFT by Ke and Sukthankar [2]). 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], [4]. 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 [5], [6] and also to some extent in outdoor environments [7]. However, outdoor environments are very different from indoor environments. 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 neighborhood 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. SURF has several descriptor types of varying length.

A variant to SIFT was proposed by Frontoni [8]. This approach is based on SIFT extractor and in particular it proposes an improvement of this feature extraction method to deal with changes in lighting conditions. This kind of adaptive vision is necessary in various application fields of vision feature based techniques, e.g. outdoor robotics, surveillance, object recognition, etc. In general, the improvement is particular useful whenever the system needs to work in presence of strong lightness variations. Also, while invariant to scale and rotation and robust to other image transforms, the SIFT feature description of an image is typically large and slow to compute. To solve this matter an approach to reduce the SIFT computational time is also presented.

In this paper, we also 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 different from one another in a translation and/or a 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. An extensive test and comparison among the mentioned approaches was conducted by Valgren [9] comparing different image datasets taken in different seasons [1], [10].

A curved mirror (Figure 4) is located over the camera to grab omnidirectional images.

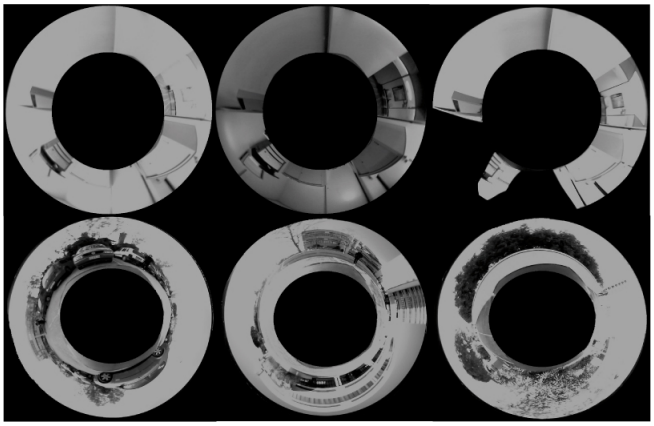

Figure 5: Example of omnidirectional images: (a), (b), (c) are from IDR1, IER1 and IDR1o, representing the same scena in different conditions; (d), (e), (f) are from ODD1

In order to use an image database for mobile robot localization, it should be considered that the probability that the position of the robot at a certain moment exactly matches the position of an image in the database is virtually zero. Accordingly, one cannot expect to find an image that exactly matches the search pattern. In our case, we therefore are interested in obtaining similar images together with a measure of similarity between retrieved images and the search pattern.

So we have to calculate, using features extraction and matching, a parameter to compute how much two different images can be considered similar. In our case, as a simply measure of similarity, we chose the ratio between the number of features matched and the maximum number of features extracted in either image . In this way two different images having a high percentage of similar features could be considered quite similar. In this way is possible to estimate how much the positions of the robot looks like one of the reference pictures.

A. Data Sets

Ten datasets were acquired in several lighting conditions and in dynamic conditions (occlusions, different car parked around, people moving, etc.), both indoor and outdoor. The indoor datasets cover a large part of DIIGA (*Dipartimento di Ingegneria Informatica Gestionale e dell'Automazione*), and outdoor datasets span a part of the area in front of Engineering Faculty.

B. Details about the Data Sets

- IDD1 consists of 218 omnidirectional images describing the two floors of DIIGA, with day-time luminosity conditions.
- IED1 consists of the same number of images than previous, but acquired with evening luminosity conditions.
- IDR1 and IDR2 consists of two routes, with 24 and 15 images respectively, acquired at daily-time at DIIGA's 170 floor.
- IER1 and IER2 are analogous to previous two, but in evening luminosity.
- IDR1o and IDR2o are analogous to IDR1 and IDR2, but images were forcedly occluded so that a part of environment is not visible.
- IDR33 consists of a route of 11 images, acquired at daily-time at DIIGA's 165 floor.
- ODD1 and ODR1 consists of 52 and 12 images acquired outdoor, in front of Engineering Faculty with full day-time luminosity: the sunlight in some cases occludes part of the image.

A "dataset" is the reference image database used as the a priori knowledge of the environment (we do not need a metric map in such an approach). Usually datasets are very large and cover, though not completely, the test area. We acquired images every few meters, even if the distance between images varies between datasets.

With "route" we refer to a series of routes followed by the robot while performing a localization task. Usually this route is relatively short and dynamic conditions around the robot are allowed. Different routes allow us to perform several localization tests without using a subset of the reference image database.

C. Data Acquisition

The datasets were acquired by an ActivMedia P3-AT robot equipped with a webcam (1280×1024 pixels). Over the camera is located a curved mirror to grab omnidirectional images, which were stored in .pgm format and resized to 640×480 pixels to reduce the number of detected features. To produce a general dataset for future tests we stored both color and gray levels images. All used datasets are available on request to authors for benchmark purposes.

D. Experiments

Two experiments were designed to prove the reliability of the proposed approaches with particular attention to image similarity rather than to probabilistic localization algorithms.

- In the first experiment all seven route datasets were compared with the two indoor datasets (daily and evening). For each image of each dataset features were extracted according to several algorithms mentioned before; then the number of feature matches between the images respect to the maximum number of extracted features is taken as measure of similarity.
  The image with the highest similarity was considered to be the winner. If the label of each pair of correspondent

images identifies the same environment, the *topological* localization is correct.
- The second experiment is similar to the first one, but using the outdoor data sets.

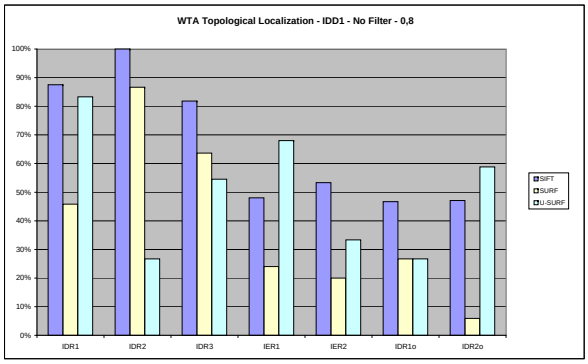

Figure 6: Topological Winner-Takes-All Localization, using IDD1 and no filter, with a threshold of 0.8

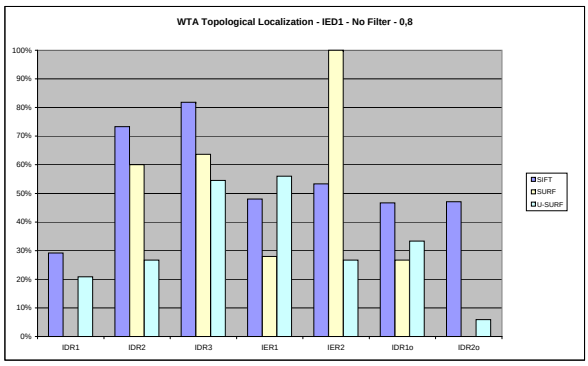

Figure 7: Topological Winner-Takes-All Localization, using IED1 and no filter, with a threshold of 0.8

E. Indoor Results

First charts show the obtained results in indoor environments with no filter applied and using the WTA approach:

SIFT and SURF outperform other algorithms, reaching respectively 65.24% and 46.94%. Using daily data set the results are better (+31%).

However, using routes with low luminosity and those with obstacles similarity is rather small (26.67%), also using adaptative SIFT .

Also using the evening dataset the worst results are obtained with routes acquired in different luminosity (13.54%).

Adaptative SIFT and U-SURF are on the average very lower than others (respectively 36.19% and 35.96%).

Using Monte Carlo Particle Filtering increases localization's percentage for 12% on the average; SIFT and SURF confirm to be the best (respectively 72,51% e 56,43%).

However using occluded datasets the improvement is lower than in other cases.

It is important to say that using Monte Carlo, localization is valuated only after 5 steps of the algorithm.

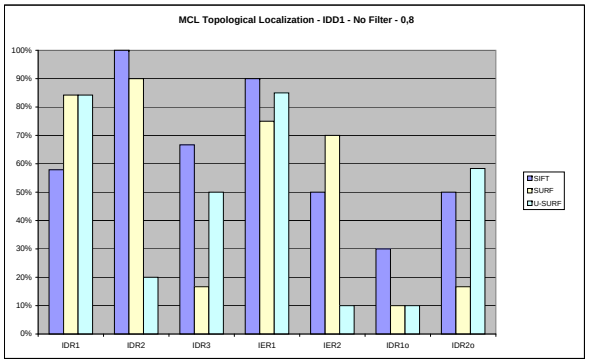

Figure 8: Topological Monte Carlo Localization, using IDD1 and no filter, with a threshold of 0.8

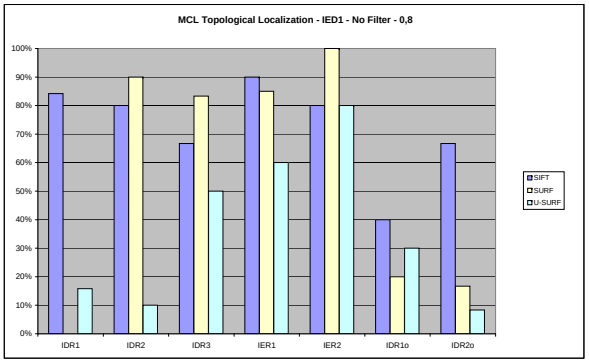

Figure 9: Topological Monte Carlo Localization, using IED1 and no filter, with a threshold of 0.8

In figure Figure 11 we report the metric localization error during the experiments; the graph shows an example of correct localization and tracking over the whole route of the robot. The minimum error is of about 20 cm. This shows that the proposed similarity metric and feature matching approach can be successfully used also in metric localization.

F. Outdoor Results

Using only one outdoor dataset and route it is not possible have an accurate analysis as in the indoor situation. However it is possible to analyze and compare algorithms in a dynamic and variable environment between buildings, cars and vegetables.

Considering WTA Localization, although all algorithms give good topological localization percentages, U-SURF outperforms others, reaching 100% versus 80.56% reached with SIFT.

Monte Carlo localization also in this case increases percentages for 8.73% on the average.

G. Result comparison and performance evaluation

The main objective of this paper remains the possibility to evaluate a benchmark quality value for each of the proposed approach. Table 1 summarizes the performances about some of the above described experiments according to the proposed evaluator.

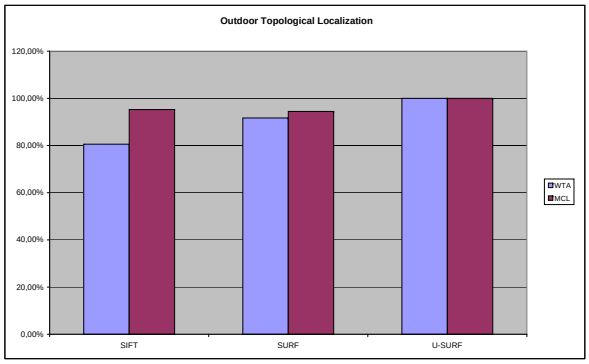

Figure 10: Topological Localization, Outdoor Environment, grouped by Feature Extractor Algorithms

Table I: Performance Evaluator

Experiment	Parameters (A L S R)				E
Indoor WTA	0.25	78	0,2	307,2	0,254
Indoor MCL	0,25	93	0,2	307,2	0,308
Outdoor WTA	0,11	63	0,07	307,2	0,206
Outdoor MCL	0,11	93	0,07	307,2	0,303

This table brings to the conclusion that is possible to compare different approaches over different dataset having a good performance evaluator. In this particular case the quality of the performance parameter is remarked by the observation that the same approach over different datasets brings to similar results. Also the proposed approach is easily applicable to metric and topological localization.

V. Conclusion

We have taken the initial steps in developing a new set of machine vision benchmarks in the areas of robot localization. The next steps involve more careful delineation of the experimental protocol, selection of the specific problems, and the gathering of imagery and other test data. We welcome comments on this new benchmarking effort. In this paper, we also addressed the issues of appearance-based topological and metric localization for a mobile robot over different lighting conditions using omnidirectional vision. Our datasets, each consisting of a large number of omnidirectional images, have been acquired over different day times both in indoor and outdoor environments. Two different types of image feature extractor algorithms, SIFT and the more recent SURF, have been used to compare the images, together with a variant of SURF, called U-SURF and a variant of SIFT called Adaptive SIFT.

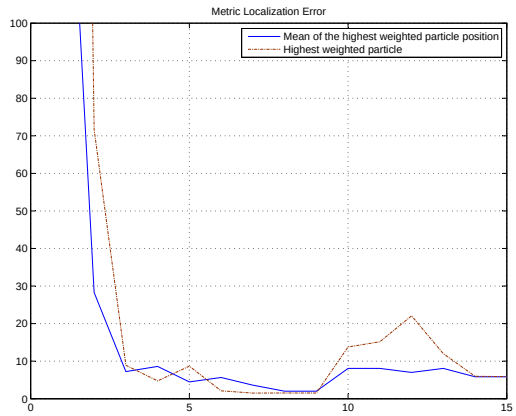

Figure 11: MCL Metric Localization Error in indoor environment during a successful test; error between the highest weighted particle and the real robot position (1) and between the weighted mean particles position (2) and the real robot pose during global localization and traking. Mean error is espressed in number of grid cells where each one is of 20 cm.

References

[1] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[2] Y. Ke and R. Sukthankar, "Pca-sift: A more distinctive representation for local image descriptors," *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 506–513, 2004.

[3] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," in *Ninth European Conference on Computer Vision*, 2006.

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

[5] H. Andreasson and T. Duckett, "Topological localization for mobile robots using omni-directional vision and local features," in *5th IFAC Symposium on Intelligent Autonomous Vehicles (IAV)*, Lisbon, 2004.

[6] B. Krose, O. Booij, and Z. Zivkovic, "A geometrically constrained image similarity measure for visual mapping, localization and navigation," in *Proc. of 3rd European Conference on Mobile Robots*, Freiburg, Germany, 2007.

[7] 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*, 2007, p. 42834288.

[8] E. Frontoni, A. Mancini, and P. Zingaretti, "Vision based approach for active selection of robots localization action," in *Proc. of the 15th Mediterranean Conference on Control & Automation*, Athens, Greece, 2007.

[9] C. Valgren and A. Lilienthal, "Sift, surf and seasons: Long-term outdoor localization using local features," in *Proc. of 3rd European Conference on Mobile Robots*, Freiburg, Germany, 2007.

[10] S. Se, D. Lowe, and J. Little, "Vision-based mobile robot localization and mapping using scale-invariant features," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) 2001*, Seoul, Korea, May 2001, pp. 2051–2058.